Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale
Physique, chimie-physique
ED 182
Université de Strasbourg

ÉCOLE DOCTORALE DE PHYSIQUE ET CHIMIE-PHYSIQUE

Observatoire Astronomique de Strasbourg (UMR 7550)

THÈSE présentée par

Wassim Tenachi

soutenue le: **16 Septembre 2024**

pour obtenir le grade de: **Docteur de l'Université de Strasbourg**
Discipline/Spécialité: Astrophysique

Symbolic Machine Learning for Physics & Astrophysics

THÈSE dirigée par:

M. IBATA Rodrigo Directeur de recherche, Observatoire Astronomique de Strasbourg

RAPPORTEURS:

Mme NECIB Lina Maîtresse de conférences, Massachusetts Institute of Technology
M. TING Yuan-Sen Professeur, The Ohio State University

AUTRES MEMBRES DU JURY:

M. d'ASCOLI Stéphane Chercheur, Facebook Artificial Intelligence Research Paris
M. GREEN Gregory Chercheur, Max Planck Institute for Astronomy
Mme LANÇON Ariane Professeure, Observatoire Astronomique de Strasbourg

Symbolic Machine Learning for Physics & Astrophysics

Wassim Tenachi

Observatoire Astronomique de Strasbourg
Université de Strasbourg

September 2024

A dissertation submitted for the degree of
Doctor of Philosophy

The present thesis outlines the research undertaken as part of a PhD program conducted at the Observatoire Astronomique de Strasbourg, Université de Strasbourg, from October 2021 to September 2024, under the supervision of Dr. Rodrigo Ibata.

The thesis was defended on September 16, 2024, in front of a jury composed of Dr. Rodrigo Ibata, Directeur de recherche at the Observatoire Astronomique de Strasbourg (supervisor); Dr. Lina Necib, Assistant Professor at the Massachusetts Institute of Technology (*rapportrice*); Dr. Yuan-Sen Ting, Professor at The Ohio State University (*rapporteur*); Dr. Stéphane d'Ascoli, Researcher at Facebook Artificial Intelligence Research Paris; Dr. Gregory Green, Researcher at the Max Planck Institute for Astronomy; and Dr. Ariane Lançon, Professeure at the Observatoire Astronomique de Strasbourg.

À la mémoire de Mohammed Djoua, un pionnier intrépide, un homme de progrès et un bâtisseur.

Remerciements

First and foremost, I would like to express my deepest gratitude to Rodrigo. Three and a half years of exchanging ideas and evolving together have created a profound intellectual connection — one where we seem to understand each other almost instantaneously. You are one of the rare people with whom I would happily converse for days on end about anything and everything.

I will always cherish the moments you taught me how to conduct astronomical observations at the INT, all while listening to Radio Paradise. You were not only an extraordinary teacher who transformed the way I work and think, but also an exceptional human being. In a word, you have been a true mentor, and for that, I will forever be grateful.

I would also like to extend my heartfelt thanks to the members of the jury. Lina N., in particular, for her thoughtful and meticulous comments on this manuscript, as well as Yuan-Sen, Stéphane, Gregory, and Ariane, whose work I deeply admire. Each of you brought a unique expertise that perfectly encompassed the multifaceted aspects of this thesis. I could not have dreamed of a better jury, and I am sincerely honored to have had your insights and guidance.

I would like to extend my gratitude to the people I have been fortunate enough to closely collaborate with, alongside Rodrigo: Alejandro, Renaud, Pierre-Antoine, Thibaut, Nicolas M., and a special thanks to Foivos for the incredible journey we shared. I am also grateful for the fortuitous encounter with Emmanuel and Victor — who, as it turns out, work just across the street — and for the enriching exchanges with other people from IRMA. I would also like to thank Clément, Louis, and Reyhaneh for the wonderful week we shared in Marseille during the summer school.

I am deeply thankful to the mentors who trusted me and provided opportunities for me to learn and grow during internships, enabling me to be where I am today: Majdi Hochlaf, Nipon Theera-Umpon, Saima Ben Hadj, and of course Rodrigo. A special thanks to Dominique, who accepted me into the astrophysics master's program in Strasbourg and gave me a chance to pursue what I do today.

I would also like to acknowledge the wonderful people I had the pleasure of meeting through conferences and seminars: Salvatore, for the endless fun and camaraderie we shared, Alexandre K., Florian L., Elizabeth M., Paula, and the inspiring members of the UniverseTBD collaboration. Each of you has made this journey all the more memorable and rewarding.

En revenant maintenant à l'Observatoire, je tiens à remercier chaleureuse-

ment Sandrine et Véronique, sans qui l'Observatoire ne pourrait pas fonctionner, pour leur patience et leur aide inestimables au fil des années. Je voudrais également exprimer ma gratitude envers Benoît, Nicolas M., Jonathan, Zhen, Simon, Rapha, Florent, Katarina, Pau, ainsi que tous les autres collègues du groupe Galaxie, avec qui j'ai eu la chance d'interagir, pour leur accueil chaleureux et leur soutien. Je remercie aussi Ada et Caroline pour leur bienveillance, qui rayonne sur l'Observatoire et contribue à en faire un lieu de travail aussi agréable.

Je tiens également à mentionner les amis que je me suis faits à l'Observatoire et qui ont rendu ces années si belles. C'est souvent dans les périodes de difficulté que l'on découvre la profondeur de certaines amitiés et la richesse humaine de ceux qui nous entourent. Cela n'a jamais été aussi vrai qu'en ce qui concerne les personnes extraordinaires que je vais mentionner ici.

Merci à Amandine, une amie qui m'est *cher* chère, et à Elisabeth S., dont nos chemins se sont croisés depuis la résidence Robert Schumann jusqu'à la thèse. Un immense merci à Pierre-Antoine$^{\triangle}$ et Thibaut$^{\square}$ (aka le Thibster), bien plus que des amis, de véritables PAWT. Cette semaine à Besac' restera à jamais gravée dans ma mémoire, tout comme nos fous rires et les épreuves que nous avons traversées ensemble. P.-A. et Rodrigo, je n'oublierai jamais l'incroyable aventure du rush Gaia que nous avons vécue tous les trois.

Merci à Lucie (aka Dustylulu), Samuel (aka Lisam el Plotlib), Thomas (c'est Thomas !) et Mathias, dont les blagues me manqueront (si si, je t'assure). Vous êtes des amis très chers à mon cœur, avec des qualités humaines inestimables, chacun à votre manière. Je garderai précieusement le souvenir des moments passés ensemble. Un grand merci également à Nicolas, avec qui nous avons partagé des sessions de planétarium pour le moins mémorables, ainsi qu'à Émilie, Julien, Yassin, Srikanth, Margot et Diana pour les instants passés en salle de pause, à Europapark, dans le jardin ou autour d'un bol de ramens.

Enfin, je tiens à exprimer ma profonde gratitude envers les personnes formidables que j'ai eu la chance de rencontrer ces dernières années à Strasbourg et qui ont tant embelli la fin de ma thèse : Paolo et Laurent, Nicolas M., Zhen, Benji, Lols, Laura, ainsi que de plus anciennes connaissances, faites ou retrouvées dans cette belle ville : Lev, Axel (aka jus de citron), Bastien et Molène.

Comme beaucoup de personnes s'orientant vers la physique, j'ai eu la chance de croiser sur ma route des professeurs qui ont profondément marqué mon parcours et auxquels je souhaite rendre hommage ici : Naceur Haddad, Olivier Contet, Valérie Vautrin et, en particulier, Éric Chevreau.

Je tiens également à saluer des figures publiques qui m'ont inspiré à suivre cette voie dès mon plus jeune âge : Brad Wright et Robert C. Cooper, à travers

leur univers de fiction, ainsi qu'Hubert Reeves, David Louapre, Isaac Arthur et Matt O'Dowd, pour leur travail remarquable de vulgarisation en physique et en astrophysique.

Une mention spéciale à Yann LeCun, qui, sans nous connaître, a été suffisamment impressionné pour partager spontanément notre travail sur Twitter. Ce geste a mis en lumière nos efforts et leur a offert une visibilité que nous n'aurions jamais osé espérer. Ce fut un moment véritablement extraordinaire.

Comment ne pas vous mentionner, Agathe (aka NidNid), Corentin (à jamais mon organomagnésien préféré), Alexandre S. (aka Alex le Lion), Quentin et Lucas, avec qui j'ai commencé ma vie d'adulte et mon aventure à Strasbourg. Méline, merci pour ton soutien et ces instants de bonheur qui ont illuminé mes années de thèse et ont participé à les rendre si belles.

Évidemment, un grand merci à Florian A. : tu es un humain extraordinaire et mon plus vieil ami (bientôt un quart de siècle !), ainsi qu'à Oumar, qui est là depuis le début. Merci pour ton amitié indéfectible et pour toutes ces centaines de soirées passées à discuter de tout et de rien.

Je tiens également à exprimer ma gratitude envers ma famille, mon père pour tout ce qu'il m'a transmis en grandissant et qui m'a conduit jusque-là ainsi qu'à Lina T., Louaye, Fedoua et Loqman, ainsi que mes grands-parents : c'est une immense fierté d'être votre descendance. Un grand merci tout spécial à mes tantes sur qui je sais que je peux toujours compter pour leur amour et leur bienveillance : Amel, Fatima, Oumaima et Roudhab. Enfin, merci à ma mère, qui m'a tant apporté, tant matériellement qu'intellectuellement. Elle a toujours été là pour moi, m'a soutenu et m'a encouragé à persévérer contre vents et marées. Toujours.

Publications

Listed below are the works[1] authored by myself, *Wassim Tenachi* [2] that were published in peer-reviewed academic journals during the duration of the present thesis. A publication list including cross-disciplinary citation counts is available here[3].

Refereed first-author publications:

2024 *Class Symbolic Regression: Gotta Fit 'Em All*
W. Tenachi, R. Ibata, T. L. François, F. Diakogiannis
ApJL **969** L26, arXiv:2312.01816

2023 *Physical Symbolic Optimization*
W. Tenachi, R. Ibata, F. Diakogiannis
NeurIPS MLPS **2024** 89, arXiv:2312.03612

2023 *Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws*
W. Tenachi, R. Ibata, F. Diakogiannis
ApJ **959** 99, arXiv:2303.03192

2022 *Typhon: A Polar Stream from the Outer Halo Raining through the Solar Neighborhood*
W. Tenachi, P.-A. Oria, R. Ibata, B. Famaey, Z. Yuan, A. Arentsen, N. Martin, A. Viswanathan
ApJL **935** L22, arXiv:2206.10405

Refereed co-author publications:

2023 *Charting the Galactic acceleration field II. A global mass model of the Milky Way from the STREAMFINDER Atlas of Stellar Streams detected in Gaia DR3*
R. Ibata, K. Malhan, **W. Tenachi**, A. Arentsen, M. Bellazzini, P. Bianchini, P. Bonifacio, E. Caffau, F. Diakogiannis, R. Errani, B. Famaey, S. Ferrone, N. Martin, P. di Matteo, G. Monari, F. Renaud, E. Starkenburg, G. Thomas, A. Viswanathan, Z. Yuan
ApJ **967** 89, arXiv:2311.17202

[1] We also include an article published through the conference on Neural Information Processing Systems acknowledging the computer science tradition of preferring conference publications to journal publications. This article similarly underwent a rigorous peer-review process involving five independent referees.

[2] https://orcid.org/0000-0001-8392-3836

[3] https://scholar.google.com/citations?user=k3yb3EgAAAAJ&hl=en

2022 *Antaeus: A Retrograde Group of Tidal Debris in the Milky Way's Disk Plane*
P.-A. Oria, **W. Tenachi**, R. Ibata, B. Famaey, Z. Yuan, A. Arentsen, N. Martin, A. Viswanathan
ApJL **936** L3, arXiv:2206.10404

Additionally, listed below are the works published through conference proceedings that did not undergo a peer review process:

In proceedings:

2024 *Generalizing the SINDy approach with nested neural networks*
C. Fiorini, C. Flint, L. Fostier, E. Franck, R. Hashemi, V. Michel-Dansac, **W. Tenachi**
ESAIM 24 **1** 1-10, arXiv:2404.15742

2024 *Symbolic regression driven by dimensional analysis for the automated discovery of physical laws and constants of nature*
W. Tenachi, R. Ibata, F. Diakogiannis
SF2A **23** 107T, hal-04325284

2023 *An end-to-end strategy for recovering a free-form potential from a snapshot of stellar coordinates*
W. Tenachi, R. Ibata, F. Diakogiannis
IAU **S379** 147, arXiv:2305.16845

Abstract

We explore the transformative potential of symbolic machine learning in physics and astrophysics, seeking to overcome the interpretability challenges of traditional methods in the era of data abundance. We introduce Φ-SO, a Physical Symbolic Optimization framework that relies on deep reinforcement learning to extract analytical symbolic expressions directly from data. This symbolic regression (SR) framework achieves state-of-the-art performance by integrating physical dimensional analysis and enabling the exploitation of diverse realizations of a singular class of phenomena — an approach we dub Class SR.

Focusing on the dark matter challenges at the galactic scale, we uncover several new stellar streams from Gaia satellite data and perform follow-up observations using the INT and VLT telescopes. Notably, we discover a polar stream from the outer halo passing through the Solar neighborhood, which we dub Typhon. Finally, we propose a first observation-driven, unsupervised learning approach to agnostically constrain the dark matter distribution of the Milky Way from a snapshot of stellar coordinates using canonical transformations.

Keywords: symbolic machine learning, deep reinforcement learning, symbolic regression, dark matter, stellar streams, Milky Way

Abstract (fr.)

Nous explorons le potentiel novateur de l'apprentissage automatique symbolique dans les domaines de la physique et de l'astrophysique, afin de surmonter les limites d'interprétabilité des méthodes traditionnelles dans cette ère caractérisée par une profusion de données. Nous présentons Φ-SO, un paradigme d'Optimisation Symbolique Physique qui exploite l'apprentissage profond par renforcement pour générer des expressions symboliques analytiques directement à partir de données. Cette approche de régression symbolique (SR) atteint des performances de premier plan en intégrant l'analyse dimensionnelle et en facilitant l'exploitation de diverses réalisations d'une unique classe de phénomènes : une approche que nous nommons Class SR.

Nous nous penchons sur les enjeux liés à la matière noire à l'échelle galactique et identifions plusieurs nouveaux courants stellaires grâce aux données du satellite Gaia, complétées par des observations de suivi effectuées avec les télescopes INT et VLT. Nous mettons en lumière l'existence d'un courant polaire émanant du halo externe traversant le voisinage solaire, que nous baptisons Typhon. Enfin, nous proposons une approche pionnière d'apprentissage non supervisé pour déterminer de manière agnostique la distribution de la matière noire dans la Voie Lactée, à partir d'un cliché des coordonnées stellaires en employant des transformations canoniques.

Mots-clés: apprentissage automatique symbolique, apprentissage profond par renforcement, régression symbolique, matière noire, courants stellaires, Voie Lactée.

Summary

Physical theories, particularly in astrophysics, stem from empirical laws. However, with the rise of deep learning and the data abundance era we are entering, many such laws have transitioned into complex neural network representations, preventing their integration into broader theories.

Despite their power and flexibility in modeling almost any physical systems, neural networks largely consist of non-interpretable black boxes. Which begs the question: how can one harness information from data while retaining their ability to interpret and connect with theory? After training a deep neural network to fit a dataset, can one open the black box, to understand the physics modeled inside?

Through the present thesis, we propose that the solution to these unique challenges in physics and astrophysics lies in the development of an innovative machine learning paradigm — one that operates through the manipulation of mathematical symbols in an unsupervised manner: one that is able to automatically distill neural networks or datasets into physical models in the form of concise symbolic analytic laws. This innovative approach complements conventional methods by introducing a valuable dimension of interpretability and has the potential of tackling the increasing challenge of connecting observations to theory in an agnostic manner.

Specifically, we introduce Φ-SO, a Physical Symbolic Optimization framework that utilizes deep reinforcement learning to train a neural network to formulate functional forms that obey specific constraints, such as fitting data points — a problem known as Symbolic Regression.

We develop an algorithm capable of conducting highly informative dimensional analyses on partially constructed equations during the expression generation process. This is useful not only in eliminating physically impossible solutions, but because the "grammatical" rules of dimensional analysis restrict enormously the freedom of the equation generator, thus vastly improving performance.

In addition, we expand our Φ-SO framework to accommodate the search for a unique functional form that fits multiple realizations of a single class of physical phenomena, allowing each realization to have (possibly) unique free parameter values — an approach particularly relevant to astrophysics. We refer to this new type of approach as Class Symbolic Regression and demonstrate its advantages over more traditional methods.

We show that Φ-SO sets a new standard in exact symbolic recovery, achiev-

ing top performances on the standardized Feynman benchmark for Symbolic Regression. We also apply it extensively across a variety of astrophysical problems to showcase its broad applicability and robustness. Additionally, we have developed Φ-SO into a fully-featured, open-source symbolic optimization software (`PhySO`) tailored for the physical sciences, which we make freely available to the community.

The main theme of this thesis is the development of novel, agnostic, unsupervised learning strategies that adhere to an observation-driven philosophy — the only approach permitting the discovery of new physics. Due to the inherently abstract nature of constructing new machine learning approaches to physics and astrophysics, we ground our research in practical challenges, particularly the dark matter problem, which presents numerous difficulties at the galactic scale.

We explore new probes of dark matter within the Milky Way by detecting stellar streams found in observational datasets from the Gaia space telescope, supplemented by dedicated follow-up observations at the Isaac Newton Telescope and at the Very Large Telescope. These structures — tidally disrupted remnants of dwarf galaxies or globular clusters —are intimately linked to the formation and evolution of the Milky Way and serve as excellent tracers of the dark matter distribution. We specifically highlight the discovery of Typhon, a stellar stream from the outer halo passing the Solar neighborhood.

We introduce a framework for agnostically recovering a free-form gravitational potential and its underlying dark matter distribution from a snapshot of stellar positions and velocities. This is the first framework capable of achieving this while leveraging canonical transformations to the space of orbits. We validate our approach on a synthetic test case and show that Φ-SO can distill the obtained neural potential into an analytic form.

Through this thesis we propose an ambitious framework and set of methodologies for extending the symbolic machine learning paradigm into the domain of physics. Our strategies draw from our experiences being confronted to concrete astrophysical challenges. The overarching statement of the present thesis being the establishment of a mutually beneficial relationship between the development of such approaches and the maximization of science returns from observational missions — and in particular the investigation of the dark matter problem, one of the most prominent challenge of physics.

Summary (fr.)

Les théories physiques, particulièrement en astrophysique, sont généralement issues de lois empiriques. Toutefois, avec l'avènement du *deep learning* (apprentissage profond) et notre entrée dans une ère d'abondance des données, de nombreuses lois se sont muées en représentations complexes au travers de réseaux neuronaux, ce qui entrave leur intégration dans des théories plus larges.

Bien que puissants et flexibles, ayant les capacités de modéliser presque tous les systèmes physiques, les réseaux de neurones demeurent largement des boîtes noires non interprétables. Cela soulève la question suivante : comment exploiter les informations sous-jacentes à un jeu de données tout en conservant notre capacité à les interpréter et à les relier à la théorie ? Après avoir entraîné un réseau de neurones, peut-on "ouvrir" la boîte noire qu'il constitue afin de comprendre la physique qu'il modélise ?

À travers cette thèse, nous proposons que la réponse à ces défis uniques en physique et en astrophysique réside dans le développement d'un nouveau paradigme de *machine learning* (apprentissage automatique) : un paradigme opérant par la manipulation de symboles mathématiques de manière non supervisée, un paradigme capable de transformer automatiquement des réseaux neuronaux ou des ensembles de données en modèles physiques sous forme de lois symboliques et analytiques concises. Cette approche novatrice, en enrichissant les méthodes conventionnelles d'une dimension d'interprétabilité précieuse, se présente comme une solution prometteuse pour relever le défi croissant de connecter de manière agnostique les observations à la théorie.

Spécifiquement, nous introduisons Φ-SO, *Physical Symbolic Optimization*, un paradigme d'optimisation symbolique physique qui respose sur le *deep reinforcement learning* (apprentissage profond par renforcement) permettant d'entraîner un réseau de neurones à formuler des formes fonctionnelles respectant des contraintes spécifiques, comme l'ajustement de points de données, une problématique connue sous le nom de *Symbolic Regression* (régression symbolique).

Nous développons également un algorithme capable de réaliser des analyses dimensionnelles informatives sur des équations n'étant que partiellement construites durant le processus de génération d'expressions. Cela nous permet d'éliminer les solutions physiquement impossibles et de restreindre considérablement la liberté du générateur d'équations grâce aux règles "grammaticales" de l'analyse dimensionnelle, améliorant ainsi considérablement les performances de notre système en matière de régression symbolique.

Nous étendons également notre paradigme (Φ-SO) afin d'inclure la recherche d'une forme fonctionnelle unique pouvant ajuster plusieurs réalisations d'une même classe de phénomènes physiques, chaque réalisation pouvant avoir des valeurs de paramètres libres (potentiellement) uniques. Nous baptisons cette nouvelle approche *Class Symbolic Regression* (régression symbolique de classe) et démontrons ses avantages par rapport aux approches plus traditionnelles.

Nous démontrons que Φ-SO établit une nouvelle norme en matière de récupération symbolique exacte de formules analytiques à partir de leur données associées, atteignant des performances de premier plan sur des tests de performance standardisés. Nous appliquons également Φ-SO à travers une variété de problèmes astrophysiques, démontrant son applicabilité large et sa robustesse. En outre, nous avons développé Φ-SO en un logiciel d'optimisation symbolique open-source (code source ouvert) et aux fonctionnalités multiples (`PhySO`) adapté aux sciences physiques, que nous rendons librement accessible à la communauté.

Le thème principal de cette thèse est le développement de nouvelles stratégies d'apprentissage machine agnostiques et non supervisées, adhérant à une philosophie de travail guidée par les données observationnelles : la seule approche permettant la découverte de nouvelles lois physiques. En raison de la nature intrinsèquement abstraite de la construction de nouvelles approches de *machine learning* pour la physique et l'astrophysique, nous ancrons notre recherche dans des défis pratiques, notamment le problème de la matière noire, qui présente de nombreuses difficultés à l'échelle galactique.

Nous explorons de nouvelles sondes observationnelles à la matière noire de la Voie Lactée en détectant de nouveaux courants stellaires dans des ensembles de données observationnels du télescope spatial Gaia, que nous complètons par des observations de suivi dédiées effectuées au Isaac Newton Telescope et au Very Large Telescope. Ces structures, reliquats de galaxies naines ou d'amas globulaires déchirés par les forces de marée, sont intimement liées à la formation et à l'évolution de la Voie Lactée et constituent d'excellents traceurs de la distribution de la matière noire. Nous mettons particulièrement en évidence la découverte d'un nouveau courant stellaire s'étendant du halo externe de notre Galaxie au voisinage Solaire que nous baptisons Typhon.

Nous introduisons également un premier paradigme pour cartographier de manière agnostique un potentiel gravitationnel de forme libre et sa distribution de matière noire sous-jacente à partir d'un instantané de positions et de vitesses stellaires, exploitant les transformations canoniques vers l'espace des orbites. Nous validons notre approche sur un cas test synthétique et montrons que Φ-SO peut distiller le potentiel neuronal résultant en une forme analytique pertinente.

À travers cette thèse, nous proposons un cadre de travail ambitieux et un ensemble de méthodes algorithmiques pour étendre le paradigme du *symbolic machine learning* (apprentissage automatique symbolique) au domaine de la physique, en nous basant sur des confrontations à des défis astrophysiques concrets. L'idée maîtresse de cette thèse est de forger une relation symbiotique entre le développement de ces nouvelles méthodologies et la maximisation des retombées scientifiques des missions observationnelles, en mettant particulièrement l'accent sur l'exploration du problème de la matière noire, l'un des défis les plus cruciaux de la physique moderne.

Contents

Chapter 1

Introduction

Summary.

We tackle the increasingly critical issue of opacity in neural networks used in physics and astrophysics, alongside the epistemological challenges posed by this new approach to science. We propose a solution through the establishment of a novel symbolic machine learning paradigm that leverages mathematical constructs for model interpretability. We then highlight, this thesis' philosophical approach which revolves around an observation-driven, unsupervised learning strategy, emphasizing agnosticism to facilitate new physical discoveries.

Physical theories, in particular in the domain of astrophysics, traditionally stem from empirical laws. Physicists typically observe natural phenomena, formulate empirical laws to describe them, and subsequently construct overarching theories that encompass these laws. For example, Newton's law of universal gravitation [Newton, 1687] elegantly explains both terrestrial object motion and Kepler's planetary motion laws [Kepler, 1609]. However, with the rise of deep learning, many empirical laws have transitioned into complex neural network representations[1], complicating their integration into broader theories.

In astrophysics, thanks to new observational missions and surveys such as Gaia [Gaia Collaboration et al., 2016a], Euclid [Laureijs et al., 2011], LSST [Željko Ivezić et al., 2019, Collaboration, 2009] and SKA [Carilli and Rawlings, 2004], we are entering a new era of (∼Petabyte) data abundance, and there is considerable excitement at the possibility of identifying new empirical laws from these unprecedentedly rich and intricate datasets that could eventually lead to the discovery of new physics. However, the colossal amount of data also presents significant conceptual challenges. Although deep learning will allow us to extract valuable information from the large surveys, it is both blessed and plagued by the underlying neural networks that are one of its most potent components.

Machine learning's opaqueness problem

Neural networks are flexible and powerful enough to model any physical system[2] and work in high dimensions, but they unfortunately largely consist of non-interpretable black boxes. Clearly, interpretability and intelligibility are of great importance in physics, which begs the question: how can one harness information from these large datasets while retaining their ability to interpret and connect with theory? After training a deep neural network to fit a dataset, can one open the black box, to understand the physics modeled inside?

A symbolic learning paradigm

Although immensely challenging, through the present thesis, we propose that the solution to these unique challenges in physics and astrophysics lies in the development of an innovative machine learning paradigm — one that operates through the manipulation of mathematical symbols in an unsupervised manner: one that is able to automatically distill neural networks or datasets into physical models in the form of concise symbolic analytic laws. This innovative approach complements conventional methods by introducing a valuable

[1] The concept of neural network will be formally detailed in in sub-section 2.1.2

[2] That can be described as a Lebesgue integrable function [Lu et al., 2017].

dimension of interpretability and has the potential of tackling the increasing challenge of connecting observations to theory in an agnostic manner.

1.1 The Need for Symbolic Approaches

Prediction vs. explanation : the epistemological dilemma posed by neural networks

This important epistemological question raised by neural networks strikes at the core of physics research. Should we be content with models that predict accurately yet offer no insight into the underlying mechanisms? Imagine, hypothetically, an all-knowing but entirely opaque neural network capable of predicting any physical outcome flawlessly. Would such a tool satisfy our scientific curiosity? Likely not, as the inherent drive for understanding — the physicist's quest for explanations — would remain unfulfilled. This scenario probes a deeper question: What is the ultimate goal of physics: is it merely to predict, or to explain? And if required to choose, which aspect would more fundamentally define the discipline?

Predictive power through unification

Historically, advancements in physics have often come through unifications in the form of simpler, yet powerful theories that explain and predict phenomena across various scales, again one can think of Newton's laws. This preference for simplicity and elegance, often encapsulated by Occam's Razor, suggests a bias towards theories with fewer parameters that still provide comprehensive explanatory power.

Predictive power through complexity

In contrast, neural network models represent a paradigm shift. They excel in making predictions within their trained scope but are typically parameter-dense and lack the explanatory simplicity of analytic models. One might argue that the sheer predictive power justifies deviating from Occam's Razor — if a model is able to predict phenomena that were previously unexplained, perhaps its complexity can be forgiven ?

Mathematics: the language of unification

However, we would argue that this is not (at least for now) a concern, as no current deep learning model can universally learn from and predict all physical phenomena. Instead, we observe a fragmentation of models across various sub-disciplines of physics, each tailored to specific datasets or phenomena. Intriguingly, hints toward new physics might already be lurking deep within one or more of these specialized neural networks, trained on vast observational or experimental datasets.

The historical method of synthesizing empirical observations into comprehensive theories has been through the universal language of mathematics. This tradition suggests that even as we leverage the power of neural networks, there remains a critical need for interpretable mathematical models. Such models are essential for facilitating the communication of physical concepts across various domains within physics.

Mathematical constructs in physics

Galileo famously intuited in Opere Il Saggiatore [Galilei, 1623] that the book of the Universe "è scritto in lingua matematica". Ever since, it has been a central concern of physics to attempt to explain the properties of nature in mathematical terms, by proposing or deriving mathematical expressions that encapsulate our measurements from experiment and observation. This approach has proven to be immensely powerful. Through trial and error over the centuries, the great masters of physics have developed and bequeathed us a rich toolbox of techniques that have allowed us to understand the world and build our modern technological civilization. But now, thanks to the development of modern deep learning networks, there is hope that this endeavor could be accelerated, by making use of the fact that machines are able to survey a vastly larger space of trial solutions than an unaided human.

Symbolic regression

This brings us to the pivotal role of "Symbolic Regression" (SR) in this thesis. Beyond traditional methods that have emerged since the onset of the computer revolution, which typically involve fitting coefficients to predefined linear or nonlinear functions (see, e.g., Press et al. 2007), SR delves deeper. It seeks not just to optimize coefficients within a given mathematical function but to discover the functional forms themselves. Specifically, SR aims to deduce a free-form symbolic analytic function $f : \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_2}$ that fits $\mathbf{y} = f(\mathbf{x})$ given $(\mathbf{x}, \mathbf{y})$ data.

1.2 Observation-Driven Modeling Philosophy

Machine learning's bias problem

Common machine learning approaches to physics and astrophysics often involve training neural networks in a supervised manner, where substantial physical assumptions are embedded into the training examples, e.g., by training on examples derived from simulations based on established physical models. While these approaches can be useful in some instances, they inherently limit the discovery of new physics by enforcing conformity of resulting models to pre-existing theoretical frameworks.

Agnosticity for scientific discovery

This thesis aims to pioneer new methodologies for scientific discovery in physics and astrophysics, advocating for an observation-driven philosophy. This philosophy, overarching the entire thesis, asserts that genuine physical discoveries cannot be achieved merely by adhering to established models but instead require an agnostic exploitation of observational data.

In pursuit of this, we develop and implement frameworks that avoid model-dependent learning in favor of unsupervised learning strategies. These strategies do not rely on predefined physical models but instead aim to construct physical models that inherently adhere to observational constraints. An illustration of this approach is the Φ-SO framework for Physical Symbolic Optimization, a centerpiece of this thesis. Φ-SO learns to formulate symbolic analytical expressions from scratch. It operates through a trial-and-error process, driven solely by the constraint to conform to empirical data without prior exposure to any symbolic expressions. This method exemplifies our commitment to discovering physical models by enforcing behavioral constraints, devoid of any presupposed models — true to the inductive bias-free approach.

Embracing the philosophy articulated by Donald Lynden-Bell [Bonaca and Price-Whelan, 2024], we endeavor to "follow the data" allowing the inherent patterns and truths within the observations to guide our theoretical developments. This approach not only fosters the potential for groundbreaking discoveries but also aligns with the fundamental objective of physics: to elucidate the underlying principles of the universe through the lens of empirical evidence.

The dark matter problem

The pursuit of methodological advances can easily lead toward abstraction, making it crucial to ground them in concrete scientific challenges. The dark

matter problem, exemplifies such a challenge as its inconsistent behavior at the galactic scale suggests potential gaps in our understanding and possible hints towards new physics, offering a fertile ground for testing new theories and methodologies.

The ultimate goal of this thesis is to foster a symbiotic relationship between the development of innovative machine learning strategies as well as symbolic learning approaches and their application in astrophysics, particularly in unraveling the mysteries of dark matter — one of the most pressing challenges in modern physics.

1.3 Overview & Outline

Our objectives are twofold: first, to expand the horizons of symbolic machine learning beyond its current confines, which primarily serve the computer science and control community, into new uncharted territories that offer significant value to physics through interpretability. Second, applying these innovative methods to respond to the contemporary challenge of dark matter, ensuring that the evolution of these methods remains firmly grounded in real science cases.

This Chapter provided a high-level introduction, setting the stage for this thesis by defining its philosophical foundation and core objectives. As outlined below, Chapters 2 and 8 will provide a more detailed contextualization of symbolic machine learning and dark matter research at the galactic scale respectively.

Outline

Chapter 2 delves into the variety of interpretable machine learning approaches applicable to physics and astrophysics, presenting key strategies from the literature that have the potential to advance these fields. It highlights the central role of symbolic learning within this spectrum and provides contextualization for these methods.

Chapter 3 explores the conceptualization of mathematical problems as graph optimization challenges and discusses the representation of formal mathematics as numerical data that can be learned on. It introduces our method for training neural networks to generate mathematical expressions that satisfy certain constraints — such as fitting a dataset (symbolic regression) — through trial and error using deep reinforcement learning.

In Chapter 4, I introduce a technique for incorporating physical dimensional analysis constraints into symbolic optimization, which I then integrate with our reinforcement learning strategy resulting in my Φ-SO framework which reports state-of-the-art performances when evaluated on a standardized SR benchmark.

Chapter 5 extends the Φ-SO framework to enable the search for a single functional form that fits multiple realizations of a particular class of phenomena, allowing each realization to possess potentially unique parameter values. An approach we dub Class SR. I demonstrate the effectiveness of my new approach by introducing and conducting a first benchmark for Class SR and by successfully deriving a synthetic Galactic potential from associated stellar streams data.

Chapter 6 provides insights into the `PhySO` software, which is our implementation of the Φ-SO framework.

Chapter 7 introduces complementary approaches to Φ-SO. These methods employ neural networks to directly capture the graph structure underlying analytic expressions, further broadening the applicability and effectiveness of our symbolic learning approaches.

Chapter 8 outlines the challenges associated with understanding dark matter at the galactic scale, with a specific focus on our own Milky Way galaxy. This chapter sets the stage for a deeper examination of dark matter's role and properties within our galaxy.

Chapter 9 details my contributions to the search for observational probes of dark matter in the Milky Way, emphasizing the discovery and analysis of new stellar streams. I notably introduce a newly identified stream, which we have named Typhon.

Chapter 10 introduces a novel method for mapping the distribution of dark matter in the Milky Way from stellar coordinates. This approach is rooted in our model-agnostic and observation-driven philosophy, utilizing unsupervised learning techniques.

Finally, Chapter 11 concludes the thesis by summarizing our findings and discussing future research directions. It emphasizes prospective methodologies aimed at revealing new constraints on dark matter and explores how the field of symbolic learning can be advanced further.

Chapter 2

Interpretable Approaches for (Astro)-Physics

Summary.

We discuss the limitations of the prevailing supervised learning paradigm stemming from engineering fields which currently dominates applications in physics and astrophysics. We outline alternative methods that may facilitate genuine discoveries in the natural sciences — along with various refreshing illustrative examples from the astrophysical literature. We then explore symbolic regression, detailing its advantages over traditional approaches and providing a brief overview of the relevant literature.

This Chapter aims at contextualizing our observation-driven approach to physical machine learning.

In Section 2.1, we discuss the limitations of the prevailing supervised learning paradigm in machine learning, which currently dominates applications in physics and astrophysics. We outline alternative methods that may facilitate genuine discoveries in the natural sciences — along with various illustrative examples from the astrophysical literature. In the goal of keeping this Chapter stimulating, we introduce essential deep learning concepts by disseminating them throughout our discussion. Specifically, we will touch on: the dense layer architecture, activation functions, and auto-differentiation [Goodfellow, 2016].

Section 2.2 delves into symbolic regression (SR), a central theme of this thesis. We highlight its advantages over traditional neural network approaches and provide a concise overview of the SR literature.

Additional Remarks

As we introduce the machine learning components of this thesis, it is pertinent to clarify a few points. Although our primary focus is on deep learning specifically (the study and usage of neural networks), we occasionally refer to the broader field of machine learning to include topics like auto-differentiation. Throughout this thesis, our discussions related to auto-differentiation and neural networks are based on implementations in `PyTorch` [Paszke et al., 2019], which is currently the most popular deep learning library in research [Papers With Code, 2023].

2.1 Scientific Discoveries in the Machine Learning Era

In sub-section 2.1.1, we analyze the prevailing computer science-centric supervised learning paradigm, discussing its limitations for discovering new physical phenomena. We also highlight specific scenarios where supervised learning remains beneficial. Sub-section 2.1.2 delves into fundamental deep learning concepts, focusing on the core of neural architectures and their ability to model non-linear phenomena. In Sub-section 2.1.3, we discuss the advantages of neural simulation emulators and their application in computational physics and astrophysics.

Sub-section 2.1.4 introduces unsupervised and observation-driven learning approaches, such as reinforcement learning and auto-differentiation, which hold

potential for groundbreaking discoveries in physics. This includes a discussion on differentiable physical simulations and innovative architectures like auto-encoders and approaches for managing uncertainties in deep learning models. Each concept is illustrated with astrophysical examples. Finally, Sub-section 2.1.5 provides a concise overview of the topics discussed in this Section.

2.1.1 The supervised learning paradigm

Deep learning, or the process of fitting a deep neural network, has evolved dramatically from its revival by LeCun et al. [1998]. Initially reignited within the engineering research, it quickly permeated industrial applications[1], owing to its capability to model or emulate almost any system, effectively supplanting entire fields like signal and image processing [Schmidhuber, 2015]. Deep learning represents a paradigm shift not just in capability but in methodology, shifting from expert-derived rules to empirical learning directly from data. This raises a profound question: can a dataset itself be considered a direct model?[2]

An engineering-centric field ?

The need for fast and accurate inference

The engineering-centric focus of machine learning reflects a distinct set of priorities: while interpretability is often secondary, the importance of rapid, accurate inference is paramount. This orientation contrasts starkly with the needs of natural sciences, where understanding and interpretability are crucial, and while rapid inference is beneficial, it is not always central. A scientific discovery being unique by definition, once a breakthrough is achieved, the immediate need for repeated inference diminishes.

Supervised learning frameworks

The engineering emphasis on accuracy and speed has popularized the supervised learning paradigm, where neural networks are trained on paired input-output examples. During inference, the model parameters are "frozen" (i.e. fixed), allowing the network to predict outcomes based on its training. This method typically performs well within the range of its training data due to the neural networks' inherent flexibility. However, this approach inherently limits the scope of discovery in physics and astrophysics, where the goal extends beyond prediction to understanding fundamental processes.

[1] This is underscored by the fact that although open-source, major deep learning frameworks like `TensorFlow` [Abadi et al., 2016] and `JAX` [Bradbury et al., 2018] are developed by large corporations like Google, while `PyTorch` [Paszke et al., 2019] is maintained by Meta.

[2] This notion parallels the definition of human languages, which are often understood through corpora rather than predefined rules [Hunston, 2006].

The limitations of supervised learning for discovering new physics

We have access to only one Universe

The fundamental limitation of applying supervised learning to discover new physics lies in our unique observational dataset — the Universe itself. Unlike other fields where data from varied sources can be used to train and validate models, physics must contend with deriving universal laws from observations constrained to a single instance. This unique situation restricts the utility of supervised learning, which traditionally relies on diverse datasets to generalize and predict outcomes in unfamiliar scenarios. To stretch the metaphor, if we had access to multiple universes, each governed by different physical laws, supervised learning could potentially "triangulate" physical laws applicable to a previously unseen Universe.[3]

Fallacious approaches

While it is feasible to train neural networks on simulations that embed certain physical assumptions, the real test comes when these models are applied to actual observational data. Ideally, simulated data should closely mimic observational data to ensure the model operates within its trained parameters. However, this approach assumes that the physical laws embedded in the simulation accurately reflect reality. There is a risk that researchers might inadvertently re-confirm the assumptions built into the simulation when applying these models to real-world data, mistaking the echo of their assumptions for a discovery. This highlights a critical pitfall in using supervised learning where the model is only as good as the assumptions of its training data and might not genuinely extend to uncovering new principles in observational data.

Sensible supervised learning approaches to (astro)-physics

Processing massive datasets

Despite its limitations in discovering new physical laws, supervised learning can serve as an invaluable tool in the preliminary analysis of vast datasets within astrophysics.

Let us illustrate our point with the analysis of stellar spectra[4]. In this context, supervised learning models are exceptionally adept at deducing key stel-

[3] This is partly why Bayesian approaches to probability are often favored over frequentist approaches in physics and astrophysics, where experimental repetition on a universal scale is impossible.

[4] A particularly interesting example since it was precisely the analysis of stellar spectra and the detection of the first patterns hinting towards stellar evolution that birthed astrophysics from astronomy.

lar characteristics, such as metallicity[5] or surface gravity, from spectral data. These models leverage large, well-characterized datasets where the properties of stars are well-understood and consistent. By training neural networks on these datasets, researchers can automate the analysis of stellar spectra, effectively standardizing this aspect of astrophysical research as exemplified by the approach used in the `APOGEE` catalog [Holtzman et al., 2018].

Utility-first modeling: When the outcome justifies the means

In some contexts, the process of model generation is less important than the utility and accuracy of the model itself. This is especially true in scenarios where the final model can be independently verified and tested, regardless of its origin. In such contexts, the method of discovery — whether through traditional methods or via a black-box providing the solution — is secondary to the model's validity and applicability.

SR exemplifies this approach. SR uses machine learning techniques to generate a physical model in the form of an analytic expression that fits observational data. The key advantage here is that the output — analytic expressions — is inherently interpretable and verifiable, standing apart from the computational method used to derive it.

Another similar application is the resolution of long-standing mathematical conjectures through the generation of formal mathematical proofs. Here, the focus is on the effectiveness of the solution provided, rather than the mechanics of the neural network that produced it. If a neural network, even a black-box model, can propose a valid proof to a mathematical problem, the proof itself can be scrutinized and validated independently of the method used to discover it.

Parameter optimization in astrophysics often involves adjusting simulation parameters so that the simulation's output aligns with observed data. While traditional methods like Markov Chain Monte Carlo (MCMC)[6] are prevalent due to their robustness in uncertainty estimation, neural networks offer a direct and potentially faster alternative. By training networks on pairs of simulation outputs and parameters, we can use them to predict parameters that produce a desired outcome, which are then easily verifiable through a single simulation run. This method provides a direct pathway to solution verification, albeit often without the uncertainty estimates provided by methods like MCMC —

[5]Metallicity in astrophysics refers to the proportion of mass in a star that is not hydrogen or helium, often measured relative to the Sun's metal content.

[6]This approach involves constructing a Markov chain within the parameter search space, where the distribution of the chain represents the underlying distribution being explored.

we will discuss solutions overcoming that limitation later in this Section.

More generally, the approach of using neural networks to emulate complex simulations represents a growing trend. As by effectively capturing the dynamics of simulations, neural networks can offer faster alternatives to running computationally expensive models.

2.1.2 Fundamental building blocks of deep learning

The effectiveness of neural networks in accurately modeling complex simulations or physical systems is often remarkable. To shed light on this phenomenon, we will discuss the most foundational architecture of deep learning — a component that persists as a sub-element in nearly all cutting-edge architectures to this day [Vaswani et al., 2017, Grathwohl et al., 2018, Ho et al., 2020]: the dense layer and the concept of activation functions.

Dense layers

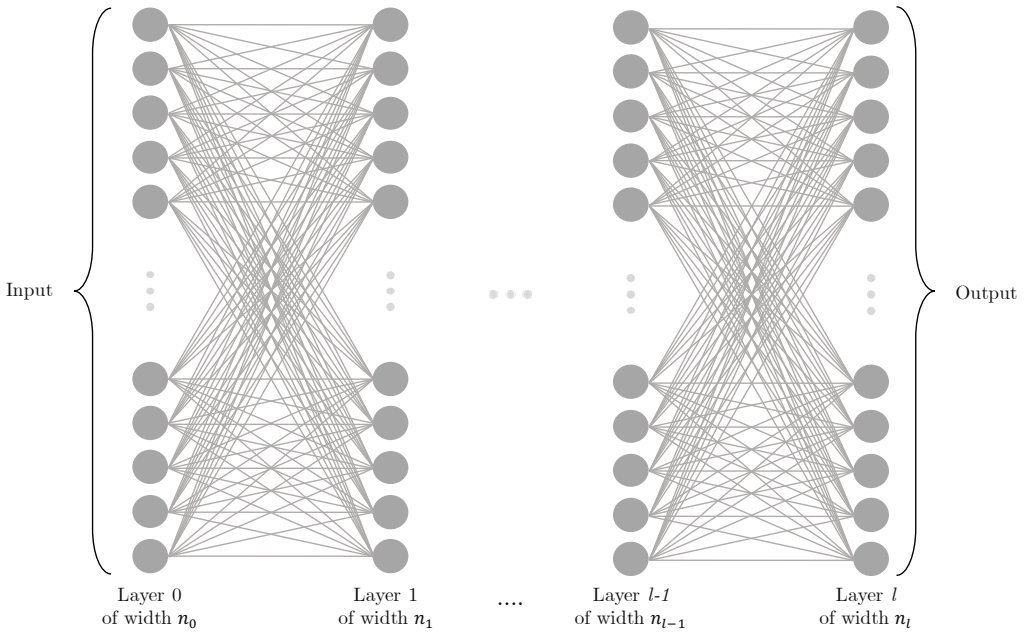


Figure 2.1: General representation of dense layers where neurons are symbolized by circles and their connections by solid lines. See 2.1.2 for a detailed description.

The dense layer, also known as a fully-connected layer or multi-layer perceptron (MLP) when multiple layers are stacked, represents a fundamental

building block of many neural networks. An MLP consists of l layers of neurons, nodes that typically hold values between 0 and 1. Conceptually, the input vector can be considered the 0-th layer, with the final layer producing the predictions.

The layers between the first layer and last layer are called hidden layers. In MLPs each neuron of a given layer i (where $i \neq 1$ and $i \neq l$) is connected to every neuron of the previous layer (i.e. layer $i-1$) and every neuron of the next layer (i.e. layer $i+1$) hence the name of fully connected. See Figure 2.1 for a general overview of this architecture. Each neuron's value is simply a sum of linear functions of the values held in the previous layer with multiplicative coefficients called weights and the offset called a bias. Specifically the activation of the i-th neuron of a given layer $j > 0$ is thus given by:

$$a_i^{(j)} = g\left(b_i^{(j)} + \sum_{k=1}^{n_{j-1}} w_{i,k}^{(j)} \cdot a_k^{(j-1)}\right) \tag{2.1}$$

Where $\left\{b_i^{(j)}\right\}_{k \leq n_{j-1}}$ and $\left\{w_{i,k}^{(j)}\right\}_{k \leq n_{j-1}}$ denote the sets of all bias and weight parameters for the j-th layer, respectively. These parameters are the trainable parameters that can be tuned or fitted to best match target values based on input values, this process is referred to as training. Here, g refers to the so-called *activation* function.

Activation functions

Activation functions are pivotal in enabling neural networks to model complex nonlinear phenomena. Though the underlying components of these networks are linear, it is the nonlinear activation functions that enables them to capture more intricate behaviors. These functions often also serve to constrain the output values within a specific range, such as $[0, 1]$.

In the context of the methodologies discussed in this thesis, we frequently use the hyperbolic tangent (tanh) and the sigmoid (σ) functions. The sigmoid function, in particular, is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

These functions are not only fundamental for introducing non-linearity but also crucial for ensuring that the outputs of neural network layers stay within manageable bounds, facilitating stable learning.

Although we will discuss later in this Section, it is important to emphasize that auto-differentiation is as fundamental a building block of deep learning as neural architectures or activation functions, if not more so [Goodfellow, 2016].

2.1.3 Neural emulation of simulations

Parameter searches

Neural emulations of simulations, commonly referred to as "simulation-based inference" (SBI), offer a powerful tool for accelerating parameter searches [Cranmer et al., 2020a]. Given that neural networks — even very large ones — are typically orders of magnitude faster to evaluate than complex astrophysics or physics simulations, they are particularly advantageous for high-dimensional parameter search problems where traditional grid searches are impractical. These approaches allow for occasional verification checks with randomly selected parameters, ensuring the emulator's accuracy without the need for exhaustive simulations across a vast parameter grid.

For instance, research in cosmology has demonstrated that such approaches can surpass Markov Chain Monte Carlo (MCMC) techniques in accurately recovering posterior distributions[7] of parameters [Zhao et al., 2022].

Furthermore, by running a series of simulations and employing neural probability distribution estimators (such as the normalizing flow models we will explore in Section 10.2) researchers can approximate the entire posterior distribution. This method circumvents the need to explore many more parameters within simulations, significantly facilitating the search process.

On the virtues of neural emulation

Beyond acceleration

While neural emulators significantly accelerate computational processes, their value extends beyond mere speed. They offer unique advantages that can make them indispensable, even in scenarios where simulations could be executed instantaneously.

Tackling inverse problems

One significant advantage is their capability to address inverse problems. By inverting the training of the neural network — training it to predict the input parameters θ from a simulation output rather than the other way around — we can simplify the process of finding initial conditions or parameters that explain observed phenomena. This approach is depicted in Figure 2.2. For instance, Chardin et al. [2019] demonstrates how a neural network distilled from simulations can predict the cosmic reionization timeline field from current day observations of the hydrogen 21 cm line.

[7] In Bayesian statistics, the likelihood of a parameter value is determined by its posterior probability, which incorporates prior knowledge through the prior distribution and includes a marginalization term that accounts for all other variables [Bayes, 1763].

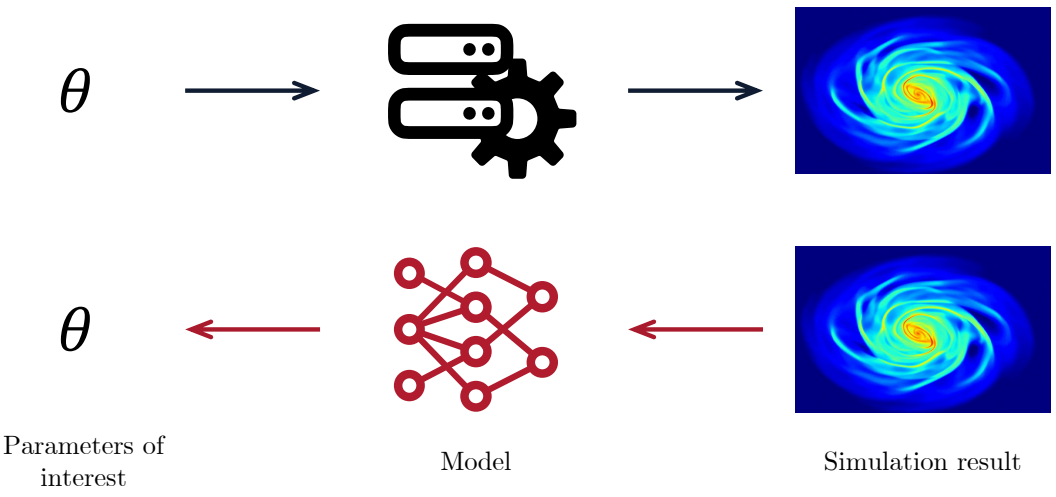


Figure 2.2: **Neural Emulator for Tackling Inverse Problems.** A trained neural emulator can reverse the typical simulation process. Instead of generating results from parameters, it infers the parameters θ that would lead to a specific simulation outcome. This facilitates solving inverse problems by predicting initial conditions or parameters that align with aribitrary outcomes.

A differentiable emulator

Another key advantage is the differentiability of neural emulators. Thanks to auto-differentiation (detailed later in 2.1.4), it is possible to obtain derivatives through neural networks, facilitating the use of gradient descent to optimize simulation parameters directly against desired outcomes, typically to align with observations.

While supervised learning techniques offer valuable tools for scientific inquiry, by construction they can not be employed to discover new physical laws. These engineering-centric approaches are intrinsically limited in the context of physics and astrophysics. As physicists, we need to employ these techniques thoughtfully, stepping beyond conventional paradigms to leverage deep learning's full potential.

2.1.4 Agnostic approaches

Framework and examples

Framework

A shift from traditional training paradigms leads us to consider methods where neural networks are not just fed training examples with known outcomes. Instead, they are tasked to make predictions by adhering to a set of constraints, typically physical or observational. This process, known as "unsupervised

learning" involves training networks through trial-and-error without predefined outcomes.

Examples

A prime example of this methodology is our approach to SR, where the network outputs an analytic expression. The requirement here is for the expression to fit observational data accurately, without the network having been exposed to prior examples of symbolic expressions.

Another common application of unsupervised learning in (that is also common to engineering) is clustering, where the objective is to identify groups of similar items within a dataset. For instance, Dodd et al. [2023] employed clustering techniques to discern Milky Way structures within a vast dataset of stellar positions and velocities near the Sun.

Another notable instance in astrophysics is the `ActionFinder` algorithm, detailed by Ibata et al. [2021]. This method learns a canonical transformation (and its underlying Hamiltonian) to the space of so-called actions — effectively orbits — in an unsupervised manner, ensuring that stars from a single stellar stream[8] have similar values in the latent space. This is achieved without prior examples or reliance on a physical dynamical model, the sole assumption being that stars from a single stellar stream approximate a single orbit.

Agnosticity

In natural sciences, particularly when new physical models are under investigation, agnosticity becomes crucial. Unsupervised approaches ensure that learning is not biased by pre-existing theories or simulations, thereby opening the door to genuine physical advancements. Unsupervised learning represents the only viable method where learning is free from biases typically introduced by training on known outcomes.

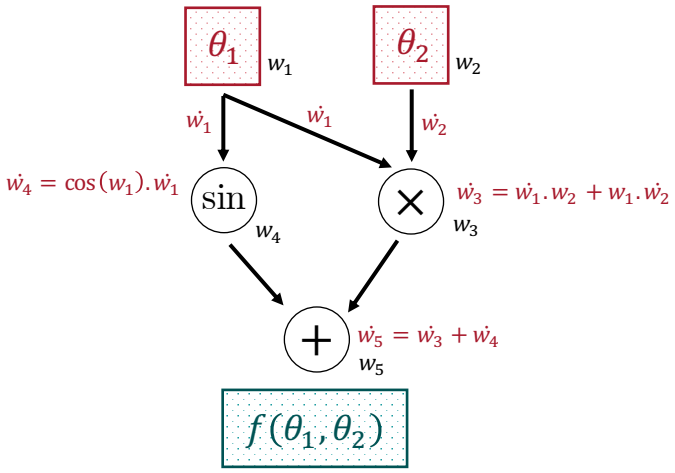
On the power of auto-differentiation

Auto-differentiation is an underappreciated yet powerful tool introduced by deep learning. Let us explore its fundamental concept and utility.

Approximation through over-parametrization

One might assume that the efficacy of deep learning stems merely from the extensive number of parameters within neural networks, enabling them to finely

[8]These elongated, thin structures are formed when celestial bodies are accreted by the Milky Way. We will explore this concept further in the contextual discussions provided in Chapter 8.



(a) Computational graph

```
θ1.grad = True
θ2.grad = True
w3 = θ1 * θ2
w4 = torch.sin(θ1)
w5 = w3 + w4
```

(b) Pytorch syntax

Figure 2.3: **Auto-differentiation illustration.** For a set of parameters $\theta = \theta_1, ..., \theta_n$, the derivatives of each computational step are stored, enabling the use of the chain rule to compute derivatives $\{\frac{\partial x}{\partial \theta_1}, ..., \frac{\partial x}{\partial \theta_n}\}$ with respect to θ for any variable x. This figure shows a computational graph (a) and a typical syntax in the `PyTorch` framework [Paszke et al., 2019] for a simple computation: $\sin(\theta_1) + \theta_1\theta_2$.

approximate a wide range of functions, similar to Taylor series. This capability is formally supported by the universal approximation theorem, which asserts that a sufficiently large Multi-Layer Perceptron (MLP) can approximate any Lebesgue integrable function [Hornik et al., 1989].

To illustrate, consider progressively complex physical models: a simple free fall equation assuming no atmospheric resistance : $z(t) = -\frac{1}{2}gt^2 + v_0t + z_0$ with three parameters, is less precise than a model considering uniform atmospheric pressure $z(t) = H \ln \frac{1+e^{-2t/T}}{2} + v_0t + z_0$ with six parameters[9], which in turn is less accurate than a neural network that may involve thousands of parameters.

A compelling observation in contemporary deep learning research is that neural networks with more parameters than data points often perform exceptionally well without overfitting, provided they are trained correctly, including the use of a distinct control test set that the network has never encountered during training [Li and Liang, 2018].

Auto-differentiation: deep learning's secret sauce

While the vast number of parameters in neural networks undeniably contributes to deep learning's success, another critical factor is backpropagation.

[9]Where z, t, g, v_0, z_0 are the altitude, time, surface gravity, initial velocity, initial altitude respectively and H and T are a scale height and a characteristic time respectively.

This process involves tracking every mathematical operation during inference — potentially amounting to millions — and recording its derivative in a computational graph. This allows for the automatic and analytical differentiation of the cost function with respect to the trainable parameters via the chain rule, greatly aiding in convergence. Without this capability, although it would be theoretically possible to set parameters allowing neural networks to emulate any function, practically finding these parameters i.e. training networks through deep and complex layers would be infeasible. Auto-differentiation make possible the propagation of cost function derivatives throughout even very deep neural networks. This foundational technique is illustrated in Figure 2.3.

Differentiable Simulations

Implementing entire simulations in an auto-differentiable framework, where every operation is differentiable or can be approximated as such, opens extraordinary possibilities. For instance, Li et al. [2022] developed a cosmological simulation entirely within this framework, allowing them to optimize initial conditions to meet any specific observational criteria. This capability to "back-propagate" through a simulation to adjust initial conditions or any variable demonstrates the profound impact of auto-differentiation, initially popularized by deep learning yet fundamentally independent from it. Such flexibility means that one could theoretically optimize a simulated universe's initial conditions to align with any desired observational outcome, showcasing the powerful utility of this approach.[10]

Deep learning techniques

Learning Methods

We have discussed unsupervised learning setups, which involve training neural networks based on any differentiable constraint. These constraints can be complex, extending to computations in physical simulations (as we will do ourselves in Chapter 10) or any process that allows for differentiation.

Auto-differentiation stands alone as a powerful tool for learning parameters within physical systems, offering a straightforward approach optimize values with respect to observational data through a physical model.

When dealing with non-differentiable objective functions, deep reinforcement learning becomes crucial. In this setup, neural networks, often referred to as policies, learn to maximize a so-called reward by adapting their strategies

[10] Li et al. [2022] demonstrate the capabilities of this approach by optimizing the initial conditions of a simulated universe so that present-day observations reveal a pattern spelling out their software's name, `pmwd`, across large-scale cosmic structures.

based on the outcomes of their actions, guided by a reward function, effectively approximating gradients. This method is pivotal to the frameworks developed in this thesis and is extensively discussed in Section 3.3. Its ability to handle non-differentiable objectives makes it particularly valuable for applications in robotics and human interactions since we obviously can not auto-differentiate reality, as well as non differentiable simulations i.e. most current day simulations[11]. Despite its utility, it remains one of the few unsupervised learning methods extensively developed within engineering fields due to its practical applications [Schmidhuber, 2015].

Auto-encoders

An important mention must go to auto-encoders employed in an unsupervised manner. In such frameworks, one aims to reproduce the input data after processing it through a highly compressed, low-dimensional bottleneck layer. This process not only reduces data dimensionality but also captures profound insights into the data's structure within the latent space.

Variational Auto-Encoders (VAEs) take this a step further by modeling the distribution of data within the latent space, learning parameters like the mean and variance. These models are invaluable for their ability to reduce complex data into more manageable forms without losing essential information.

A notable application of this approach can be seen in the work by Laroche and Speagle [2024], which demonstrated that entire stellar spectra could be effectively encoded using just six scalar values through this method. This example highlights the potential of VAEs to significantly condense vast amounts of data while retaining critical information, a technique that has profound implications given how it parallels physics research.

Addressing uncertainties

In the realm of natural sciences, accounting for uncertainties is paramount. The technique known as dropout [Srivastava et al., 2014], initially designed to prevent overfitting by randomly disabling a fraction of neurons during training, also facilitates uncertainty estimation [Gal and Ghahramani, 2016]. This approach effectively trains multiple variants of the model simultaneously, each operating with a different subset of neurons. As a result, the variability in the network's predictions can be interpreted as an uncertainty measure, providing a range of possible outcomes instead of a single fixed prediction.

Building on this concept, one can envision each neuron not merely as a deterministic unit but as a mini-distribution governed by its mean and variance. This notion forms the foundation of Bayesian Neural Networks [Goan

[11] This also includes e.g., video games.

and Fookes, 2020], where exploiting these distributions allows for a quantifiable uncertainty in predictions, offering a deeper insight into the reliability of the neural outputs.

Towards symbolic learning

Throughout this discussion, we have explored various ways deep learning can contribute to scientific endeavors, occasionally offering a level of interpretability. Neural networks are invaluable in fields like image processing or complex systems modeling, where such "soft" models have the ability to capture subtle nuances-qualities. We will explore how this may concern endeavors to map the potential of the Milky Way in Chapter 10.

Yet, when our objective shifts towards uncovering fundamental physical laws, the necessity for symbolic interpretability becomes apparent. The language of mathematics provides a clearer, more definitive description of natural phenomena. The following Section introduces symbolic regression, setting the stage for Chapter 3 where we delve deeper into how symbolic approaches represent and manipulate formal mathematics.

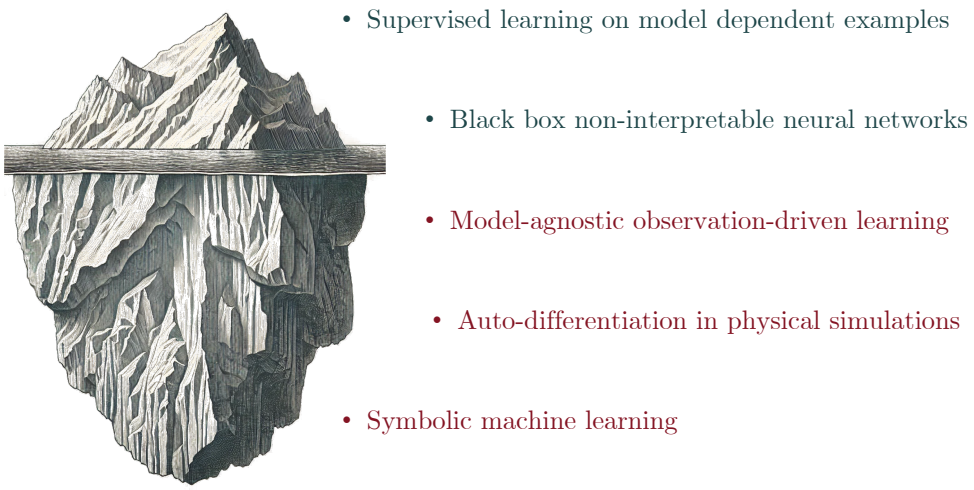


Figure 2.4: An iceberg of machine learning approaches to physics & astrophysics

2.1.5 Beyond the surface

The key takeaway from this Section is that beneath the conventional surface of machine learning applications in physics and astrophysics — most of

which are directly borrowed from engineering — lies a vast 'iceberg' of innovative approaches. These methodologies, deeply rooted in interpretability, hold the potential to drive genuine scientific discoveries. This concept is visually summarized in Figure 2.4, which depicts an iceberg of machine learning approaches, illustrating the substantial yet under-developed opportunities for groundbreaking research in physics and astrophysics.

2.2 Symbolic Approaches

Since the beginning of the scientific revolution, researchers have tried to find repeatable regularities in experiments and observations. Mathematical structures were used in this exploration, and many new ones including functions and differential equations were developed to respond to this need to model nature. Perhaps because of shared symmetries between nature and mathematics, these abstract structures have often been found to work exceedingly well in reproducing and predicting properties of the world, to the point where some have even considered whether the universe is actually mathematical at heart [Tegmark, 2008].

Symbolic Regression (SR) which is central to the present thesis is has a long pedigree. Perhaps its most famous application was by Kepler to planetary ephemerides, thereby finding the fitting law that bears his name [Kepler, 1609]. This empirical law gave the observational basis upon which Newton was able to build the physical theories developed in his Principia Mathematica [Newton, 1687].

In this Section, we introduce modern SR which aims to use the immense computational resources at our disposal to search through possible analytic descriptions in terms of a set of functions and operators (e.g. x, $+$, $-$, $\times$, $/$, sin, cos, exp log, ...) to best fit some numerical dataset $(\mathbf{x}, y)$ we wish to model. Concretely, one seeks some analytic function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ that fits $y = f(\mathbf{x})$ given those data.

Sub-section 2.2.1 introduces keys aspects of SR, namely its potential for compactness, generalization, intelligibility & interpretability and sub-section 2.2.2 offers a brief review of modern SR approaches. Further exploration of novel and more sophisticated symbolic learning approaches, which extend beyond existing SR methods, will be discussed in the perspectives outlined in Section 11.2.

2.2.1 Symbolic regression

Symbolic regression addresses the aforementioned opaqueness of machine learning methods by producing compact, interpretable and generalizable models. Indeed, the goal is to find very simple prescriptions such as Newton's law of universal gravitation that can explain well a vast number of experiments and observations. There are many advantages to discovering physical laws in the form of succinct mathematical expressions rather than large numerical models:

Compactness

SR methods can produce extremely compact models, e.g., with expressions of containing $\sim 10^1$ symbols [La Cava et al., 2021] which is on par with the typical length of expressions in the Feynman Lectures on Physics [Feynman et al., 1971] for example which is of 16 (with the higher end of SR methods producing expressions well below a length of 10^3). In contrast numerical models such as neural networks typically rely on many more parameters. This makes the models computationally inexpensive to run and in principle also enables SR to correctly recover the exact underlying mathematical expression of a dataset using much less data than traditional machine learning approaches [Wilstrup and Kasak, 2021] and with a robustness towards noise even for perfect model recovery [Reinbold et al., 2021, La Cava et al., 2021].

Generalization

In addition, unless the target equations consist of arbitrarily long polynomials, the compact expressions produced by SR are less prone to overfitting on measurement errors and are much more robust and reliable outside of the fitting range provided by the data than large numerical models, showing overall much better generalization capabilities as demonstrated in [Sahoo et al., 2018, Kamienny et al., 2022, Kamienny and Lamprier, 2022, Wilstrup and Kasak, 2021] (we will provide an example of this in Section 4.4.5). This makes SR a potentially powerful tool to discover the most concise and general representation of the measurements.

Intelligibility & interpretability

Since the models produced by SR consist of mathematical expressions, their behavior is intelligible to us, unlike large numerical models. This is of enormous value in physics [Wu and Tegmark, 2019] as SR models may enable one to connect newly discovered physical laws with theory and make subsequent theoretical developments. More broadly, this approach fits into the increasing push towards intelligible [Sabbatini and Calegari, 2022], explainable [Arrieta

et al., 2020] and interpretable [Murdoch et al., 2019] machine learning models.[12]

2.2.2 A brief survey of modern symbolic regression

Traditional approaches to SR

Genetic programming

SR has traditionally been tackled using genetic programming where a population of candidate mathematical expressions are iteratively improved through operations inspired by natural evolution such as natural selection, crossover, and mutation. This type of approach includes the well known `Eureqa` software [Schmidt and Lipson, 2009, 2011] (see Graham et al. 2013, Thing and Koksbang 2025 for benchmarks evaluating the capabilities of `Eureqa`-type algorithms on astrophysical test cases), as well as more recent works [Cranmer, 2023, de Franca and Aldeia, 2021, La Cava et al., 2019, Cava et al., 2019, Virgolin et al., 2019, Cranmer et al., 2020b, Virgolin et al., 2021, Stephens, 2015, Kommenda et al., 2020].

Other traditional approaches

In addition, SR has been implemented using various methods ranging from brute force to (un-)guided Monte-Carlo, all the way to probabilistic searches [McConaghy, 2011, Kammerer et al., 2020, Bartlett et al., 2023a, Brence et al., 2021, Jin et al., 2019], as well as through problem simplification algorithms [Luo et al., 2022, Tohme et al., 2023].

Deep learning

Main approaches

Given the great successes of deep learning techniques in many other fields, it is not surprising that they have now been applied to symbolic regression, and now challenge the reign of `Eureqa`-type approaches [La Cava et al., 2021, Matsubara et al., 2022]. Multiple methods for incorporating neural networks into SR have been developed, ranging from powerful problem simplification schemes [Udrescu and Tegmark, 2020, Udrescu et al., 2020, Cranmer et al., 2020b], to end-to-end symbolic regression methods where a neural network is trained in a supervised manner to map the relationship between datasets and their corresponding symbolic functions [Kamienny et al., 2022, Lalande et al., 2023, Biggio et al., 2020, 2021, Vastl et al., 2022, d'Ascoli et al., 2022,

[12] This is particularly important in fields where such models can affect human lives [European Commission, 2021, 117th US Congress, 2022].

Kamienny et al., 2023, Bendinelli et al., 2023, Holt et al., 2023, Li et al., 2024a,b, Chen et al., 2024a, Meidani et al., 2024, Becker et al., 2022, Shojaee et al., 2024, Alnuqaydan et al., 2022, Aréchiga et al., 2021], all the way to incorporating symbols into neural networks and sparsely fitting them to enable interpretability or to recover a mathematical expression [Fiorini et al., 2024, Scholl et al., 2023, Martius and Lampert, 2017, Brunton et al., 2016, Zheng et al., 2022, Sahoo et al., 2018, Valle and Haddadin, 2021, Kim et al., 2020, Panju and Ghodsi, 2020, Ouyang et al., 2018], an approach often refer to as neuro-symbolic. See [La Cava et al., 2021, Makke and Chawla, 2022, Angelis et al., 2023], for recent reviews of symbolic regression algorithms.

Deep reinforcement learning

While some of the aforementioned algorithms excel at generating very accurate symbolic approximations, the reinforcement learning based deep symbolic regression framework proposed by Petersen et al. [2021a] is the new standard for exact symbolic function recovery, particularly in the presence of noise [La Cava et al., 2021, Matsubara et al., 2022]. This has resulted in a number of studies in the literature built on this framework [Tenachi et al., 2023a,b, 2024, Landajuela et al., 2021a,b, Kim et al., 2021, Petersen et al., 2021b, Landajuela et al., 2022, Faris et al., 2024, He et al., 2024a, Bastiani et al., 2024, Du et al., 2022, Tian et al., 2024, Michishita, 2024, DiPietro and Zhu, 2022, Zheng et al., 2022, Landajuela et al., 2021b, Usama and Lee, 2022].

Overview

Finally, we highlight `PySR` [Cranmer, 2023], an open source attempt at implementing the `Eureqa` software [Schmidt and Lipson, 2009, 2011] that has similar performances. Although `PySR` does not utilize deep learning techniques, it has gained significant traction in the astrophysics community. We show that the approach proposed as part of this thesis significantly outperforms `PySR` in subsection 4.3.2.

A comparative analysis of major SR methodologies, including our own deep reinforcement learning approach — which is the subject of Chapters 3-6 — will be presented through the standard Feynman benchmark [La Cava et al., 2021] in Figure 4.3. Our approach is notable as the only one to date where a neural network manipulates mathematical symbols developed within any field of physics or astrophysics.

We will also explore a problem simplification approach in Section 7.1 and a neuro-symbolic approach in Section 7.2.

Chapter 3

Learning Symbolic Mathematics

Portions of the content presented in this Chapter have been previously discussed in the following publications:

2023 *Physical Symbolic Optimization*
W. Tenachi, R. Ibata, F. Diakogiannis
NeurIPS MLPS **2024** 89, arXiv:2312.03612

2023 *Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws*
W. Tenachi, R. Ibata, F. Diakogiannis
ApJ **959** 99, arXiv:2303.03192

Summary.

We explore how the symbolic language of mathematics can be numerically encoded using graph representations, enabling neural networks to generate and learn formal mathematical concepts. We detail the automated generation of analytic expressions, ensuring their validity through systematic principles.

We then introduce a deep reinforcement learning strategy that trains neural networks to formulate functional forms that meet specific constraints, such as fitting data points in the case of symbolic regression.

The essence of interpretability and intelligibility in science traditionally hinges on the language in which most scientific models have historically been written: mathematics. A language that some argue to be the language of the Universe itself [Galilei, 1623].

In this Chapter, we explore how mathematical language, traditionally represented in symbolic form, can be translated into numerical data — a translation process commonly referred to as *embedding*. We then show how such representations can be exploited to enable statistical models and neural networks to generate and learn on formal mathematical representations.

In Section 3.1, we explain how mathematical knowledge can be represented through graphs and how both formal mathematical problems and analytical modeling problems can be approached as combinatorial graph optimization problems. In Section 3.2, we elaborate on how these concepts facilitate the description of analytic expressions as mere probabilistic distributions and detail how they can be sequentially generated by algorithms employing the so called *prefix* notation. Lastly, Section 3.3 shows how one can use deep Reinforcement Learning (RL) to solve such combinatorial problems within the domain of formal mathematics. In particular, we present a RL framework designed for generating analytic expressions that obey specific constraints, such as accurately fitting data points — a process known as *Symbolic Regression* (SR).

This Chapter is integral to one of the overarching themes of this manuscript: the Φ-SO framework for Physical Symbolic Optimization that was developed as part of the present thesis. At the core of this framework is the automated generation and management of arbitrary analytic mathematical expressions, designed to enable interaction with machine learning techniques — a subject that is thoroughly explored within this Chapter. Additionally, a critical component of the Φ-SO framework is its reinforcement learning algorithm, which enables the optimization of symbolic expressions to meet specified objectives, and in particular SR — a process that is also detailed here.

3.1 Encoding Symbolic Mathematics as Graphs

This section explores the creation of interpretable models in physics in the form of mathematical expressions. However, the principles discussed here can be applied more broadly, encompassing automated computer program generation, automated proof generation and even the automated creation or extension of

mathematical theories themselves such as the Zermelo–Fraenkel set theory[1] This interdisciplinary applicability is particularly relevant given the theoretical aspect of physics.

To represent these structures as numerical data, one might initially contemplate encoding them as plain text using conventional text-encoding techniques, much like current generation Large Language Models, which processes information as sequences of string characters [OpenAI, 2023]. However, this approach overlooks a crucial element: the inherently hierarchical information structure common to these applications. Such structures are poorly represented by plain text encoding, yet they are critical for effectively generating and learning mathematical structures — a process we will elaborate on further in this Chapter. To prevent the loss of this vital information, we will demonstrate how graph representation can be effectively utilized to encode these structures.

In sub-section 3.1.1, we demonstrate how formal theories can be encoded as graphs. This discussion leads into the encoding challenges specific to physical theories, particularly regarding analytic expressions, which we explore further in sub-section 3.1.2. Finally, in sub-section 3.1.3, we show how these challenges fall into the category of NP-hard graph optimization problems.

3.1.1 Encoding formal theories

Mapping theories with directed graphs

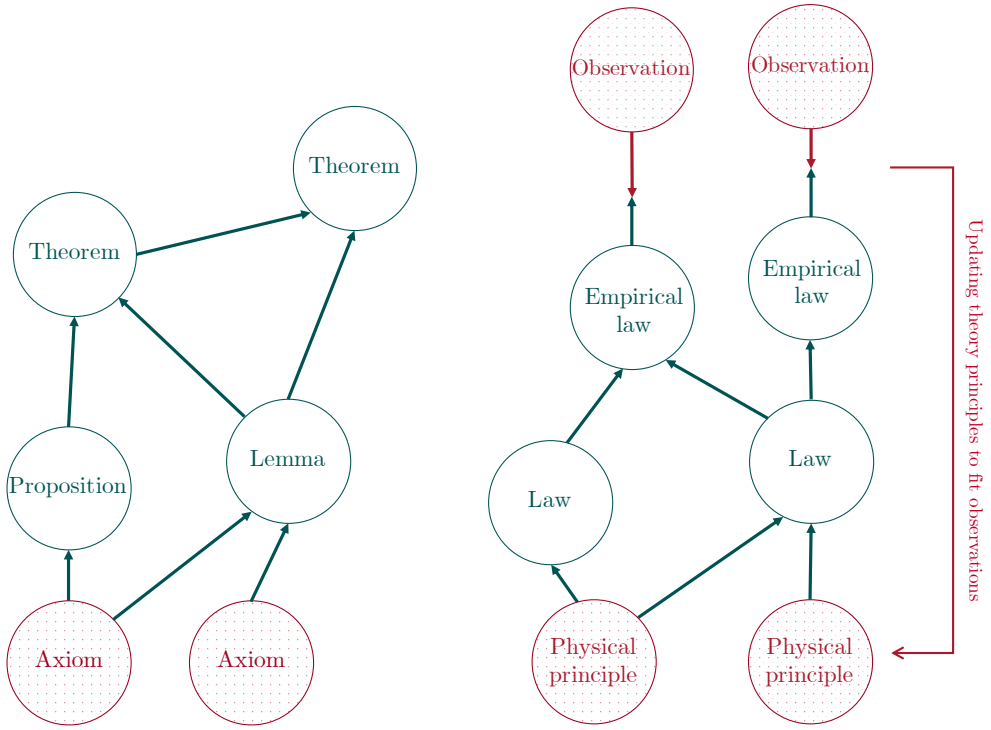
Mathematical theories can be effectively represented as directed graphs, where axioms serve as root nodes. From these axioms, theorems and propositions emerge as subsequent nodes, connected by edges that represent proofs. This directed, computable representation can be used to interface statistical methods to formal mathematical theories.

In this graph structure, each proof, represented as an edge, can itself be viewed as a separate graph if examined more closely. This detailed graph begins with hypothesis nodes necessary for proving a theorem and progresses through a series of derived true statements until the conjecture is verified.

In practice, the application of these graph-based concepts to existing mathematical knowledge has culminated in the creation of the `MLFMF` dataset (*Machine Learning for Mathematical Formalization*) [Bauer et al., 2023a], which currently stands as the largest collection of formalized mathematical knowledge in a machine-readable format. This dataset includes over 250,000 theorem

[1] Zermelo–Fraenkel set theory with the Axiom of Choice (ZFC) constitutes the foundational axiom system predominantly employed in the construction of modern mathematics.

and proposition nodes interconnected by proof edges. Additionally, other notable initiatives include the `CoqGym` environment [Yang and Deng, 2019] which provides a platform for training automated proof agents, featuring more than 71,000 proofs.



(a) Mathematical Theory (b) Physical Theory

Figure 3.1: **Representing mathematical and physical theories as graphs.** An illustration of how mathematical and physical theories can be represented as graphs. While mathematical theories can grow freely from their axioms, physical theories derived from fundamental principles must adhere to experimental observations — discrepancies between theoretical predictions and observations often triggering updates to the underlying physical principles.

Automated theorem proving

This type of computerized representation of mathematics — applicable across a variety of contexts[2] — is typically utilized to develop and refine proof assistants capable of automatically identifying nodes (premises in the form of

[2]For example such representations are also relevant given the ongoing debates [Ochigame, 2024] regarding the use of computers to verify the correctness of mathematical proofs during the peer-review process. Such verification can prevent erroneous proofs from going undetected for years, as seen in the case of the notorious Four Colour Conjecture Kempe [1879].

axioms, propositions or theorems) critical for theorem proving or even suggest proofs. Examples of such systems include `Coq` [The Coq Development Team, 2024], `Lean` [De Moura et al., 2015] and `Minimo` [Poesia et al., 2024].

Recent literature has introduced numerous powerful methods in this field Li et al. [2024c], including automated premise selection which can be treated as a classification problem. Here, premises can be classified as either useful or not, with the knowledge processed sequentially in seminal works using e.g., Recurrent Neural Networks (RNNs) [Irving et al., 2016] and leveraging the graph structure of mathematical theories with Graph Neural Networks (GNNs) as proposed in more recent works [Olšák et al., 2020].

Furthermore, there has been significant progress in automatically generating proof steps and even in fully automated theorem proving. For instance, some studies have employed large transformer models trained in a supervised manner — an approach similar to Large Language Models training [Polu and Sutskever, 2020, Han et al., 2022, Polu et al., 2023]. However, the state-of-the-art performance has been achieved using deep reinforcement learning strategies reminiscent of `AlphaZero` [Silver et al., 2018], where a player and a critic engage in self-play using automatically generated problems to learn efficient exploration and demonstration strategies Lample et al. [2022]. Additionally, the notable `AlphaGeometry` [Trinh et al., 2024], specifically designed for solving geometry problems was proposed. This system was trained in a supervised manner on synthetic problems and approaches the performance levels of an average International Mathematical Olympiad (IMO) gold medalist on previously unseen problems.

Toward a computational paradigm in formal mathematics

With the rise of automated conjecture generators and theorem provers, one can envision the growth of our mathematical knowledge base in an unsupervised yet verifiable manner through a computable embedding. By decomposing each proof into a detailed graph, and recursively applying this process to the sub-proofs, one can ensure that each link between nodes becomes sufficiently simple for verification by deterministic formal mathematics software. Such methods suggest a potential paradigm shift toward a computer science-oriented approach of formal mathematical research.

Exploring physical theories with computational methods

This approach is particularly significant for physics, not only because it can advance its foundational language — mathematics — but also because it allows the computerization of physics theories in a similar manner. However, unlike mathematical theories, which can expand indefinitely by elaborating upon

their axioms, physics operates under a fundamentally different paradigm: understanding and predicting natural phenomena through empirical validation. Although physics theories have fundamental principles analogous to axioms, from which empirical laws are derived, these derived laws must ultimately predict and align with experimental and observational data.

Significant discrepancies between observed phenomena and theoretical predictions often necessitate theoretical shifts or the formulation of new theories. For instance, one can think of the development of quantum mechanics which addresses observations stemming from the atomic scale, or the introduction of dark matter through the standard Λ Cold Dark Matter (ΛCDM) cosmological model prompted by observations of the Cosmic Microwave Background (CMB) and galaxy rotation curves [Bullock and Boylan-Kolchin, 2017]. This crucial distinction between mathematical theories and physics theories is illustrated in Figure 3.1.

3.1.2 Encoding analytic expressions

As illustrated in Figure 3.1, the development of interpretable analytic physical theories involves a critical step where theories are confronted to observational data : empirical laws. Given the reliance of physical theories on analytic expressions, this subsection addresses the crucial question of how to effectively encode these expressions.

In this subsection, and throughout the rest of this Chapter, we will explore the encoding of analytic expressions — a process drawing upon principles from both *computational symbolic mathematics* (often referred to as computer algebra) [Davenport et al., 1993], and *natural language processing* [Manning and Schutze, 1999].

Tokenization

In the field of natural language processing, tokenization traditionally involves enumerating all possible words in a dataset, typically numbering around $\sim 10,000$, and assigning each a unique categorical label. For instance, with a vocabulary of size $n_{vocab} = 4$, such as $\{\text{beaver}, \text{tree}, \text{galaxy}, \text{planet}\}$, 'beaver' would be encoded as category #1 with a one-hot encoded vector or categorical distribution $e_1 = (1, 0, 0, 0)$, 'tree' as category #2 with $e_2 = (0, 1, 0, 0)$, and so forth.

Recent advancements, however, have refined this approach. Tokens may now represent sub-words, which are identified and encoded using statistical methods to better capture linguistic nuances. For example, with state-of-the-art tokenization techniques [OpenAI, 2023], the word 'tokenization' would be split into two tokens: 'token' and 'ization'. This allows statistical models to learn the functions of base words and their suffixes separately, enhancing the

granularity and effectiveness of language processing[3].

Similarly, in encoding mathematical expressions, the process mirrors that of text. Given a library of mathematical symbols, such as a 6 symbol set $\{+, /, \log, \cos, a, b\}$, each symbol is tokenized into a categorical vector. For example, '+' would be represented as $\tau_1 = e_1 = (1, 0, 0, 0, 0, 0)$, and '/' as $\tau_2 = e_2 = (0, 1, 0, 0, 0, 0)$, allowing statistical models to effectively process and analyze symbolic mathematical data.

Representing expressions with DAGs

Unlike simple text processing, mathematical expressions inherently contain hierarchical information that can also be encoded. Consider the toy example of expression $a + \cos(b)$. Here, '+' functions as a mathematical operation that takes two arguments: 'a' and '$\cos(b)$', with '$\cos(b)$' further decomposed into 'cos' and its argument 'b'. This hierarchical relationship can be structured into a tree, where each node represents a token and its children (or leaves) represent the arguments of the operation. This translation is illustrated between panels (a) and (b) of Figure 3.2.

The concept of arity, or the number of arguments a token can accept, is crucial in defining this hierarchical structure. For instance, in our example, the tokens $\{+, /\}$ are binary and take two arguments each, while $\{\log, \cos\}$ are unary, taking a single argument. The tokens $\{a, b\}$ serve as terminal nodes, representing values that do not take any arguments.

Assuming no multiple references to a single subtree, such a structure can be effectively described as a *Directed Acyclic Graph* (DAG). Furthermore, in simple scenarios where a maximum of two arguments is assumed for any mathematical operation, the structure can be encoded as a so called *binary tree*.

3.1.3 NP-hard graph optimization problems

Many of the challenges discussed earlier, such as automated theorem proving, automated analytic expression generation satisfying a given constraint, and computer program generation, share commonalities as they all involve optimizing directed graph structures to meet specific constraints. While in most cases, it is relatively straightforward and quick to verify if a solution is correct, exploring the vast potential solutions is considerably more difficult[4]. This is

[3] See Tat Dat Duong's tiktokenizer for a live demonstration. (`https://tiktokenizer.vercel.app/`)

[4] A notable exception lies in automated computer program generation, where the complexity and feasibility of solution verification can vary significantly depending on the specific

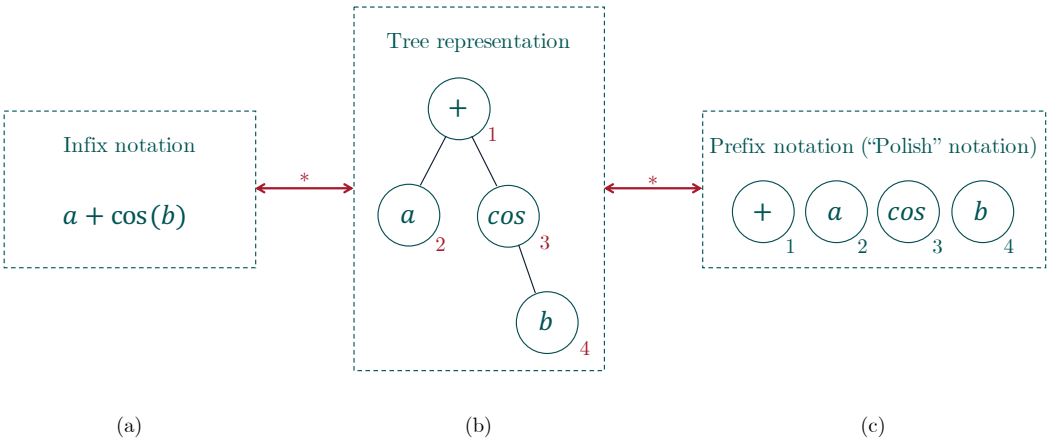


Figure 3.2: **Illustration of various representations of a toy symbolic expression.** Panel (a) displays the familiar infix notation. Panel (b) shows the tree representation, where arguments are positioned as leaf nodes under their respective operators. Panel (c) presents the prefix notation, which can be computed by taking nodes first by depth and then from left to right. Prefix notation eliminates the need for parentheses. In prefix notation, expression validity can be verified by counting the total number of arguments required by all tokens. Terminal tokens may be appended as needed to guarantee the expression's correctness. Here, * indicates a one-to-one equivalence among all three types of representations. For this example, we use the toy expression $a + \cos(b)$.

notably true for automated theorem proving, where solutions can be readily validated using deterministic theorem checkers like `Coq` and `Lean`. Similarly, for analytic expression generation, particularly in SR (a key focus of this thesis) — which focuses on finding expressions that fit data — the verification process is simple, but the search space is immense.

SR is NP-hard

To grasp the complexity of SR, it is instructive to consider the challenges this problem presents when approached naively. Imagine attempting to generate trial analytic expressions with a length of 35 symbols, selecting from 15 different possible variables or operations (e.g., , x, $+$, $-$, $\times$, $/$, sin, log, ...), as we will do in subsequent Chapters. A brute-force attempt to match these expressions to a dataset would require evaluating up to $15^{35} \approx 1.5 \times 10^{41}$ trial solutions which is obviously vastly beyond our computational means to test against the data at the present day or at any time in the foreseeable future. Although it has long been suspected, it was recently formally proven by Virgolin and Pissis [2022] that SR, like many similar problems such as automated theorem prov-

problem.

ing, falls into the category of "NP-hard" (nondeterministic polynomial time) problems. This categorization does not even take into account the additional complexity introduced by optimizing free constants within the expressions.

NP-hardness

This classification means that the problem cannot be solved within a polynomial time, though solutions can be verified within such a polynomial time. This is exemplified in the well-known *Traveling Salesman Problem*, which involves finding the shortest route that visits a list of cities and returns to the origin. This problem is a classic example of an NP-hard challenge in combinatorial optimization [Gavish and Graves, 1978]. Notably addressed by some researchers using reinforcement learning to develop models capable of generating tours by sequentially selecting cities Stohy et al. [2021].

Cross-polination

The techniques developed to address this family of problems — NP-hard directed graph combinatorial challenges — offer substantial opportunities for cross-pollination, presenting a range of intriguing possibilities. In this Chapter, we will focus specifically on reinforcement learning, a strategy widely employed across various problems within this category, including automated theorem proving, as highlighted by studies such as Lample et al. [2022]. Further discussions (in sub-section 3.3.3) will explore how SR can leverage methodologies from other areas that tackle similar combinatorial challenges.

3.2 Computational Symbolic Mathematics

This section explores the computational approaches used to manipulate symbolic mathematics effectively. We begin by showing how one can sample mathematical symbols from a model in the form of tokens, as detailed in sub-section 3.2.1. We then examine how expressions can be sequentially sampled and their validity and termination ensured by employing the so-called *prefix* notation, in sub-section 3.2.2. Lastly, we explore how this method of sequential sampling facilitates the formulation of deterministic priors that can impose constraints on the arrangement of symbols, as presented in sub-section 3.2.3.

3.2.1 Sampling tokens

Distribution function across the library

Given a library of possible tokens representing n_{lib} mathematical symbols $\{\tau_i\}_{i \leq n_{\text{lib}}}$, e.g., $\{a, b, +, -, \times, /, \cos, \exp, \log, \square^2\}$, one can use a statistical model or a neural network (as we will describe in the Section 3.3) to generate a categorical distribution across the space of mathematical symbols in the library $\{p(\tau_i)\}_{i \leq n_{\text{lib}}}$.

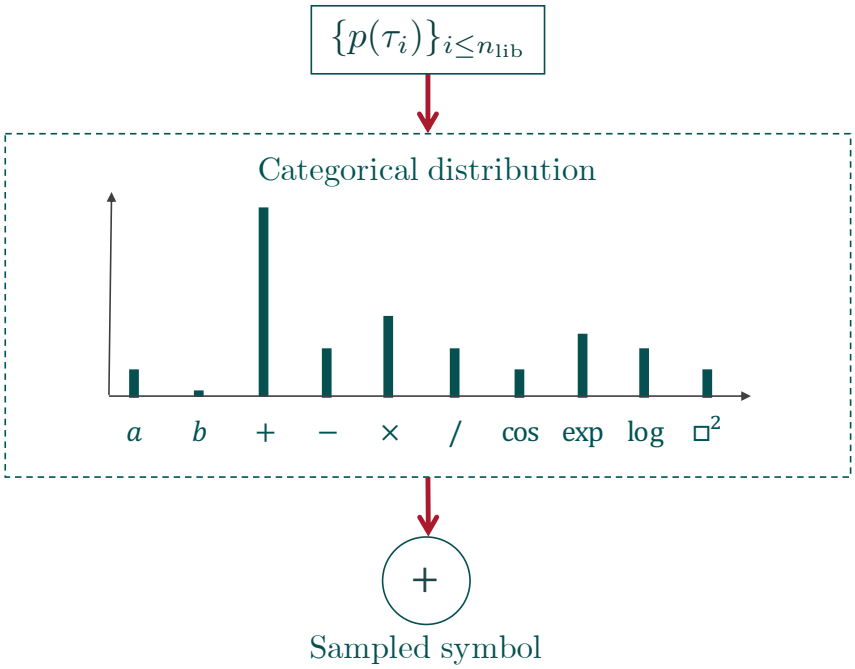


Figure 3.3: **Token sampling illustration.** Illustration of the process by which a token representing a mathematical symbol is sampled from a distribution generated by a model covering the space of tokens available in the library. For this example, we use the library of possible symbols $\{\tau_i\}_{i \leq 10} = \{a, b, +, -, \times, /, \cos, \exp, \log, \square^2\}$ from which $\tau_3 = e_3 = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$ representing '+' is sampled.

Softmax

It is useful to apply a softmax function to the output vector $\{z_i\}_{i \leq n_{\text{lib}}}$ of the model, transforming it into a probabilistic distribution. This transformation is mathematically expressed as:

$$p(\tau_i) = \frac{e^{z_i}}{\sum_{j=1}^{n_{\text{lib}}} e^{z_j}} \tag{3.1}$$

From this distribution, one can then sample to select a specific token, τ, effectively choosing a mathematical symbol, as illustrated in Figure 3.3.

3.2.2 Sampling symbolic expressions

Prefix notation

As explored in sub-section 3.1.2, symbolic expressions can be effectively encoded as DAGs. By processing each node first in depth and then from left to right, it is possible to transform these graphs into a one-dimensional vector, or prefix notation, $\left(\tau^{\langle t \rangle}\right)_{t \leq n_{\text{expr}}}$, where $\tau^{\langle t \rangle}$ denotes token at position t in the expression and n_{expr} represents the size of the expression. In this format, operators precede their corresponding operands, thereby eliminating the need for parentheses. This notation, often referred to as "Polish" notation, can seamlessly be converted back to a tree representation or to the more familiar "infix" notation, given their one-to-one correspondence. For instance, the toy expression $a + \cos(b)$ would be represented in prefix notation as $(+, a, \cos, b)$. This transcription is depicted in Figure 3.2 between panels (b) and (c).

As demonstrated in sub-section 3.1.2, any token can be encoded as a categorical vector. Thus, employing prefix notation allows any analytic expression to be rendered as a sequential array of categorical vectors. This one-to-one relationship between the categorical representations and the expressions ensures that any analytic expression can be directly converted into numerical form and vice versa. This process is further illustrated in Figure 3.4.

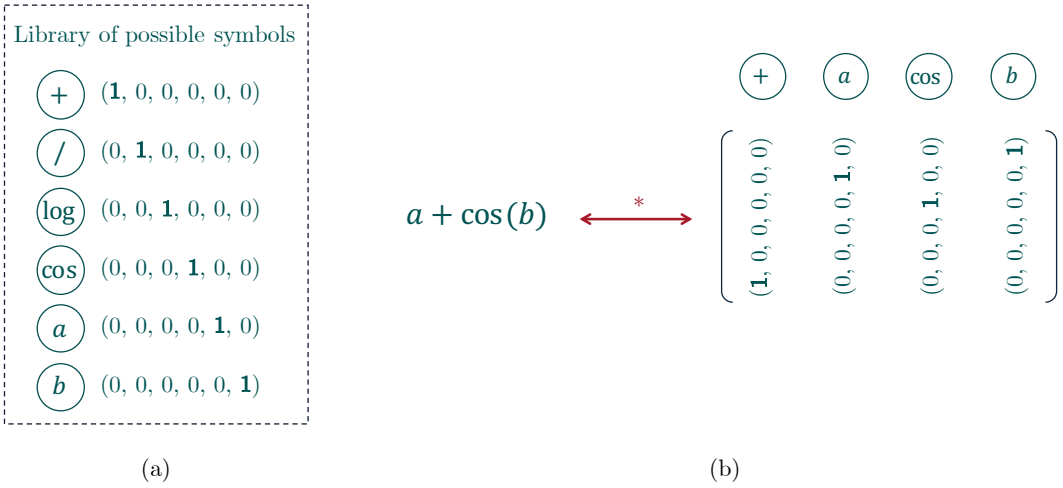
Given a library of available tokens $\{+, /, \log, \cos, a, b\}$ our toy expression $a + \cos(b)$, that is $(+, a, \cos, b)$ in prefix notation, can be encoded as:

$$\left(\tau^{\langle t \rangle}\right)_{t \leq 4} = (\tau^{\langle 1 \rangle}, \tau^{\langle 2 \rangle}, \tau^{\langle 3 \rangle}, \tau^{\langle 4 \rangle}) = (\tau_1, \tau_5, \tau_4, \tau_6) = (e_1, e_5, e_4, e_6) \tag{3.2}$$

As for simplicity, we consider a token and its one hot encoding across the library as equivalent e.g., in this example the token encoded at position $t = 3$ in the sequence ('cos') can be noted as $\tau^{\langle 3 \rangle} = \tau_4$, as it represents the 4-th token in the library, and with a one hot encoding $\tau^{\langle 3 \rangle} = e_4 = (0, 0, 0, 1, 0)$.

Expression validity & termination condition

The use of prefix notation eliminates the need for parentheses, enabling continuous sampling of new tokens from a model to extend the sequence representing



(a) (b)

Figure 3.4: **Illustration of how an analytic expression can be encoded into numerical form.** This diagram illustrates how analytic expressions can systematically be represented as mere numerical arrays rendering them machine-readable by encoding mathematical symbols with categorical distributions and arranging them using prefix notation. For this example, we use the toy expression $a+\cos(b)$ with a library $\{+, /, \log, \cos, a, b\}$. This expression is encoded as $(+, a, \cos, b)$ i.e. $\left(\tau^{\langle t \rangle}\right)_{t \leq 4} = (\tau^{\langle 1 \rangle}, \tau^{\langle 2 \rangle}, \tau^{\langle 3 \rangle}, \tau^{\langle 4 \rangle}) = (\tau_1, \tau_5, \tau_4, \tau_6) = (e_1, e_5, e_4, e_6)$.

an analytic expression. A key advantage of prefix notation is its inherent flexibility. Regardless of the mathematical symbols added, the expression maintains the potential for validity. As tokens are added, the corresponding tree or graph representation expands, though some leaf nodes may initially lack arguments. The count of these dangling, or unconnected, nodes can be determined by summing the arities of each token within the expression:

$$n_{\text{dangling}} = n_{\text{expr}} - \sum_{t=1}^{n_{\text{expr}}} \text{arity}(\tau^{\langle t \rangle}) \tag{3.3}$$

This principle is illustrated using our toy expression in Figure 3.5. An expression is valid and its associated tree is complete as long as there remains exactly 1 dangling node:

$$\text{Expression is valid} \iff n_{\text{dangling}} = 1 \tag{3.4}$$

A count greater than one ($n_{\text{dangling}} > 1$) indicates an excess of arguments, while a count less than one signifies insufficient arguments. In practice, when sampling expressions, any tokens drawn after achieving $n_{\text{dangling}} = 1$ can be disregarded. Furthermore, the validity of any expression can be ensured by appending terminal nodes-tokens that represent values and require no arguments.

To prevent indefinite sampling without achieving $n_{\text{dangling}} = 1$, the model's output probabilities can be adjusted to favor the selection of terminal nodes once a certain length of the expression is reached. This adaptive probability tuning leverages the sequential generation of tokens, enabling the formulation of deterministic priors that significantly enhance the sampling process, as we will explore in the next sub-section.

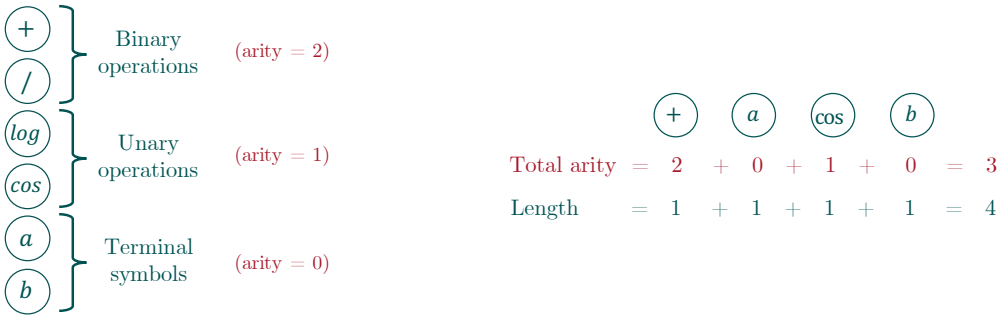


Figure 3.5: **Illustration of expression validity assessment.** An expression is deemed valid if and only if the difference between its length and the sum of the arity of its components equals 1. For this example, we use the toy expression $a + \cos(b)$ with a library $\{+, /, \log, \cos, a, b\}$.

3.2.3 Incorporating priors through sequential sampling

In situ priors

It is crucial to highlight that the categorical distribution generated during the sampling process can be deterministically tuned to integrate prior knowledge directly (*in-situ*) as expressions are being sampled. For instance, probabilities of certain tokens can be selectively nullified based on the contextual information encoded within the developing expression tree, thereby significantly narrowing the search space [Petersen et al., 2021a,b]. This process is illustrated in Figure 3.6.

The ability to formulate such effective priors stems from the sequential nature of the sampling process. This approach is particularly advantageous compared to traditional methods like genetic algorithms used in symbolic expression optimization. While mutations in genetic algorithms can somewhat adhere to predetermined rules, the crossover process often complicates compliance, potentially necessitating numerous trials to identify successful crossovers that respect these deterministic rules. The fundamental distinction here lies in our method's reliance on a tunable model actively generating new expressions, as opposed to genetic programming approaches that manipulate already existing expressions stemming from previous iterations.

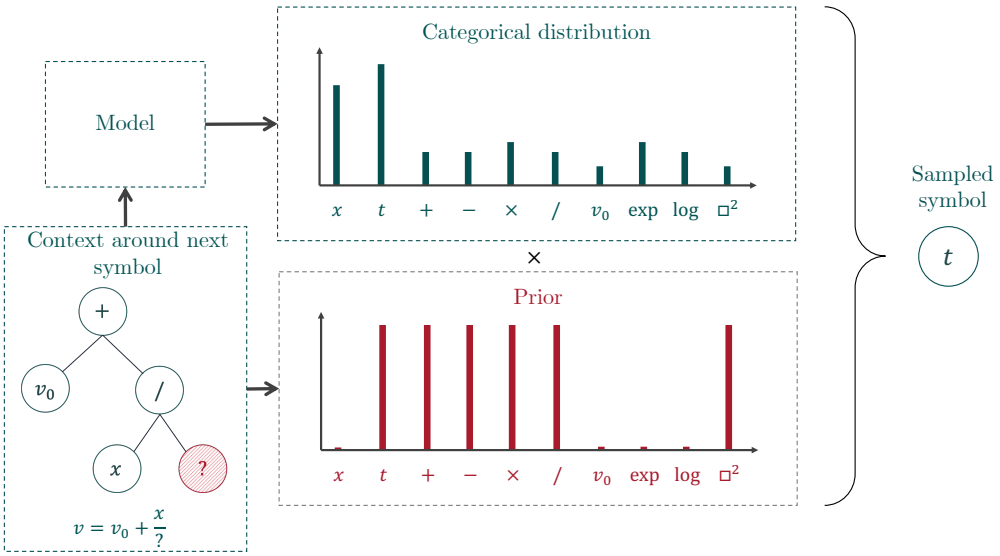


Figure 3.6: **Incorporation of prior knowledge in symbolic arrangement.** This illustration shows how prior knowledge about the available tokens can tune the distribution emitted by a model to facilitate the selection of an appropriate token. In the example shown, an expression is being generated where a velocity v_0 is summed to a length x divided by a node currently being selected, denoted as '?'. Given the context, this node must represent time t or a more complex expression that eventually equates to a time, stemming from a parent node such as $\{+, -, \times, \square^2\}$, but it cannot be another length x, a velocity v_0, or dimensionless operators like $\{\exp, \log\}$. Utilizing this prior information alongside the model's output helps guide the selection of the most suitable token.

Substantial literature has explored the development of techniques that incorporate prior knowledge into symbolic optimization. Notably, some studies have formulated priors based on typical symbolic arrangements found in academic publications [Guimerà et al., 2020, Bartlett et al., 2023b] or even from resources like the *Wikipedia* encyclopedia [Kim et al., 2021].

In practical applications, we adjust the probability distribution emitted by the model using the following equation:

$$p(\tau_i) = \frac{e^{z_i + \log(\text{prior}_i)}}{\sum_{j=1}^{n_{\text{lib}}} e^{z_j + \log(\text{prior}_j)}} \tag{3.5}$$

Here, prior_i denotes the prior probability of selecting the token τ_i from a library of n_{lib} possible tokens. Below, we will elaborate on the specific priors employed within the context of this thesis.

The majority of these priors are dependent on specific hyper-parameters, for which we provide the values utilized in our studies. These values are used

consistently throughout the work presented in this thesis, unless specified otherwise. However, it is important to note that these parameters are adjustable within our implementation, allowing for flexibility and adaptation to different scenarios.

Length priors

We implement priors designed to maintain reasonable expression sizes, which is particularly critical when expressions are sampled from a learnable model, such as a neural network as in the initial learning phases, the distribution emitted by the model may resemble a near-random pattern. Thus leading to a scenario where a significant fraction of trial expressions might extend indefinitely without satisfying the expression termination condition.

To mitigate this, we constrain expression sizes to a maximum of $\leq N$ with $N = 35$ tokens. This is achieved by preventively zeroing out the probability of selecting non-terminal tokens, ensuring that the length of the expression remains within this limit throughout the sampling process:

$$n_{\text{expr}} + n_{\text{dangling}} \leq N \tag{3.6}$$

Moreover, to promote brevity within expressions, we apply a soft length prior modeled as a Gaussian distribution with a variance of $\sigma^2 = 5$ centered around a length of 8. This approach subtly biases the selection process towards more "concise" expressions, aligning with our objective of generating intelligible expressions.

Other priors

We also implement a set of more specific priors, designed to preclude atypical symbolic arrangements that are rarely seen in physics. These priors' relevance has been empirically validated through assessments of our methods on physical test cases.

One such prior restricts expressions to no more than two levels of nested trigonometric functions. For example, expressions like $\cos(f \cdot t + \sin(x/x_0 + \tan(\square)))$ are prohibited, whereas $\cos(f \cdot t + \sin(x/x_0))$ remains permissible. Additionally, we prevent the self-nesting of exponential and logarithmic functions, such as forbidding expressions like $e^{e^{\square}}$.

Furthermore, we eliminate redundant inverse unary operations, such as $e^{\log \square}$, which unnecessarily consume symbolic space and learning resources that could be more effectively allocated towards exploring meaningful symbolic structures.

Summary & conflicts

Although not utilized in the SR experiments presented in this work, we have implemented additional priors, which will be described in detail in sub-section 6.1.2. Beyond these relatively straightforward priors that leverage local tree structure, one can formulate much more sophisticated priors, for example based on dimensional analysis. This is the subject of Chapter 4.

It is important to note that the combination of different priors can sometimes result in conflicts. For example, the physical units constraints prior detailed in Chapter 4 may require a certain number of additional tokens to comply with the rules of dimensional analysis, potentially conflicting with the length prior that would enforce premature termination of the expression. In such cases, the conflicting candidate is discarded.

3.3 Learning Analytic Models

In this Section, we explore how neural networks can be utilized to automatically generate symbolic expressions that conform to a set of constraints. While these constraints can be varied, our focus here is on producing analytic functions that fit given data i.e. *symbolic regression*. We will therefore illustrate these techniques specifically in the context of this problem.

Considering the success of deep reinforcement learning methods in accurately recovering not only accurate but exact symbolic expressions (as indicated in Section 2.2.2 and shown in Figure 4.3), which is particularly important in the field of physics where precise physical law recovery is crucial we have chosen to incorporate this methodology into the machine learning component of our SR approach.

In sub-section 3.3.1, we describe how we generate analytic expressions using a neural network, in sub-section 3.3.2 we show how we train our neural network to produce accurate expressions by trial and error using deep reinforcement learning and finally in sub-section 3.3.3 we discuss this approach, its limitations and potential improvements. Results and applications of our method are presented later on in Chapters 4 and 5.

3.3.1 Generating symbolic expressions with a neural network

As in previous SR studies employing a neural network [e.g., Petersen et al., 2021a, Landajuela et al., 2021a, 2022, Kamienny et al., 2022, Lalande et al., 2023, Biggio et al., 2020, 2021, Vastl et al., 2022], and as detailed in earlier sections, treating mathematical expressions as sequences allows us to utilize techniques similar to those used in natural language processing. This approach is particularly relevant during the sampling process described in Section 3.2.

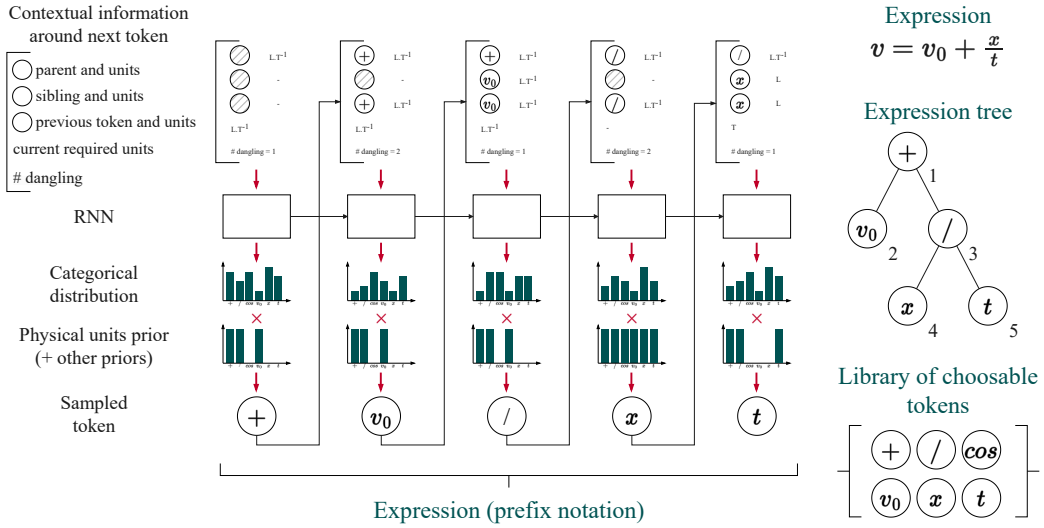


Figure 3.7: **Sampling an expression from a recurrent neural network.** The process starts at the top left RNN block. For each token, the RNN is given the contextual information regarding the surroundings of the next token to generate, namely: the parent, sibling and previously sampled token along with their physical units, the required units for the token to be generated and the dangling number (i.e. the minimum number of tokens needed to obtain a valid expression). Although the integration of dimensional analysis into analytic expression generation is illustrated here, it is further explored in Chapter 4. Based on this information, the RNN produces a categorical distribution over the library of available tokens (top histograms) as well as a state which is transmitted to the RNN on its next call. The generated distribution is then masked based on local units constraints (bottom histograms), forbidding tokens that would lead to nonsensical expressions. The resulting token is sampled from this distribution, leading to the token '+' in this example. Repeating this process, from left to right, allows one to generate a complete physical expression, here $(+, v_0, /, x, t)$ which translates into $v_0 + x/t$ in the infix notation we are more familiar with.

Recurrent neural networks

Token sequences are generated by using an RNN, which in essence, is a neural network that can be invoked multiple times to create a logical chain of similar

operations. At each invocation $t < N$ (N representing the maximum number of steps), the RNN generates a time-dependent output and a corresponding memory state $S^{\langle t \rangle}$. The RNN takes as input some time dependent observations[5] $O^{\langle t \rangle}$ as well as the state of the previous call $S^{\langle t-1 \rangle}$. In practice, we use the RNN to generate a categorical probability distribution over the library of available tokens, which we then simply sample to draw a definite token. Once a token is generated, we feed the minimum number of tokens still needed to obtain a valid analytic expression (i.e. the number of dangling nodes n_{dangling}), the token's properties and the properties of its surroundings in graph representation as observations for the next RNN call. Namely, we give the nature of the token which was sampled at the previous step $\tau^{\langle t-1 \rangle}$ (since the RNN does not have access to this information[6] which is derived from a stochastic process), the sibling (if any at this step) $\tau^{\langle s \rangle}$ and parent $\tau^{\langle p \rangle}$ tokens of the token to be generated in a tree representation (s and p here respectively referring to the position of the sibling and parent tokens in prefix notation). That is:

$$O^{\langle t \rangle} = \{n_{\text{dangling}}, \tau^{\langle t-1 \rangle}, \tau^{\langle s \rangle}, \tau^{\langle p \rangle}\} \tag{3.7}$$

This allows the inner mechanisms of the neural network to take into account the local structure of the expression for generating the next token. The process described above can be repeated multiple times until a whole token function is generated in prefix notation, as illustrated in Figure 3.7.

It is important to note that while expressions are treated sequentially, their intrinsic graph structure is utilized in multiple aspects of the model's operation. This utilization includes the integration of priors that depend on the graph structure, informing the RNN about local graph structure directly through observations, and through the RNN's training approach. Specifically, the RNN interacts with its environment via a trial-and-error process during training, which fundamentally depends on an evaluation process involving fit quality assessment, a process that is directly representative of the graph structure[7].

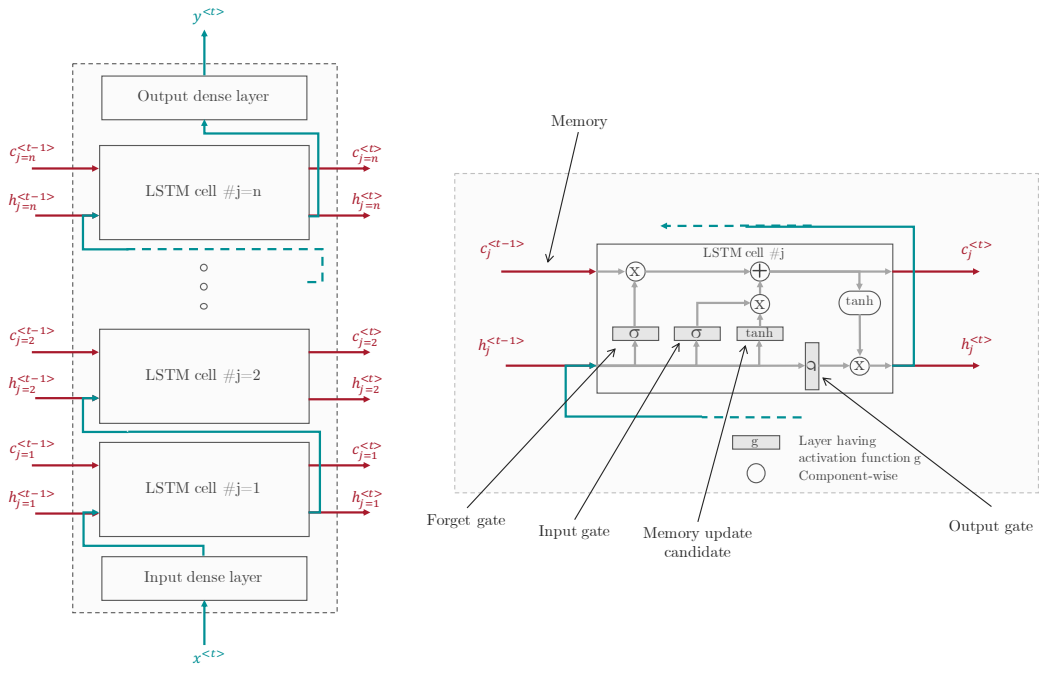
[5] We refer to "observations" in the context of RL, here pertaining to contextual analytic information related to the expression being generated, rather than to the scientific data being fitted in the context of symbolic regression.

[6] Not providing this information typically hinders performance.

[7] Although this indirect interaction with graph structure through the evaluation of functional forms that depends on symbolic combinations expressed by the model might seem intuitive, it marks a significant departure from the conventional approaches used in traditional supervised learning approaches where datasets are plainly mapped to corresponding sequences of tokens.

Long-short term memory

In our RNN configuration, we employ a stack of Long-Short Term Memory (LSTM) cells [Hochreiter and Schmidhuber, 1997], with dense layers positioned both before and after the LSTM layers. This architecture is designed to effectively produce and analyze sequences of vectors. Each LSTM cell is called upon multiple times during the sequence processing, allowing it to maintain and transmit a state in latent space between calls, denoted as $S^{\langle t \rangle}$. Specifically, our LSTM setup is tasked with analyzing a sequence of input vectors $x^{\langle t \rangle}$, corresponding to the time-dependent observations of the symbolic graph, $x^{\langle t \rangle} = O^{\langle t \rangle}$, as previously described, and generating a sequence of output vectors $y^{\langle t \rangle}$ — referred to as 'actions' in the context of RL. These output vectors represent a probability distribution over the library of tokens available for selection, $y^{\langle t \rangle} = \{\{z_i\}_{i \leq n_{\text{lib}}}\}^{\langle t \rangle}$. From this distribution, a specific token $\tau^{\langle t \rangle}$ is then sampled. The architecture and its operational details are illustrated in Figure 3.8.



(a) RNN cell (b) LSTM cell

Figure 3.8: **RNN and LSTM diagrams.** Panel (a) illustrates an RNN framework composed of stacked LSTM cells. Panel (b) details the information flow within a single LSTM cell. Refer to the descriptions provided in 3.3.1 for details.

What makes the LSTM cell particularly effective is its ability to maintain

both a long-term memory, $c^{\langle t \rangle}$, and a short-term memory, $h^{\langle t \rangle}$, which directly influences the output. Thus, the overall state of the LSTM at any time step t is given by $S^{\langle t \rangle} = \{c^{\langle t \rangle}, h^{\langle t \rangle}\}$. The LSTM cell consists of multiple dense layers functioning as trainable logic gates. Upon receiving an input, the LSTM employs a forget gate to selectively remove information from the long-term memory. It then uses the input gate to integrate new information into the long-term memory. Finally, based on the current input and the updated long-term memory, the LSTM generates an output.

Multiple LSTM layers are stacked within our model to achieve greater depth and enhance the model's capability to capture more subtle and complex patterns. The process is orchestrated as follows: the observation $x^{\langle t \rangle}$ is initially processed through a dense layer, which prepares it for input into the LSTM sequence. The prepared input is first passed to the initial LSTM layer, which utilizes its preceding state $S_{j=1}^{\langle t-1 \rangle} = \{c_{j=1}^{\langle t-1 \rangle}, h_{j=1}^{\langle t-1 \rangle}\}$ to generate an output. This output then serves as the input to the subsequent LSTM layer, which similarly accesses its prior state $S_{j=2}^{\langle t-1 \rangle} = \{c_{j=2}^{\langle t-1 \rangle}, h_{j=2}^{\langle t-1 \rangle}\}$, and so on, through the stack until the final LSTM layer. The output from the last LSTM layer is then conveyed to an output dense layer, which synthesizes the final result $y^{\langle t \rangle}$.

About transformers

Our embedding framework is compatible with multiple neural network architectures, including the transformers architecture, the current state-of-the-art in sequence processing [Vaswani et al., 2017]. Although we are in the process of transitioning to this advanced architecture, our current system utilizes the traditional LSTM approach. We will therefore only provide a brief overview of transformers here.

Transformers operate on a different principle compared to recurrent neural networks (RNNs). Instead of sequentially processing elements and maintaining a state across iterations, transformers employ a mechanism where queries are emitted for all elements of a sequence simultaneously. These queries represent specific information needs relevant to the task at hand. In response, the model generates keys based on these queries, facilitating cross-communication across the entire sequence. This method allows for the detection and utilization of subtle and distant patterns within the data.

Given the operational mechanics of transformers, they depend heavily on a so-called 'positional encoding' — typically a position dependent function that is applied to inputs before they are fed to the network. This technique is critical for preserving the sequence order within their representations. While the graph structure inherent in the observations $O^{\langle t \rangle}$ is not directly maintained through this form of embedding, it is possible to learn this structure empirically. Moreover, it is conceivable to develop an encoding process that explicitly

incorporates graph structure, potentially enhancing the model's ability to recognize and utilize complex relational information.

Transformers' relevance to SR, was previously demonstrated in works such as : Kamienny et al. 2022, Lalande et al. 2023, Biggio et al. 2020, 2021, Vastl et al. 2022 which employ transformers within a supervised learning framework.

3.3.2 Learning

One might imagine that SR problems could be solved by directly optimizing the choice of symbols to fit the problem, using the auto-differentiation capabilities of modern machine learning frameworks[8] in an unsupervised manner. Unfortunately this approach cannot be used for general SR footnoteWe will moderate that point in Section 7.2 which explores such approaches. because the cost function is non differentiable (the choice of selecting say the sin function over log is not differentiable with respect to the data), which prevents one from using gradient descent. A practical solution is to use a neural network as a "middle man" to generate a categorical distribution from which we can sample symbols. One can then optimize the parameters of this neural network whose task is to generate these symbols according to fit quality and physical units constraints.

Reinforcement learning

The training of the network that generates the distribution of symbols relies on the "reinforcement learning" (RL) strategy [Sutton and Barto, 2018], which is a common method used to train artificial intelligence agents to navigate virtual worlds such as video games[9], or master open-ended tasks [Bauer et al., 2023b]. In the present context, the idea is to generate a set (usually called a "batch" in machine learning) of trial symbolic functions, and compute a scalar reward for each function by confronting it to the data. We can then require the neural network to generate a new batch of trial functions, encouraging it to produce better results by reinforcing behavior associated with high reward values, approximating gradients and applying them to a so-called "policy" here our neural network model. The hope is that, by trial and error, the learnable parameters of the network will converge to values that are able to generate a symbolic function that fits the data well.

Following the insight by Petersen et al. [2021a], we adopt the risk-seeking policy gradient strategy along with the entropy regularization scheme found

[8]Most machine learning tasks use the differentiability of the implemented model with respect to the data to implement a (stochastic) gradient descent towards an optimal model solution that fits the data best.

[9]See, e.g., `https://youtu.be/igZ6IPQimjQ`

by Landajuela et al. [2021a]. In essence, we only reinforce on the best 5 % of candidate solutions, not adjusting the neural network based on the 95 % of other candidates, therefore maximizing the reward of the few best performing candidates rather than the average reward. With our chosen batch size of 10k, this strategy reinforces the leading 500 candidates. This enables an efficient exploration of the search space at the expense of average performance, which is of particular interest in SR as we are often mostly concerned in finding the very best candidates in particular if the goal is exact symbolic recovery and do not care if the neural network performs well on average. This is contrary to many other applications of RL (e.g., robotic automation) which can even sometimes require risk-adverse gradient policies (e.g., self driving cars) [Rajeswaran et al., 2017]. This novel risk-seeking policy, inspired by Rajeswaran et al. [2017] and first proposed by Petersen et al. [2021a], has significantly boosted performance in SR.

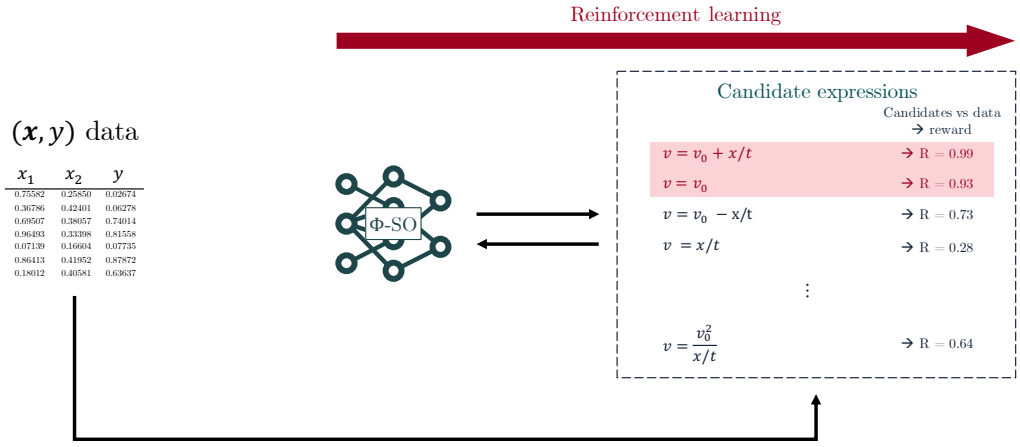


Figure 3.9: **Reinforcement learning-based symbolic regression framework.** The data itself is not directly utilized, instead, it informs the computation of a reward that reflects the quality of fit. This reward creates an environment in which the neural network policy operates, learning to maximize the fit quality over iterations. This framework employs a risk-seeking policy, where only the top-performing expression candidates (highlighted in red) are used to reinforce and train the neural network.

Annealing temperature parameter

Among various enhancements presented throughout this thesis, we extend the framework of Petersen et al. [2021a] by introducing an adjustable temperature parameter, θ_T. This parameter modulates the distribution over the space of tokens derived from the model's output vector, $\{z_i\}_{i \leq n_{\text{lib}}}$, and is trainable alongside other model parameters during the learning process. The probability

of selecting a token τ_i from a library of n_{lib} possible tokens is computed as follows:

$$p(\tau_i) = \frac{e^{z_i + \log \theta_T}}{\sum_{j=1}^{n_{\text{lib}}} e^{z_j + \log \theta_T}} \tag{3.8}$$

A higher value of θ_T increases the exploration by flattening the probability distribution, making the model more likely to sample less probable tokens. Conversely, a lower value of θ_T sharpens the distribution, prioritizing exploitation by focusing on the highest-probability tokens [Sutton and Barto, 2018].

By making the temperature parameter θ_T adjustable, the sharpness of the probability distribution can be dynamically tuned during training. This so-called "annealing"[10] approach [Kirkpatrick et al., 1983], is particularly advantageous in reinforcement learning contexts. Specifically:

1. The level of exploration can be empirically learned based on the task requirements, allowing the model to autonomously balance exploration and exploitation depending on the SR task.

2. During the initial stages of training, when the agent is learning the structure of the environment, higher exploration can lead to the discovery of novel or effective solutions. As the model improves, it can automatically reduce exploration and concentrate on refining the most promising candidates.

Evaluating expressions

We allow the candidate functions f to contain "constants" with fixed physical units specified by the user, but with free numerical values. These free constants allow us the possibility to model situations where the problem has some unknown physical scales. A (somewhat contrived) example from galactic dynamics could be if we were provided a set of potential values Φ, and cylindrical coordinate values (R, z) of some mystery function that was actually a simple logarithmic potential model:

$$\Phi = \frac{1}{2} v_0^2 \ln \left(R_c^2 + R^2 + \frac{z^2}{q^2} \right), \tag{3.9}$$

whose parameters are the velocity parameter v_0, the core radius R_c and the potential flattening q. Of course, we will generally not know in advance either

[10] The term annealing originates from metallurgy, where materials are gradually cooled to achieve a stable structure. In machine learning, annealing refers to the gradual adjustment of a temperature parameter, which controls the randomness (or entropy) of an algorithm's decisions.

the number of such parameters that the correct solution requires, or their numerical values. Yet to be able to evaluate the loss of the trial functions f, we need to assign values to all such free "constants" they may contain. We accomplish this task by processing each trial function, with the L-BFGS [Zhu et al., 1997] optimization routine in `PyTorch` [Paszke et al., 2019] (optimizing over 20 steps and using a mean squared error metric), leveraging the fact that we can encode the symbols of f using `PyTorch` functions. Since `PyTorch` has in-built auto-differentiation, finding the optimal value of the constants via gradient descent is extremely efficient.

Then, for each candidate f, we compute a reward r that is representative of fit quality:

$$r = \frac{1}{1 + \text{NRMSE}} \tag{3.10}$$

Where NRMSE is the root mean squared error normalized by the deviation of the target (σ_y) :

$$\text{NRMSE} = \frac{1}{\sigma_y}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2} \tag{3.11}$$

R^2 and reward r

To assess fit quality, a commonly used metric is the coefficient of determination, R^2, defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.12}$$

Where $\bar{y}$ is the mean of the target values. The relationship between R^2 and the reward r is given by:

$$R^2 = \frac{2}{r} - \left(\frac{1}{r}\right)^2, \tag{3.13}$$

Evaluating policy gradients

Having computed the rewards, we next approximate gradients to optimize the parameters of our neural network, θ. First, we use these rewards to identify the top 5 % of candidates -those with rewards $r > r_\epsilon$, where r_ϵ represents the ϵ-th quantile and $\epsilon = 5\%$. These top-performing candidates are then used as targets [Petersen et al., 2021a] to guide the neural network towards generating similar expressions, effectively treating the problem like a classification task at the token level.

$$\mathcal{L}_\theta = \mathcal{L}_{\theta,\text{risk}} + \alpha.\mathcal{L}_{\theta,\text{entropy}} \tag{3.14}$$

Risk-seeking loss:

The primary component of our loss function, $\mathcal{L}_{\theta,\text{risk}}$, is modeled after a cross-entropy loss that is modulated by the rewards of the candidates. This configuration ensures that higher-quality candidates exert a stronger influence on the model than those with lower rewards. Specifically, for a given candidate expression encoded as $\left(\tau^{\langle t\rangle}\right)_{t\leq n_{\text{expr}}}$ with an associated reward r, we define the main component of the loss as:

$$\mathcal{L}_{\theta,\text{risk}} = -(r - r_\epsilon) \sum_{t=1}^{n_{\text{expr}}} \sum_{i=1}^{n_{\text{lib}}} e_{\tau^{\langle t\rangle},i}. \log p_\theta^{\langle t\rangle}(\tau_i) \tag{3.15}$$

Here, $p_\theta^{\langle t\rangle}(\tau_i)$ represents the probability assigned by our model for selecting the i-th token from the library at the t-th position in the sequence. The term $e_{\tau^{\langle t\rangle}}$ denotes the token at position t from the target expression that received the reward r, encoded as a one-hot categorical vector — this vector assigns a probability of one to the actual token position in the library and zero elsewhere. The notation used her corresponds to the one outlined in 3.2.2.

Entropy loss:

The secondary component of our loss function, $\mathcal{L}_{\theta,\text{entropy}}$, implements an entropy regularization scheme [Landajuela et al., 2021a]. This component is designed to enhance the diversity of the generated symbolic expressions by incorporating a decay factor, γ, applied along the sequence dimension to variably weight the importance of nodes across the expression. This decay prioritizes root nodes to discourage the model from converging too quickly around the same initial nodes and encourages exploration of varied starting points:

$$\mathcal{L}_{\theta,\text{entropy}} = -\sum_{t=1}^{n_{\text{expr}}} \gamma^t \sum_{i=1}^{n_{\text{lib}}} p_\theta^{\langle t\rangle}(\tau_i). \log p_\theta^{\langle t\rangle}(\tau_i) \tag{3.16}$$

Overview:

In practice, we optimize the neural network by calculating the loss across the top 500 candidates and applying gradients derived from this loss using an Adam optimizer [Kingma and Ba, 2015]. Essentially, this process involves

Learning parameters	
Batch size	10 000
Learning rate	0.0025
Risk factor (ϵ)	5 %
Entropy coefficient (α)	0.005
Entropy decay (γ)	0.7
Annealing parameter ($\log \theta_T$) initial value	1.54

Table 3.1: Reinforcement learning hyper-parameters

comparing the neural network's outputs with those of the 500 leading candidates and minimizing the discrepancies. Thus, the neural network does not directly access the data; rather, it 'learns' indirectly through trial and error, being adjusted based on the rewards associated with the leading candidates. These rewards are themselves derived from the data through a fit quality assessment. This learning process is depicted on Figure 3.9. The values for hyper-parameters pertaining to RL used throughout this thesis, unless otherwise specified, are listed in Table 3.1.

For a live demonstration of our system applied to a symbolic regression task, refer to [SR demo][11]. In this example, the system aims to derive a model fitting data points corresponding to a damped harmonic oscillator. The video illustrates the iterative process, displaying the curves associated with trial candidate expressions over successive iterations, highlighting their progressive improvement in fit quality until convergence.

Non-differentiable reward:

It is worth noting that in the RL framework, the the reward function can be considered as as a black box, which does not have to be differentiable, therefore one could use anything as the reward[12]. For example, we can also include the complexity of the symbolic function in the reward function, so as to have a criterion akin to Occam's razor. But actually one could in principle implement many ideas into the reward function: symmetries, constraints on primitives or derivatives, fitness in a differential equation, the results of some symbolic computation using external packages such as `Mathematica` [Wolfram, 2003] or `SymPy` [Meurer et al., 2017], behavior of the function when implemented an n-body simulation, and so on. Note that in the context of this work,

[11] https://youtu.be/wubzZMkoTUY

[12] As long as a positive correlation between symbolic arrangement and the reward metric exists.

although there are more sophisticated schemes to define complexity (see e.g., Vladislavleva et al. 2009) we simply define it as length i.e. the number of tokens appearing in the expression excluding parentheses or the number of nodes in a tree representation.

3.3.3 Discussion

Incorporating prior knowledge in RL-based SR

Our approach is based on a deep reinforcement learning methodology, where the neural network is reinitialized at the start of each SR task. It is therefore trained independently for each specific problem, and so does not benefit from past experience nor is it pre-trained on a dataset of well known physical functional forms. One could argue that this makes our approach in principle "unbiased" akin to unsupervised learning setups and therefore well suited for discovering new physics [Karagiorgi et al., 2022]. However, this also intrinsically limits SR capabilities as exploiting such prior knowledge is of great value for resolving the curse of accuracy guided SR described below. One can exploit such prior knowledge by formulating it as an *in situ* prior [Kim et al., 2021, Guimerà et al., 2020] or by learning on it in a supervised manner using transformers learning techniques [Kamienny et al., 2022, 2023, Bendinelli et al., 2023, Biggio et al., 2021, Vastl et al., 2022]. However, although state-of-the-art supervised SR methods, as of now, shine in providing accurate approximations, they show poorer exact symbolic expression recovery rates than other methods (see e.g., the performances of `NeSymReS` in Figure 4.3 or the ablation study conducted by Landajuela et al. [2022]).

While the combination of supervised and RL may seem promising, Landajuela et al. [2022] demonstrated that such a combination offers only marginal enhancements in exact symbolic recovery. Nonetheless, in the age of large language models, there is potential to harness vast internet-scale knowledge (see e.g., Valipour et al. 2021). By learning the association between data points and mathematical expressions in realistic scenarios, and aligning with domain-specific assumptions using supervised learning techniques, it is conceivable to integrate this knowledge into a RL framework, as exemplified by Fan et al. [2022]. This approach might allow the recovery of expressions of substantially greater complexity than those we explore in Section 4.4.

Furthermore, one might envision using supervised approaches as a preliminary step to detect which non-linear tokens (e.g., cos, log, etc.) should appear in the expression underlying a dataset, leaving the RL framework with the task of properly combining them. This would be similar to how a physicist conducting SR might identify periodicity and damping in the data and then attempt to manually assemble them.

Data exploitation in accuracy-guided SR

Existing RL-based SR frameworks [Petersen et al., 2021a, Landajuela et al., 2021a, Petersen et al., 2021b, Landajuela et al., 2022, Faris et al., 2024, Michishita, 2024, Tian et al., 2024, He et al., 2024a, Du et al., 2022] and most other SR methods primarily focus on maximizing fit quality. This approach often results in limited constraints on symbol arrangement driven only by a non-differentiable (with respect to symbolic arrangement) scalar value indicative of fit quality. Unfortunately, the pathways that lead to optimal fit quality and those leading to accurate symbolic arrangement do not necessarily align. This misalignment can result in the "curse of accuracy-guided SR" [Grindle, 2021] where minor improvements in fit quality may mask significant deviations in the functional form of solutions, and vice versa. Consequently, enhancing the fit quality of candidates across learning iterations might inadvertently distance them from the correct symbolic arrangement.

To address this issue, a more nuanced approach to data exploitation could be beneficial. Currently, since the data remains constant, directly feeding it into the neural network proves ineffective, as the network would be unable to discern correlations between the static data and symbolic arrangements. A potential solution could involve a hybrid learning strategy that combines reinforcement learning with supervised learning. Specifically, this approach would involve using supervised learning to map the search space locally. This could be executed by training the neural network forward with reinforcement learning and backward with supervised learning, using expressions generated by our model to help it associate data and expressions locally. Both approaches would rely on a single neural network, which is a technique referred to as hard parameter sharing.

Practically, this method would entail maintaining our current setup but introducing synthetic data generation at each output of an expression by the neural network. Between each cycle of reinforcement learning, a round of supervised learning would be conducted using these synthetic datasets and their corresponding expressions. It is hoped that given these associations, once confronted with data of interest, the neural network would be able to effectively utilize it, having mapped the local correlations between data and symbolic arrangements in a supervised manner. This combined SR and supervised learning approach aims for a less ambitious, thus potentially more manageable, solution compared to approaches attempting to map all possible expressions to all datasets as seen in other supervised learning strategies for SR. This strategy is depicted in Figure 3.10 and is somewhat reminiscent of strategies used in automated theorem proving, such as those employed in AlphaZero.

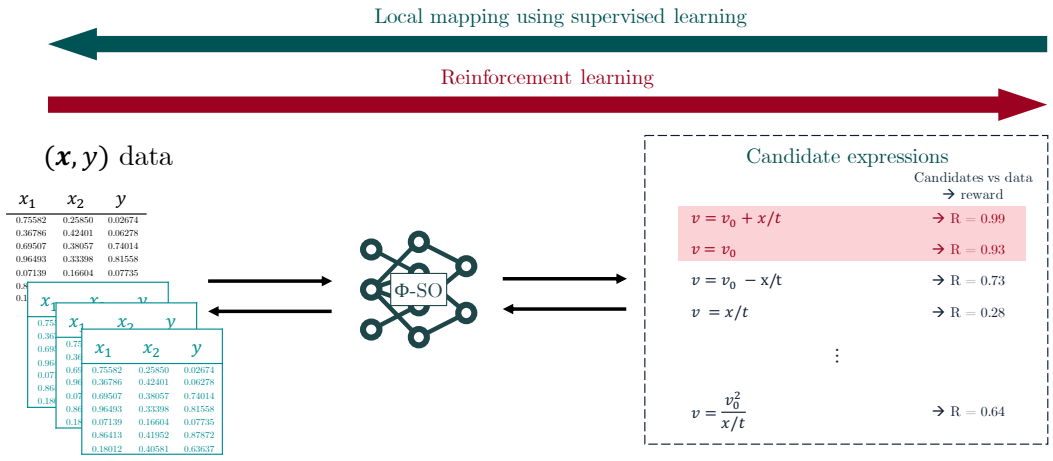


Figure 3.10: Perspective setup for going beyond accuracy guided SR. See 3.3.3 for a detailed description.

Complexity-accuracy metric

In addition, while our approach generates a Pareto front that gives accuracy-complexity trade-offs, future enhancements that integrate both complexity and accuracy into a singular metric (as in Bartlett et al. 2023a) could potentially enhance SR performances and address model selection challenges. However, one should note that our approach incorporates complexity requirements through several mechanisms: the entropy component of the loss function, the length priors and the Pareto-front which filters the results returned by the system to favor low complexity expressions at a similar accuracy.

Problem simplification schemes

As we will show in Section 4.4.3, it is straightforward to improve our method by combining it with the powerful problem simplification schemes devised in [Udrescu and Tegmark, 2020, Udrescu et al., 2020, Luo et al., 2022, Tohme et al., 2023, Cranmer et al., 2020b] exploiting separabilities, symmetries and more to divide SR problems into simpler sub-problems. The results of the separability procedures implemented in the `AI Feynman` [Udrescu et al., 2020] algorithm are conveniently recorded in separate datafiles, which makes it completely straightforward to use their approach as a pre-processing step for our approach. We anticipate that integrating their method within our algorithm, following the approach of Landajuela et al. [2022], should enhance the performance of our method. A preliminary framework combining an improved version of `AI Feynman` style simplification schemes with our framework will be discussed in Section 7.1.

Chapter 4

Physical Symbolic Regression

Portions of the content presented in this Chapter have been previously discussed in the following publication:

2023 *Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws*
W. Tenachi, R. Ibata, F. Diakogiannis
ApJ **959** 99, arXiv:2303.03192

Summary.

Research on symbolic regression methods has not been focused on physics, where we have important additional constraints due to the units associated with our data. We present a framework for recovering analytical symbolic expressions from physics data using deep reinforcement learning techniques by learning units constraints. Our system is built, from the ground up, to propose solutions where the physical units are consistent by construction.

This is useful not only in eliminating physically impossible solutions, but because the "grammatical" rules of dimensional analysis restrict enormously the freedom of the equation generator, thus vastly improving performance. We test our machinery on a standard benchmark of equations from the Feynman Lectures on Physics, achieving state-of-the-art performances and showcase its abilities on a panel of examples from astrophysics.

As established in the previous Chapter, although the prospect of using symbolic regression (SR) for discovering new analytical physical laws from data may be very appealing, it is also a very challenging combinatorial problem requiring one to develop highly efficient strategies to prune poor guesses

In this Chapter we build upon principles established in Chapter 3 and show how dimensional analysis can be exploited for this purpose and how our Φ-SO — Physical Symbolic Optimization — method combining deep reinforcement learning and dimensional analysis performs. This Chapter is primarily based on [Tenachi et al., 2023a], which led to a press release included in Appendix A.

The layout of this Chapter is as follows. We first detail our motivations in Section 4.1, in Section 4.2 we give our methodology for performing *in situ* dimensional analysis in partially written expressions and show how we exploit it to effectively learn the rules of dimensional analysis. In Section 4.3, we apply our method to a benchmark of 120 equations from the Feynman Lectures on Physics and other physics textbooks and compare it to 17 other popular SR algorithms, reporting state-of-the-art performances. In Section 4.4 we showcase Φ-SO's capabilities on a panel of astrophysical test cases and perform an ablation study. In Section 4.5 we show that our method can also be used to discover both physical laws and fundamental constants of nature from empirical data. Finally in Section 4.6 we discuss our results.

Remark on physical validity:

In the context of this Chapter, "physically valid" refers specifically to equations where the units are balanced, ensuring their coherence with dimensional analysis. We further discuss the enforcement to other physical principles in Section 11.2.

4.1 Motivations

This section outlines the rationales for the development of our method. We highlight its utility in ensuring the physical correctness of expressions in terms of unit dimensions (sub-section 4.1.1) and its effectiveness in significantly reducing the search space, thereby enhancing performance (sub-section 4.1.2). Additionally, we discuss the originality of our approach compared to existing literature (sub-section 4.1.3).

4.1.1 Ensuring the physicality of expressions

There are multiple approaches to SR (detailed in Section 2.2.2) which are capable of generating accurate analytical models. However, in the context of physics, we have the additional requirement that our equations must be balanced in terms of their physical units, as otherwise the equation is simply non-sensical, irrespective of whether it gives a good fit to the numerical values of the data. Although powerful, to the best of our knowledge, all of the available SR approaches spend most of their time exploring a search space where the immense majority of candidate expressions are unphysical in terms of units and thus often end up producing unphysical models (with the exception of approaches in which variables are rendered dimensionless beforehand as discussed in 4.2.3).

A very simple solution to this problem would have been to use an existing SR code, and check *post hoc* whether the proposed solutions obey that constraint. But not only does that constitute an immense waste of time and computing resources, which could render many interesting SR tasks impossible, it also makes a significant fraction of the resulting "best" analytical models unusable and uninterpretable. We note that for the sake of clarity, throughout this Chapter we refer to a system of *unique* quantities such as physical dimensions $\{L, M, T, I, \Theta, N, J\}$ i.e. with physical units $\{\mathrm{m}, \mathrm{kg}, \mathrm{s}, \mathrm{A}, \mathrm{K}, \mathrm{mol}, \mathrm{cd}\}$ a subset thereof, or problem-specific quantities such as $\{L, V, \rho, P, v\}$ i.e. with physical units $\{\mathrm{m}, \mathrm{m}^3, \mathrm{kg/m}^{-3}, \mathrm{Pa}, \mathrm{m.s}^{-1}\}$ as "units" [1].

4.1.2 Search space reduction

At first glance, one could think of the units constraints as severe restrictions that limit the capabilities of SR as they would prevent the generation of unphysical intermediary expressions. However, in this work we show that respecting physical constraints actually helps improve SR performance not only in terms of interpretability but also in accuracy by guiding the exploration of the space of solutions towards exact analytical laws. This is consistent with the studies of [Petersen et al., 2021a,b, Kammerer et al., 2020] who found that using *in situ* constraints during analytical expression generation is much more efficient as it vastly reduces the search space of trial expressions (though we note that incorporating such constraints in those frameworks would not be straightforward as one would need to recompute the whole relational graph representing an analytical expression and its underlying units constraints each time a new symbol is added).

[1] Although this can also be extended to systems with non-physical quantities, such as $\{\mathrm{scalar}, \mathrm{vector}, \mathrm{matrix}\}$ or even $\{\mathrm{dollars}, \mathrm{capita}, \mathrm{annum}\}$.

Here we present our Physical Symbolic Optimization framework (Φ-SO) which was designed from the outset to incorporate and take full advantage of physical units information during SR by storing and managing information related to dimensional analysis. This addresses in part the combinatorial challenge discussed in sub-section 3.1.3. Our Φ-SO framework includes the units constraints *in situ* during the equation generation process, such that only equations with balanced units are proposed *by construction*, thus also greatly reducing the search space as illustrated in Figure 4.1.

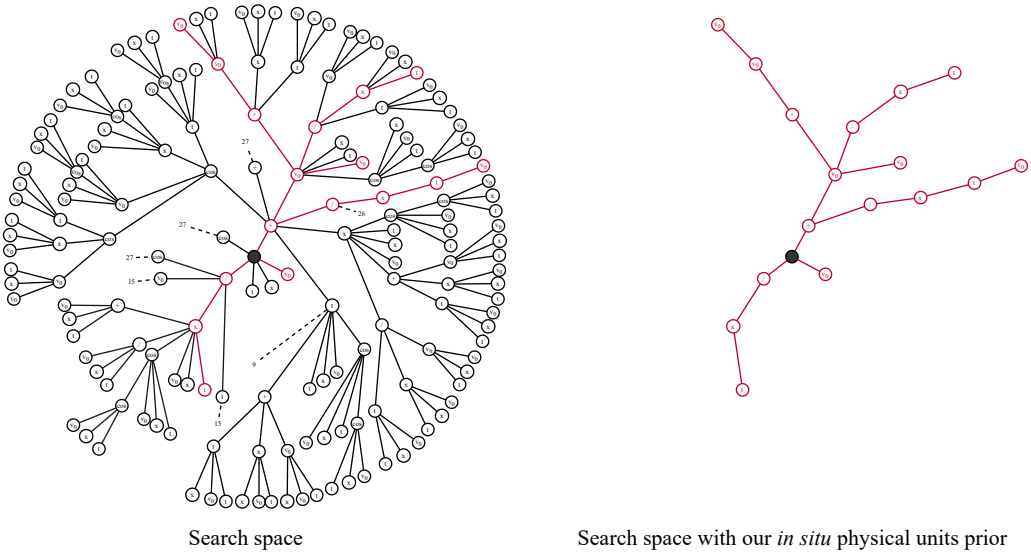


Figure 4.1: **Illustration of the symbolic expression search space reduction enabled by our *in situ* physical units prior.** We represent paths (in prefix notation) leading to expressions with physically-possible units (in red), a sample of the paths that lead to expressions with unphysical units (in black) with other unphysical paths redacted for readability summarized with dotted lines and their total number. Here we consider the recovery of a velocity v using a library of symbols $\{+, /, \cos, v_0, x, t\}$ where v_0 is a velocity, x is a length, and t is a time (limiting ourselves to 5 symbol long expressions for readability). This reduces the search space from 268 expressions to only 6. Note that the performance gain should scale exponentially with the expression length we allow the system to survey.

4.1.3 Innovativity

In the present study, we develop a foundational symbolic embedding for physics that enables the entire expression tree graph to be tackled, as well as local units constraints. Unlike previous attempts to consider units in which datasets were rendered dimensionless before applying standard SR techniques [Udrescu and Tegmark, 2020, Matchev et al., 2022, Keren et al., 2023], our approach allows us to anticipate the required units for the subsequent symbol to be generated

in a partially composed mathematical expression. By adopting this approach, we not only focus on training a neural network to generate increasingly precise expressions, as in Petersen et al. [2021a], but we also generate labels of the necessary units and actively train our neural network to adhere to such constraints. In essence, our method equips the neural network with the ability to learn to select the appropriate symbol in line with local units constraints.

To the best of our knowledge such a framework was never built before. This constitutes a first step in our planned research program of building a powerful general-purpose symbolic regression algorithm for astrophysics and other physical sciences. The purpose of this Chapter is to present our algorithm and show its workings and its potential.

4.2 Exploiting *In situ* Units Constraints

Our work is part of the broader field of grammar-guided SR [Ali et al., 2022, Brence et al., 2021, Crochepierre et al., 2022, Korns, 2011, Hoai et al., 2002, Manrique et al., 2009, Worm and Chiu, 2013] which aims at constraining the symbolic arrangement of mathematical expressions based on domain specific rules. Specifically and as discussed above, in physics we already know that some combinations of tokens are not possible due to units constraints. For example, if the algorithm is in the process of generating an expression in which a velocity (v_0) is summed with a length (x) divided by a token or sub-expression which is still to be generated ($\square$):

$$v_0 + \frac{x}{\square}\,, \tag{4.1}$$

then based on the expression tree (this is illustrated in Figure 3.7), we already know that that $\square$ must be a time variable or a more complicated sub-tree that eventually ends up having units of time, but that it is definitely not a length or a dimensionless operator such as the log function.

Sub-section 4.2.1 provides details about the algorithm we use to generate *in situ* units constraints, which are used to teach the neural network dimensional analysis rules and help to reduce the search space. In sub-section 4.2.2, we describe the reinforcement learning strategy we adopted to make our neural network not only produce accurate expressions but also physically meaningful ones. And finally in sub-section 4.2.3 we discuss other resembling approaches from the literature.

4.2.1 *In situ* dimensional analysis

Dimensional analysis in (in)-complete expressions

Computing such constraints *in situ* i.e. in incomplete, only partially sampled trees (containing empty placeholder nodes) is much harder than simply checking *post hoc* if the units of a given equation make sense, because in some situations it is impossible to compute such constraints until later on in the sequence, leaving the units of some nodes *free* (i.e. compatible with any units at this point in the sequence). For example, it is impossible to compute the units requirement in the left child node of a (binary) multiplication operator token $\square \times \triangle$, as any units in the $\square$ left child node could be compensated by units in the $\triangle$ right child node. Following the dimensional analysis rules summarized in Table 4.1, we devised Algorithm 1.

Dimensional analysis rules

Expression	Units
$\tau_A \pm \tau_B$	Φ_A or Φ_B
$-\tau_A$	Φ_A
$\tau_A \times \tau_B$	$\Phi_A + \Phi_B$
τ_A / τ_B	$\Phi_A - \Phi_B$
${\tau_A}^n$	$n \times \Phi_A$
$\mathrm{op}_{\mathbf{0}}(\tau_A)$	$\mathbf{0}$

Units requirements rules

Expression	Requirement
$\tau_A \pm \tau_B$	$\Phi_A = \Phi_B$
$y = \tau_A$	$\Phi_y = \Phi_A$
$\mathrm{op}_{\mathbf{0}}(\tau_A)$	$\Phi_A = \mathbf{0}$

Table 4.1: Dimensional analysis prescriptions to enforce. With τ_A, τ_B, y, Φ_A, Φ_B, Φ_y referring to two nodes, the output variable and the powers of their units vectors, $\mathrm{op}_{\mathbf{0}}$ denoting a dimensionless operation (e.g., $\{\cos, \sin, \exp, \log\}$) and ${\tau_A}^n$ representing any power operation (including e.g., $1/\tau_A = {\tau_A}^{-1}$, $\sqrt{\tau_A} = {\tau_A}^{\frac{1}{2}}$)

Maximally informing algorithm

This algorithm gives the pseudo-code of the procedure we devised to compute the required units whenever possible and leaving them as *free* otherwise. The procedure is applied to a token at position $t < N$ in an incomplete or complete sequence of tokens $\{\tau^{\langle t \rangle}\}_{t<N}$ of size N, knowing the units of terminal nodes

and of the root node (e.g., respectively $\{v_0, x, t\}$ and $\{v\}$ in the example of Figure 3.7). The sequence may be partially made up of placeholder tokens of yet undetermined nature (representing dangling nodes). Running algorithm 1 before each token generation step allows one to have a maximally informed expression tree graph in terms of units.

Algorithm 1: *In situ* units requirements algorithm

Input: (In)-complete expression $\{\tau^{\langle t \rangle}\}_{t<N}$, Position of token t
Output: Required physical units $\Phi^{\langle t \rangle}$ of token at t
Function ComputeRequiredUnits($\{\tau^{\langle t \rangle}\}_{t<N}, t$)
  $p \leftarrow$ PositionOfParent(t)
  $s \leftarrow$ PositionOfSibling(t)
  $\Phi^{\langle p \rangle} \leftarrow$ Units($\tau^{\langle p \rangle}$)
  $\Phi^{\langle s \rangle} \leftarrow$ Units($\tau^{\langle s \rangle}$)
  NodeRank $\leftarrow$ 1 if left side node and 2 if right node
  AdditiveTokens $\leftarrow \{+, -\}$
  MultiplicativeTokens $\leftarrow \{\times, /\}$
  PowerTokens $\leftarrow \{1/\square, \sqrt{\square}, \square^n\}$
  PowerValues $\leftarrow \{1/\square : -1, \sqrt{\square} : 1/2, \square^n : n\}$
  DimensionlessTokens $\leftarrow \{\cos, \sin, \tan, \exp, \log\}$
  if $\tau^{\langle p \rangle}$ is in AdditiveTokens **and** $\Phi^{\langle s \rangle}$ is known **then**
    $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle s \rangle}$
  else if $\tau^{\langle p \rangle}$ is in AdditiveTokens **and** $\Phi^{\langle p \rangle}$ is free **and** NodeRank is 2 **and** $\Phi^{\langle s \rangle}$ is free **then**
    BottomUpUnitsAssignement(start $= s$, end $= t - 1$)
    $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle s \rangle}$
  else if $\Phi^{\langle p \rangle}$ is free **and** $\tau^{\langle p \rangle}$ is **not** in MultiplicativeTokens **and** $\tau^{\langle s \rangle}$ is **not** a placeholder **then**
    $\Phi^{\langle t \rangle} \leftarrow$ free;
  else if $t = 0$ **then**
    $\Phi^{\langle t \rangle} \leftarrow$ Units(root)
  else if $\tau^{\langle p \rangle}$ is in AdditiveTokens **then**
    $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle p \rangle}$
  else if $\tau^{\langle p \rangle}$ is in PowerTokens **then**
    $n \leftarrow$ PowerValues$[\tau^{\langle p \rangle}]$
    $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle p \rangle}/n$
  else if $\Phi^{\langle p \rangle} = \mathbf{0}$ **or** $\tau^{\langle p \rangle}$ is in DimensionlessTokens **then**
    $\Phi^{\langle t \rangle} \leftarrow \mathbf{0}$
  else if $\tau^{\langle p \rangle}$ is in MultiplicativeTokens **then**
    if $\tau^{\langle t \rangle}$ is a placeholder **and** $\tau^{\langle s \rangle}$ is a placeholder **then**
      $\Phi^{\langle t \rangle} \leftarrow$ free
    else if NodeRank is 1 **then**
      $\Phi^{\langle t \rangle} \leftarrow$ free
    else if $\Phi^{\langle p \rangle}$ is free **then**
      $\Phi^{\langle t \rangle} \leftarrow$ free
    else
      BottomUpUnitsAssignement(start $= s$, end $= t - 1$)
      if $\tau^{\langle t \rangle}$ is $\{\times\}$ **then**
        $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle p \rangle} - \Phi^{\langle s \rangle}$;
      else if $\tau^{\langle t \rangle}$ is $\{/\}$ **then**
        $\Phi^{\langle t \rangle} \leftarrow \Phi^{\langle s \rangle} - \Phi^{\langle p \rangle}$;
  return $\Phi^{\langle t \rangle}$

4.2.2 Learning from physical units

Having access *in situ* to the (required) physical units of tokens allows us to not only to inform the recurrent neural network (RNN) of our expectations in terms of units as well as to feed it units of surrounding tokens, thus allowing the model to leverage such information, but also to express a prior distribution over the library. This is illustrated on Figure 4.2 which is a dimensional analysis centric version of Figure 3.7 highlighting the flow of physical units information.

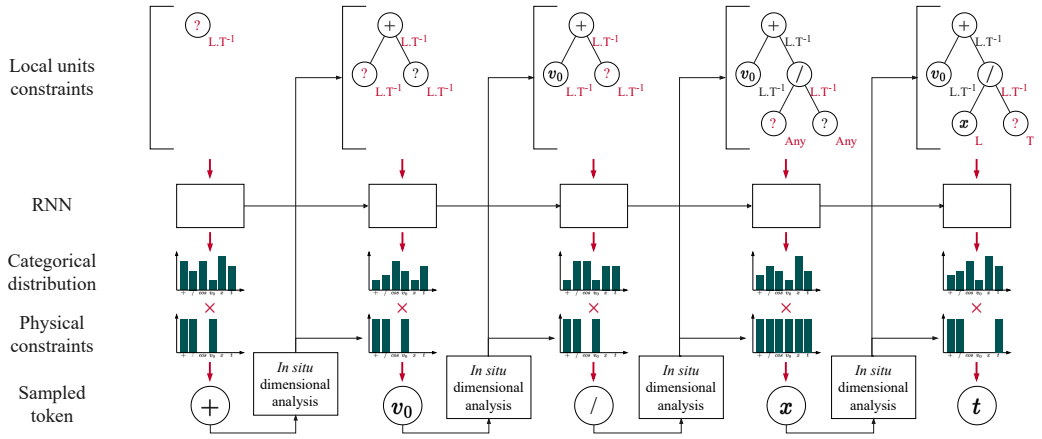


Figure 4.2: **Dimensional analysis centric expression generation sketch highlighting the flow of physical units information.** The process starts at the top left RNN block. For each token, the RNN is given the contextual information regarding the surroundings of the next token to generate in the graph, including physical units information. Based on this information, the RNN produces a categorical distribution over the library of available tokens (top histograms) as well as a state which is transmitted to the RNN on its next call. The generated distribution is then masked based on local units constraints (bottom histograms), forbidding tokens that would lead to nonsensical expressions. The resulting token is sampled from this distribution, leading to the token '+' in this example. Based on the new symbolic graph, the *in situ* dimensional analysis updates physical units information and constraints which are used to inform the RNN and emit the prior distribution. Repeating this process, from left to right, allows one to generate a complete physical expression, here $(+, v_0, /, x, t)$ which translates into $v_0 + x/t$ in the infix notation we are more familiar with.

Implicit guarantee of units consistency

This enables the algorithm to zero-out the probability of forbidden symbols that would result in expressions that violate units rules. Combining this prior distribution with the categorical distribution given by the RNN while expressions are being generated results in a system where *by construction* only correct expressions with correct physical units can be formulated and learned on by the neural network.

This is rendered possible by the fact that in addition to the straightforward priors described earlier in sub-section 3.2.3 whose formulation depends on the local tree structure (parent, sibling, ancestors), our method is able to accommodate any priors that take into account the entire tree structure without having to recompute it from scratch at each step. This is possible thanks to the fact that contrary to other deep learning based SR algorithms, in the Φ-SO framework we compute and keep track of the full graph of the tree representation and its underlying grammatical information (such as units, symbol types like functions, free parameters, fixed constants or the number of arguments a symbol requires) — as detailed in sub-section 6.1.1 — while the expression is being generated, as it is an essential ingredient to compute units constraints as detailed in the following sub-section. Note that this also enables Φ-SO to accommodate any future prior relying on such information.

Informing the neural network of physical units constraints

As detailed in Section 3.2, symbolic expressions can be regarded as binary trees where each node represents a symbol of the expression in the library of available symbols, i.e., an input variable (e.g., x, t), a constant (e.g., v_0) or an operation (e.g., $+$, $-$, $\times$, $/$, sin, log, ...). In this representation, input variables and constants can be referred to as terminal nodes or symbols (having no child node), operations taking a single argument (e.g., sin, log, ...) are unary symbols (having one child node) and operations taking two arguments (e.g., $+$, $-$, $\times$, $/$, ...) are binary symbols (having to child nodes that can be considered sibling nodes). By considering each node first in depth and then left to right, one can compute a one dimensional list i.e. a prefix[2] notation in which operators are placed before the corresponding operands in the expression, alleviating the need for parentheses. Using the prefix notation and treating symbols, referred to as tokens, as categories allows us to treat any expression as a mere sequence of categorical vectors. E.g., considering short toy library of tokens $\{+, \cos, x\}$, the operator $+$ can be encoded as $(1, 0, 0)$, the function cos as $(0, 1, 0)$ and the variable x as $(0, 0, 1)$.

In the framework of our dimensional analysis approach, the RNN is provided with an enriched set of observations beyond those previously described in 3.3.1. In addition to the number of dangling nodes n_{dangling}, the nature of the token sampled at the previous step $\tau^{\langle t-1 \rangle}$, and the sibling $\tau^{\langle s \rangle}$ and parent $\tau^{\langle p \rangle}$ tokens at the current step, the RNN is also informed about the physical units relevant to each token. Specifically, we augment the observations with

[2]This is also called "Polish" notation, and can be converted to a tree representation or the "infix" notation which we are more familiar with, as there is a one-to-one relationship between them.

the units of the previously generated token $\Phi^{\langle t-1\rangle}$, the sibling $\Phi^{\langle s\rangle}$, and the parent $\Phi^{\langle p\rangle}$, as well as the required units $\Phi^{\langle t\rangle}$ for the token being generated such that as to ensure compliance with dimensional analysis rules. The observations provided to the RNN at each step t are therefore as follows:

$$O^{\langle t\rangle} = \{n_{\text{dangling}}, \tau^{\langle t-1\rangle},\ \tau^{\langle s\rangle}, \tau^{\langle p\rangle}, \Phi^{\langle t-1\rangle},\ \Phi^{\langle s\rangle}, \Phi^{\langle p\rangle}, \Phi^{\langle t\rangle}\} \tag{4.2}$$

Where $\tau^{\langle t\rangle}$ refers to the token at position t in prefix notation, p and s respectively denote the position of the parent and sibling tokens of $\tau^{\langle t\rangle}$ and $\Phi^{\langle t\rangle}$ refers to the physical units of token at t.

This allows the inner mechanisms of the neural network to take into account not only the local structure of the expression for generating the next token, but also to take into account the local units constraints. The process described above can be repeated multiple times until a whole token function is generated in prefix notation, as illustrated in Figure 4.2.

Deep learning considerations

As detailed in Chapter 3 our framework could be applied to virtually any one of the SR approaches described in Section 2.2.2 where tokens are sampled sequentially, we chose to implement our algorithm in `PyTorch` (Paszke et al. 2019), building our method from scratch yet using some of the mathematical principles and key strategies pioneered in the state-of-the-art Deep Symbolic Regression framework proposed in Petersen et al. [2021a] and Landajuela et al. [2021a] which rely on reinforcement learning via a risk-seeking policy gradient (which is based on Rajeswaran et al. 2017).

Our learning hyper-parameters were given in Table 3.1. It is worth noting that the empirically tuned batch size we found (10k) is larger than the one found by Petersen et al. [2021a] which was of 1k. We attribute this to the very strong constraints offered by our Φ-SO setup in symbolic arrangements which require a strong exploration counterpart to avoid getting stuck in local minima. This helps ensure that the model does not prematurely converge by continuously reinforcing a locally optimal expression, but rather seeks more solutions until identifying the most favorable one.

It is worth noting that our approach reinforces candidates which are sampled based on not only the output of the RNN, but also the local units constraints derived from the units prior distribution, which ensures the physical correctness of token choices. As a result, our approach effectively trains the RNN to make appropriate symbolic choices in accordance with local units constraints, in a quasi supervised learning manner. This combined with the general reinforcement learning paradigm enables us to produce both accurate and physically relevant symbolic expressions.

4.2.3 Comparison to other approaches in the literature

We acknowledge a previous attempt by [Udrescu and Tegmark, 2020] in the `AI Feynman` algorithm to consider units in the context of SR. The approach adopted by `AI Feynman` addresses SR problems by first transforming the variables to make them dimensionless, often leading to a reduction in the number of variables and allowing the generation of physically balanced expressions. However, if this method fails, the algorithm reverts to the original problem setup. This results in `AI Feynman` resorting to fitting high-order polynomials or complicated expressions that although very accurate lack physical meaning from a dimensional analysis perspective most of the time when it is not able to find a perfect fit solution. For instance, even in the shorter range of expressions it proposes, one can find equations such as $K = \arcsin(0.169e^{-3.142m+w})$ where K, m and w denote an energy, a mass and a velocity for Feynman problem I.13.4 (details about the Feynman symbolic regression problems can be found in sub-section 4.3.1). In contrast, Φ-SO is designed to yield only physically plausible expressions (in terms of their units) by construction all of the time. Contrary to `AI Feynman`, Φ-SO works on dimensional data by leveraging constraints on the functional forms while generating expressions as outlined in Table 4.1. It is worth noting, however, that making problems dimensionless, as implemented in `AI Feynman`, is a valuable approach that can work in pair with any SR method to ensure outputs are not non-sensical.

Indeed, it could be argued that we could have tackled the physical units validity of expressions in SR by taking advantage of the Buckingham Π theorem [Buckingham, 1914], with variables and constants rendered dimensionless by means of multiplicative operations amongst them. Such an approach can actually be adopted as a preliminary step in conjunction with any SR framework (see, e.g., Matchev et al. 2022, Keren et al. 2023). However, although working with so called Π groups ensures the generation of physically valid expressions (since all terms become dimensionless), it simultaneously removes constraints imposed by dimensional analysis, complicating the SR process. It is interesting to note that nature (or at least physics) is not dimensionless, so information is lost during the process of making variables and constants dimensionless, preventing us from leveraging the powerful constraints on the functional form associated with this dimensional information. Drawing from the example presented in Udrescu and Tegmark [2020], let us consider a dataset associated with the target expression

$$F = \frac{Gm_1m_2}{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}\,. \tag{4.3}$$

When rendered dimensionless, the target expression becomes

$$y = \frac{\frac{m_2}{m_1}}{(\frac{x_2}{x_1} - 1)^2 + (\frac{y_2}{x_1} - \frac{y_1}{x_1})^2 + (\frac{z_2}{x_1} - \frac{z_1}{x_1})^2} \,. \tag{4.4}$$

While this transformation decreases the number of input variables to $\{\frac{m_2}{m_1}, \frac{x_2}{x_1}, \frac{y_2}{x_1}, \frac{z_1}{x_1}\}$, it simultaneously nullifies the inherent dimensional analysis constraints. Consequently, the SR algorithm could potentially produce expressions such as $\frac{m_2}{m_1} - \frac{x_2}{x_1}$ or $(\frac{x_2}{x_1} - 1)^2 + \frac{y_2}{x_1}$. In contrast, with our *in situ* constraints, lengths could only be summed with lengths terms, similarly, squared lengths could only be summed with squared lengths and having Gm_1m_2 in the numerator would be enforced by the requirement of the expression being homogeneous to a force. In essence, while rendering variables dimensionless ensures physicality of the expressions, it simultaneously relinquishes valuable constraints on their functional forms.

Finally, we note that since the initial submission of our work, three similar approaches have been introduced, further underscoring the utility and relevance of our method in the field. The first approach integrating dimensional analysis with a sparsity-fitting method [Purcell et al., 2023], the second employing a probabilistic search strategy [Brence et al., 2023], and the third leveraging a genetic programming framework [Reissmann et al., 2024]. Notably, the latter approach was benchmarked directly against our method using the same dataset and protocol as described in Section 4.3, with our approach demonstrating superior performance, as documented in [Reissmann et al., 2024].

4.3 Feynman Benchmark

To validate the efficacy of our Φ-SO method, we conducted benchmark tests using the widely-recognized Feynman symbolic regression benchmark. This set of challenges, first introduced by Udrescu and Tegmark [2020] and subsequently formalized in `SRBench` [La Cava et al., 2021], encompasses 120 equations including 100 sourced from the renowned Feynman Lectures on Physics [Feynman et al., 1971] with the other 20 sourced from other textbooks: Goldstein et al. 2002, Jackson 2012, Weinberg 1972, Schwartz 2014. The primary objective is to retrieve these equations using only the provided data points at various levels of noise.

Although this benchmark has inherent limitations, such as treating constants of nature (e.g., G, c, $\hbar$) and discrete physical values from quantum

mechanics as continuously varying input variables (which places a higher emphasis on the implementation of the problem simplification schemes developed in Udrescu and Tegmark [2020]), it offers a comprehensive representation of the diversity of physical functional forms and remains a valuable standard for comparison as most SR methods have been thoroughly benchmarked on it (see La Cava et al. 2021).

Details on the benchmarking procedure can be found in 4.3.1. Results on exact symbolic recovery are provided in 4.3.2, while findings regarding fit quality are presented in 4.3.3. Finally, we provide training curves in 4.3.4.

4.3.1 Benchmarking procedure

Benchmarking rules

We meticulously adhered to the established protocol delineated in `SRBench` by La Cava et al. [2021], setting our `PhySO` algorithm to identify expressions that fit 10,000 data points corresponding to each Feynman benchmark equation. `PhySO` was only allowed to evaluate a maximum of one million expressions during each run and exact symbolic recovery was assessed by ensuring the difference between the expression generated by `PhySO` and the target expression reduced to a constant or that the fraction simplified to a constant using the `SymPy` library for symbolic mathematics [Meurer et al., 2017]. In addition, fit quality was assessed on 100,000 noiseless test data points using the R^2 metric defined as :

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.5}$$

As per benchmark rules, in order to ensure robustness, for each equation, this procedure was repeated multiple times (opting here for 5 trials over 10 due to the considerable computational demands associated with such benchmarks), each with a unique random seed, and the recovery rates were subsequently averaged. In alignment with `SRBench` stipulations, equations I.26.2, I.30.5, and test_10 (containing arccos and arcsin functions) as well as II.11.17 were excluded from our results. The whole benchmark tests were conducted across four noise levels: 0%, 0.1%, 1% and 10%, leading to the evaluation of 2,320,000,000 expressions.[3]

[3] Additional details about the implementation of this protocol can be found in 6.1.3 and a straightforward way to technically reproduce the results presented here is given in paragraph 6.3.

Method	Technique(s)	Description	Reference
PhySO	RL, DA	Physical Symbolic Optimization	Tenachi et al. [2023a]
uDSR	RL, GP, Simp., Sup.	A Unified Framework for Deep Symbolic Regression	Landajuela et al. [2022]
AIFeynman 2.0	Simp., DA	Symbolic regression exploiting graph modularity	Udrescu et al. [2020]
AFP_FE	GP	AFP with co-evolved fitness estimates, Eureqa-esque	Schmidt and Lipson [2009]
DSR	RL	Deep Symbolic Regression	Petersen et al. [2021a]
AFP	GP	Age-fitness Pareto Optimization	Schmidt and Lipson [2011]
gplearn	GP	Koza-style symbolic regression in Python	Stephens [2015]
GP-GOMEA	GP	GP-Optimal Mixing Evolutionary Algorithm	Virgolin et al. [2021]
ITEA	GP	Interaction-Transformation EA	de Franca and Aldeia [2021]
EPLEX	GP	ϵ-lexicase selection	La Cava et al. [2019]
NeSymReS	Sup.	Neural Symbolic Regression that Scales	Biggio et al. [2021]
Operon	GP	SR with Non-linear least squares	Kommenda et al. [2020]
SINDy	NeuroSym	Sparse identification of non-linear dynamics	Brunton et al. [2016]
SBP-GP	GP	Semantic Back-propagation Genetic Programming	Virgolin et al. [2019]
BSR	MCMC	Bayesian Symbolic Regression	Jin et al. [2019]
FEAT	GP	Feature Engineering Automation Tool	Cava et al. [2019]
FFX	Rand.	Fast function extraction	McConaghy [2011]
MRGP	GP	Multiple Regression Genetic Programming	Arnaldo et al. [2014]

Table 4.2: **Baseline SR methods.** Summary of baseline symbolic regression methods along with the the underlying techniques they rely on: reinforcement learning (RL), genetic programming (GP), problem simplification schemes (Simp.), end-to-end supervised learning (Sup.), dimensional analysis (DA), neuro-symbolic / auto-differentiation based sparse fitting techniques (NeuroSym), Markov chain Monte Carlo (MCMC) and random search (Rand.).

PhySO evaluation

We ran PhySO using the hyper-parameters and reward metric given in Section 3.3 (with the notable exception of the trigonometric prior which was set to a maximum nesting of one) and allowing the use of $\{+, -, \times, /, 1/\square, \sqrt{\square}, \square^2, -\square, \exp, \log, \cos, \sin\}$ as well as two dimensionless adjustable free constants and a constant equal to one $\{c_1, c_2, 1\}$. After each run, the first few expressions (in accuracy) of the Pareto front were inspected, which proved beneficial for cases where SymPy faced simplification challenges only and making a marginal difference of approximately 1% in recovery rate. Notably, while the Feynman dataset includes unit information for each variable, PhySO is the only method that capitalizes on this feature since its introduction in Udrescu and Tegmark [2020], a testament to its unique physics specific design. For the sake of reproducibility, we provide all the code required to execute the benchmark using PhySO as well as the detailed SRBench-style results regarding each run.

Comparison notes

We compare the performance of our Φ-SO approach to other SR algorithms with documented exact symbolic recovery rates, as reported in [La Cava et al.,

2021] and [Landajuela et al., 2022]. These algorithms are summarized in Table 4.2.

Remarkably, this includes `AFP_FE` a `Eureqa`-like method, by the same authors combining `AFP` with `Eureqa`'s method for fitness estimation [La Cava et al., 2021] and which we denote as `AFP_FE` ($\sim$ `Eureqa`).

In La Cava et al. 2021, `DSR` [Petersen et al., 2021a] was not permitted to use any free parameters when generating expressions, greatly hindering its capabilities; we therefore also consider the performance of the latest version of `DSR` [Landajuela et al., 2021a] self-reported in the ablation study of Landajuela et al. [2022] which relies on more suitable hyper-parameters as a baseline. However, we note that is important to exercise caution when interpreting this additional `DSR` performance data-point as well as the performances of `SINDy`, `NeSymReS`, and `uDSR` as our available data only offers their final scores on a composite dataset, which encompasses both the Feynman benchmark and the Strogatz benchmark [La Cava et al., 2016] — the latter accounting for approximately 5% of the total score. This aggregated score is what we illustrate in our figures throughout this Section. In addition, it is worth noting that the exact conditions under which `SINDy` and `NeSymReS` were benchmarked are unknown and that in the case of `uDSR` and the additional `DSR` data-point, the benchmarking respectively permitted an evaluation of up to 2 million and 0.5 million expressions respectively, in contrast to the 1 million limit set for other methods.

Furthermore, detailed results for these methods, in particular those regarding the specific expressions they identified, are unavailable, preventing their inclusion in our comparative analysis when concerning expression metrics (complexity or number of free parameters). Although, per `SRBench` rules, we permitted our method to evaluate up to 1 million expressions compared to `DSR`'s 0.5 million, `PhySO` typically identifies the correct expression well before reaching this limit or not at all. Additionally, while `DSR`'s 42% score is influenced by another benchmark, the impact is very low, accounting for only 5%. This external benchmark is relatively straightforward, with `DSR` achieving around 25% even without free parameters [La Cava et al., 2021], indicating its limited effect on the overall score. Thus, we believe a direct comparison between `PhySO`'s score and `DSR`'s from Landajuela et al. [2022] is valid especially considering the gap in performance as detailed in the next sub-section.

4.3.2 Exact symbolic recovery

Performances on noisy data

Figure 4.3 presents the performance of `PhySO` against baseline algorithms from Table 4.2 on the Feynman benchmark. This includes the average exact symbolic recovery rate, accurate expression rate (defined as those with a fit

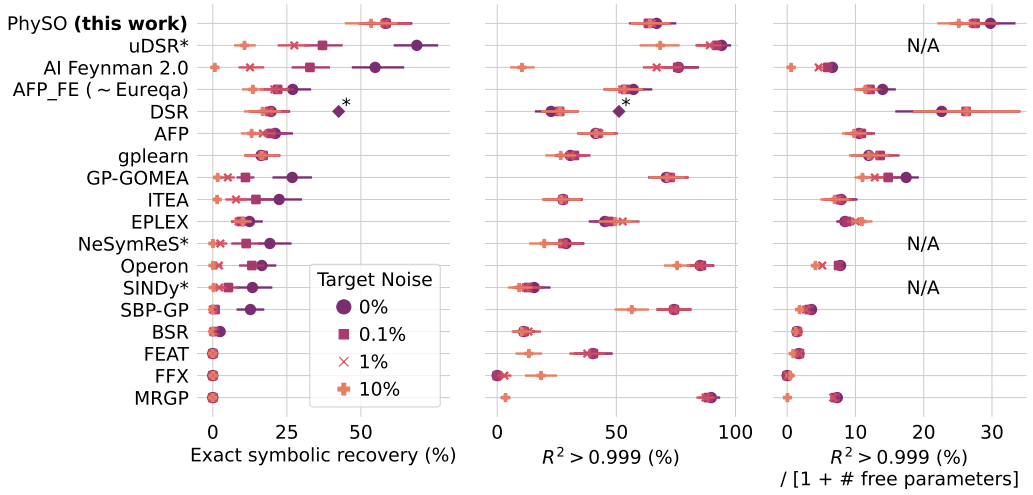


Figure 4.3: **Performances of `PhySO` against baseline methods on the Feynman benchmark.** Exact symbolic recovery rates, rates of accurate expressions (having $R^2 > 0.999$) and same rates normalized by the number of free parameters appearing in expressions for `PhySO` and other baseline methods on the Feynman benchmark. `PhySO` vastly outperforms all other methods in symbolic recovery in the presence of even minimal levels of noise ($> 0.1\%$). In addition, the effectiveness of the dimensional analysis schemes of our Φ-SO approach are clearly visible when comparing `DSR` (a purely RL method) with our implementation: `PhySO` (combining RL with dimensional analysis). Error-bars indicate a 95% confidence interval, ♦ denotes performances of `DSR` [Landajuela et al., 2021a] reported in Landajuela et al. [2022] on noiseless data with free constants allowed and * denotes that benchmarking conditions may vary and scores are polluted by approximately 5% of results from another benchmark.

coefficient $R^2 > 0.999$), and normalized accurate expression rate considering the number of free parameters in the expressions, across different noise levels. Compared to `DSR`, which strictly relies on reinforcement learning, `PhySO` utilizes both reinforcement learning and dimensional analysis. With `DSR`'s score at roughly 42%, our method's 58.5% score highlights the significant benefits of incorporating dimensional analysis. In the realm of physics, the exact symbolic recovery rate is a paramount metric and given that real-world physics data is often noisy, the resilience of an algorithm to noise is also crucial. However, with a minor noise level of 0.1%, many high-performing methods see their recovery rates almost halved. In contrast, `PhySO` maintains consistent performances. Remarkably, at a 10% noise level, where most methods' recovery rates dip below 20%, and even high performers like `uDSR` and `AI Feynman 2.0` score only 10.7% and 0.7% respectively, `PhySO` continues to accurately recover expressions over 53% of the time.

Performances on noiseless data

In noiseless scenarios PhySO is only surpassed by uDSR which relies on a cocktail of five of the most potent SR techniques: reinforcement learning for iterative adjustments, genetic programming for enhanced randomization and exploration, supervised learning to leverage existing knowledge, neuro-symbolic style sparse coefficient fitting for its linear symbolic modules and powerful simplification strategies, similar to those utilized by AI Feynman 2.0, which narrowly lags behind PhySO. These techniques rely on the exploitation of separability (e.g., simplifying the search of $f(x_1, x_2)$ to the search of the simpler functions $f_1(x_1)$ and $f_2(x_2)$ with $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$), symmetry (e.g., simplifying the search of $f(x_1, x_2)$ to the search of $f_1(x_1, x_2)$ and $f_2(x_2)$ with $f(x_1, x_2) = f_1(x_1, f_2(x_2))$), and many other schemes to circumvent the intricate functional forms in the benchmark. Despite relying solely on reinforcement learning and dimensional analysis, on noiseless data PhySO rivals uDSR and surpasses AI Feynman 2.0, demonstrating the effectiveness of our approach.

Overview

It is worth noting that while the aforementioned AI Feynman-style "divide and conquer" simplification strategies are effective, they are extremely noise sensitive, a scenario where PhySO's approach remains stable. In summary, this benchmark shows that incorporating dimensional analysis constraints into SR significantly bolsters performance. Given the improvements shown from PhySO over DSR thanks to the inclusion of units constraints, and given uDSR's impressive performances in noiseless scenarios, we believe combining Φ-SO with uDSR could elevate outcomes even further.

Performance comparison with PySR

PySR, an open-source implementation of the Eureqa software, which has gained popularity in the astrophysics community was shown to exhibit performance levels comparable to those of Eureqa [Cranmer, 2023].

Had PySR [Cranmer, 2023] been included in this benchmark alongside the methods evaluated in [La Cava et al., 2021], we anticipate that our PhySO approach would have demonstrated superior performance in both noisy and noiseless scenarios. Given that PySR is essentially a reimplementation of Eureqa, we estimate that PhySO would achieve approximately double the exact symbolic recovery rate of PySR. This advantage would persist regardless of the use of dimensional analysis features, since even without this feature, our performance would remain at least marginally superior to that of DSR's

self reported performance.

Noise level	0%	0.1%	1%	10%
PySR	-	59	-	-
PhySO	**71**	**70**	**69**	**67**

Table 4.3: **Comparison of `PySR` and `PhySO` performance.** The table reports the number of expressions exactly recovered by `PySR` and `PhySO` from their associated data across different noise levels, out of the bulk 100 Feynman benchmark problems. The single data point for `PySR` at 0.1% noise is sourced from [Grayeli et al., 2024], while `PhySO` results are computed under the same benchmarking protocol.

However, a recent study has directly reported the performance of `PySR` on the Feynman benchmark [Grayeli et al., 2024][4] Unfortunately, this study evaluated `PySR` at a single noise level of 0.1%, which emphasizes the system's performance, as it notably impacts one of the top-performing systems, `uDSR`[5]. The study did not extend to higher noise levels that would likely pose more significant challenges for `PySR` and might reveal behavior akin to that of `Eureqa`. Additionally, the comparison did not include other state-of-the-art systems known for greater noise resilience, such as the current `PhySO` approach [Tenachi et al., 2023a] or `DSR` [Petersen et al., 2021a] at its full potential[6].

The authors self-benchmarked their `PySR` system on the 100 bulk challenges (excluding the 20 "hard" challenges) from the Feynman benchmark. Surprisingly, the reported performance surpasses that of `Eureqa`, as assessed by an independent benchmark in `SRBench`. To the best of our knowledge, no explanation has yet been provided for this inconsistency. There is currently no straightforward way to replicate these experiments and the study does not offer problem-by-problem results or a detailed list of the specific problems solved, as is in other SR studies. Instead, it presents only an aggregate count of successfully recovered expressions. Furthermore, the authors did not conduct multiple runs to calculate a recovery rate, opting instead to report whether their system could solve each problem in a single attempt.

[4]The reported performances are available through an ablated version of a system combining `PySR`. with an additional component designed to perform SR from data supplemented by user-provided text hints about the dataset revealing its features (e.g., , indicating the presence of dampening or periodicity etc.).

[5]`uDSR` is particularly sensitive to noise when its `AI Feynman`-style "divide and conquer" scheme is activated.

[6]Allowing the system to use free constants.

Case	Expression
I.11.19	$x_1 y_1 + x_2 y_2 + x_3 y_3$
I.50.26	$x_1 \left(\alpha \cos^2(\omega t) + \cos(\omega t)\right)$
II.6.15b	$3 p_d \cos(\theta) \sin(\theta) / (4\pi\epsilon r^3)$
III.15.14	$\left(\frac{h}{2\pi}\right)^2 / 2 E_n d^2$

Table 4.4: **Examples of Feynman benchmark expressions recovered by `PhySO` but not by `PySR`.** This table provides a non-exhaustive sample of expressions from the Feynman benchmark where `PySR` fails to recover the exact solution, but `PhySO` succeeds. Information derived from [Grayeli et al., 2024].

To facilitate a direct comparison, we adapted our benchmarking protocol to align with the one described in the referenced study and computed corresponding performance figures for our system. The self-reported performance metrics of `PySR` from [Grayeli et al., 2024] are juxtaposed with the results of `PhySO` under this specific Feynman benchmark setup in Table 4.3. At a noise level of 0.1%, `PhySO` demonstrates superior performance, successfully recovering 11 additional expressions compared to `PySR`. We further hypothesize that this performance gap would increase at higher noise levels, given the tendency of `Eureqa`-style systems to exhibit substantial performance degradation under elevated noise conditions, as illustrated in Figure 4.3.

Although the authors do not provide a comprehensive list of expressions that `PySR` can or cannot solve, they do present a sample. Using this information, we identify examples of expressions that `PySR` fails to recover but `PhySO` successfully solves. These cases are detailed in Table 4.4.

4.3.3 Fit quality

Regarding the fraction of expressions with an $R^2 > 0.999$, many methods achieve high scores by incorporating an extensive number of free constants, resulting in intricate expressions that often lack interpretability and are nonsensical from a dimensional analysis standpoint. For example, `AI Feynman 2.0`, when not identifying the precise symbolic expressions, tends to generate complex expressions comprising, post-simplification, an average of 147 symbols and 18 free constants due to its brute-force polynomial fitting approach. Similarly `Operon` [7] and `MRGP` expressions contain on average respectively 17 and 88 free constants post-simplification at a 10% noise level. This is not a problem in many fields where human-interpretability is not a priority. However, given the importance of this criterion in physics we also show in Figure 4.3 the rate of

[7] It should be noted that a recent improvement of Operon (see Burlacu [2023]) allowing it to produce simpler expressions was introduced after the publication of our method. We expect this improved version to perform better on the Feynman benchmark.

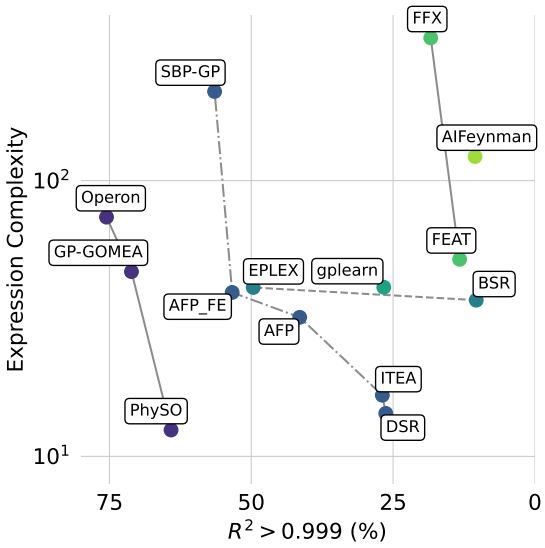


Figure 4.4: **Complexity vs. accuracy performances of `PhySO` against baseline methods on the Feynman benchmark.** Complexity versus rate of expressions having $R^2 > 0.999$ at a 10% noise level for `PhySO` and other symbolic regression methods from the literature on the Feynman benchmark [La Cava et al., 2021]. Lines and colors denote the 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th} and 6^{th} Pareto fronts following the `SRBench` algorithm comparison framework. `PhySO` is a Pareto optimum producing simple yet effective expressions.

accurate expressions normalized by the number of free constants plus 1. `PhySO` emerges as the leading method in generating succinct, physically coherent, and interpretable expressions that best approximate a dataset, that is when it is not able to recover the exact underlying expression all together.

This is further illustrated in Figure 4.4, where we show Pareto frontiers of expression complexity versus fit quality at a 10% noise level for all benchmarked methods with available output expression information. On this plot `PhySO` is a Pareto optimum demonstrating its ability to produce simple yet good-fitting expressions.

4.3.4 Learning curves

Due to its very constraining nature, using a yet untrained neural network, our *in situ* units prior often conflicts with the length prior which is essential to avoid the expression generation phase going on forever. This typically results in the majority of expressions being discarded due to this conflict during the first iterations of the training process. However, enabling the neural network to learn on physically correct expressions, and enabling it able to observe local units constraints, allows it to actively learn dimensional analysis rules.

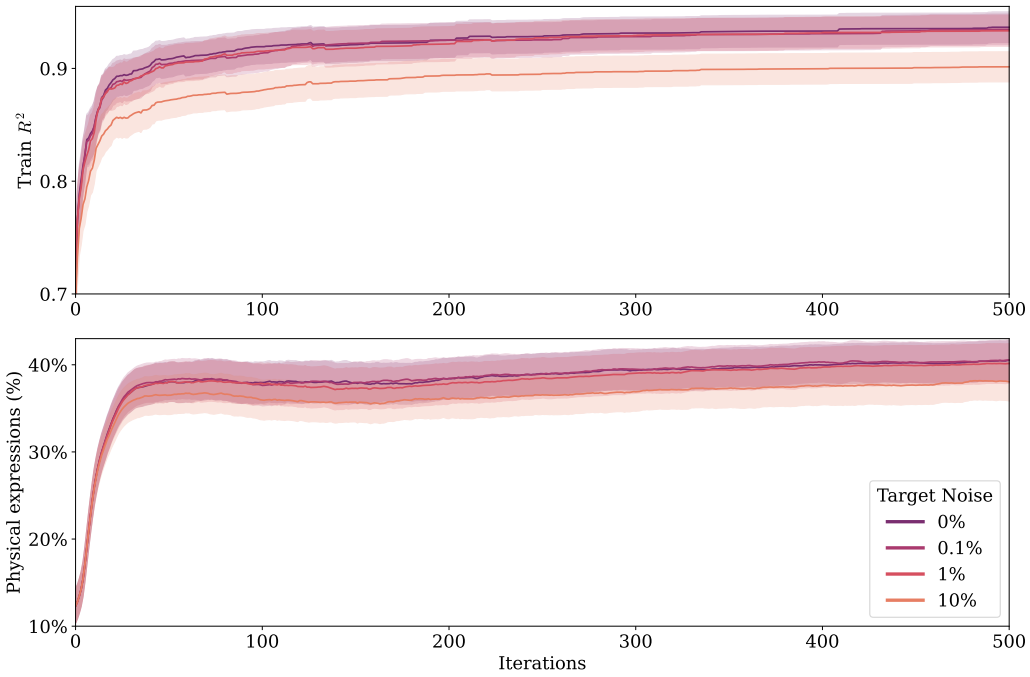


Figure 4.5: **Learning curves.** R^2 value on training data and percentage of expressions natively proposed by the neural network that have balanced physical units averaged across the Feynman benchmark with error regions indicating a 95% confidence interval. Φ-SO's neural network learns to produce not only good fitting expressions but also physically meaningful ones.

This is shown in Figure 4.5, which gives the fraction of physical expressions successfully generated over iterations of learning averaged over all runs of the Feynman benchmark at each level of noise.

Moreover, Figure 4.5 presents the evolution of the R^2 fit coefficient on training data for the best expression identified at each iterations. The figure demonstrates that as the iterations progress, the neural network not only improves in generating expressions with better fits but also refines its capacity to produce expressions that are physically meaningful.

In our observations, while Φ-SO occasionally escapes local minima through stochastic variations, convergence is typically characterized by the neural network mostly producing identical expressions. This state of convergence is typically reflected by both average fit quality and rate of physical expression remaining static, as well as by the reward distribution peaking. The rate of convergence is dependent on the difficulty of the case, the level of noise and the chosen hyper-parameters. As depicted in Figure 4.5, under the hyper-parameters detailed in this study, Φ-SO typically reaches convergence well within several hundred iterations. Note that since it is operating in a rein-

forcement learning framework, Φ-SO is trained on 'moving targets' as its targets consist of expressions generated by itself during the last iteration which is characterized by the loss not decreasing during training except when it starts consistently producing similar equations while converging.

4.4 Astrophysical Case Studies

In this Section, we showcase our Φ-SO method on a panel of astrophysical test cases: the relativistic energy of a particle is examined in sub-section 4.4.1, the law describing the expansion of the Universe in sub-section 4.4.2, the isochrone action from galactic dynamics in sub-section 4.4.3 and additional toy test cases given in 4.4.4. We give the results along with an ablation study, disabling specific components our system to determine their impact on performance, in 4.4.5. We perform this ablation study in a noiseless scenario using mock data but still demonstrate Φ-SO's abilities on observational noisy data for the case detailed in sub-section 4.4.2, showing that the method can successfully recover physical laws and relations from real or synthetic data. Mock data generation details are given in sub-section 4.4.6 along with units of all variables and constants involved.

Protocol

Note that for each of these showcases, we explicitly add the free constants described in sub-section 4.4.6 along with their units in Φ-SO's library of available tokens. We use the hyper-parameters and and reward metric detailed in 3.3 and limit ourselves to the exploration of 10 million trial expressions which roughly takes $\sim$ 4 hours (using the parallelization feature and the computational systems examined in paragraph 6.2) and is only necessary for the most difficult case (the relativistic energy). In addition, for the relativistic energy showcase, we give a Pareto front which shows the most accurate expression based on RMSE (root mean squared error) for each level of complexity. Moreover, similarly to the benchmarking in Section 4.3, we define the successful exact symbolic recovery of an expression by its symbolic equivalence using the `SymPy` symbolic simplification subroutine [Meurer et al., 2017]. Finally, we agnostically rely on the same library of choosable tokens for all test cases: $\{+, -, \times, /, 1/\square, \sqrt{\square}, \square^2, \exp, \log, \cos, \sin, 1\}$ to which we only add input variables and free or fixed constants depending on the test cases.

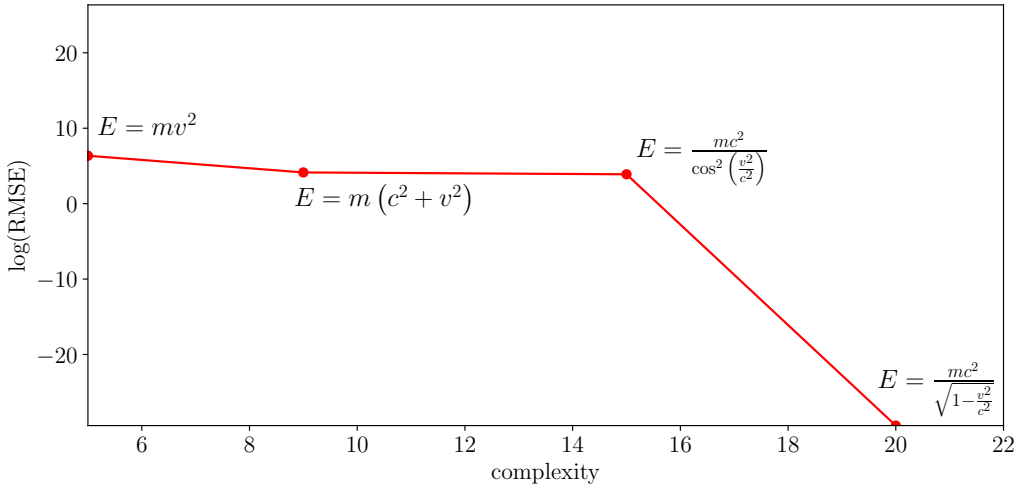


Figure 4.6: **Pareto-front encoding accuracy-complexity trade off of recovered physical formulae typically recovered using our Φ-SO method when applied to data for the relativistic energy of a particle.** We recover the relativistic expression as well as the classical approximation. Note that although the exact classical expression $\frac{1}{2}mv^2$ is encountered by Φ-SO it is Pareto dominated by the simpler mv^2 expression.

4.4.1 Relativistic energy of a particle

Let us consider the expression for the relativistic energy of a particle:

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}\,, \tag{4.6}$$

where m, v and c are respectively the mass of the particle, its velocity and the speed of light.

Using the aforementioned library of tokens as well as the $\{m, v\}$ input variables and a free constant $\{c\}$, Φ-SO is able to successfully recover this expression 100% of the time. Figure 4.6 contains the Pareto front of recovered expressions where similarly to Udrescu et al. [2020], we showcase that we are able to recover the relativistic energy of a particle as well as the classical approximation which has a lower complexity.

However, we note that our system is able to recover the exact expression for the relativistic energy test case without any of the powerful simplification on which relies the `AI Feynman 2.0` approach proposed in [Udrescu et al., 2020] (in particular, the identification of symmetries as well as the identification of additive and multiplicative separability), nor by simplifying the problem further by treating c (a constant of nature) as a variable taking a range of different values as in [Udrescu et al., 2020]. Neither `DSR` [Landajuela et al.,

2021a] nor `AI Feynman` [Udrescu et al., 2020] are able to crack this case under these more stringent conditions.

4.4.2 Expansion of the Universe

The next case study we examine is the Hubble Diagram of supernovae type Ia, namely the change in the observed luminosity of these important standard candles as a function of redshift z. This is one of the major pieces of evidence that indicates that the Universe is experiencing an accelerating expansion, and it is also one of the observational pillars underlying Λ Cold Dark Matter (ΛCDM) cosmology in which Dark Energy dominates the energy-density budget of the Universe.

We will use the so-called *Pantheon* state-of-the-art compilation dataset [Scolnic et al., 2018], shown in Figure 4.7. We use a similar calibration and follow an almost identical methodology as Bartlett et al. [2023a], to find the Hubble parameter $H(z)$ from the measured supernova magnitude and redshift pairs. Following Bartlett et al. [2023a], we use the auxiliary function

$$y(x \equiv 1 + z) \equiv H(z)^2 \,, \tag{4.7}$$

which for ΛCDM in a flat Universe with negligible radiation pressure is

$$y_{\Lambda\mathrm{CDM}}(x) = H_0^2(\Omega_m x^3 + (1 - \Omega_m)) \,, \tag{4.8}$$

where Ω_m is the matter density parameter and H_0 is the Hubble constant. In a flat Universe model the cosmological luminosity distance is

$$d_L(z) = (1 + z) \int_0^z \frac{c\, dz'}{H(z')} \,, \tag{4.9}$$

where c is the speed of light.

We adapt our machinery to the Hubble diagram problem by integrating numerically the $H(z \equiv x - 1) = \sqrt{y(x)}$ functions proposed by the algorithm under Equation 4.9 to derive the implied luminosity distance d_L. These are then trivially converted into a distance modulus $\mu(z) = 5 \log_{10}(d_L(z)/10\,\mathrm{pc})$, which we compare to the *Pantheon* data following the procedure given in subsection 3.3.2.

This Hubble Diagram example showcases the capability of the software to include free "constants" (here we include one having the units of H_0 and the other being dimensionless as Ω_m) in the expression search, whose values are found thanks to auto-differentiation via L-BFGS optimization, as mentioned in Section 3.3.2. The optimal values of these constants need to be calculated after being passed through the numerical integration step (integrating Eqn. 4.9 via `PyTorch` differentiable cumulative trapezoids), which turns out to be the

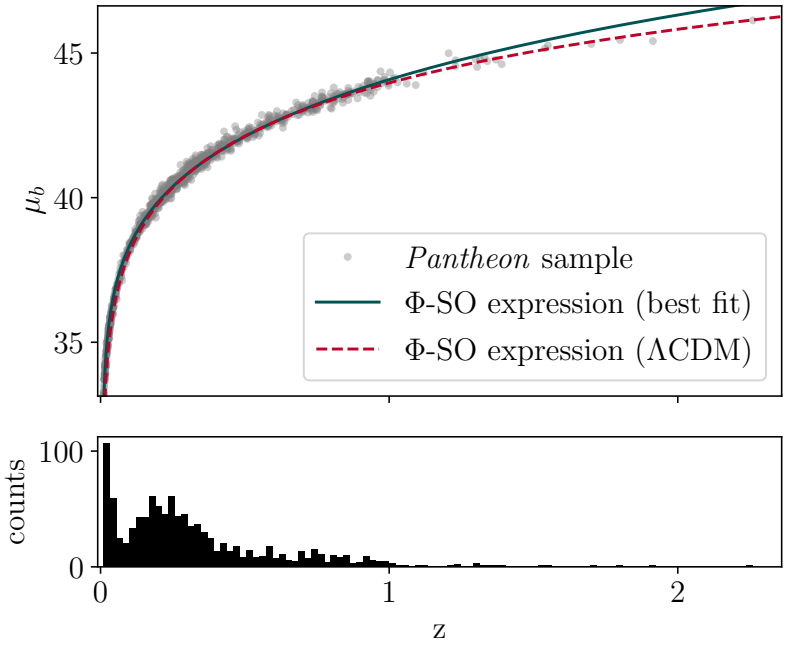


Figure 4.7: **Fit found applying Φ-SO to supernovae Ia data.** SR results when applying the Φ-SO algorithm (allowing two free parameters) to the Hubble Diagram of supernovae Ia from the *Pantheon* sample. Φ-SO rediscovers the ΛCDM relation (in red) as well as another relation (in blue) which has a slightly better fit than ΛCDM when solely considering *Pantheon*'s observational constraints due to the over-abundance of low z SNe.

Expression	Complexity	$\widetilde{H}$	c	R^2
$\widetilde{H}^2\sqrt{c^2+\log(c+x)}$	14	5.175	-0.01	0.9955
$\widetilde{H}^2\sqrt{c+x}$	9	4.692	-1.01	0.9946
$\widetilde{H}^2\log(x)$	6	7.499	-	0.9627
$\widetilde{H}^2\log^2(x)$	8	28.276	-	0.9523
$\widetilde{H}^2(cx^3-c+1)$ [a]	14	73.3	0.315	0.9166

[a] ΛCDM expression for reference.

Table 4.5: **Accuracy vs. complexity results applying Φ-SO to supernovae Ia data.** Pareto accuracy-complexity trade-off expressions (for the auxiliary function $y(x \equiv 1+z) \equiv H(z)^2$) applying the Φ-SO algorithm (allowing two parameters) to the Hubble Diagram of supernovae Ia from the *Pantheon* sample. Although Φ-SO generates the ΛCDM expression, it is not a Pareto optimum when solely considering *Pantheon*'s observational constraints due to the over-abundance of low z SNe. We include it for reference as the last line of this Table.

main bottleneck of the problem in terms of computational cost. However, this also shows that the algorithm allows one to derive expressions that are subsequently passed through complicated operations (such as an integral in this example) before being compared to data.

The Pareto front is given in Table 4.5 alongside the ΛCDM expression. Although we are able to recover it using synthetic data, we note that as Bartlett et al. [2023a], using observational data our system finds more accurate solutions at lower complexities than the ΛCDM model. Although this could signify that the ΛCDM theory is inaccurate, here we refrain from jumping to this conclusion because our system is only given the chance to confront its trial model of $H(z)$ to a relatively noisy dataset of standard candles where there is an over abundance of low z events, and is not provided other observational constraints such as the cosmic microwave background which might tilt the balance in favor of ΛCDM as the most accurate model at its level of complexity. However, although the ΛCDM expression is not the global minimum with this set of observational constraints, while exploring a space of increasingly accurate expressions our system recognizes it as an intermediate step, recording it in its history, before eventually converging to a different expression. In addition, we note that it is not surprising that our system recovers the ΛCDM expression as we allowed a maximum of two free parameters since the main goal was simply to demonstrate our system's capabilities. We defer multi-parameter studies to future contributions. Finally we are able to recover this expression by typically exploring < 50k expressions (which takes less than a minute on the computational systems examined 6.2), the same order of magnitude as in the exhaustive symbolic regression approach proposed in Bartlett et al. [2023a] but allowing more functions $(\cos, \sin, \exp, \log)$.

4.4.3 Isochrone action from galactic dynamics

Another interesting application of symbolic regression is to derive perfect analytical properties of analytical models of physical systems. To this end, we chose to attempt to find the radial action J_r of the spherical isochrone potential.

$$\Phi(r) = -\frac{GM}{b + \sqrt{b^2 + r^2}}\,, \tag{4.10}$$

where G is the gravitational constant, M is the mass of the model, b is a length scale of the model, and r is a spherical radius [Binney and Tremaine, 2011]. Action variables are special integrals of motion in integrable potentials which can be used to describe the orbit of an object in a system, and they are of particular interest in Galactic Archaeology as they are adiabatic invariants,

so they are preserved if a galaxy or stellar system has evolved slowly. The isochrone is the only potential model to have actions known in analytic form in terms of elementary functions[8]. For the case of the isochrone model, the radial component of the action of a particle can be expressed as

$$J_r = \frac{GM}{\sqrt{-2E}} - \frac{1}{2}\left(L + \frac{1}{2}\sqrt{L^2 - 4GMb}\right), \tag{4.11}$$

where E and L are, respectively, the particle energy and total angular momentum [Binney and Tremaine, 2011].

We provide our algorithm numerical values of J_r (which has units of angular momentum) given L and E, and leave b as a free scaling parameter. Since we expect each occurrence of M to be accompanied by an occurrence of the gravitational constant, we provide the algorithm with GM as a single variable.

This expression (Eqn. 4.11) could not be solved either by the standard `DSR` algorithm [Landajuela et al., 2021a], or by the `AI Feynman 2.0` algorithm [Udrescu and Tegmark, 2020]. Our algorithm was also not able to identify the equation in 10 million guesses. However, one of the steps of the `AI Feynman 2.0` algorithm is a test for additive and multiplicative separability of the mystery function, and it creates new datasets for each separable part. For the case of additive separability, the units remain unchanged, and so it is trivial to simply provide our Φ-SO algorithm separated data generated by `AI Feynman 2.0` to be fitted in turn, one at a time. Thus the first term of the right hand side of Eqn. 4.11 (with an E dependence) was easily solved together with a fitted additive free constant. We then subtracted the fitted constant from the second dataset, and Φ-SO correctly recovered the second term on the right hand side of Eqn. 4.11 (with an L dependence).

4.4.4 Supplementary cases

In addition to the cases above, we consider the following set of textbook equations for the ablation study in 4.4.5. We include Newton's law of universal gravitation:

$$F = \frac{Gm_1m_2}{r^2}, \tag{4.12}$$

where G is the universal gravitational constant, m_1 and m_2 are the masses of the attracting bodies and r is the distance separating them. For this test case, we use $\{m_1, m_2, r\}$ as input variables and leave G as a free constant.

We also include a damped harmonic oscillator which appears in a wide range of (astro)-physical contexts:

$$y = e^{-\alpha t}\cos(\omega t + \Phi), \tag{4.13}$$

[8]It has recently been shown that actions can be calculated numerically from samples of points along orbits in realistic galaxy potentials using deep learning techniques [Ibata et al., 2021].

Ablation configuration		A[a]	B	C	D[b]	E	F[c]
Physical units prior		✓		✓		✓	
Physical units informed neural network		✓	✓				
Neural network enabled		✓	✓	✓	✓		
Expression	**# expressions**						
$E = \frac{mc^2}{\sqrt{1-v^2/c^2}}$	10M	100 %	0 %	60 %	0 %	20 %	0 %
$J_r = \frac{GM}{\sqrt{-2E}} - \frac{1}{2}\left(L + \frac{1}{2}\sqrt{L^2 - 4GMb}\right)$	4M	100 %	0 %	80 %	0 %	60 %	0 %
$\rho = \rho_0 / \left(\frac{r}{R_s}(1+\frac{r}{R_s})^2\right)$	2M	100 %	100 %	40 %	100 %	20 %	100 %
$y = e^{-\alpha t}\cos(\omega t + \Phi)$	1M	100 %	0 %	0 %	0 %	0 %	0 %
$F = \frac{Gm_1m_2}{r^2}$	100K	100 %	80 %	100 %	20 %	80 %	0 %
$H^2(x \equiv 1+z) = H_0{}^2(\Omega_m x^3 + (1-\Omega_m))$	100K	100 %	100 %	100 %	100 %	40 %	40 %
	Average	**100 %**	**47 %**	**63 %**	**37 %**	**37 %**	**23 %**

[a] Full Φ-SO method.

[b] Similar to Landajuela et al. [2021a].

[c] Solely relying on a random number generator.

Table 4.6: **Ablation study.** Exact symbolic recovery rate summary and ablation study on our panel of astrophysical examples using noiseless synthetic data, averaged across 5 runs. By studying the performance in combinations of ablations of the *in situ* units prior, the neural networks's ability to be informed of local units constraints, and of the neural network itself (i.e. replaced by a random number generator when not marked as enabled), we show that all three are essential ingredients of the success of our Φ-SO method. Input variables and free parameters are colored in red and blue respectively with fixed constants left in black.

where α and ω are respectively the damping parameter and the angular frequency of oscillations (both homogeneous to the inverse of a time) and Φ is the (dimensionless) phase. We leave these three parameters as free constants and use t as our input variable.

Finally, we consider a Navarro–Frenk–White (NFW) halo profile [Navarro et al., 1996] which is an empirical relation that describes the density profile $\rho(r)$ of halos of collisionless dark matter in cosmological N-body simulations:

$$\rho = \frac{\rho_0}{\frac{r}{R_s}\left(1+\frac{r}{R_s}\right)^2}\,, \tag{4.14}$$

where r is the radius which we use as an input variable and ρ_0 and R_s are respectively the density and radius scale parameters which we leave as free constants.

4.4.5 Ablation study

In physics, we often seek to build approximate models, such as might be obtained via a polynomial function or a Fourier series fit to some data. In those instances, the root mean square error is usually the criterion of relevance to determine whether the procedure worked well or not. However, here we wish

to recover the "true" underlying model, in which case the recovery rate should be the criterion of success.

The performance of Φ-SO on noiseless mock data from the test cases detailed above is summarized in the ablation study reported in Table 4.6. There we also report SR performance after disabling the units prior (only using the units informed RNN), disabling the RNN's ability to be informed of local units units constraints (only using the units prior and a standard SR RNN), disabling both the units prior and units information (only using a standard RNN which is similar to the Landajuela et al. 2021a setup), doing a units guided random search by using a random number generator *in lieu* of the RNN, and finally doing a purely random search.

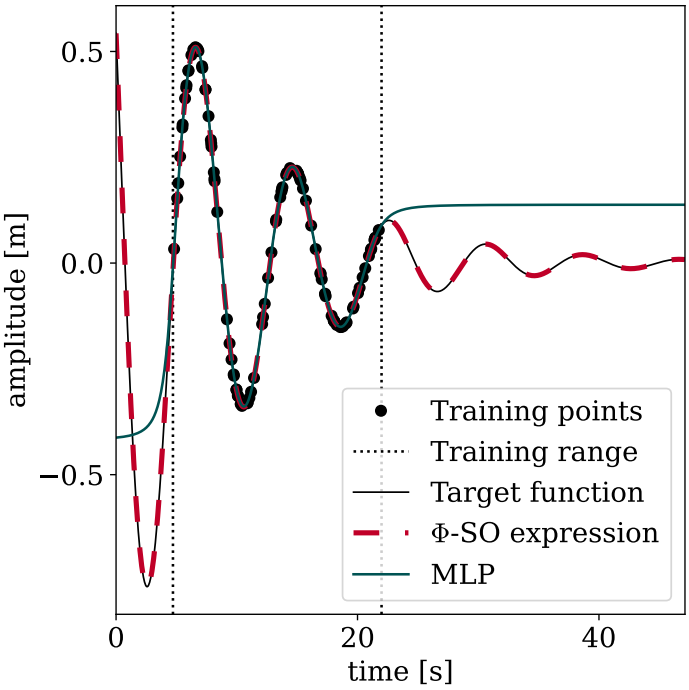


Figure 4.8: **Generalization capability illustration.** Example of the generalization capability of SR. Here we show randomly drawn data points (black dots) from the damped harmonic oscillator model given in Eqn. 4.13 (black line). The data are well fitted by an MLP (green line), which however fails in regions beyond the range of the training data (vertical dotted lines). In contrast, our SR algorithm Φ-SO (red dashed line) manages to provide much more reliable extrapolation.

We show that merely constraining the choice of symbols using the external units prior distribution scheme (described in 4.2) is not enough to ensure perfect symbolic recovery of physical laws, but that informing the RNN of local units constraints (as described in 4.2.2) is essential as it allows the RNN

to actively learn units rules. In addition, we show that our system does not only rely on a mere brute force approach combined with units constraints, but that the deep reinforcement learning setup described in 4.2.2 is an essential ingredient of the success of Φ-SO.

It should be noted that in the NFW test case, simply expressing the inverse of a third-degree polynomial is sufficient to solve the problem. However, using the units prior without enabling the RNN to observe local units constraints or utilizing the units prior in conjunction with a random number generator can result in a lower recovery rate compared to the use of a standalone random number generator. This is due to the highly restrictive nature of the units prior which in a simple case like this can actually slow down the convergence toward the solution.

Finally, we also illustrate the generalization capabilities offered by virtue of finding the exact analytical expression[9] underlying a dataset compared to a good approximation in Figure 4.8, where we show that such analytical expressions, as expected, vastly outperform a multilayer perceptron (MLP) neural network (here a 5 layers of 32 units MLP having sigmoid activations and being trained until convergence on a test set, following a mean squared error loss function at $10^{-}3$ learning rate using an Adam optimizer, Kingma and Ba 2015).

4.4.6 Datasets details

This sub-section gives details regarding the synthetic datasets for the astrophysical examples. For each case, we generate 1000 noiseless data points following a random uniform law using arbitrary scales for the mock data. Table 4.7 gives the target expressions and Table 4.8 and 4.9 give details regarding the variables and constants appearing in those expressions.

[9] We will also show that symbolic approximations also tend to outperform neural networks in generalization capabilities in 7.2.

Case	Expression
Relativistic Energy	$E = \frac{mc^2}{\sqrt{1-v^2/c^2}}$
Isochrone Action	$J_r = \frac{GM}{\sqrt{-2E}} - \frac{1}{2}\left(L + \frac{1}{2}\sqrt{L^2 - 4GMb}\right)$
NFW Profile	$\rho = \rho_0 / \left(\frac{r}{R_s}(1+\frac{r}{R_s})^2\right)$
Damped Harmonic Oscillator	$y = e^{-\alpha t}\cos(\omega t + \Phi)$
Classical Gravity	$F = \frac{Gm_1m_2}{r^2}$
Expansion Law	$H^2(x \equiv 1+z) = {H_0}^2(\Omega_m x^3 + (1-\Omega_m))$

Table 4.7: **Astrophysical examples target expressions.** Input variables and free parameters choosable by Φ-SO as symbols are colored in red and blue respectively, with fixed constants left in black.

Output		Variable 1			Variable 2			Variable 3		
Name	Units	Name	Range	Units	Name	Range	Units	Name	Range	Units
E	$M.L^2.T^{-2}$	m	[-10,10]	M	v	[-9,9]	$L.T^{-1}$			
J_r	$L^2.T^{-1}$	L	[2.3, 3]	$L^2.T^{-1}$	E	[-4, -6]	$M.L^2.T^{-2}$			
ρ	$M.L^{-3}$	r	[0.2, 3]	L						
y	1	t	[1.5π, 7π]	T						
F	$M.L.T^{-2}$	m_1	[0,1]	M	m_2	[0,1]	M	r	[1,4]	L
H^2	T^{-2}	z	[0.01, 2.5]	1						

Table 4.8: Data range and units of the output and input variables appearing in the astrophysical examples.

Constant 1			Constant 2			Constant 3		
Name	Value	Units	Name	Value	Units	Name	Value	Units
c	10	$L.T^{-1}$						
GM	0.467	$L^3.T^{-2}$	b	1.234	L			
r_s	1.391	L	ρ_0	0.984	$M.L^{-3}$			
ω	0.784	T^{-1}	α	0.101	1	ϕ	0.997	1
G	1.184	$L^3.M^{-1}.T^{-2}$						
H_0	1.072	T^{-1}	Ω	1.315	1			

Table 4.9: Target value and units of constants appearing in the astrophysical examples.

4.5 Discovering Both Analytical Laws & Constants of Nature

We note that for new scientific discovery, there are instances where the appropriate free parameters and their corresponding units are not immediately evident. In such situations, we propose a protocol wherein Φ-SO is allowed one free parameter for each input variable, sharing the same units, and

another free parameter reflecting the units of the output variable. Specifically, for an SR problem consisting in the deduction of y from $\{x_1, ...x_n\}$, we would permit the inclusion of $\{c_y, c_{x_1}, ...c_{x_n}\}$ as free constants. This grants Φ-SO the flexibility to selectively combine or omit these free parameters to construct new parameters that align with dimensional analysis constraints. In light of these combinations, we adjust the center of the soft length prior to a length of 12, facilitating longer expressions.

In this more demanding setup, we demonstrate that Φ-SO can adeptly resolve the SR challenges outlined in Table 4.10 (with dataset details given in Table 4.11 and 4.12), yielding both the precise symbolic expressions and their corresponding physical constants with accurate units. The scripts employed for these experiments are accessible in our repository.

Case	Expression
Ideal Gas Law	$P = \frac{nRT}{V}$
Free Fall Terminal Velocity	$v_t = \sqrt{\frac{2mg}{\rho A C_d}}$
Classical Gravity	$F = \frac{Gm_1m_2}{r^2}$
Black Body Photon Count	$n = 1/(e^{\frac{h\nu}{k_bT}} - 1)$
Wave Interference	$E = E_1 + E_2 + 2\sqrt{E_1E_2} + \cos\Delta\Phi$

Table 4.10: **Target expressions.** Input variables are colored in red.

Output		Variable 1			Variable 2			Variable 3		
Name	Units	Name	Range	Units	Name	Range	Units	Name	Range	Units
P	$L^{-1}.T^{-2}.M$	n	[1,5]	N	T	[1,5]	Θ	V	[1,5]	L^3
v_t	$L.T^{-1}$	m	[1, 10]	M	ρ	[1, 6]	$M.L^{-3}$	A	[1, 5]	L^2
F	$L.T^{-2}.M$	m_1	[1, 5]	M	m_2	[1, 5]	M	r	[1,5]	L
n	1	ν	[1, 5]	T^{-1}	T	[1, 5]	Θ			
E	L	E_1	[1, 5]	$L^2.T^{-2}.M$	E_2	[1, 5]	$L^2.T^{-2}.M$	$\Delta\Phi$	[-5, 5]	1

Table 4.11: Data range and units of the output and input variables appearing in the examples.

For illustration, Φ-SO successfully derives the equation describing the equation of state of an ideal gas $P = C\frac{nT}{V}$ with $C = \frac{c_P c_V}{c_n c_T}$ having units $M.L^2.T^{-2}.K^{-1}.N^{-1}$ effectively rediscovering the ideal gas constant usually denoted by R. Similarly, Φ-SO is able to recover the expression for the terminal velocity of a free falling object as a function of its mass m, its surface area A and the density of the medium it traverses ρ as $v_t = \sqrt{C\frac{m}{\rho A}}$ by unveiling its proportionality to the square root of an acceleration $\sqrt{C}$,

Constant 1			Constant 2		
Name	Value	Units	Name	Value	Units
R	8.314	$L^{-2}.T^{-2}.M.N^{-1}.\Theta^{-1}$			
g	9.807	$L.T^{-2}$	C_d	0.470	1
G	6.674	$L.T^{-2}.M$			
h	6.626	$L^2.T^{-1}.M$	k_b	1.123	$L^2.T^{-2}.M.\Theta^{-1}$
-	-	-			

Table 4.12: Target value and target units of constants appearing in the examples.

formulated by Φ-SO as $\sqrt{c_{v_t}/\sqrt{c_A}}$, corresponding to the Earth surface gravity $\sqrt{g}$ and other scale factors. Furthermore, Φ-SO identifies the gravitational force in relation to the involved masses m_1, m_2 and distance r as $F = Cm_1m_2/r^2$ discovering the need for a constant C having units $L^3.T^{-2}.M$ formulated by Φ-SO as $C = c_F{c_r}^2/{c_{m_1}}^2$, effectively rediscovering the gravitational constant G in the process. In an other scenario, deriving the number density of photons recovered from a black body at any given temperature T and frequency ν, Φ-SO is able to recover $n = 1/(e^{\nu C/T} - 1)$ where C represents the quotient $C = h/k_b$, h and k_b denoting the Planck and Boltzmann constants, respectively. In most aforementioned cases, Φ-SO judiciously combined a subset of the available free parameters to pinpoint the precise constants needed to resolve the SR problems through a physically consistent physical law. In this last example, we show that Φ-SO recognizes scenarios where free parameters are largely redundant as it is able to derive the energy E resultant from the interference of two waves, given their energies E_1, E_2 and their phase shift $\Delta\Phi$ without the need for any of $\{c_E, c_{E_1}, c_{E_2}\}$.

4.6 Discussion & Conclusions

Overcoming the curse of accuracy-guided SR by constraining symbolic arrangement

Since the Deep Symbolic Regression framework [Petersen et al., 2021a] and most other SR methods work by maximizing fit quality, there are few constraints on the arrangement of symbols. However, the paths in fit quality and the paths in symbol arrangement toward the global minima (perfect fit quality and perfect symbol arrangement) are not necessarily correlated. This results in the curse of accuracy guided SR, as small changes in fit quality

can hide dramatic changes in functional form and vice-versa. In essence, one can improve fit quality of candidates over learning iterations while getting further away from the correct solution in symbolic arrangement. Therefore strong constraints on the functional form, such as the one we are proposing in our setup, are of great value for guiding SR algorithms in the context of physics. This is an advantage that physics has and that Φ-SO leverages by: (i) reducing the search space and (ii) enabling the neural network to actively learn dimensional analysis rules and leverage them to explore the space of solutions more efficiently. Although the possibility of making a physical units prior was hinted by Petersen et al. [2021b], to the best of our knowledge such a framework was never built before.

Dimensional analysis in SR

The guidance offered by the units constraints gives Φ-SO an edge over other methods for finding the exact symbolic solutions, improving performance from a purely predictive standpoint. This makes Φ-SO a potentially useful tool for opening up black-box physics models such as neural networks fitted on data of physical phenomena. In addition, we note that in the context of physics, components of our Φ-SO framework can not only be used to improve the performance of algorithms built upon Petersen et al. [2021a]'s framework [Landajuela et al., 2022, 2021a, DiPietro and Zhu, 2022, Du et al., 2022], but can also be used in tandem with other approaches. For instance, our *in situ* units prior can be used to reduce search space in the context of probabilistic or exhaustive searches [Bartlett et al., 2023a, Kammerer et al., 2020, Brence et al., 2021, Jin et al., 2019], by severing physically impossible symbolic links in neuro-symbolic approaches [Martius and Lampert, 2017, Brunton et al., 2016, Zheng et al., 2022, Sahoo et al., 2018, Valle and Haddadin, 2021, Kim et al., 2020, Panju and Ghodsi, 2020], during the seeding or mutation phases of genetic programming algorithms [Schmidt and Lipson, 2009, 2011, de Franca and Aldeia, 2021, La Cava et al., 2019, Cava et al., 2019, Virgolin et al., 2019, Cranmer, 2023, Cranmer et al., 2020b, Virgolin et al., 2021, Stephens, 2015, Kommenda et al., 2020, Landajuela et al., 2022] or for making a physically motivated dataset of expressions, which in conjunction with enabling the RNN to be informed of local units constraints, could improve the performance of supervised approaches [Kamienny et al., 2022, Biggio et al., 2021, 2020, Vastl et al., 2022, Becker et al., 2022, Kamienny et al., 2023, Landajuela et al., 2022].

Physical units of free parameters

We recognize that in its current form, Φ-SO needs to be provided the physical units of the free parameters it is allowed to use. Although this is typically not an issue for SR problems that tend to fall on the more theoretical side as constants that can appear in expressions if any are usually well known, in scenarios of novel empirical scientific exploration, the appropriate selection and units of free parameters may not be immediately evident. In such scenarios, we suggest the inclusion of one free parameter for each variable, matching their units. This approach grants Φ-SO the flexibility to combine these parameters, or a subset thereof, to derive the most coherent combination that seamlessly integrates into the expression from a units perspective. As detailed in Section 4.5, utilizing this protocol enables Φ-SO to accurately deduce formulae and the physical constants appearing in those. Examples include the recovery of expression for the terminal velocity during free fall, and its proportionality with the square root of an acceleration, by adeptly combining a velocity with an area to derive the acceleration parameter. In other examples, we show that Φ-SO is able to effectively rediscover the universal gravitational constant or the ideal gas constant along with their units in addition to the expressions they intervene in.

Arguably, permitting a multitude of free parameters of various physical units, could inadvertently expand the search space. While this is a valid observation, it is worth noting that the algorithm remains significantly constrained, both by the limited assortment of these parameters and by the inherent units constraints between input variables, especially when considering dimensionless operations like $\cos, \exp$ and so forth. Moreover, given that the algorithm combines parameters based on the units of the variables and prioritizes solutions of lower complexity, the units of new physical constants typically align closely with the family of units of the problem, rather than assuming arbitrary values. Finally, it is worth noting that in addition to dimensional analysis constraints, another key finding of our study is that making the neural network able to observe units of symbols and currently required units in partially written expressions while they are being generated typically improves the recovery rate even without enforcing constraints directly. However, resolving SR problems without knowing *a priori* the units of the free parameters that can appear in the expressions is typically more difficult. We acknowledge this limitation and are actively considering future enhancements to Φ-SO that would enable it to intelligently and autonomously ascertain the units of its free parameters.

Conclusion

We have presented a new symbolic regression algorithm, built from the ground up to make use of the highly restrictive constraint that we have in the physical sciences that our equations must have balanced units. The heart of the algorithm is an embedding that generates a sequence of mathematical symbols while cumulatively keeping track of their physical units. We adopt the very successful deep reinforcement learning strategy of Petersen et al. [2021a], which we use to train our RNN to not only produce accurate expressions but physically sound ones by making it learn local units constraints.

The algorithm was benchmarked and compared to 17 other baseline SR approaches on 120 cases from the Feynman Lectures on Physics and other textbooks. The results demonstrated the usefulness of constraints arising from dimensional analysis compared to Petersen et al. [2021a], a purely reinforcement learning based baseline approach. In addition our approach achieved state-of-the-art leading performances in the presence of even minimal levels of noise (exceeding 0.1%) and showing consistent performances up to 10% noise levels.

The algorithm was applied to several test cases from astrophysics. The first was a simple search for the energy of a particle in Special Relativity (Section 4.4.1), which our algorithm was able to find, yet is a problem that the standard Petersen et al. [2021a] code fails on. The second test case applied the algorithm to the famous Hubble diagram of supernovae of type Ia. While the form of the Hubble parameter $H(z)$ in standard ΛCDM cosmology was indeed recovered, the algorithm finds that other simpler solutions fit the supernova data (in isolation) better. This result is consistent with the findings of Bartlett et al. [2023a]. Another test examined a relatively complicated function in galactic dynamics, where we searched for the functional form of the radial action coordinate in an isochrone stellar potential model. This is an equation that neither the Petersen et al. [2021a] nor the Udrescu et al. [2020] methods are able to find. Although our algorithm initially fails in this test, we managed to recover the correct equation by first splitting the dataset using the additive separability criterion as implemented by Udrescu and Tegmark [2020].

These tests have demonstrated the applicability of the algorithm to model data of the real world as well as to derive non-obvious analytic expressions for properties of perfect mathematical models of physical systems. Although we realise that the physical laws potentially discovered by our method will depend on data range, choice of priors, etc, this is a step toward a full agnostic method for connecting observational data to theory. Future contributions in this research program will extend the algorithm to allow for differential and integral operators, potentially permitting the solution of ordinary and partial differential equations with physical units constraints. However, our primary goal will be to use the new machinery to discover as yet unknown physical relationships

from the state-of-the-art large surveys that the astrophysical community has at its disposal.

Chapter 5

Class Symbolic Regression

Portions of the content presented in this Chapter have been previously discussed in the following publication:

2024 *Class Symbolic Regression: Gotta Fit 'Em All*
W. Tenachi, R. Ibata, T. L. François, F. Diakogiannis
ApJL **969** L26, arXiv:2312.01816

Summary.

We introduce Class Symbolic Regression a first framework for automatically finding a single analytical functional form that accurately fits multiple datasets — each realization being governed by its own (possibly) unique set of fitting parameters. This hierarchical framework leverages the common constraint that all the members of a single class of physical phenomena follow a common governing law.

Additionally, we introduce the first Class SR benchmark, comprising a series of synthetic physical challenges specifically designed to evaluate such algorithms. We demonstrate the efficacy of our novel approach by applying it to these benchmark challenges and showcase its practical utility for astrophysics by successfully extracting an analytic galaxy potential from a set of simulated orbits approximating stellar streams.

Our modern computational abilities have allowed us to examine nature in unprecedented quantitative detail, with cameras, spectrographs and other detectors amassing vast quantities of numerical data. It is likely that the clues to next-generation physics and understanding lie therein, and so we are tasked to devise methodologies capable of handling this wealth of information and translating it into coherent, interpretable and intelligible physical models. Symbolic regression (SR) — which is defined as the search of an analytic description of that best fits a dataset — may allow us in part to answer this need to find accurate and intelligible empirical laws in giant datasets to best capitalize on the community's observational investments.

However, as pointed out in sub-section 3.1.3 the search space becomes exponentially larger the longer the analytic expression is that we seek to find. Hence the key to SR is to develop efficient schemes to search through the possibilities, and most importantly, to prune out poor choices.

While SR has been extensively applied in scientific research, its focus has largely been on single dataset analysis, overlooking the rich potential in examining multiple datasets linked to a specific physical phenomenon. The present Chapter extends our Φ-SO framework for Physical Symbolic Optimization (presented in Chapters 3 and 4) further by allowing the search for a functional form that can simultaneously fit several datasets at once, each realization having (possibly) different fitting parameters. This "Class Symbolic Regression" (Class SR) approach opens up the new possibility of implementing a functional search on the properties of a *class* of objects. This hierarchical framework leverages the common constraint that all the members of a single class of physical phenomena follow a common governing law.

This approach is relevant across various natural sciences, but it particularly shines in astrophysics, where multiple observations of a single phenomenon are often available, providing a rich multi-dataset setup enabling us to devise 'universal' laws that apply to a range of celestial objects of interest.

In particular, we apply this new framework to the recovery of a Milky Way-like analytic galactic potential from simulated orbits that can be inferred from stellar streams. Specifically, our approach recovers a single analytical form for the energy of stellar stream members, incorporating a 'universal' term that encapsulates the dark matter distribution alongside a nuisance term that accounts for the specifics of individual streams — containing parameters allowed to have object-specific values. Unlike traditional black-box deep learning methods, such as auto-encoders, our method generates a physically meaningful, low-dimensional model in the form of an analytical model.

The layout of this Chapter is as follows: In Section 5.1, we present the methodology of our approach. Section 5.2 details a first benchmark for Class

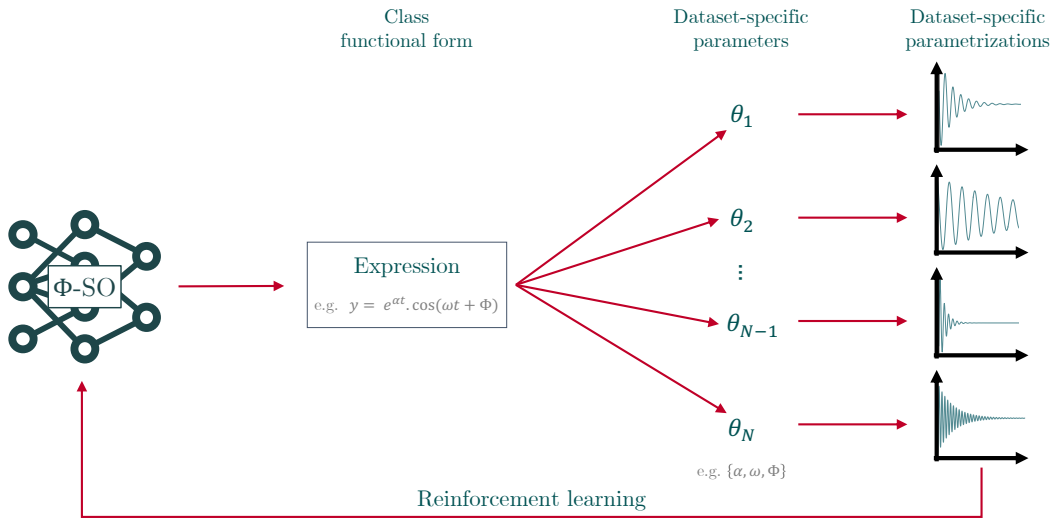


Figure 5.1: **Class Symbolic Regression framework sketch.** Searching for a unique functional form simultaneously fitting multiple datasets. The process starts at the left hand side, a batch of trial class analytical expressions are generated using our Φ-SO framework [Tenachi et al., 2023a]. The free parameters appearing in those expressions are then optimized in a dataset-specific manner i.e. allowing each dataset to have its own unique associated values for each parameter. The neural network used to generate the trial expressions is then reinforced based on the fit quality of the trial symbolic functions. This process is repeated until convergence.

SR, consisting of a series of physics problems designed to assess the performance of Class SR systems; here, we also evaluate our method against these benchmarks. In Section 5.3, we illustrate the practical application of our method in the more complex scenario of a Milky Way-like potential recovery from orbits. Finally, Section 5.4, offers a discussion.

5.1 Method

We build our Class SR on Φ-SO which combines deep reinforcement learning with *in situ* dimensional analysis constraints to construct solutions that avoid physically nonsensical combinations of units. The algorithm currently achieves state-of-the-art performance on physics datasets, and significantly outperforms competitors on the standard Feynman SR benchmark [La Cava et al., 2021] in exact symbolic recovery in the presence of even slight levels of noise ($\geq 0.1\%$).

Figure 5.1 gives an overview of our Class SR framework. In sub-section

5.1.1 we discuss the implication of such a system on free parameters and in sub-section 5.1.2 we detail the specifics of our Class SR learning strategy.

5.1.1 Free parameters

Class vs realization-specific parameters

Using Φ-SO we generate a batch of analytical expressions via a recurrent neural network (RNN). In these expressions, class-parameters ($\mathbf{c}$) — which are shared across the entire class and have consistent values across all datasets — can appear alongside realization-specific parameters ($\mathbf{k}$). Subsequently, we optimize the free parameters appearing in each expression $(\mathbf{c}, \mathbf{k})$, assigning unique values to realization-specific parameters $\{\mathbf{k}_i\}_{i<N_r}$ for each of the N_r datasets.

Free parameters optimization

This optimization is conducted using the L-BFGS nonlinear optimization routine [Zhu et al., 1997]. Encoding our mathematical symbols with `PyTorch` [Paszke et al., 2019], enables us to use `PyTorch`'s implementation of the L-BFGS routine, which benefits from `PyTorch`'s auto-differentiation capabilities to efficiently and simultaneously optimize both class and realization-specific parameters employing a mean squared error (MSE) cost function:

$$\text{MSE} = \frac{1}{N_r . \sum_{i=1}^{N_r} N(i)} \sum_{i=1}^{N_r} \sum_{j=1}^{N(i)} \left(y_{ij} - f(\mathbf{c}, \mathbf{k}_i, \mathbf{x}_{ij})\right)^2 \tag{5.1}$$

Where $\mathbf{x}_{ij}$ are the input variables, y_{ij} are the target values and $N(i)$ is the number of samples which depends on the dataset[1].

5.1.2 Generating expressions

Learning

We then use reinforcement learning to update the RNN's parameters following a risk-seeking gradient policy [Petersen et al., 2021a], as detailed in 3.3. This update is based on a reward $R = (1 + \text{NRMSE})^{-1}$ that is representative of the fit quality of the trial functional form f across all datasets — evaluated using

[1] Our implementation accommodates this variability while maintaining the use of vectorization, ensuring no compromise in computational efficiency.

a normalized root mean squared error (NRMSE):

$$\text{NRMSE} = \frac{1}{\sigma_y} \sqrt{\frac{1}{N_r . \sum_{i=1}^{N_r} N(i)} \sum_{i=1}^{N_r} \sum_{j=1}^{N(i)} (y_{ij} - f(\mathbf{c}, \mathbf{k}_i, \mathbf{x}_{ij}))^2} \tag{5.2}$$

Where σ_y is the standard deviation of target values evaluated across all datasets. We repeat this process until the RNN converges to a unique high quality expression and its associated parameter values simultaneously fitting all datasets.

For a live demonstration of our system applied to a class symbolic regression task, refer to [▶ Class SR demo][2]. In this example, the system aims to derive a model fitting data points corresponding to the stellar stream challenge that will be the subject of Section 5.3. The video illustrates the iterative process, displaying the curves associated with trial candidate expressions for each realization over successive iterations, highlighting their progressive improvement in fit quality until convergence.

Priors

Furthermore, the sequential nature of expression generation in our Φ-SO framework enables the incorporation of various priors regarding the resulting expressions as demonstrated in [Tenachi et al., 2023a, Bartlett et al., 2023b, Petersen et al., 2021b, Kim et al., 2021]. This allows for customized constraints on the generated expressions such as adherence to the rules of dimensional analysis (which was one of the focal points of Chapter 4) but also simpler priors such as constraints on the number of occurrences of given parameters, the length of the expression and more.

5.2 Multi-Dataset Symbolic Regression Challenges

Despite existing research efforts to establish benchmarks for SR [La Cava et al., 2021, Matsubara et al., 2022, Marinescu et al., 2023, Graham et al., 2013,

[2] https://youtu.be/Mu51K9EKMms

Thing and Koksbang, 2025], a benchmark tailored specifically for Class SR has yet to be developed, reflecting the innovative nature of this approach. To address this, we introduce our own Class SR challenges, designed to evaluate a system's ability to analyze multiple datasets. These datasets represent varied observations of a similar phenomenon occurring at different scales but governed by a consistent functional form. Table 5.1 outlines these challenges, each focusing on accurately recovering the symbolic expression from synthetic datasets having varied scale parameter values. To heighten the challenge, we include multiple scenarios incorporating class parameters that are common to all realizations in addition to other realization-specific parameters. We detail our benchmarking protocol in sub-section 5.2.1 and give the performances of our method in sub-section 5.2.2.

#	Challenge	Formula	Variables	Realization-specific free parameters
1	Harmonic Oscillator	$A\cos(\Phi+\omega t)$	$t \in [0.0, 9.4]$ -	$A \in [0.6, 1.2]$ $\omega \in [0.2, 1.5]$ $\Phi \in [0.9, 1.1]$
2	Radioactive Decay	$n_0 e^{\frac{-t}{T}}$	$t \in [0.5, 6.0]$ -	$n_0 \in [0.4, 2.0]$ $T \in [0.9, 1.4]$
3	Free Fall	$\frac{1}{2}9.81t^2 + tv_0 + z_0$	$t \in [0.0, 1.0]$ -	$v_0 \in [-2.0, 8.0]$ $z_0 \in [-3.0, 3.0]$
4	Damped Harmonic Oscillator A	$e^{-kt}\cos(\Phi+1.389t)$	$t \in [0.0, 9.4]$ -	$k \in [0.2, 1.0]$ $\Phi \in [-0.2, 0.3]$
5	Damped Harmonic Oscillator B	$e^{-0.345t}\cos(\Phi+\omega t)$	$t \in [0.0, 9.4]$ -	$\omega \in [0.6, 1.4]$ $\Phi \in [-0.2, 0.3]$
6	Black Body Photon Count	$\frac{1}{e^{5.9\nu/T}-1}$	$\nu \in [1.0, 5.0]$ -	$T \in [1.0, 5.0]$ -
7	Ideal Gas Law	$\frac{n8.314T}{V}$	$T \in [1.0, 5.0]$ $V \in [1.0, 5.0]$	$n \in [1.0, 5.0]$ -
8	Free Fall Terminal Velocity	$\sqrt{\frac{2m9.807}{0.47A\rho}}$	$m \in [1.0, 10.0]$ $A \in [1.0, 5.0]$	$\rho \in [1.0, 6.0]$ -

Table 5.1: **Class Symbolic Regression challenges.** Each row details a distinct challenge, with the objective being the exact symbolic recovery of the designated target expression using multiple synthetic datasets. Each dataset being generated using unique realization-specific parameter sets sampled from the given parameter ranges by sampling from the target expression within the given variable ranges.

5.2.1 Benchmarking protocol

Benchmark settings

We evaluate our algorithm by randomly sampling 10 datasets of 10^2 samples for each of the 8 challenges described in Table 5.1 and allowing a maximum

of 200,000 expressions to be explored during each run. In order to ensure robustness, for each challenge, the procedure was repeated 5 times, each time with a unique random seed, and the recovery rates were subsequently averaged. The whole benchmark tests were conducted across four noise levels: 0%, 0.1%, 1% and 10% with noise being added individually to each dataset as per the `SRBench` [La Cava et al., 2021] standardized SR benchmarking protocol :

$$y_{\text{noise}} = y + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \gamma\sqrt{\frac{1}{N}\sum_i y_i^2}\right) \tag{5.3}$$

Where γ is the level of noise. We conduct runs having access to a single dataset (SR) and having access to all 10 datasets (Class SR), leading to the total evaluation of 64,000,000 expressions through 320 runs.[3]

Method settings

We run our algorithm using the hyper-parameters detailed in Table 3.1, with dimensional analysis disabled to ensure a fair comparison with other algorithms (as a consequence the batch size is lowered to 2000). This adjustment allows future comparisons with our system to be focused solely on the machine learning technique used (here reinforcement learning), rather than the problem simplification achieved through dimensional analysis. We allow the use of the following operations: $\{+, -, \times, /, 1/\square, \sqrt{\square}, \square^2, -\square, \exp, \log, \cos, \sin\}$, a constant equal to one $\{1\}$, two adjustable realization specific free constants $\mathbf{k} = \{k_1, k_2\}$ allowed to have dataset-specific values and one adjustable class free constant $\mathbf{c} = \{c_1\}$. The recovery rate is evaluated by examining each expression in the Pareto front, which contains optimum expressions found in conciseness / accuracy i.e. : best fitting expressions at each level of complexity generated by our algorithm. Successful recovery is noted if an expression on the Pareto front is symbolically equivalent to the target expression. Exact symbolic recovery is assessed by formally comparing these expressions with the target expression using the `SymPy` library for symbolic mathematics [Meurer et al., 2017], following a methodology similar to the one in the `SRBench` [La Cava et al., 2021]. Specifically, expressions are deemed equivalent if their fraction is symbolically equivalent to 1 or a constant or if their difference is symbolically equivalent to 0 or a constant.

[3] Additional details about this benchmark and the one presented in the next Section regarding implementation of their protocols and procedures to technically reproduce the results presented can be found in 6.1.3 and in 6.3 respectively.

5.2.2 Performances

Figure 5.2 presents a comparison of exact symbolic recovery rates between our Class SR framework and the traditional SR approach under both noiseless and noisy conditions using an `SRBench`-style benchmarking pipeline, with detailed challenge-by-challenge results published online (see sub-section 6.3). Our results demonstrate the superiority of Class SR over traditional SR in exact symbolic recovery, particularly in noisy scenarios where noise overfitting is generally an important concern [La Cava et al., 2021].

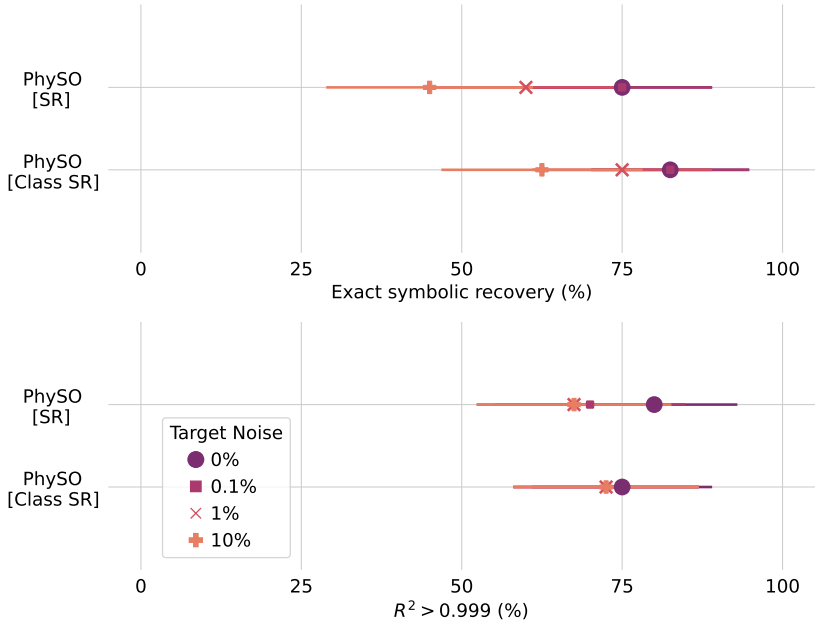


Figure 5.2: **Performances of `PhySO` on our Class SR benchmark.** Comparison of exact symbolic recovery rates and rate of accurate expressions (having $R^2 > 0.999$) between Class SR and standard SR on our Class SR challenges using an `SRBench`-style benchmarking pipeline [La Cava et al., 2021]. This figure demonstrates the enhanced effectiveness of Class SR in identifying common underlying functions across multiple datasets with varying scale parameters, resulting in a higher success rate compared to the traditional SR method exploiting only one dataset at a time — especially in the presence of noise.

Exact symbolic recovery

While one might consider employing traditional SR individually on each dataset and subsequently aggregating the results, this approach would not only be substantially more computationally demanding, but it would also fail to differentiate class constants from realization-specific scale parameters,

thus yielding a less interpretable model. Furthermore, our analysis uncovers several instances where traditional SR did not successfully identify the correct expression in any of the 5 attempts but in which Class SR effectively discovered the correct expressions. This concerns Problem #3 and #6 at 10% noise level scenarios, as well as Problem #5 across all noise levels. These findings highlight the superior robustness and efficiency of Class SR over traditional methods.

Accuracy

Following the `SRBench` protocol, we also include, on Figure 5.2, the rate of accurate expressions (having $R^2 > 0.999$) with the R^2 metric defined as :

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5.4}$$

We evaluate fit quality by refitting all constants of candidate expressions on newly generated previously unseen test datasets. This approach ensures a fair comparison between Class SR expressions, whose numerical parameters must accommodate multiple observations, and expressions derived from traditional SR, which only fit a single observation. Our results demonstrate that Class SR is not only more efficient at recovering the exact expressions but also more effective at deriving accurate approximations than traditional SR, in scenarios with noise levels exceeding 0.1%.

5.3 Recovering an Analytic Potential form Stellar Streams

We now turn to an astrophysical application of the algorithm: to try to find the underlying potential of a gravitational system from a set of orbit segments within it. Sub-section 5.3.1 contextualizes this problem, sub-section 5.3.2 details the testing protocol, and sub-section 5.3.3 presents the results.

5.3.1 Context

This could be practically applicable for finding an analytic potential model of a galaxy from a set of stellar streams. These linear structures form from

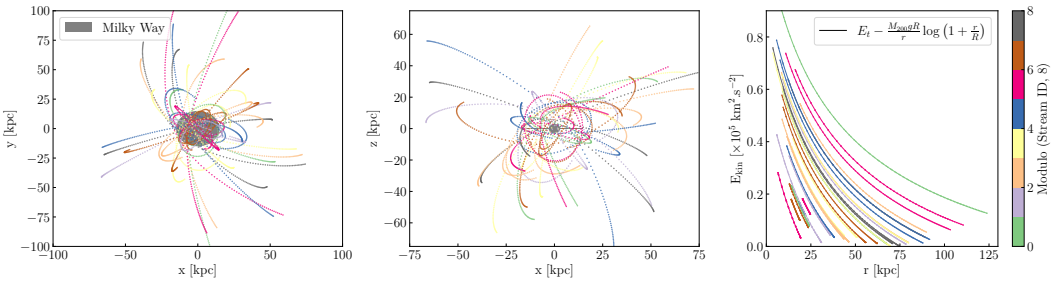


Figure 5.3: **Synthetic stellar stream data utilized by our algorithm to recover the galactic potential.** The left and middle panels display the spatial positions of stream members relative to the Milky Way, while the right panel illustrates the kinetic energy of these members as a function of their radial distance from the galactic center.

the tidal dissolution of globular clusters and dwarf satellite galaxies. When their progenitors are of low mass, the escaping stars have similar energy to the progenitor, and therefore follow a similar orbit. Hence stellar streams approximate orbits in the host galaxy. As has recently been shown by Ibata et al. [2021], for many real streams one can calculate a "correction function" to convert an orbit model into a stream track, and these functions are relatively insensitive to the adopted potential. This procedure can be inverted to give the orbit from the stream.

For this test we imagine having access to full 6-dimensional phase-space measurements of a sample of streams. For each structure i, the kinetic energy per unit mass $E_{i,\mathrm{kin}}(\mathbf{x})$ is simply:

$$\frac{1}{2}\mathbf{v}^2 = E_{\mathrm{t}}^i - \Phi(\mathbf{x})\,. \tag{5.5}$$

The total energy per unit mass E_{t}^i, which is constant, but different, for each stream, can be considered to be nuisance terms in our search for the underlying potential Φ.

5.3.2 Testing protocol

Method settings

We run our algorithm with the objective of recovering the analytic form for $E_{i,\mathrm{kin}}(\mathbf{x})$. We use the the hyper-parameters detailed in Tenachi et al. [2023a], allowing the use of the following operations: $\{+, -, \times, /, 1/\square, \sqrt{\square}, \square^2, -\square, \exp, \log, \}$, a constant equal to one $\{1\}$, one adjustable realization specific free constant (having units of energy) and three adjustable class free constants (one having units of energy, one having length units and the other being dimensionless).

Benchmark settings

Again we conduct runs at four noise levels (0%, 0.1%, 1% and 10%), having access to a single orbit (SR), 25% of the orbits, 50% of the orbits and 100% of the orbits (Class SR), repeating experiments 16 times with different random seeds and allowing a maximum of 250,000 expressions to be explored during each run, leading to the total evaluation of 64,000,000 expressions through 256 runs.

Stellar streams

For the present analysis we generated a sample of artificial orbit data (shown in Figure 5.3) that approximates the sample of 29 thin and long streams studied by Ibata et al. [2021]. To this end we used the present day progenitor positions estimated by Ibata et al. [2021], and integrated orbits within a universal (NFW) dark matter halo model [Navarro et al., 1997] that very roughly approximates the large-scale mass distribution in the Milky Way. The adopted potential [Łokas and Mamon, 2001] is

$$\Phi_{NFW} = -M_{200}.g.\frac{R}{r}.\log\left(1+\frac{r}{R}\right) \tag{5.6}$$

where M_{200} is the virial mass of the halo, $g \equiv (\ln(1+c) - c/(1+c))^{-1}$ is a function of the halo concentration c and R is the scale radius. We chose $M_{200} = 10^{12}\,\mathrm{M}_\odot$, $c = 10$ and $R = 20.0\,\mathrm{kpc}$. The orbits consist of 100 phase space points at locations between $\pm 1\,\mathrm{Gyr}$ from the current progenitor location.

5.3.3 Results

Figure 5.4 presents the results of our analysis in terms of exact symbolic recovery and fit quality, evaluated using the R^2 metric. This metric was determined by refitting candidate expressions on noiseless test data and computing the median across various random seeds.

Exact symbolic recovery and accuracy

As anticipated, our results underscore that utilizing more realizations during the SR process significantly enhances model accuracy and the likelihood of exact symbolic recovery. This trend is particularly evident as noise levels rise, reinforcing our findings of Section 5.2. Notably, at a 1% noise level, none of the 16 runs that analyzed stellar stream individually succeeded in recovering the correct functional form. In contrast, when all 29 stellar streams were utilized,

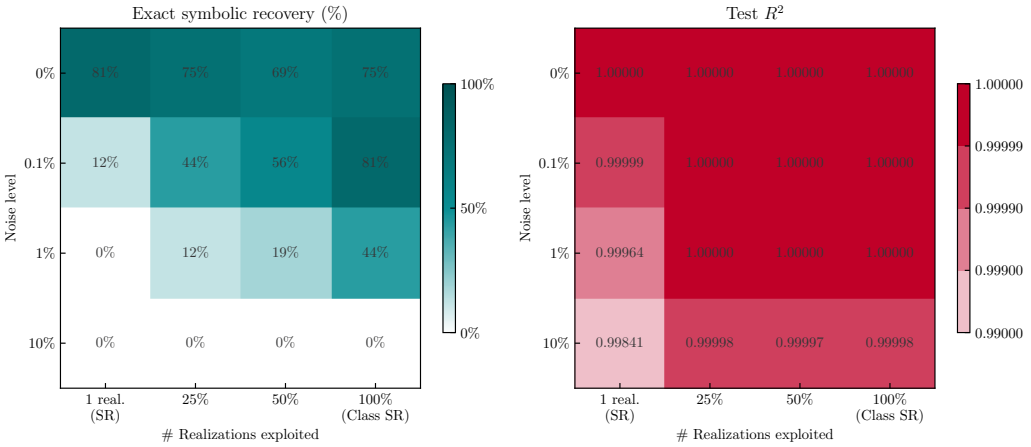


Figure 5.4: This figure presents the exact symbolic recovery rate and median R^2 achieved by our Class SR algorithm in the task of recovering an NFW dark matter halo model [Navarro et al., 1997] from synthetic datasets of stellar stream positions and velocities. The performance metrics are displayed as functions of noise levels and the number of realizations exploited. The edge case, in which a single realization is used, corresponds to the conditions of traditional SR. The results distinctly demonstrate that Class SR substantially outperforms traditional SR, particularly in noisy environments.

the correct functional form was identified nearly half of the time, showcasing the advantages of Class SR under noisy conditions.

Degeneracies in the space of functional forms

We observe that the inability of our algorithm to recover the exact symbolic expression in the presence of 10% noise can be attributed to the fact that, under such high noise conditions, the difference in fit quality between the expressions typically identified by our algorithm and the true solution yields only a minimal improvement in terms of reward, $\Delta R \sim 10^{-5}$. This minute improvement, which corresponds to a difference in R^2 of approximately (10^{-6}), is the sole metric available to guide the algorithm, as it operates on a trial-and-error basis. Unfortunately, such a small difference often remains undetected due to it falling below the tolerance threshold of the free constants optimization procedure. This scenario highlights a known intrinsic limitation of purely empirical SR, where degeneracies in the space of functional forms can go undetected.

Overview

Excluding scenarios where noise levels render the numerically found expression indistinguishable from the true solution, our Class SR algorithm typically converges toward the correct functional form by exploring under $250,000$ expressions, despite the presence of multiple alternative functional forms that provide a near-perfect fit to individual streams. Φ-SO identifies an offset parameter specific to each stream (corresponding to E_t^i) and a functional form parameterized by class-parameters common to all streams corresponding to Φ_{NFW}. These results show that our algorithm can effectively recover a concise intepretable model for a Milky-Way like potential in the form of an analytic expression based solely on stellar positions and velocities without any prior information about the system.

5.4 Discussion and Conclusions

We presented a first framework for discovering symbolic analytical functions that simultaneously fit multiple datasets by allowing for (possibly) unique dataset-specific parameter values. This new framework which we dub "Class Symbolic Regression" is built upon our earlier Φ-SO framework which already delivers state-of-the-art performances in symbolic recovery in the presence of noise.

We demonstrated the efficacy of Class SR through simple textbook physics examples which we compiled into a first Class SR benchmark, finding better performance in exact symbolic recovery over traditional SR, especially in noisy situations. Additionally, we applied our method to a more complex astrophysical scenario, successfully rediscovering an NFW galaxy potential model from orbits approximating stellar streams.

Class SR's edge

Regular SR, when applied to a single dataset, often risks overfitting to specific characteristics of an observation, such as observational biases or transient events, and noise. In contrast, our Class SR framework should facilitate the finding of *universal* analytical laws that apply to a range of observations within a single class of physical phenomena. This enables our framework to model the underlying physics rather than the specifics of individual observations, with dataset-specific free parameters modeling the unique aspects of each observation. For instance, an application within galactic dynamics

that we intend to explore in a future contribution is the analysis of galactic rotation curves. Here, a universal law derived through Class SR could provide insights into the general behavior of dark matter, whereas traditional SR, if applied to a single galaxy, might merely find the specific attributes of that galaxy.

It should be noted that while Class SR might superficially resemble regular SR applied to unbalanced datasets with dataset-specific parameters being akin to additional input variables, this comparison is not entirely accurate. In Class SR, these additional degrees of freedom represent unknown values that must be determined, differentiating it as a distinct problem with its own unique challenges.

A persistent issue in SR is model selection as the correct expression can often be overlooked in favor of those that fit better or are less complex (these concerns led to e.g., the development of single objective criterion Bartlett et al. 2023a). Our framework, by searching for expressions that fit multiple datasets, effectively utilizes information about the physical phenomena's class structure. This approach significantly mitigates model selection challenges, helping avoid incorrect model choices influenced by dataset-specific peculiarities. In addition, exploiting multiple datasets with regular SR techniques would require fitting the individual datasets independently, and then identifying the solutions in common between the objects, which may not be possible if the measurements are uncertain, would be computationally inefficient and would result in lower performances in exact symbolic recovery and fit quality alike in the presence of noise.

Perspectives

Finally, we note that after our approach was released, another Class SR approach built on `Operon` [Kommenda et al., 2020] — a genetic algorithm approach to SR — was applied to supernovae photometry in Russeil et al. [2024] and that [Cranmer, 2023] was upgraded to incorporate this feature as well as our free constants fitting method.

In future work, we intend to improve on the machine learning aspects of our method to more effectively leverage multiple datasets. As each dataset might distinctly highlight certain symbolic terms or sub-expressions more prominently than others, a promising strategy could be to periodically shift the neural network's training emphasis between datasets. This technique could potentially refine the performance of Class SR by sequentially learning different segments of the expression, rather than attempting to learn the entire expression simultaneously, thereby facilitating the learning process.

Chapter 6

PhySO : A Physical Symbolic Optimization software

Summary.

We give details about the `PhySO` software for Physical Symbolic Optimization developed during this thesis. We outline its features, particularly those not utilized in the experiments presented here, and showcase how its distinctive capabilities position it as an ideal tool for future research projects.

We give its computational performances and detail its parallelization feature and unique vectorized handling of batches of symbolic expressions. Additionally, we discuss the benefits of adopting an open-source approach, which has allowed `PhySO` to become a foundational framework relied upon by multiple research projects. We show that through this openness we provide benchmarking tools to the SR community, ensure transparency, and promote reproducibility within the community.

Notably, we include technical details necessary to reproduce the outcomes presented in our benchmarking experiments in this Chapter.

This Chapter is dedicated to the PhySO — Physical Symbolic Optimization — software that was developed as part of the present thesis. This software notably integrates the Φ-SO framework for symbolic regression (SR)[1] that was described in Chapters 3-5.

The PhySO software is publicly available on GitHub at github.com/WassimTenachi/PhySO with a comprehensive documentation deposited on the same repository and live at physo.readthedocs.io. It is important to note that PhySO is a live software that may evolve in the future. For the purposes of this discussion, references will be made to the most recent stable release as of July 2024, available under the release tag v1.1.0.

PhySO is a symbolic optimization package built for physics. It takes the form of a Python package relying on PyTorch [Paszke et al., 2019] for its deep learning and auto-differentiation component. Its SR module employs deep reinforcement learning to derive analytical laws from datasets, exploring the space of functional forms through a process of trial and error. This software is engineered to be as fast as technically feasible, while also being user-friendly and straightforward to install.

PhySO is unique in its capability to leverage :

- Physical units constraints : using the rules of dimensional analysis to constrain the search space. [Tenachi et al., 2023a,b, Tenachi et al., 2023a]
- Class constraints : to infer a single analytical functional form that accurately fits multiple datasets, each governed by its own (possibly) unique set of fitting parameters. [Tenachi et al., 2024]

In Section 6.1, we delineate the functionalities of PhySO. Following that, Section 6.2 provides detailed insights into the implementation of the software, including its computational performances. Lastly, Section 6.3 discusses the advantages of making the software open source. This section also emphasizes the enhanced reproducibility of our experiments — a direct result of the software's accessibility.

[1] SR involves searching for an analytical function that best fits a given dataset.

6.1 Capabilities & Features

PhySO is designed to be a versatile symbolic optimization software with the potential for organic growth to meet various future needs, including those discussed in Section 11.2. It aims to engage and stimulate the machine learning enthusiasts within the astrophysics and physics communities.

In the following sub-sections, we delve deeper into the structure and capabilities of PhySO. Sub-section 6.1.2 outlines its current features, while sub-section 6.1.3 details the benchmarking systems developed to evaluate our method. These systems are integrated within PhySO not only to assess our own methodology but also to offer a means for evaluating other symbolic systems, ensuring fairness and transparency in comparative analyses.

6.1.1 Symbolic graph

The core of PhySO — its encoding and management of symbolic expressions — was developed to support a variety of projects related to physics and symbolic methods. The symbolic management system operates independently of its machine learning framework, allowing it to potentially integrate with any system, although it integrates many features that are particularly relevant for systems generating expressions in a sequential manner.

Graph management system

PhySO continuously monitors the full graph structure of any symbolic expression during its generation. For each node within the graph, PhySO tracks its parent, sibling, and child nodes, ensuring comprehensive connectivity without incurring additional computational overhead. This capability is crucial for efficiently conducting the *in situ* dimensional analysis described in Chapter 4. Each time a new token is introduced into the expression graph, the physical units constraints are dynamically updated. This ensures the graph always contains the most complete and current information possible. This is illustrated in Figure 6.1.

This entire process is executed in a fully vectorized manner across the batch of expressions, which is non-trivial due to the unique graph structure of each expression. This vectorization is particularly crucial for future projects that may require managing significantly larger batches — potentially more than the current $\sim 10^4$ expressions — simultaneously.

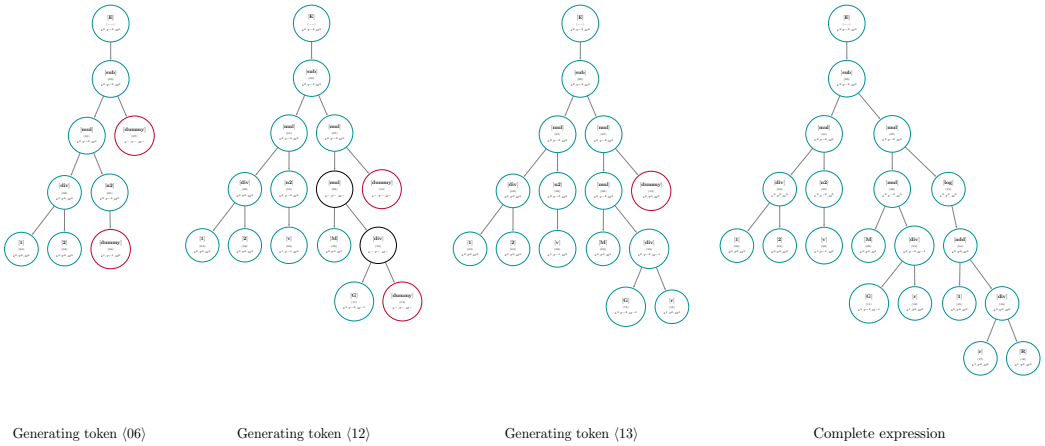


Figure 6.1: **Illustration of the symbolic graph encoded within PhySO at various stages of its generation.** Each node's physical units are denoted using L, T, M (length, time, mass) dimensions. Nodes with established physical units are highlighted in blue, placeholder nodes for dangling branches are marked in red, and nodes with yet-to-be-determined units are shown in black. This graph representation, easily accessible via the `expressions.show_tree()` function, demonstrates the dynamic evolution of the graph structure during the generation process. The graph representation of the expression $E = \frac{1}{2}mv^2 - \frac{GM}{r}\log(1 + \frac{r}{R})$, expressed in prefix notation as $\left(-, \times, /, 1, 2, \square^2, v, \times, \times, M, /, G, r, \log, +, 1, /, r, R\right)$, is depicted at various stages — specifically during the generation of tokens $\langle 06\rangle$, $\langle 12\rangle$, $\langle 13\rangle$, and upon completion of the generation process.

Computational graph

The evaluation of symbolic expressions is a critical component of many symbolic approaches. In PhySO, the computational graph for these expressions is constructed using PyTorch functions, which facilitate rapid evaluation and auto-differentiation (see sub-section 2.1.4 for a detailed discussion of auto-differentiation). This capability is crucial, as it allows PhySO to perform fittings of free constants efficiently, leveraging the computational graph to gain an advantage over other SR implementations. Typically, fitting arbitrary constants in arbitrary functional forms rapidly (a necessity given time constraints per equation) yields suboptimal results with conventional methods that do not utilize such a graph. To date, PhySO is the only SR framework to exploit a differentiable graph structure of candidate expressions for its free constants optimization component.

Moreover, PhySO enhances the efficiency of this process through the parallelization of the free constants fitting process. To further enhance robustness and prevent numerical issues during evaluations outside their definition ranges, PhySO employs protected variants of functions, e.g., using logabs instead of log, though this feature can be disabled as needed.

Interfaces

`PhySO` offers an interface with `SymPy` [Meurer et al., 2017] — a well established library for symbolic mathematics in Python, comparable to Mathematica [Wolfram, 2003]. This integration allows users to seamlessly interact with any equation generated by `PhySO`. Additionally, at the conclusion of the exploration within the equation space, `PhySO` archives optimal solutions on the Pareto front-those that balance fitness and simplicity-including the overall best-fitting equation and the best-fitting equation across all iterations. Furthermore, `PhySO` maintains a comprehensive log of all equations generated during the process, ensuring detailed documentation and reproducibility of the computational experiments.

6.1.2 Symbolic optimization

Arbitrary symbolic optimization task

`PhySO` is a general-purpose symbolic optimization software, designed to optimize symbolic expressions for various objectives beyond the typical data fitting objective associated with SR. The software can manage any objective quantifiable by a scalar reward $r \in [0, 1]$, derived from trial expressions generated by the neural network. This versatility enables users to define custom objectives via the `rewards_computer` parameter. However, it is essential to acknowledge that the success of this approach depends on a positive correlation between the symbolic arrangement and the reward.

Complementary features

As previously mentioned, `PhySO`'s main features are the dimensional analysis (the subject of Chapter 4) and Class SR functionalities (the subject of Chapter 5). Beyond these complex features, the software also incorporates several simpler yet useful features.

One of the prominent features is the use of priors. In addition to the complex dimensional analysis prior and the various priors detailed in 3.2.3 — including hard length prior, soft length prior, nesting priors, and the prior forbidding inverse unary operations — `PhySO` offers multiple other priors.

Uniform arity prior

We implement a uniform arity prior, which applies a uniform probability distribution over tokens based on their arities. This prior aims to balance the representation of tokens by normalizing token probabilities according to the

number of tokens sharing the same arity, thereby promoting under-represented arities and discouraging over-represented ones.

Relationship constraints

Another important prior is the general relationship constraint prior, which ensures that expressions adhere to specified relationship constraints. For instance, it enforces rules preventing user-defined target tokens from being descendants, children, or siblings of user-defined effector tokens.

Occurrences constraints

Additionally, the occurrences prior restricts the frequency of certain target tokens, ensuring they do not appear more than a specified maximum number of times within expressions.

Symbolic constraints

Lastly, the symbolic prior allows for the enforcement of specific tokens at designated positions within the symbolic expressions, e.g., enforcing that the expression should start with $\log x + \square$ or contain $\cos(x)/x$.

Weighted data points

`PhySO` incorporates a straightforward method for applying weights to data points, which can reflect experimental or observational errors. This weighting can be applied on a point-by-point basis or realization-wise within Class SR frameworks, allowing for the prioritization of one realization over another.

Custom symbolic functions

Furthermore, `PhySO` facilitates the inclusion of new symbolic operation tokens. Users can easily integrate custom tokens by defining their name and providing an auto-differentiable function associated with them. Additionally, users can specify how these tokens behave during dimensional analysis and define any protected variants.

Arbitrary symbolic optimization

Given data pairs $\{x, y\}$, the `PhySO` software's `wrapper` feature allows for more complex optimizations than the standard SR approach of optimizing a function f such that $y = f(x)$. Instead, users can perform indirect optimizations

by fitting $y = g_{\text{wrap}}(f, x)$, where g_{wrap} is a wrapper function that takes f and x as arguments and returns a prediction after executing arbitrary intermediate steps before comparison to the target data y. If g_{wrap} is auto-differentiable, it even allows for the optimization of free constants within such frameworks, highlighting the versatility of our software. This approach was crucial in applying our algorithm to the expansion of the Universe scenario, which involved indirect comparisons to data from cosmological standard candles as detailed in Section 4.4.2.

6.1.3 Benchmarking sub-module

As part of our research efforts in SR, we have developed and implemented a comprehensive set of benchmarking protocols. These protocols are designed to rigorously evaluate SR systems and have been made available to the broader SR community. Details on these protocols and how they were applied to evaluate our own system can be found in Section 4.3 for the Feynman benchmark and Section 5.2 for the Class SR benchmark.

This sub-section introduces the benchmarking utilities incorporated within `PhySO`. We offer user-friendly access to benchmarking challenges, ensuring reproducibility and adherence to standardized practices (e.g., , using standard data ranges, constant values and more). For each challenge, we describe methods to generate data and evaluate how candidate expressions compare to the ground truth. Our protocols align with the standard benchmarking practices outlined in `SRBench` [La Cava et al., 2021].

Moreover, we have enhanced the benchmarking process by automatically incorporating assumptions about variable characteristics (such as positivity) within these challenges, thus facilitating fair comparisons of symbolic equivalence. For example, we ensure that $\sqrt{a.b}$ is recognized as equivalent to $\sqrt{a}.\sqrt{b}$ by properly encoding the assumption that $a > 0$. Additionally, we have developed custom-built trigonometric tools to address issues like $\cos(x - 1.5708)$ not being recognized as equivalent to $\sin(x)$ due to slight errors between 1.5708 and $\frac{\pi}{2}$.

Feynman benchmark

The Feynman benchmark is designed to assess symbolic regression systems, particularly those aimed at scientific discovery. That is methods able to produce compact, predictive, and interpretable expressions from potentially noisy data. The benchmark was initially introduced by Udrescu and Tegmark [2020] and later formalized by La Cava et al. [2021]. It consists of 120 unique challenges, each associated with a distinct ground truth expression. `PhySO`'s interface is based on the original files (`FeynmanEquations.csv`, `BonusEquations.csv`, `units.csv`) from Udrescu and Tegmark [2020], with

adjustments to correct errors in units and other discrepancies found in the original dataset.

We offer to the community a convenient interface for using our implementation of the Feynman benchmark, running: `pb = FeynmanProblem(i)` will instantiate challenge number $i \in \{0, 1, ..., 119\}$ of the Feynman benchmark ($\{0, ..., 99\}$ corresponding to bulk challenges and $\{100, ..., 119\}$ to bonus challenges) of the Feynman benchmark. This interface offers simple ways to generate data points (via `pb.generate_data_points`) and compare a candidate expression to the target (via `pb.compare_expression`).

Class SR benchmark

The purpose of the Class benchmark is to evaluate Class symbolic regression systems, that is: methods for automatically finding a single analytical functional form that accurately fits multiple datasets — each governed by its own (possibly) unique set of fitting parameters. See 5.2 [Tenachi et al., 2024] in which we introduce this first benchmark for Class SR methods.

We similarly offer to the community a convenient interface for using our Class SR benchmark, running: `pb = ClassProblem(i)` will instantiate challenge `i` $\in \{0, 1, ..., 7\}$ of the Class benchmark presented in Table 5.1. This interface offers simple ways to generate data points (via `pb.generate_data_points`) and compare a candidate expression to the target (via `pb.get_sympy`).

6.2 Implementation

Code Structure

Our codebase is methodically organized into several distinct sub-modules, adhering to the interface segregation principle [Martin, 2009]. The core components of our software are contained in the `physo.physym` and `physo.learn` sub-modules, which contain the algorithms for managing symbolic expressions and the machine learning algorithms, respectively. The `physo.task` sub-module provides high-level functions that users can employ to perform standard Symbolic Regression (`physo.SR`) and Class Symbolic Regression (`physo.ClassSR`). Additionally, the `physo.config` sub-module offers pre-set hyper-parameters

configurations, while the `physo.benchmark` sub-module includes the aforementioned benchmarking utilities aimed at the SR community.

Each sub-module is equipped with multiple unit test scripts, designed to test the individual components of the code independently. This testing framework is not only crucial for ensuring the robustness and reliability of the software during the development of new features but also supports community engagement by facilitating modifications and enhancements. `PhySO` currently achieves a coverage score[2] of 81%, indicating that 81% of the code lines are executed during unit tests.

Computational performances

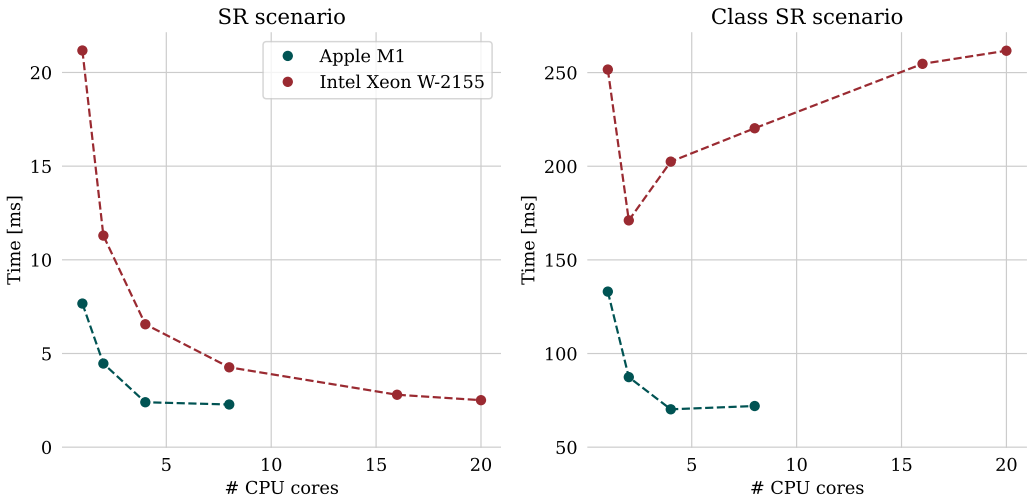


Figure 6.2: **Computational performances of `PhySO` in SR and Class SR scenarios.** (SR scenario) Computational time optimizing free constants $\{c_1, c_2\}$ in $y = c_1 \sin(c_2.x) + e^{-x}$ over 20 iterations using 10^3 data points when running this task 10 000 times in parallel. (Class SR scenario) Computational time optimizing class free constants $\{c_1, c_2\}$ and realization-specific free constants $\{k_1, k_2, k_3\}$ in $k_1 e^{-k_2 t} \cos(c_1 t + k_3) + c_2 x$ over 50 iterations using 100 data points per realization over 30 realizations running this task 10 000 times in parallel. Performances are assessed on an Apple M1 laptop (a machine with 4 fast CPU cores) and an Intel Xeon W-2155 CPU (a machine with a high number of cores).

Due to the substantial number of trial expressions that need to be evaluated at each iteration during symbolic regression tasks, and considering that each expression must undergo multiple evaluations to optimize its free constants, the optimization step constitutes a significant performance bottleneck in the `PhySO` algorithm. To address this, we have parallelized this step across batches,

[2] https://coveralls.io/github/WassimTenachi/PhySO

achieving optimization times for free constants of a given expression typically on the order of 1 ms. The efficiency of this process in realistic scenarios is illustrated in Figure 6.2.

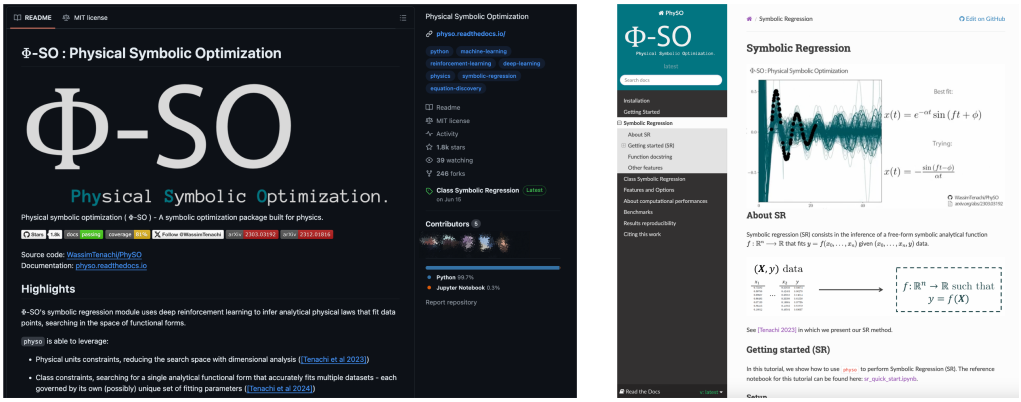
Testing was conducted on an Apple M1 machine, which features 4 high-performance cores and 4 energy-efficient cores, explaining the observed performance plateau when the core count is increased to 8. While parallel execution generally enhances performance, in Class SR scenarios, no improvements are seen beyond the use of 2 CPUs on most machines. With the notable exception of machines with exceptionally fast CPU cores. Note that here our manual parallelization strategy imposes parallel processing across the batch of expressions, whereas `PyTorch` natively parallelizes across the dimension of data points.

In addition, the management of symbolic information necessary for computing priors and providing contextual data to the neural network also occupies a considerable portion of computational time. In `PhySO`, these operations are efficiently vectorized across both expression lengths and batches, enhancing overall computational efficiency.

Currently, `PhySO` is capable of running in parallel on a single computer node but does not yet support multi-node execution for a single run. A common workaround involves running multiple instances of the same problem with varying seeds, hyper-parameters, and setups to enhance the probability of discovering an effective model. Looking forward, an interesting development would be enabling `PhySO` to utilize multiple nodes simultaneously for a single symbolic regression task.

6.3 An Open Source Software

Our objective with `PhySO` is to embrace the open source ethos fully. Our entire codebase is publicly available on GitHub (see Figure 6.3). Each function and object within the code is accompanied by a comprehensive documenting header. We provide extensive documentation, which is also maintained as open source on GitHub and is automatically compiled with each update. To further support user engagement and facilitate the integration of new features, we offer tutorials, demonstration notebooks, and even release our unit tests as open source. This approach not only enhances the transparency of our process but also encourages active community participation and contribution.



(a) Repository (b) Documentation

Figure 6.3: **Repository (a) and documentation (b) of `PhySO`.**

Community interactions

Embracing an open source model has proven mutually beneficial. By making `PhySO` freely available, we have observed its adoption across various scientific fields. Instances of `PhySO` being utilized for research — as of December 2024 — are detailed in Table 6.1.

Conversely, the community has actively contributed to the enhancement of `PhySO` by identifying and correcting bugs through *pull requests*[3]. A particularly notable contribution was recently made by He et al. [2024b], where our code was forked to incorporate a transformer-based neural network architecture [Vaswani et al., 2017], aligning with our research plan as outlined in 3.3.2. Additionally, Li et al. [2024d] made an outstanding contribution by leveraging `PhySO`'s capabilities to express *in situ* priors and perform live dimensional analysis. This innovation enables the enforcement of prior knowledge about equation structures, ensuring the use of sub-equations with known units and dependencies on specific subsets of variables, in line with the research plan outlined in Section 7.3.3.

This exemplifies the powerful synergies enabled by our open source approach.

Reproducibility

Reproducibility stands as a cornerstone of the open source philosophy and is especially pivotal in scientific research. To this end, we have made it extremely straightforward to replicate the results presented in our experiments. The following paragraphs give the specific steps to reproduce the experiments

[3] A mechanism in the Git protocol allowing users to suggest changes to a repository.

Title	Ref.
A universal crack tip correction algorithm discovered by physical deep symbolic regression	Melching et al. [2024]
Advancing symbolic regression for earth science with a focus on evapotranspiration modeling	Li et al. [2024d]
An Efficient and Generalizable Symbolic Regression Method for Time Series Analysis	Xie et al. [2024]
Channel Modeling Based on Transformer Symbolic Regression for Inter-Satellite Terahertz Communication	He et al. [2024b]
Class Symbolic Regression: Gotta Fit 'Em All	Tenachi et al. [2024]
Constraining Genetic Symbolic Regression via Semantic Backpropagation	Reissmann et al. [2024]
Deep symbolic regression for numerical formulation of fundamental period in concentrically steel-braced RC frames	Rahman et al. [2024]
Discovery of physically interpretable wave equations	Cheng and Alkhalifah [2024]
Enhancing Symbolic Regression And Universal Physics-informed Neural Networks With Dimensional Analysis	Podina et al. [2024]
From inflation to dark matter halo profiles: the impact of primordial non-Gaussianities on the central density cusp	Stahl et al. [2024]
Function Class Learning with Genetic Programming: Towards Explainable Meta Learning for Tumor Growth Functionals	Sijben et al. [2024]
Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl	Cranmer [2023]
Machine Learning in Proton Exchange Membrane Water Electrolysis – Part I: A Knowledge-Integrated Framework	Chen et al. [2024b]
Machine learning the governing principle of strong coupling constant across the global energy scale	Wang et al. [2024]
Physics-constrained robust learning of open-form PDEs from limited and noisy data	Du et al. [2024]
Physics Education and Symbolic Regression	Shin et al. [2024]
Recent advances in the SISSO method and their implementation in the SISSO++ code	Purcell et al. [2023]
Reinforced Symbolic Learning with Logical Constraints for Predicting Turbine Blade Fatigue Life	Li et al. [2024e]
Revisiting Disparity from Dual-Pixel Images: Physics-Informed Lightweight Depth Estimation	Kurita et al. [2024]
Unit-Constrained Data-Driven Turbulence Modeling for Separated Flows Using Symbolic Regression	Zhang and Lei [2024]

Table 6.1: Research papers relying on the Φ-SO framework as of December 2024 — 21 months after its release in March 2023.

described in Chapter 4 on Physical Symbolic Regression and Chapter 5 on Class Symbolic Regression.

Physical SR:

A frozen version related to our work on Physical Symbolic Regression [Tenachi et al., 2023a] is released under tag v1.0.0 and deposited on zenodo: 10.5281/zenodo.8415435.

For the sake of result reproducibility, we offer a simple method to replicate the outcomes presented in Figure 4.3 by simply executing the following command: `python feynman_run.py --equation i --noise n`. This command will run `PhySO` on challenge number `i` $\in \{0, 1, ..., 119\}$ of the Feynman benchmark, employing a noise level of `n` $\in [0, 1]$.

In addition, we include challenge-by-challenge and run-by-run performances results tables at `PhySO/benchmarking/FeynmanBenchmark/results`.

Class SR:

A frozen version related to this work on Class Symbolic Regression is released under tag v1.1.0 and deposited on zenodo: 10.5281/zenodo.11663147.

For the sake of result reproducibility, we similarly offer a simple method to replicate the outcomes presented in Figure 5.2 by simply executing the following command: `python classbench_run.py --equation i --noise n --n_reals Nr`. This command will run `PhySO` on challenge number `i` $\in \{0, 1, ..., 7\}$ of the Class benchmark presented in Table 5.1, employing a noise level of `n` $\in [0, 1]$ and exploiting `Nr` $\in \mathbb{N}$ realizations. We also include the script we used to estimate performances post-run : `classbench_results_analysis.py`.

Similarly, we offer a straightforward method to replicate the outcomes presented in Figure 5.3 by simply executing the following command: `python MW_streams_run.py --noise n --frac_real fr`. This command will run `PhySO` on the stellar stream problem described in Section 5.3, employing a noise level of `n` $\in [0, 1]$ and exploiting a fraction of `fr` $\in [0, 1]$ realizations. Again, we include our results analysis script: `MW_streams_results_analysis.py`.

In addition, we include challenge-by-challenge and run-by-run performances results tables: see `PhySO/benchmarking/ClassBenchmark/results` for results pertaining to the Class SR benchmark and `PhySO/demos/demos_class_sr/demo_milky_way_streams/results` for results pertaining to the stellar stream problem.

Chapter 7

Neural Networks as Symbolic Graph Representations

Portions of the content presented in this Chapter have been previously discussed in the following publication:

2024 *Generalizing the SINDy approach with nested neural networks*
C. Fiorini, C. Flint, L. Fostier, E. Franck, R. Hashemi, V. Michel-Dansac, **W. Tenachi**
ESAIM 24 **1** 1-10, arXiv:2404.15742

Summary.

We examine methods that leverage neural networks to directly capture and embody the graph structure of a dataset, reflecting its underlying analytical representation.

We introduce a method that identifies additive and multiplicative separabilities within data. This approach incrementally deconstructs a dataset into simpler components, progressively uncovering its graph structure to facilitate interpretability and subsequent symbolic regression.

We explore how incorporating non-linear basis functions into a neural network — while promoting sparsity can gradually transform the network into an interpretable, compact symbolic expression. We discuss promising cross-pollination avenues in symbolic regression.

In previous Chapters (3-5), our discussion centered on the use of neural networks to generate and learn from symbolic graph representations of analytical expressions.

In this Chapter, we delve into the use of neural network models to directly embody symbolic expressions within their architectures. This method is pivotal in the field of symbolic regression (SR) — aimed at deriving compact analytical functions that accurately fit datasets — as such direct integration enables neural networks to effectively encapsulate data in an intelligible, symbolic graph format. As we will demonstrate, this approach facilitates the extraction of functional analytical forms, with the representations being learned directly from data points.

An unsupervised learning approach

It is important to note that the methods previously discussed involve sampling expressions from neural networks that tokenize symbols in a manner akin to natural language processing frameworks — although adapted to handle the complexities of computational symbolic mathematics. These methods can be employed within both reinforcement learning (RL) frameworks, where neural networks learn through trial and error to generate expressions that satisfy certain constraints (e.g., fitting data in SR), and supervised learning frameworks, where neural networks learn from a vast number of examples to generate expressions that meet specific conditions (e.g., matching a dataset in SR).

Unsupervised learning has traditionally been considered unfeasible in this context, as selecting one mathematical symbol over another is not a differentiable operation. In RL methods, the neural network indirectly interacts with the data through a non-differentiable scalar reward that reflects fit quality. However, the promise of unsupervised learning lies in its potential for data to directly influence graph structure through gradients, enabling a more direct and potentially insightful learning process.

Nevertheless, as we will demonstrate in this Chapter, representing the expression directly within a neural network architecture enables the application of such unsupervised learning techniques. Unsupervised learning here uniquely facilitates the exploitation of gradients for tuning graph structure with respect to data. This capability allows for the optimization of functional forms in a manner that can complement the reinforcement learning strategies implemented in our Φ-SO framework.

Outline

In Section 7.1, we introduce a method that breaks down a dataset into a graph of simpler sub-functions by exploiting separabilities and symmetries, significantly enhancing approaches akin to the `AIF` (i.e. `AI Feynman`) [Udrescu

and Tegmark, 2020, Udrescu et al., 2020]. Section 7.2 discusses a technique for refining a dataset into an increasingly sparse neural network structure, incorporating polynomial components and non-linear basis functions such as $\{\exp, \log, \cos, \sin, \sqrt{\square}, ...\}$ thereby facilitating the derivation of an analytical function. Finally, Section 7.3 outlines how these methods can be integrated into our Φ-SO framework to broaden its capabilities.

The content in this chapter is the result of collaborative efforts led by Alejandro M. Illescas Gimenez[1] with Rodrigo Ibata and myself (Section 7.1), as well as a collaborative effort equally led by Clément Flint, Louis Fostier, and Reyhaneh Hashemi with Camilla Fiorini, Emmanuel Franck, Victor Michel-Dansac and myself[2] (Section 7.2).

7.1 Uncovering Structures using Differential Precision Learning

Modularity

In this Section, we explore how datasets can be represented within neural networks to detect inherent hierarchical structures related to the input variables, such as separabilities and symmetries. For example, a complex function $f(x_1, x_2)$ can be decomposed into simpler functions $f_1(x_1)$ and $f_2(x_2)$ with $f(x_1, x_2) = f_1(x_1) + f_2(x_2))$, Similarly, symmetries might simplify $f(x_1, x_2)$ into $f_1(x_1, f_2(x_2))$.

Recursive structure search

Once these sub-functions are identified, they can be modeled as distinct neural networks. This modularization allows for an iterative process where each sub-function is further analyzed and simplified if possible. By repeating this process, a complex dataset can be methodically simplified into an intelligible graph of simpler sub-models. This technique is especially valuable in physics, where datasets often display an underlying graph structure [Udrescu and Tegmark, 2020]. This iterative disentangling process is illustrated in Figure 7.1.

[1] As part of Alejandro M. Illescas Gimenez's master research project.

[2] As part of the research project undertaken by Clément Flint, Louis Fostier, and Reyhaneh Hashemi at the Scientific Machine Learning CEMRACS 2023 summer school, supervised by Camilla Fiorini, Emmanuel Franck, Victor Michel-Dansac, and myself hosted at CIRM, Marseille during July 17 — August 25, 2023

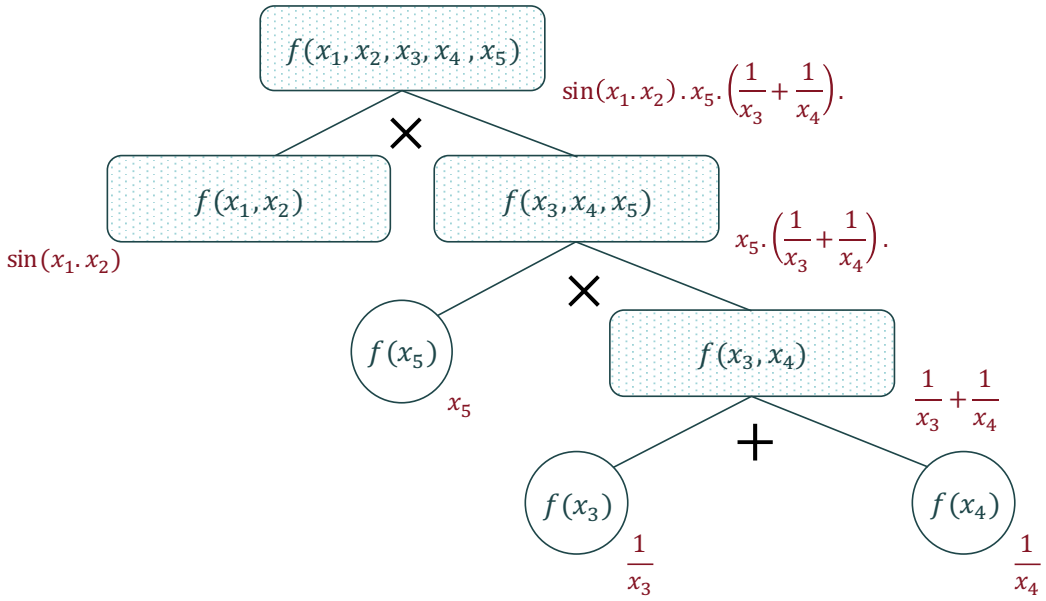


Figure 7.1: **Unveiling graph structures by detecting separabilities in data.** This figure illustrates the process of disentangling a complex data representation into a graph of simpler sub-models through the detection of multiplicative (denoted by '×') and additive (denoted by '+') separabilities. Without prior knowledge of the underlying mathematical expression, our method identifies that the function $f(x_1, x_2, x_3, x_4, x_5) = \sin(x_1 x_2) x_5 \left(\frac{1}{x_3} + \frac{1}{x_4}\right)$ can be decomposed into two distinct sub-functions: $\sin(x_1 x_2) x_5$ and $\left(\frac{1}{x_3} + \frac{1}{x_4}\right)$, which are multiplicatively separable. This iterative process is repeated on each newly identified component until a comprehensive graph structure is obtained, with each node (or leaf) represented by a distinct instantiation of our `GradNet` architecture. Each leaf's underlying functional form is denoted on its side in red.

Symbolic regression

This approach is not only advantageous for enhancing interpretability through structured representations but also significantly simplifies SR processes. By breaking down complex functions into sub-function models that can later be translated into analytical representations, we facilitate a more manageable SR workflow. This "divide and conquer" strategy, which aids in streamlining the entire SR process, was initially introduced by Luo et al. 2017, 2022 and has been further explored in subsequent works [Cranmer et al., 2020b, Tohme et al., 2023, Landajuela et al., 2022], including the popular `AI Feynman` approaches: `AIF` [Udrescu and Tegmark, 2020] and `AIF` [Udrescu et al., 2020].

A first multiplicative separability detection scheme

In this section, we introduce a novel method capable of detecting not just additive separabilities but also multiplicative separabilities — e.g., , identifying structures within $f(x_1, x_2)$ such that it can be expressed as $f_1(x_1).f_2(x_2)$. While methods for identifying multiplicative separabilities have been suggested by Udrescu et al. [2020], they have not been successfully exploited or implemented to date. As we will see, effective exploitation of multiplicative separabilities necessitates exceptionally precise gradients, which in turn require unconventional neural network architectures.

Outline

We introduce the innovative `GradNet` — Gradient Network — architecture, which is pivotal for differential precision learning in sub-section 7.1.1. We then outline our methodology for detecting separabilities in sub-section 7.1.2. Finally, we discuss the results of these techniques in sub-section 7.1.3, highlighting their effectiveness and implications.

7.1.1 The `GradNet` architecture

Physics-informed neural networks (PINNs) have garnered significant attention in the scientific community. Despite the implication of their name, PINNs do not represent a specific architecture; rather, they refer to a training methodology that enforces the physical consistency of models by evaluating their gradients and enforcing that they fit relevant differential equations. This approach allows for the embedding of prior theoretical knowledge directly into the neural network's framework. In contrast, the `GradNet` architecture we present offers a more defined structure that provides perfectly accurate gradients, making it an ideal foundation for PINN-like applications.

The `GradNet` architecture[3] features a shallow structure paired with an accuracy-focused optimizer. This design is pivotal: the shallow architecture ensures the availability of analytically precise gradients, which are crucial for accurate modeling and the innovative[4] optimization approach allows the network to maintain a substantial size through width despite its shallow depth, effectively overcoming the typical challenges associated with shallow networks in complex modeling scenarios.

[3]This architecture is currently being developed by Rodrigo Ibata and collaborators at the Observatoire Astronomique de Strasbourg.

[4]Innovative in the context of deep learning.

A shallow architecture

The `GradNet` utilizes a notably shallow architecture, consisting of a mere single dense layer. This foundational neural network component was previously detailed in 2.1.2.

Given input data $x \in \mathbb{R}^{n_x}$, the output of the model $y \in \mathbb{R}^{n_y}$ is computed as follows:

$$y = a \log(1 + \exp(Kx + b)) \tag{7.1}$$

Where K represents the weight matrix with dimensions (h, n_x), b is a bias vector of size h, and a is a scale matrix of dimensions (n_y, h). The tunable parameters of the model, denoted by θ, include $\theta = \{K, b, a\}$.

This streamlined architecture allows us to derive the Jacobian of the entire neural network model analytically — a distinct advantage over deeper networks, where gradients must be obtained through the iterative application of the chain rule through auto-differentiable frameworks, often leading to inaccuracies in higher-order derivatives. The Jacobian and Hessian of our architecture are computed as follows:

$$\nabla y = K^T a.\sigma(Kx + b) \tag{7.2}$$

$$\frac{\partial^2 y}{\partial x_i \partial x_j} = Ka(1 - \sigma(Kx + b))\sigma(Kx + b)K^T \tag{7.3}$$

Despite the ability to access precise analytical gradients, such shallow neural network architectures typically face limitations in complex modeling tasks, which prompted the development of deeper learning models [LeCun et al., 2015].

Interestingly, despite these practical limitations, the universal approximation theorem [Hornik et al., 1989] suggests that increasing the width — here variable h — of even a single-layer neural network could, in theory, enable it to model any function. Indeed, theoretically, with a sufficiently wide layer an ideal parametrization in a, K, and b can approximate any desired smooth function. This principle highlights the potential of shallow architectures under ideal conditions, although practical constraints often necessitate deeper configurations.

Optimization

In practice, even with a sufficiently large layer, finding the optimal parameter values using standard optimization procedures such as Adam [Kingma and Ba, 2015] or Stochastic Gradient Descent (SGD) [Robbins and Monro, 1951] can be infeasible. These conventional methods rely on minimizing a scalar

loss function, $\mathcal{L}_\theta$, which depends on the model parameters θ. They rely on derivatives of this cost function with respect to each parameter, obtained via auto-differentiation (this concept was detailed in 2.1.4): $\left\{\frac{\partial \mathcal{L}}{\partial \theta_1}, \frac{\partial \mathcal{L}}{\partial \theta_2}, \ldots\right\}$.

To enhance our optimization capability, we implement the sophisticated second-order Levenberg-Marquardt algorithm [Levenberg, 1944, Marquardt, 1963]. This approach uses the Jacobian matrix $J = \partial f(x_i)/\partial \theta_j$ of the model $f(x_i)$ together with the Hessian matrix approximated as $J^T J$ to solve a simultaneous system of M equations (one for each of the M parameters of the problem) for the parameter update vector $\Delta\theta$ that will minimize χ^2. The method thus directly derives an estimate of the ideal update vector, instead of simply deriving the direction in which the update should move to reduce the cost function (like SGD, for instance). For a detailed review of this method, see Ranganathan [2004].

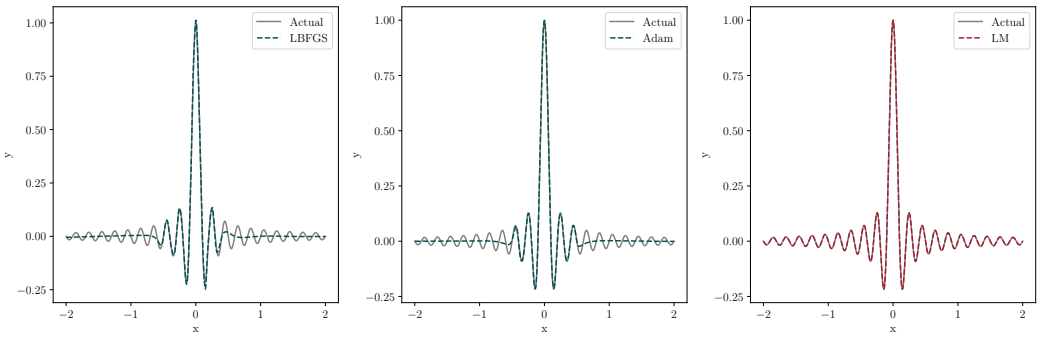


Figure 7.2: **Comparative efficacy of optimization algorithms: Adam, LBFGS, and LM.** This figure demonstrates the performance of the Levenberg-Marquardt (LM) algorithm [Levenberg, 1944, Marquardt, 1963], comparing it against the Adam optimizer [Kingma and Ba, 2015] and LBFGS [Zhu et al., 1997] on a toy multi-layer perceptron modeling the function $\sin(10\pi x)/(10\pi x)$. The LM algorithm shows enhanced optimization capabilities, achieving more accurate fits and faster convergence, illustrating its superior handling of complex model adjustments compared to traditional first-order methods like Adam and specialized methods like LBFGS designed for non-linear optimization tasks.

This optimization technique has proven significantly more effective at accurately capturing low scales of data, as demonstrated in Figure 7.2. It also reduces the number of necessary parameters and achieves faster convergence due to its capacity to utilize more comprehensive information [Taylor et al., 2022]. While the Levenberg-Marquardt (LM) algorithm is highly effective, it can be resource-intensive, primarily due to the matrix inversion step it involves, which typically limits the size of neural networks it can optimize. However, this constraint is manageable in physical systems, which are fundamentally less complex, and we further mitigate this issue by implementing parameter batching.

Implementing the LM algorithm in popular frameworks like `PyTorch` or `TensorFlow` is non-trivial due to its requirement for a matrix of gradients rather than a simple vector[5]. In our work, we address this challenge by leveraging `PyTorch`'s custom vectorization features and by batching along parameters, enabling efficient execution of the LM optimization process within these frameworks.

7.1.2 Detecting separabilities

In our methodology, we employ the `GradNet` to model any given dataset $\{\mathbf{x}, y\} = \{x_1, ..., x_n, y\}$, where $y = f(\mathbf{x})$. This approach provides us with access to accurate analytical gradients, which are crucial for detecting separabilities in the data. Our system not only enhances efficiency beyond that of `AIF 2`, but it also facilitates the implementation of multiplicative separability approaches, as suggested by Udrescu et al. [2020]. This capability significantly broadens the potential applications and effectiveness of our model in identifying hierarchical structures within datasets.

Additive separability

We evaluate additive separability by determining if the dataset can be decomposed into sub-functions f_a and f_b, which operate on distinct subsets of the input variables $\mathbf{x_a}$ and $\mathbf{x_b}$, respectively, such that $y = f_a(\mathbf{x_a}) + f_b(\mathbf{x_b})$.

To assess the additive separability between two given variables x_i and x_j $(i \neq j)$, we verify the following condition:

$$\frac{\partial^2 y}{\partial x_i \partial x_j} < \epsilon_{\text{add}} \tag{7.4}$$

where ϵ_{add} is a small numerical threshold empirically determined to indicate negligible interaction between the variables.

After testing separability across all variable pairs, in cases where multiple separable configurations are possible, we prioritize the configuration that minimizes the number of variables in $\mathbf{x_b}$. This approach aims to simplify the function f_b as much as possible. This process is recursive, allowing for further separations to be explored in subsequent iterations.

Multiplicative separability

We assess multiplicative separability by identifying if the dataset can be decomposed into sub-functions f_a and f_b, operating on distinct subsets of input variables $\mathbf{x_a}$ and $\mathbf{x_b}$, such that $y = f_a(\mathbf{x_a}) \cdot f_b(\mathbf{x_b})$.

[5]The `JAX` framework is a promising alternative as it inherently supports operations with matrices of gradients.

To establish multiplicative separability, we leverage the property that a function's logarithm transforms multiplicative separability into additive separability. Specifically, this transformation allows y to be represented as $y = \log f_a(\mathbf{x_a}) + \log f_b(\mathbf{x_b})$. Therefore, to test multiplicative separability between two variables x_i and x_j $(i \neq j)$, we evaluate the following criterion:

$$\frac{\partial^2 \log y}{\partial x_i \partial x_j} < \epsilon_{\text{mul}} \tag{7.5}$$

where ϵ_{mul} is an empirically determined numerical threshold. The second derivative of the logarithm of y with respect to x_i and x_j is given by:

$$\frac{\partial^2 \log y}{\partial x_i \partial x_j} = \frac{1}{y}\frac{\partial^2 y}{\partial x_i \partial x_j} - \frac{1}{y^2}\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j} \tag{7.6}$$

This analytical approach helps us ascertain whether interactions between x_i and x_j can be expressed in a multiplicative manner.

Addressing the offset problem

While exploring multiplicative separability, we must consider potential offsets in the functional relationship, such as $y = f_a(x_1) \cdot f_b(x_2) + d$. Direct application of logarithmic transformation on such functions, where d is non-zero, disrupts the additive separability. To address this, we adopt an initial assumption of multiplicative separability and concurrently estimate an offset d that satisfies:

$$(y - d)\frac{\partial^2 y}{\partial x_i \partial x_j} = \frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j} \tag{7.7}$$

across all $(\mathbf{x}, y)$ pairs in the dataset.

In practice, we compute d for each pair and evaluate its median. If the median of d is less than a predefined threshold ϵ_{off}, we consider that $d = 0$ and proceed with the condition described in Equation 7.5. If the offset proves significant, we calculate the Median Root Squared Relative Deviation (MRSRD) of d:

$$\text{MRSRD} = \frac{1}{\text{med}(d)} \cdot \text{med}\left(\sqrt{(d_i - \text{med}(d))^2}\right)_{i<N} \tag{7.8}$$

where med denotes the median. The separability is accepted if:

$$\text{MRSRD} < \epsilon_{\text{MRSRD}} \tag{7.9}$$

This intricate methodology hinges on the availability of highly accurate derivative values, a requirement not met by standard multi-layer perceptron architectures (such architectures are detailed in 2.1.2) used in approaches like `AIF 2`, making an architecture such as the `GradNet` indispensable.

7.1.3 Performances

We evaluate the effectiveness of our graph structure detection method, which capitalizes on the detection of separabilities, using the Feynman benchmark as proposed by Udrescu and Tegmark [2020]. This benchmark was initially tailored to assess approaches like ours that focus on structural discovery in data as it was first proposed to evaluate the `AIF` algorithm (refer to Section 4.3 for a detailed discussion on this benchmark).

Hyper-parameters

The hyper-parameter values, which we have empirically determined to yield optimal results on the Feynman benchmark, are documented in Table 7.1. These values have been fine-tuned to enhance the performance and accuracy of our separability detection procedure.

Criterion	Value
ϵ_{add}	10^{-4}
ϵ_{mul}	5×10^{-3}
ϵ_{off}	5×10^{-3}
ϵ_{MRSRD}	10^{-1}

Table 7.1: Hyper-parameter values for our separability detection procedure.

Results

Upon applying our approach to the 44 cases that our Φ-SO algorithm for SR could not resolve in Section 4.3, we achieved notable results. Our method accurately judged separabilities in 31 out of the 44 cases, culminating in a 70% success rate. Success was categorized as follows: complete simplification to the simplest form possible in 14 cases (32%), partial simplification in 9 cases (20%), and correct identification of non-separability in 8 cases (18%). Despite these positive outcomes, there remain 13 instances (29%) where the algorithm misjudged separability, either through false positives or by failing to detect existing separabilities due to training limitations.

These results are particularly impressive, given that they address the most challenging and complex problems within a benchmark specifically designed to test such capabilities. Our approach demonstrates the potential of directly modeling physical phenomena within a neural network structure, which allows for the recursive application of additive and multiplicative separability detection, thereby uncovering the underlying graph structure of the data. We

anticipate that integrating our Φ-SO framework with this approach — a development currently in progress — will result in exceptional performance on the Feynman benchmark.

7.2 Nested SINDy

This Section explores an alternative approach for encoding hierarchical symbolic structures within neural networks from data, which can complement the methods previously discussed.

Supervised learning approaches to SR offer rapid inference but lack a self-correction mechanism. If the generated expression is suboptimal, there are little means of correction. In contrast, unsupervised approaches enable iterative correction based on fit quality. However, they often rely on reinforcement learning frameworks (as we did in Chapters 3-5) to approximate gradients because direct optimization using auto-differentiation is infeasible due to the discrete nature of the problem, which involves discrete symbolic choices.

However, other unsupervised methods include neuro-symbolic approaches, wherein mathematical symbols are integrated into neural network frameworks. The goal being to sparsely fit the neural network to enable interpretability, generalization or even recover a compact mathematical expression. Prominent examples include SINDy [Brunton et al., 2016], which stands for Sparse identification of non-linear dynamics and is central to this study, and others such as Martius and Lampert 2017, Scholl et al. 2023, Sahoo et al. 2018, Valle and Haddadin 2021, Kim et al. 2020, Panju and Ghodsi 2020, Ouyang et al. 2018.

SINDy-like approaches are the only type of unsupervised techniques capable of directly utilizing gradients from data to iteratively refine function expressions as they effectively render the discrete symbolic optimization problem continuous. Moreover, SINDy-like frameworks possess the advantage of being well-suited for exact symbolic recovery by enabling the creation of concise, intelligible analytical expressions through the promotion of sparse symbolic representations while yielding highly accurate and general expressions when exact symbolic recovery is unsuccessful or impossible. However, a limitation of the current SINDy framework is its inability to handle nested symbolic functions, which often results in suboptimal performances, especially in more complex problems as shown by our comparative benchmark in 4.3. This is the primary motivation for our study, where we introduce a Nested

SINDy approach.

We outline our nested SINDy architecture in sub-section 7.2.1, describe our fitting procedure and sparsity enforcement methods in sub-section 7.2.2, and present our findings in sub-section 7.2.3.

7.2.1 Architecture: going deeper

The standard SINDy approach involves fitting polynomial combinations of differentiable non-linear basis functions. This framework can theoretically incorporate an extensive array of non-linear functions, provided they are differentiable. For the purposes of this study, we consider the following basis functions: $\{\mathrm{id}, \square^2, \arctan, \sin, \cos, \exp, \sqrt{\square}, e^{-\square^2}, \log(1+e^{\square})\}$.

Given N data points (x, y), our objective is to discover a function f such that $y = f(x)$ — a SR challenge. We define $\mathcal{F} = \{f_1, \ldots, f_l\}$ as the library of possible basis functions. The space of linear combinations of these functions, denoted by $L(\mathcal{F})$, is defined as follows:

$$f \in L(\mathcal{F}) \iff \exists \theta \in \mathbb{R}^l, \text{ such that } f = \sum_{i=1}^{l} \theta_i f_i. \tag{7.10}$$

Recognizing the limitations of shallow architectures in capturing complex relationships, our goal is to explore deeper, more structured models to enhance the expression capability and accuracy of SR solutions.

Polynomial-Radial model

The Polynomial-Radial (PR) model introduces a structured layering approach by incorporating a polynomial transformation prior to the standard SINDy 'radial' non-linear operations. This dual-layer architecture enhances the basic SINDy framework by first applying a polynomial layer that manipulates the input variables into a series of monomial forms.

Specifically, the polynomial layer is defined as:

$$f_{\text{poly}}(x) = \sum_{i=0}^{d} \omega_i x^i,$$

where d denotes the highest degree of polynomial allowed, and ω_i represents the coefficients. The coefficient ω_0 serves as the constant component, analogous to the bias in traditional neural network layers.

For multidimensional inputs, this layer extends to handle multivariate polynomials, enabling intricate combinations of input variables. For instance, for

two variables x and y, the polynomial becomes:

$$f_{\text{poly}}(x, y) = \sum_{i=0}^{d} \sum_{j=0}^{d} \omega_{i,j} x^i y^j,$$

with $\omega_{0,0}$ representing the polynomial's constant term. To manage complexity and the number of terms, the summation can be constrained so that $i + j \leq d$, forming bivariate polynomials up to degree d. This formulation not only broadens the modeling capacity but also embeds a deeper interaction between variables right from the initial transformation stage.

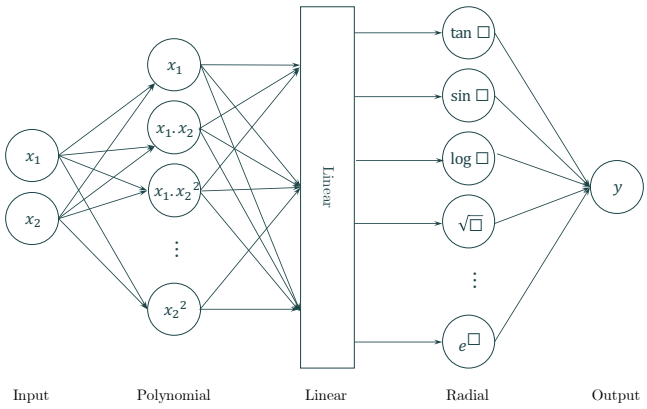


Figure 7.3: **Polynomial-Radial architecture of our Nested SINDy approach.** See paragraph 7.2.1 for a detailed description.

As an example, given one input variable, if $\mathcal{F} = \{\sin, \cos\}$ and $d = 2$, then the PR model can express functions such as:

$$\lambda \cos(a_1 + b_1 x + c_1 x^2) + \mu \sin(a_2 + b_2 x + c_2 x^2).$$

This opens up much more combination than the standard SINDy approach, where functions such as $\cos(2\square)$ or $\sin(1 + \square^2)$ would have to be manually added to the library to be able expres this combination.

This PR layer can also be seen as a pure polynomial layer combined with a linear layer. We, hence, consider the PR model to have four layers: a polynomial layer, a linear layer, a radial layer, and a final linear layer. The last two layers are identical to the standard SINDy model, while the first two layers are new additions. 7.3 illustrates this structure.

The entire expression for the PR model can be formalized as follows:

$$f_{\theta,\mathrm{PR}}(x) = \sum_{j=1}^{l} c_j f_j \left(\sum_{i=0}^{d} \omega_{i,j} x^i \right) + B, \tag{7.11}$$

where θ encapsulates all trainable parameters of the model. Where f_j are functions selected from the predefined library $\mathcal{F}$ and $\omega_{i,j}$ represent the coefficients of the polynomial transformations for each radial basis function f_j. c_j and B denote the coefficients and bias of the output linear layer, respectively.

This architecture introduces a significant enhancement in expressivity by allowing for the integration of complex polynomial transformations prior to applying the radial functions. The PR model is particularly effective in handling multivariate data by constructing and leveraging intricate inter-variable interactions before further transformations. In contrast to conventional SINDy approaches, this model reduces the dependency on a large function library by enabling discovery and optimization of linear combinations of functions directly from the data.

Our experimental results show that the PR model can efficiently converge to accurate solutions even in complex SR tasks, demonstrating substantial improvements over traditional methods. The efficacy of this approach in solving problems will be detailed in 7.2.3, where we discuss the model's performance in a few test scenarios.

Polynomial-Radial-Polynomial model

The Polynomial-Radial-Polynomial (PRP) model builds upon the PR model by incorporating an additional polynomial layer following the radial function layer. This advancement significantly broadens the model's capability to capture complex data interactions that were previously unattainable. For instance, the PRP model is capable of expressing functions such as $f(x) = \arctan(x)\sin(x)$, provided that arctan and sin are included in the function set $\mathcal{F}$, surpassing the PR model's limitations in representing such multiplicative interactions between functions.

In the PRP model, the outputs from the radial layer pass through a linear transformation layer, typically constrained in size (in our implementation, we fix it to 2 units), acting as an intermediary that reshapes these outputs into new variable forms. These transformed outputs are subsequently processed by another polynomial layer, which facilitates the creation of diverse monomial forms by combining these intermediate variables in various ways, such as squaring or product terms.

This layered architecture not only enhances the model's expressivity but also provides a flexible framework capable of approximating complex non-linear phenomena within datasets more effectively. Such structural complexity allows

the PRP model to achieve a higher degree of representation accuracy and adaptability across varied SR tasks.

The expression for the PRP model can be formalized as follows:

$$f_{\theta,\mathrm{PRP}}(x) = \sum_{1\leq|i|\leq d} \omega^{\mathrm{PR}}_{i_1,i_2,\ldots,i_l} f_{\theta_1,\mathrm{PR}}(x)^{i_1} f_{\theta_2,\mathrm{PR}}(x)^{i_2} \ldots f_{\theta_l,\mathrm{PR}}(x)^{i_l} + B', \quad (7.12)$$

Where $|i| = i_1 + i_2 + \ldots + i_l$ is the length of the multi-index $(i_1, \ldots, i_l)$, θ includes all the trainable parameters of the model, $\omega^{\mathrm{PR}}_{i_1,i_2,\ldots,i_l}$ are the weights of the final linear layer, B' is the bias of the final linear layer, l is the size of the output chosen for the intermediate linear layer (set to 2 in our experiments), and $f_{\theta_1,\mathrm{PR}}(x)$, $f_{\theta_2,\mathrm{PR}}(x)$, $\ldots$, $f_{\theta_l,\mathrm{PR}}(x)$ correspond to the output of the PR model given in 7.11. The coefficients $\theta_1, \theta_2, \ldots, \theta_l$ correspond to the parameters of the PR models, which are the same as θ, except for the weights of the final linear layer (c_j and B in 7.11). 7.4 gives a graphical representation of this model.

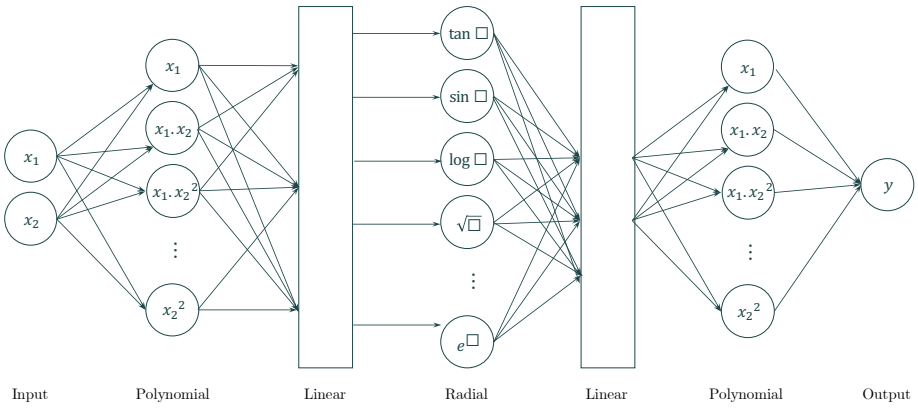


Figure 7.4: **Polynomial-Radial-Polynomial architecture of our Nested SINDy approach.** See paragraph 7.2.1 for a detailed description.

The inclusion of a second polynomial layer significantly enhances the PRP model's capacity to express complex mathematical relationships. This additional layer allows the model to capture more nuanced interactions within the dataset, proving particularly advantageous for analyzing complex datasets where simpler models might not suffice. The PRP model's ability to handle intricate interactions between variables makes it exceptionally effective for such scenarios.

In conclusion, the PRP model, with its dual polynomial layers, represents an advanced evolution of the SINDy approach. It offers a robust framework

capable of modeling sophisticated systems, adept at uncovering higher-order interactions and non-linear dynamics prevalent in complex datasets. This enhanced expressivity makes the PRP model a valuable SR approach.

7.2.2 Fitting method

Enforcing sparsity

Given the potentially complex architectures described earlier, which can include hundreds of parameters and given our goal of obtaining compact and interpretable models that can be written as analytical expressions, we employ specific strategies to prune as many network parameters as possible. To achieve this, sparsity is enforced during the model fitting process. We incorporate a lasso regularization term [Tibshirani, 1996] into the loss function, which encourages the reduction of parameter magnitudes, effectively pruning negligible weights during training when they fall below a defined threshold. The overall loss function for our model is articulated as follows:

$$\mathcal{L}_\theta = \mathcal{L}_{\text{MSE},\theta} + \mathcal{L}_{\text{Lasso},\theta} \tag{7.13}$$

$$\mathcal{L}_\theta = \sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 + \sum_{i=1}^{l}|\theta_i| \tag{7.14}$$

Here, λ is the regularization parameter that balances the mean squared error (MSE) and the lasso penalty, ensuring that complexity is directly managed through the optimization of the loss function, promoting a balance between accuracy and model simplicity.

In practice, we remove a parameter θ_i from the active set if it remains below a threshold ϵ_{prune} over n_{prune} consecutive epochs, provided that the overall mean squared error (MSE) is below a predefined threshold $\text{MSE}_{\text{prune}}$. This approach allows us to learn the model's structure in an unsupervised manner that directly leverages the information within the data, promoting an optimal balance of model complexity and explanatory power.

Learning strategy

Our Nested SINDy approach introduces more local minima than the standard SINDy approach, primarily due to the addition of linear layers which, when combined with non-linear layers, can create complex optimization landscapes. To mitigate the risk of being trapped in these local minima, we have implemented several empirically effective learning strategies.

Firstly, we dynamically adjust the importance of the Lasso component in the loss function across iterations (it): $\lambda(\text{it}) = \lambda_0(1 + 0.4\sin(\text{it}/10))$, where λ_0 is the initial Lasso coefficient. This oscillation between $0.6\lambda_0$ and $1.4\lambda_0$ is facilitated by a sine function, allowing for periodic adjustments to the regularization strength, which helps escape local minima.

In addition, we opt for the LBFGS optimizer [Zhu et al., 1997], an unconventional choice in neural network training due to it being designed for optimizing constants in non-linear analytic functions. However, given its efficacy and given that here the optimization of neural network parameters shares similarities with constant optimization in non-linear setups — although with many more constants in this case — its application here is particularly beneficial.

7.2.3 Results & discussion

Results

Employing the PR and PRP frameworks, we successfully recovered very simple models such as $-\sin(x^2 + 1)$ directly from the data, demonstrating the frameworks' capability to prune from a large parameter space efficiently. The PRP architecture proved particularly effective, recovering more complex models such as $2\cos(x_1)\cos(x_2)$, showcasing its ability to handle multiplicative relationships between functions not explicitly included in the library — a significant improvement over traditional SINDy methods. Typically, a naive SINDy approach would require a significantly larger function library to accommodate potential products of functions for multi-dimensional inputs.

In a distinct experiment involving the perimeter of ellipses, defined by the parametric equations $x = a\cos(\alpha)$ and $y = b\sin(\alpha)$ for $\alpha \in [0, 2\pi)$, we explored the limitations of elementary functions in describing ellipse perimeters. The perimeter, traditionally computed using an integral[6], has been approximated through various formulas, including Ramanujan's well-known approximation: $P(a, b) \approx \pi(3(a + b) - \sqrt{(3a + b)(a + 3b)})$. Utilizing our PRP model, we developed a new approximation which shows increased accuracy for larger values of a $(a > 5)$[7]:

$$\begin{aligned} P_{\text{PRP}}(a) = {} & 0.535a + 0.966(0.394a + 0.721\arctan(0.278a^2 + 0.393) + 1 \\ & + 0.111\exp(-0.063a^4))^2 + 0.978\arctan(0.278a^2 + 0.393) + 1.36 \\ & + 0.15\exp(-0.063a^4), \end{aligned}$$

[6] The perimeter is given by $P(a, b) = 4\int_0^{\frac{\pi}{2}} \sqrt{a^2\cos^2(\alpha) + b^2\sin^2(\alpha)}, d\alpha$

[7] Assuming $b = 1$ for simplicity, since scaling does not alter the elliptical nature.

These results highlight the practical efficacy of our model in generating accurate and complex approximations that challenge existing models.

Discussion

Conclusion

While challenges applying this approach to more diverse SR tasks remains challenging due to its inability to express compositions of non-linearites, the results demonstrate its potential to derive new combinations of basis functions. These novel combinations can serve as effective basis functions in various physical contexts and assist SR methodologies, as further explored in the next Section. Moreover, our approach has shown a capacity at inferring complex but highly accurate and computationally efficient approximations, extending its utility beyond mere interpretability.

Perspective

Another promising direction for extending SINDy-like approaches involves integrating the basin-hopping algorithm, which combines the LBFGS optimizer with global search techniques to circumvent local minima, as suggested by Scholl et al. [2023]. Additionally, incorporating a supervised learning component could further enhance the framework's efficacy. As proposed by Scholl et al. [2023], pre-training a model to identify relationships between datasets and sparse SINDy-like patterns could provide high-quality initial guesses that can subsequently be fine-tuned to specific cases, thereby streamlining the process of deriving analytical expressions. This hybrid strategy could substantially improve the initial condition setup, allowing for more focused and effective refinement phases.

7.3 Synergies & Perspectives

This Section explores the integration of the aforementioned unsupervised approaches utilizing a neural network to represent and discover hierarchical structures, with the Φ-SO reinforcement learning framework detailed in Chapters 4-6. We will examine how these methodologies can complement each other, enhancing the overall efficacy and scope of SR applications within our framework.

Sub-section 7.3.1 discusses the potential integration of a nested SINDy model as a token within our reinforcement learning framework. Sub-section 7.3.2 explores how leveraging the Levenberg-Marquardt optimizer one can enhance neuro-symbolic SINDy-like approaches. Finally, sub-section 7.3.3 examines the incorporation of these approaches within the Φ-SO framework.

7.3.1 A nested-SINDy token

Token

The nested-SINDy approach, with its capability to compose relevant higher level functions from simple basis functions by exploiting data directly, offers a unique opportunity for integration into the Φ-SO framework. Considering that nested-SINDy essentially functions as a comprehensive sub-expression containing numerous constants to optimize, it aligns well with the constant optimization problems typical in our Φ-SO framework. By treating nested-SINDy as a token, it can complement existing tokens which represent simple functions. This special token would not only contain multiple free parameters requiring optimization but also employ a unique pruning process to encourage sparsity. Such a token would be particularly useful in complex expressions where sub-components demand intricate modeling, allowing Φ-SO to utilize a nested-SINDy token for these parts, optimizing it alongside other constants during the fitting phase.

Example

For instance, consider the challenge of performing SR on a dataset whose underlying function is $e^x - \sin(x^2 + 1)$. While this complexity might be beyond the individual capabilities of nested-SINDy, integrating it as a τ_{PRP} token within Φ-SO could simplify the task. By representing the expression as $e^x + \tau_{\text{PRP}}$ (i.e., in prefix notation $\{+, \exp, x, \tau_{\text{PRP}}\}$), the τ_{PRP} token could autonomously model $-\sin(x^2 + 1)$ in an unsupervised manner. This approach significantly enhances Φ-SO's efficiency by abstracting a complex sub-function into a manageable token, rather than requiring the framework to deduce the entire function from scratch, which would conventionally be represented as $\{+, \exp, x, -, \sin, +, \square^2, x, 1\}$).

Comparison with existing literature

This proposed approach extends the concept of the *linear token* found in `uDSR` [Landajuela et al., 2022], which allows for a linear combination of input variables within the symbolic regression process. However, our approach offers enhanced capabilities for two primary reasons:

First, unlike uDSR, which does not utilize auto-differentiation for constant optimization and thus struggles with fitting numerous constants, our method leverages auto-differentiation. This reliance on auto-differentiation avoids the need to computationally invert sub-functions across their ancestral graph to resolve linear relationships, simplifying the optimization process.

Second, our specialized token transcends the mere linear combinations of uDSR, incorporating a more versatile modeling capacity. It can represent complex polynomial relationships involving non-linear functions of input variables. This capacity for handling sophisticated mathematical structures significantly augments the expressive power of our symbolic regression toolkit, enabling it to tackle more complex and nuanced problems.

7.3.2 Enhancing neuro-symbolic methods with LM optimization

Considering future developments, the Levenberg-Marquardt (LM) optimizer [Ranganathan, 2004] presents a promising enhancement for neuro-symbolic approaches, particularly for unsupervised learning in physics-related problems as outlined in 7.1.1. In our reinforcement learning discussions in 3.3.2, the optimization centered on a scalar objective function-referred to as a reward-which did not directly utilize data-driven gradients and thus required reinforcement learning to approximate these gradients. Conversely, in 7.2, we leveraged scalar objective functions with direct data-derived gradients, allowing for more data-informed constraints on the functional forms.

Building on this, integrating the LM optimization process could further refine these approaches by utilizing the Jacobian, which expresses the derivatives of each adjustable parameter relative to each data point. This method can capture a richer set of data-specific information, potentially transforming the landscape of unsupervised learning by enhancing model accuracy and reducing dependency on large data samples. The ability of the LM optimizer to handle complex, non-linear relationships with high precision makes it a compelling choice for advancing neuro-symbolic integration, where precise gradient information is crucial.

7.3.3 Incorporating graph structure priors from AIF

AIF as a pre-processor

As demonstrated in Section 7.1, identifying the underlying graph structure of a dataset through the detection of separabilities offers significant advantages. This approach enhances SR by decomposing complex SR tasks into simpler, manageable sub-tasks. Each sub-function discovered within the graph structure can be independently resolved through SR, and the resulting expressions

can subsequently be integrated to construct the comprehensive model. This method not only streamlines the SR process but also potentially increases the accuracy and interpretability of the results. Such a strategy of using `AI Feynman`-style pre-processing has been effectively employed in `uDSR` [Landajuela et al., 2022], yielding impressive outcomes, particularly on the Feynman benchmark, which is specifically designed to test the capabilities of such approaches.

`AIF` as a prior

While `AIF`-style approaches are transformative in revealing potential graph structures within datasets, they are also susceptible to producing false positives, leading to partially incorrect graphs. To mitigate against this, such findings could be utilized as preliminary priors. As previously outlined in 3.2.3, it is feasible to integrate deterministic priors into methods where expressions are sequentially generated. Consequently, we propose enhancing the Φ-SO framework to leverage an initial graph structure provided by our approach described in Section 7.1. This integration would enable Φ-SO to iteratively assess the reliability of these priors and adaptively adjust or discard them if they are found to be erroneous.

This approach not only increases safety by reducing reliance on potentially incorrect graph structures but also enhances the optimization process. By optimizing all constants collectively rather than independently for each subfunction potentially fraught with noise induced inaccuracies, it ensures a more cohesive and robust learning environment. Furthermore, allowing a single recurrent neural network to generate all sub-functions within the discovered graph structure encourages cross-pollination of features and insights across different parts of the model.

Moreover, integrating an `AIF`-like discovery tool with a comprehensive SR approach enhances the overall capability of the system. While `AIF` methods excel at simplifying the SR problem, they typically do not extend to solving it, often resorting to elementary brute-force methods as seen in Udrescu and Tegmark [2020]. By coupling these methods with a robust SR framework such as Φ-SO the resulting combined approach has the potential to effectively solve complex SR challenges.

Comparison with existing literature

Encoding an entire expression directly into a neural network while imposing and discovering separabilities and symmetries is the foundation of Kolmogorov-Arnold Networks (KANs) [Liu et al., 2024], which adopt a representation strategy similar to SINDy, but replace basis functions with free-form spline functions. While this adds flexibility, it also introduces challenges in terms of

interpretability and complexity, which may impede the extraction of an analytical symbolic representation. Furthermore, KANs impose a condition that each spline function should operate independently on one dimension at a time, facilitating the detection of separabilities and promoting sparse, simple representations akin to those in SINDy-like methods and `AI Feynman` approaches.

However, there is no guarantee that the learned spline functions learned will be sufficiently simple to translate into analytical forms, thus limiting full SR capabilities. Although utilizing free-form spline functions simplifies the modeling compared to using predefined basis functions, it shares many of the training difficulties inherent in SINDy-like approaches, such as achieving effective learning and convergence.

KANs represent a compelling midpoint between our approach, which prioritizes interpretability with a secondary focus on accuracy, and traditional black-box neural networks, which prioritize accuracy at the expense of transparency. They could be described as "grey boxes" that occasionally can be converted into "transparent boxes", offering a balance between understanding and performance.

Chapter 8

Dark Matter at the Galactic Scale

Summary.

We examine the prevailing cold dark matter paradigm that dominates cosmology, focusing particularly on the challenges it encounters at smaller, galactic scales. We discuss alternative models that may offer solutions to these discrepancies. We then review key galactic dynamics concepts that are critical for galactic dark matter studies.

Special emphasis is placed on the Milky Way and the data provided by the Gaia space telescope. We specifically highlight how studying stellar streams — long, thin tidal debris formed by the accretion of smaller structures into our host Galaxy — can help constrain the nature of dark matter.

In building entirely new machine learning methods for physics and astrophysics one could easily drift towards abstractions, which underscores the need to ground these innovations in tangible physics & astrophysics research. The enigma of dark matter (DM) [Bullock and Boylan-Kolchin, 2017] is a perfect example of such a challenge, presenting a unique opportunity to explore potential new physics, while also serving as an ideal proving ground for these methods.

In this context, we delve into the issue of constraining the nature of DM. Despite its success in explaining cosmic phenomena on large scales, the prevailing DM models encounter significant obstacles at smaller scales — a situation that often signals the need for new physical theories. The Galactic scale, in particular, allows for an intimate examination of DM, supported by a wealth of high-resolution astronomical data.

This Chapter sets the foundation for further discussions exploring the constraining on DM through the study of its effects at the galactic scales, setting the stage for subsequent Chapters focused on this theme. In Section 8.1, we provide an overview of the contemporary DM paradigm and highlight its challenges at smaller scales. Section 8.2 examines the influence of DM on galactic dynamics, introducing essential concepts. Lastly, Section 8.3 explores how our own galaxy, the Milky Way (MW) aligns with these theories and discusses methodologies to further constrain the characteristics of DM through galactic studies.

8.1 Dark Matter

This section offers a concise overview of the Cold Dark Matter (CDM) paradigm, which currently serves as the dominant model in cosmology (sub-section 8.1.1). We will also discuss the challenges this model faces at smaller scales (sub-section 8.1.2) and briefly explore potential resolutions to these tensions (Sub-section 8.1.3).

8.1.1 The cold dark matter paradigm

Dark matter (DM) is a fundamental component of the standard cosmological model, accounting for approximately 85% of the universe's total matter content, or about 27% of its total energy budget. In contrast, dark energy consti-

tutes about 68%, and baryonic matter-comprising protons and neutrons-makes up only about 5% [Planck Collaboration, 2016]. Within our own MW, DM is estimated to constitute roughly 90% of the Galaxy's mass, with estimates ranging from 0.90×10^{12} to $1.40 \times 10^{12} M_{\odot}$ [Cautun et al., 2020]. This mass predominantly forms a halo extending about 500 kpc, while luminous matter contributes only about $5 \times 10^{10} M_{\odot}$ within a stellar halo of approximately 50 kpc [Callingham et al., 2019, Cautun et al., 2020].

Despite not having been directly observed, multiple lines of evidence support the existence of DM. These include the rotation curves of galaxies [Rubin and Ford Jr, 1970], gravitational lensing effects [Bartelmann, 2010], the distribution of mass during galactic collisions [Clowe et al., 2004], and constraints from the cosmic microwave background [Smoot et al., 1992]. Given its overwhelming abundance relative to baryonic matter, DM predominantly governs gravitational interactions both at galactic scales and at larger scales.

The quest for the dark matter particle

Direct search:

In both astrophysics and particle physics, determining the nature of DM is pivotal as particle physics aims at modeling all matter and elementary interactions via a set of particles, yet DM eludes direct observation, detectable only through its gravitational influences. It is hypothesized to be non-interactive through electromagnetic forces and minimally interactive, if at all, through weak nuclear forces. Current experimental constraints place the cross-section for DM interaction via the weak nuclear force at less than $10^{-42}\,\mathrm{cm}^2$, assuming a DM particle mass greater than 8 GeV [Agnese et al., 2014]. This is exceedingly small compared to even the most unlikely neutron capture events, such as hydrogen neutron capture at 10^5 eV, which has a cross-section greater than $10^{-31}\,\mathrm{cm}^2$ [Kopecky et al., 1997].

Direct searches for DM involve constructing large-scale astro-particle detectors on Earth, designed to detect potential DM particles directly. These experiments are crucial for advancing our understanding of DM's weak interactions, if any [Gascon, 2015]. However, as experimental constraints on the cross-section of DM particles become progressively lower, the prevailing view is shifting towards a consensus that DM may not interact via the weak force at all [Oks, 2021]. Recent results from the XENONnT experiment [Aprile et al., 2012] have further constrained the interaction cross-section of dark matter particles, approaching the sensitivity limits posed by the neutrino floor[1].

[1]The neutrino floor refers to a limit in direct dark matter detection experiments where the signal from neutrinos, ubiquitous subatomic particles, becomes indistinguishable from potential dark matter signals. This background noise creates a fundamental threshold that

Beyond the standard model:

Numerous extensions to the standard model of particle physics have been proposed to accommodate DM, each suggesting various hypothetical particles [Bullock and Boylan-Kolchin, 2017]. Notably, the concept of weakly interactive massive particles (WIMPs), with mass-energy around $m.c^2 \sim 100\,\mathrm{GeV}$[2], emerged from theories predicting a freeze-out production/destruction process active until the Universe cooled to a temperature T such that $k.T \sim m.c^2$. This scenario has been famously dubbed the *WIMP miracle*. Potential WIMP candidates include particles predicted by extensions such as supersymmetry [Sohnius, 1985] and the Little Higgs models [Schmaltz and Tucker-Smith, 2005]. Additionally, more conservative extensions of the standard model suggest other candidates like axions, which could also resolve the strong interaction charge/-parity symmetry problem [Kim and Carosi, 2010], and sterile neutrinos-right-handed neutrinos with $m.c^2 \sim 1\,\mathrm{eV}$ that interact predominantly through gravity. Beyond these, there exists a "zoo" of other particle candidates, each primarily interacting via gravity and varying widely in mass. For a comprehensive review, see Feng [2010].

Cold dark matter

The majority of particle candidates align with the ΛCDM framework, where Λ represents dark energy, and CDM denotes cold dark matter. In this cosmological model, DM is categorized as *cold*, meaning its velocity was significantly less than the speed of light at the epoch of radiation-matter equivalence, and *dark*, indicating that it interacts solely through gravitational forces. This implies that DM does not engage in electromagnetic interactions and has minimal, if any, interactions through nuclear forces, rendering it effectively dissipationless and collisionless.

The ΛCDM model excels in explaining large-scale phenomena (over 1 Mpc), such as the formation and structure of the cosmic microwave background (CMB), the distribution of matter on large scales, the observed abundances of hydrogen, helium, and lithium, and the accelerating expansion of the universe [Bullock and Boylan-Kolchin, 2017]. ΛCDM gives robust predictions for the current counts of DM halos as well as their structure. In addition, the ΛCDM paradigm as a backbone of galactic formation theory is also able to predict

complicates the detection of weakly interacting massive particles, necessitating more advanced detection technologies or methods to differentiate between neutrino interactions and genuine dark matter signals.

[2]Throughout this Section, k and c refer to the Boltzmann constant, and the speed of light in vacuum, respectively.

macroscopic properties of galaxies that form within DM halos with an increasing accuracy (counts, colors, morphologies, evolution over time) [Vogelsberger et al., 2014, Schaye et al., 2014].

8.1.2 Cold dark matter at small scales

Within the CDM framework, smaller objects are theorized to collapse first, later merging hierarchically to form larger, more massive structures in a bottom-up hierarchical formation. This pattern is due to the shape of the power spectrum of the cosmic microwave background [Durrer, 2020].

Nevertheless, the CDM model encounters significant challenges when compared with observational data at smaller scales-specifically, lengths smaller than $\sim 1\,\mathrm{Mpc}$ and masses below $\sim 10^{11} M_\odot$. These discrepancies raise questions about the sufficiency of the CDM model.

Missing satellite problem

A significant challenge for the CDM paradigm is the discrepancy known as the missing satellite problem. Observationally, there are only about 50 galaxies with stellar masses $M_* > 300 M_\odot$ within 300 kpc of the MW. In contrast, CDM predicts the presence of approximately 1000 dark subhalos with masses exceeding $10^7 M_\odot$ around MW-like galaxies. However, accounting for our current detection limits through completeness corrections, many studies [see, for example, Doliva-Dolinsky et al., 2022] have suggested that this missing satellite problem may be overstated.

In addition, several mechanisms have been proposed to reconcile these predictions with a potentially lower count of observed dwarf galaxies. For instance, it is suggested that many of the DM halos do exist but remain devoid of luminous matter due to baryonic feedback effects, which inhibit galaxy formation in lower mass halos. Additionally, it is possible that interactions with larger galaxies could have stripped some dwarf galaxies, preventing their detection [Bullock and Boylan-Kolchin, 2017]. Despite the lack of direct evidence for subhalos in general, their existence is not ruled out, and their detection is an active area of research, as will be discussed in subsequent sections of this chapter.

Cusp core discrepancy

The cusp-core discrepancy represents a more significant challenge to the CDM paradigm. In the central regions of DM-dominated galaxies, particularly dwarf galaxies, CDM predicts a DM density profile that is considerably cusped, modeled as $\rho(r) \sim r^{-\alpha}$ with α typically between 0.8 and 1.4.

However, observational data from galaxy rotation curves suggest a much flatter density profile, with α values ranging from 0 to 0.5 (ρ and r representing the density and the radius respectively). While it is proposed that baryonic feedback might modify the DM distribution in these galaxies, such mechanisms are estimated to account for the observed profiles in only about 50% of DM-dominated dwarf galaxies [Battaglia and Nipoti, 2022, Bullock and Boylan-Kolchin, 2017]. This discrepancy remains a critical issue, pointing towards potential modifications needed within the CDM framework or alternative explanations.

Too big to fail problem

The "too big to fail" problem presents another formidable challenge within the CDM framework. This issue arises from the assumption that baryonic feedback mechanisms prevent galaxy formation in smaller subhalos (with virial masses less than $10^{10} M_{\odot}$), which could reconcile some observations with the predictions of ΛCDM. Under this assumption, only larger subhalos would facilitate galaxy formation, aligning with the relatively small number of observed galactic satellites. Contrary to this, studies have found that galaxy formation is feasible in even smaller halos, highlighting a contradiction in this assumption [Boylan-Kolchin et al., 2011]. Some theories propose that galaxies formed in these smaller halos might be so small that baryonic feedback mechanisms, which could disrupt halo structure and impede star formation, have little to no effect, rendering these galaxies anomalies to the feedback mechanism narrative [Ogiya and Burkert, 2015]. Furthermore, while some argue that the scarcity of observed dwarf galaxies could be attributed to interactions with larger galaxies, which absorb or disrupt their smaller counterparts, this does not account for the absence of field dwarfs unbound to any major galaxy group [Jiang and van den Bosch, 2015].

Cold dark matter tensions

The three aforementioned challenges — missing satellites, the cusp-core discrepancy, and the too big to fail problem — have received significant attention in the literature. However, it is important to acknowledge that additional significant tensions have also emerged, further complicating the ΛCDM narrative at these scales.

Rotation curves:

One notable issue is the diverse rotation curves observed in dwarf galaxies. These galaxies display a wide variety of curve shapes in their inner regions. While recent simulations suggest that baryon-induced core formation in dwarf galaxies of medium to high mass is possible and even common, reproducing the diversity of these rotation curves across different mass ranges remains a challenge for all current models [Santos-Santos et al., 2020]. This issue is further complicated by strong correlations observed between features in the baryonic mass distribution and features in the rotation curves, contrary to expectations under the DM hypothesis. This correlation, often referred to as "Renzo's rule", is notably evident in several spiral galaxies [Famaey and McGaugh, 2012].

Satellite planes:

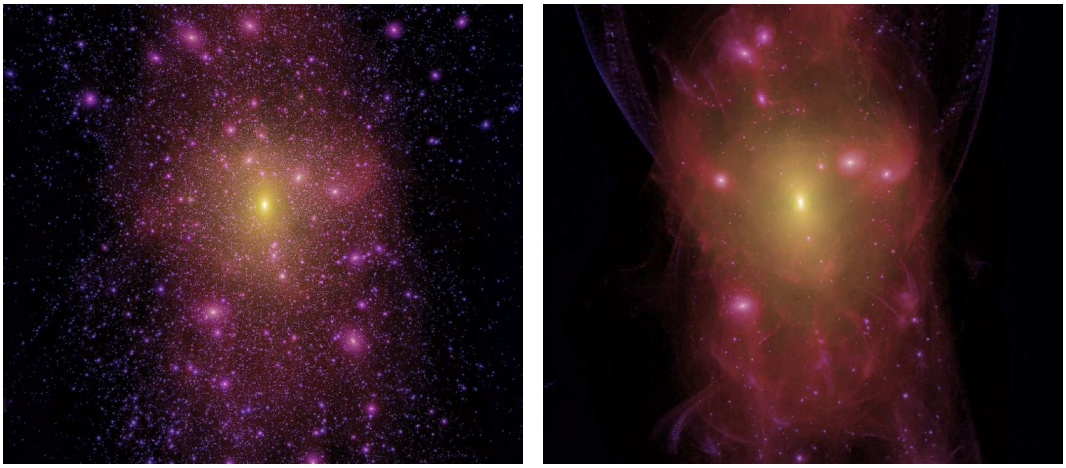
Enhanced observational capabilities in the Local Group have led to another intriguing discovery: nearly all of the satellites orbiting the MW [Kroupa et al., 2005] and about half of those orbiting our neighboring Andromeda galaxy (M31) [Ibata et al., 2013] are arranged in kinematically coherent, thin planar distributions.

This phenomenon could be attributed to specific accretion events involving massive satellites such as the Large Magellanic Cloud (LMC) for the MW or M33/M32 for Andromeda, which can significantly influence the planarity of their respective satellite populations by accreting multiple satellites on similar orbits and enhancing the planarity of existing satellites [Samuel et al., 2021, Li and Helmi, 2008, Garavito-Camargo et al., 2021]. In addition, the alignment within the Local Group as a whole may also play a role in explaining this observed planarity [Pawlowski and McGaugh, 2014].

However, recent observations of satellite systems beyond the Local Group, particularly around dozens of MW analogs that are more cosmologically representative than just the MW and M31, indicate that this planar distribution of satellites is not an isolated phenomenon but a more widespread occurrence [Geha et al., 2017, Mao et al., 2021] that is not explained by current CDM simulations. Nevertheless, some studies suggest that such planar structures may be transient phenomena. For instance, Sawala et al. [2023] argue that these planes are short-lived and not necessarily in conflict with the standard ΛCDM model.

A comprehensive review of the small-scale challenges faced by the ΛCDM model, including these satellite planes observations, was provided by Bullock and Boylan-Kolchin 2017, Sales et al. 2022.

8.1.3 CDM alternatives



Cold Dark Matter Warm Dark Matter

Figure 8.1: **Cold Dark Matter vs. Warm Dark Matter Halos.** This figure, adapted from Lovell et al. [2014], displays results from N-body simulations comparing Cold Dark Matter (CDM) and Warm Dark Matter (WDM) configurations. The image intensity represents the projected squared density of dark matter, while the hue indicates the density-weighted mean velocity dispersion. The simulations reveal that, in contrast to the lumpy dark matter distribution predicted by CDM models for MW-like halos, WDM models (here with a particle mass of 1.456 keV) predict fewer, yet more massive, dark matter lumps.

Dark matter alternatives

The standard ΛCDM model, while phenomenologically successful in various scenarios, faces challenges at small scales, prompting consideration of alternative DM models. One such alternative, the Warm Dark Matter (WDM) model, proposes that DM particles have non-negligible velocities that smooth out distributions on small scales, potentially addressing the "too big to fail" problem. This is illustrated on Figure 8.1. In contrast, Hot Dark Matter (HDM) would facilitate smoothing only at supra-galactic scales due to its top-down formation mechanism-forming superclusters first and galaxies later-unlike the bottom-up approach of CDM where smaller structures form first [Peter, 2012, Bullock and Boylan-Kolchin, 2017]. Figure 8.1 summarizes key parameters distinguishing these models.

Another innovative model is the self-interacting dark matter (SIDM) model, which introduces interactions between DM particles akin to a screened Coulomb potential. This interaction mechanism is designed to resolve the

Parameter	Hot Dark Matter	Warm Dark Matter	Cold Dark Matter
$v_{th}^{z=0} = k_B.T/m$	$30\,\mathrm{km}.s^{-1}$	$0.03\,\mathrm{km}.s^{-1}$	$\sim 0\,\mathrm{km}.s^{-1}$
Mass $(m.c^2)$	1eV	1keV	100GeV

Table 8.1: **Mass and $v_{th}^{z=0}$ for Hot, Warm and Cold Dark Matter models.** Mass and thermal velocity (at redshift z=0) $v_{th}^{z=0}$ of dark matter in the Cold Dark Matter, Warm Dark Matter and Hot Dark Matter paradigms [Bullock and Boylan-Kolchin, 2017].

cusp-core discrepancy by allowing DM to self-interact and thereby alter its distribution within galaxy cores [Spergel and Steinhardt, 2000].

Modified Newtonian Dynamics

In response to unresolved tensions within the ΛCDM model, some researchers propose that our understanding of gravity itself might be incomplete. Modified Newtonian Dynamics (MOND) introduces an empirical law that seeks to explain galactic phenomena without the need for DM [Milgrom, 1983, Bekenstein and Milgrom, 1984]. It incorporates a fundamental acceleration scale, a_0 effectively substituting the degrees of freedom introduced by DM with this new parameter.

MOND modifies the Newtonian dynamics at low accelerations. Specifically, the acceleration g under MOND is described by $g = \sqrt{a_0 g_N}$, where g_N is the Newtonian gravitational acceleration. This modification allows MOND to explain various observational phenomena, especially the rotation curves of galaxies, using only their baryonic matter [Famaey and McGaugh, 2012]. The theory is now being tested within the Local Group [Oria et al., 2021]. However, MOND remains an empirical framework tailored primarily to fit galaxy rotation curves and faces significant challenges in explaining phenomena at larger scales, such as galaxy clusters, gravitational lensing, and cosmological observations [Angus et al., 2006].

A challenge for small-scale physics ?

Although these models often require additional parameters when compared to ΛCDM they allow for the resolution of some small-scale paradoxes posed by ΛCDM. However, it is not clear yet if these small-scale issues can simply be accommodated by a better understanding of small-scale physics such as galactic dynamics, feedback phenomena or if they will instead require a radical revision of the standard cosmological model and of our understanding of the nature of DM. In any case, any correct description of the Universe must at least match ΛCDM on large scales ($>$ 1Mpc). Understanding which of these options is correct is a pressing matter for both astrophysics and physics in general.

Insights

This thesis contributes to ongoing efforts to resolve the challenges associated with understanding DM at small scales. Specifically, this work is aligned with broader initiatives that aim to derive a reliable and high-resolution map of the spatial distribution of DM in galaxies, based on observations from the local universe. Here, "reliable" implies minimal assumptions about the nature of DM, while "high resolution" suggests a level of detail sufficient to potentially reveal DM substructures, should they exist.

Such a map would be invaluable, potentially confirming or refuting ΛCDM predictions at galactic scales and clarifying whether discrepancies at small scales arise from the ΛCDM model itself or from difficulties of accurately modeling baryonic physics on scales smaller than 1 Mpc. For instance, ΛCDM posits that the dense centers of small halos should withstand the hierarchical merging process, implying that DM halos should host numerous substructures [Ghigna et al., 1998]. The review by Bullock and Boylan-Kolchin [2017] underscores the importance of observational projects aimed at constraining DM distributions, categorizing them as one of the most critical — yet achievable — objectives in this domain over the coming decade. Since essential properties of non-linear small scale structures are governed by the particle nature of DM, this type of approach might provide valuable constraints (such as the particle mass, for instance) [Bullock and Boylan-Kolchin, 2017].

8.2 A Dynamical Picture of Galaxies

This section delves into the dynamics of galaxies, reviewing foundational notions critical for understanding the dynamical clues about dark matter discussed in this thesis.

Sub-section 8.2.1 discusses various halo profile models that describe the distribution of DM in galaxies. Sub-section 8.2.2 explores how galactic orbits can be characterized through so called *actions*, utilizing Jeans' Theorems as a theoretical framework. Finally, sub-sections 8.2.3 and 8.2.4 respectively examine the hierarchical structure formation of galaxies and how merging fragments can be stripped into tidal streams.

8.2.1 Halo models

In the CDM model, DM forms extensive halos that envelop galaxies. Through N-body simulations of CDM several empirical DM density profiles have been identified. A well-established model is the Navarro, Frenk, and White [1997] (NFW) profile, described by:

$$\rho_{\mathrm{NFW}}(r) = \frac{\rho_0}{\frac{r}{r_s} + (1 + \frac{r}{r_s})^2} \tag{8.1}$$

where ρ represents the DM density, r is the radial distance from the center, ρ_0 is the central density, and r_s is the scale radius. These parameters vary from one halo to another.

Another significant profile is the Einasto profile, which is often favored in dark-matter-only simulations [Einasto, 1965]:

$$\rho_{\mathrm{Ein}}(r) = \rho_0 \exp\left[-\frac{2}{\gamma}\left(\left(\frac{r}{r_s}\right)^{\gamma} - 1\right)\right] \tag{8.2}$$

Here, γ is a shape parameter that determines the steepness of the profile's slope, and r_s is the scale radius.

Contrasting these "cuspy" profiles, where density increases sharply towards the center, is the Burkert profile, which is often applied to dwarf galaxies and suggests a more "cored" inner structure — accommodating a flatter density distribution at their centers. [Burkert, 1995]:

$$\rho_{\mathrm{Burk}}(r) = \frac{\rho_0}{\left(1 + \frac{r}{r_s}\right)\left(1 + \left(\frac{r}{r_s}\right)^2\right)} \tag{8.3}$$

The gravitational potential Φ related to these density profiles can be derived using the Poisson equation given by:

$$\nabla^2\Phi = 4\pi G\rho \tag{8.4}$$

The diverse array of halo profiles, often employed as empirical tools in such investigations, present an ideal scenario for the application of symbolic regression (SR) —- which aims at automatically finding analytical functions fitting data. This approach which was the subject of Chapters 4-7 can efficiently navigate through numerous potential analytical combinations to identify the most relevant models.

This methodology is particularly valuable here, where halo profiles are expected not only to align with simulation data but also to yield analytically tractable expressions for integrated mass and potential, which must be physically valid at infinity. Incorporating these criteria into the objective function of a SR analysis can further enhance the relevance of such approaches. Additionally, these profiles should correspond to observable stellar distributions to facilitate direct comparisons with empirical data.

8.2.2 Actions

Let us consider a probability density function f of stars in a phase-space coordinate (which includes both position $\mathbf{x}$ and velocity $\mathbf{v}$) evolving within a gravitational potential Φ. According to Jeans' weak theorem [Binney and Tremaine, 2011], the dynamical state of such a system can be elegantly described using integrals of motion:

Jeans' weak theorem:

Theorem 1. *The probability density function of orbits in phase-space f can fully be described by integrals of motion such that $f = f(J_1, ..., J_n)$, $n \in \mathbf{N} \Leftrightarrow$ The probability density function of orbits in phase-space f is a solution of the steady-state collisionless Boltzmann equation (CBE).*

The CBE being given by:

$$\frac{df}{dt} = \mathbf{v}.\frac{\partial f}{\partial \mathbf{x}} - \frac{\partial \Phi}{\partial \mathbf{x}}.\frac{\partial f}{\partial \mathbf{v}} = 0 \tag{8.5}$$

Each unique orbit can in that context be associated with a unique value of $\mathbf{J}$. $\boldsymbol{\theta}$ specifies the position of the star on this orbit. Thus, for each star, $\mathbf{J}$ is conserved along the orbit. Hence, $\mathbf{J}$ vectors are named integrals of motion. An integral of motion I is described as a quantity which does not change along the orbit $\frac{dI}{dt}(\mathbf{x}, \mathbf{v}) = 0$. In certain scenarios, Jeans' strong theorem can be applied to further simplify this representation:

Jeans' strong theorem:

Theorem 2. *Almost all orbits of the system are regular (i.e. non chaotic) with non-resonant frequencies $\Rightarrow$ The probability density function of orbits in phase-space f can fully be described by only three integrals of motion $\mathbf{J} = (J_1, J_2, J_3)$ such that $f = f(J_1, J_2, J_3)$ which can be named actions.*

Hereafter, when referring to actions, we denote them as the vector $\mathbf{J}$. In cylindrical coordinates, which are typically used to describe the MW, action coordinates are expressed as $\mathbf{J} = (J_r, J_\phi, J_z)$, representing the amplitude of orbital motion in the radial, azimuthal, and vertical directions, respectively. This is visually depicted in Figure 8.2. Actions are especially valuable for galactic modeling due to their nature as adiabatic invariants, they remain constant if the galactic potential evolves slowly over time [Binney and Tremaine, 2011].

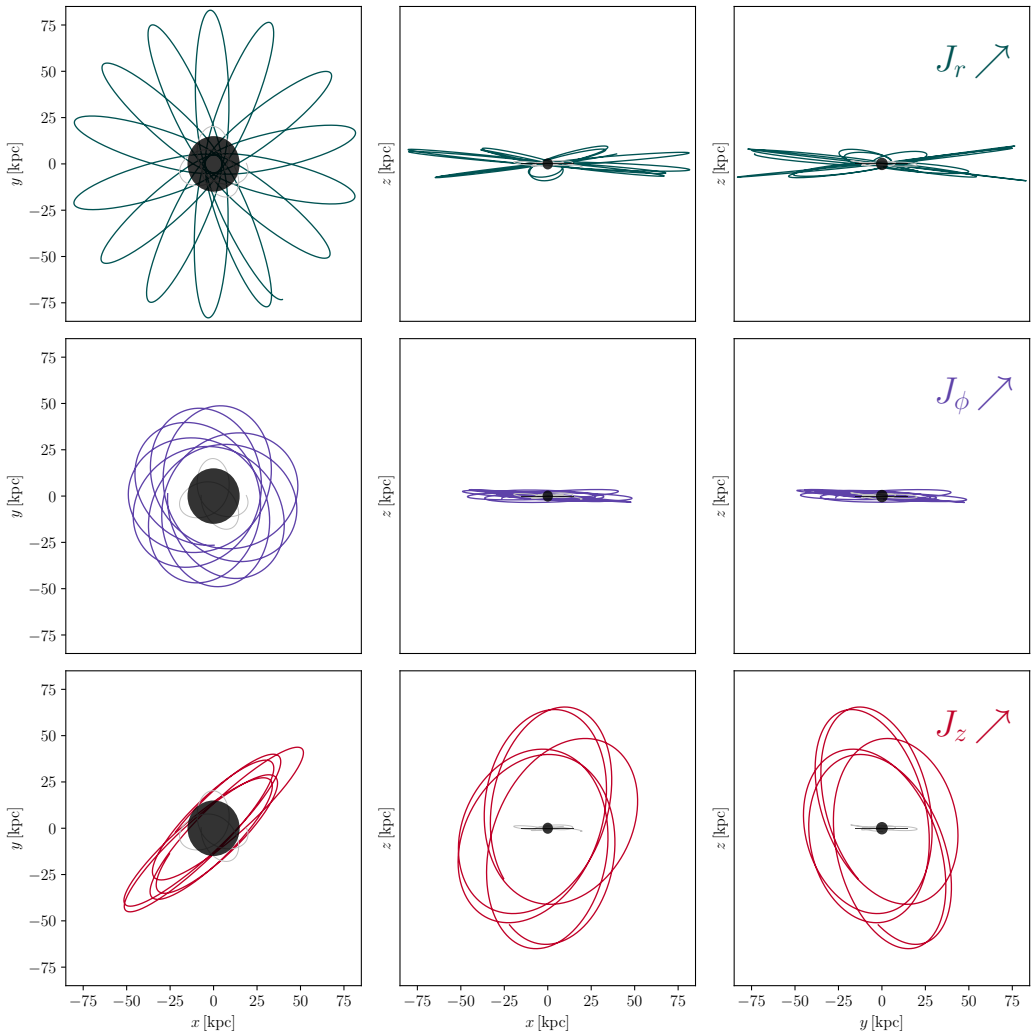


Figure 8.2: **Illustrating the physical significance of actions in galactic dynamical models.** When Jeans' strong theorem is satisfied, one can describe orbits by three actions-integrals of motion-expressed in cylindrical coordinates as $\mathbf{J} = (J_r, J_\phi, J_z)$. Displayed here are the effects of each component of these action coordinates within a McMillan [2017a] MW-like potential. The Milky Way is represented in black. A toy orbit encoded by $\mathbf{J_0} = (500, 20, 2500)\,\mathrm{kpc} \cdot \mathrm{km} \cdot \mathrm{s}^{-1}$, is shown in light grey. Successive panels illustrate the orbital changes induced by individually increasing each action component: $\Delta\mathbf{J} = (5000, 0, 0)\,\mathrm{kpc} \cdot \mathrm{km}{\cdot}\mathrm{s}^{-1}$ (top panel), $\Delta\mathbf{J} = (0, 5000, 0)\,\mathrm{kpc}{\cdot}\mathrm{km}{\cdot}\mathrm{s}^{-1}$ (middle panel), and $\Delta\mathbf{J} = (0, 0, 8000)\,\mathrm{kpc}{\cdot}\mathrm{km} \cdot \mathrm{s}^{-1}$ (bottom panel) computed using the `AGAMA` [Vasiliev, 2019] software for action-based modeling. These adjustments reveal how J_r, J_ϕ, J_z specifically influence the radial, azimuthal, and vertical components of orbital motion, respectively.

8.2.3 Hierarchical structure formation

Hierarchical formation

In the ΛCDM cosmology, galaxies form hierarchically — a process where massive galaxies grow by accreting smaller mass structures, including low-mass galaxies [Mo et al., 2010]. The first evidence for proto-galactic fragments was actually discovered in the MW and provided empirical support for this theoretical framework, demonstrating that galaxies grow, at least partly, by merging with these smaller constituents [Searle and Zinn, 1978]. One should note that an intriguing prediction of ΛCDM that is not yet confirmed observationally is the absorption of starless DM halos [Deason and Belokurov, 2024].

Galactic archaeology

This concept of hierarchical assembly paved the way for the emergence of the field of *Galactic Archaeology*, which investigates the formation and evolution of galaxies. Fueled by extensive astronomical surveys, Galactic Archaeology has rapidly advanced in the past two decades, yielding significant insights from both observational and theoretical studies. Despite its progress, as we will see this field still grapples with fundamental challenges, triggering an never ending appetite for more data [Bland-Hawthorn and Gerhard, 2016].

Tidal streams

One of the promising methods in galactic archaeology involves studying the tidal streams formed from the disruption of infalling low-mass fragments by their larger hosts. If these disruptions occur gradually — where the energy imparted is just sufficient to unbind the outermost stars — these stars trace out orbits closely resembling that of their progenitor fragment, forming what is known as a "tidal stream." These streams offer a direct method to map the gravitational potential of the host galaxy, providing valuable insights into its mass distribution and therefore the underlying nature of DM [Combes et al., 1999].

8.2.4 Stellar streams

About sellar streams

These stellar streams represent the elongated trails of tidal debris that originate from disrupted dwarf galaxies and globular clusters (GCs) — dense and gravitationally bound systems 10^3-10^6 stars. These structures often endure significant tidal forces exerted by the gravitational field of a host galaxy, leading to their gradual disintegration [Lynden-Bell and Lynden-Bell, 1995].

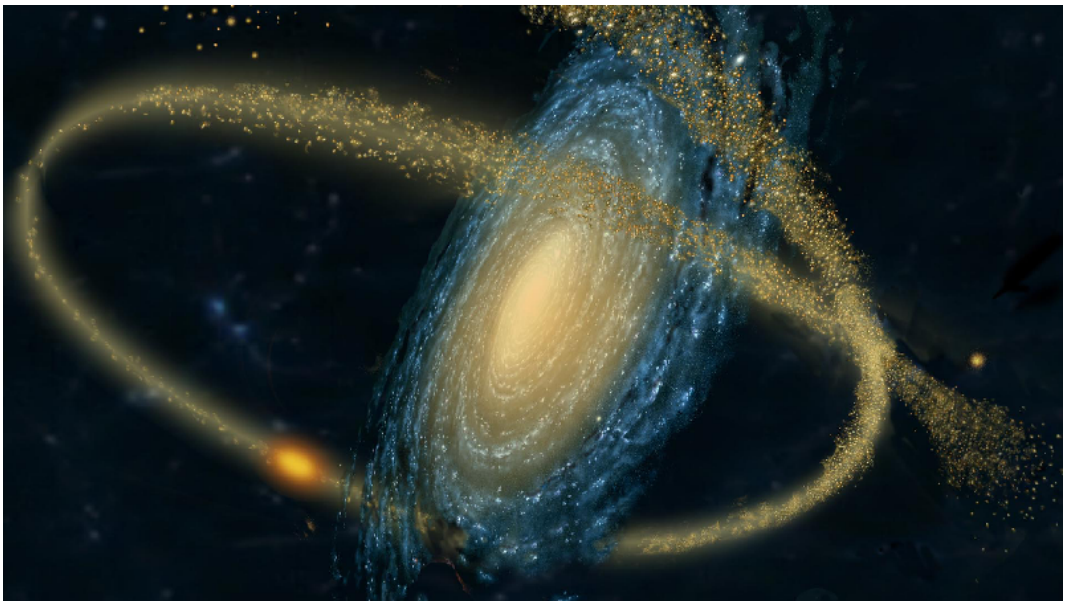


Figure 8.3: **Illustration of a stellar stream.** Artistic view of a dwarf galaxy being ripped apart by tidal forces of a galaxy and forming a stellar stream. Illustration courtesy of Jon Lomberg.

Formation mechanism

The mechanism of stream formation involves tidal forces pulling stars away from their parent bodies at Lagrange points, where the gravitational forces of the satellite and the galaxy balance the centripetal force of orbiting stars. This dynamic results in the creation of two distinct features: the leading tidal tail, composed of stars that orbit faster and move ahead of the progenitor, and the trailing tidal tail, which consists of stars that orbit slower and thus lag behind [Binney and Tremaine, 2011]. See Figure 8.3 for an illustration of a stellar stream. These tails trace the orbital path of the original cluster or dwarf galaxy, providing a unique probe into the distribution of matter.

About progenitors

However, streams originating from dwarf galaxies tend to be longer, wider, and smoother compared to those from GCs. This difference stems not only from their initial mass but also from their star density. Dwarf galaxies, possessing higher initial masses and velocity dispersions, produce streams that are more substantial, wider, and develop more rapidly [Simon and Geha, 2007].

Dark matter and accretion history

Stellar streams serve as both fossil records of galactic histories and as probes of the galactic acceleration field, reflecting the distribution of dark matter that dominates the mass of most galaxies [Moster et al., 2010].

Streams beyond the Milky Way

Beyond the MW, numerous stellar streams have been observed throughout the Local Universe, referred to as extragalactic streams. These streams present a rich diversity of merger histories across various host galaxies. However, current observations typically capture only their 2D positions[3], making their analysis particularly challenging [Nibauer et al., 2023].

Moreover, recent surveys have uncovered many low surface-brightness features across the local Universe, revealing that what were thought to be isolated features are often extensive networks of tidal streams and shells [Bílek et al., 2020, Sola et al., 2022]. Similarly infalling structures have been reported within the Local Group, although these observations often lack velocity data and are only projections, complicating their use in dynamic modeling [Doliva-Dolinsky et al., 2022, 2023].

8.3 The Milky Way

This section reviews why the Milky Way (MW) is an ideal laboratory for studying DM. We begin by discussing how near-field cosmology can be leveraged to study DM within the MW (sub-section 8.3.1), followed by a concise overview of the structural components of the MW (sub-section 8.3.2). We then highlight the transformative impact of Gaia observations on our understanding of galactic archaeology (sub-section 8.3.3), and explore methodologies for using these observations to constrain the properties of DM. This includes analyzing stellar streams (sub-section 8.3.4) and utilizing snapshots of phase-space stellar coordinates (this will be the subject of Section 10.1).

[3]We will nuance that point in sub-section 11.2.1.

8.3.1 A dark matter laboratory

Data availability

Given our residence within the MW, referencing its 'proximity' is quite the understatement. This unique positioning grants us access to exceptionally high-quality data, establishing the MW as an unparalleled laboratory for constraining DM models and deepening our understanding of galaxy formation. The ability to resolve individual stars allows for precise measurements of velocities via spectral redshift analysis and distances through parallax methods[4].

Archaeology

This exceptional data quality facilitates in-depth studies of stellar populations, including analyses of metallicity that can be correlated with dynamical properties to reveal patterns in the galactic structure and history. Such rich local data triggered a profound interest in the history of the MW itself for its own purpose, akin to the geological study of Earth.

Sagittarius

The discovery of the Sagittarius stream within the MW marked a significant milestone in galactic archaeology [Ibata et al., 1994]. It was crucial in providing early constraints on the shape of the MW's halo, suggesting it could be prolate, mildly oblate, or even triaxial [Ibata et al., 2001, Helmi, 2004, Johnston et al., 2005, Law and Majewski, 2010]. Its detailed features continue to be a subject of study, revealing more about our galactic environment [see e.g., Oria et al., 2022a].

The Sagittarius stream's radial velocities have been instrumental in developing accurate models of its dynamics, facilitated by the abundance of bright stars in such a massive satellite [Majewski et al., 2004] — a key advantage of studying local objects.

Near-field cosmology

This focus on our immediate neighborhood is termed *near-field cosmology*, a field that leverages detailed observations of the MW and its surroundings to infer universal physical processes that govern cosmic time scales [Freeman and Bland-Hawthorn, 2002, Bland-Hawthorn and Gerhard, 2016].

[4]This method involves measuring the apparent positional shift of a star observed from opposite ends of the Earth's orbit around the sun and applying basic trigonometry to determine distance.

8.3.2 Milky Way structure

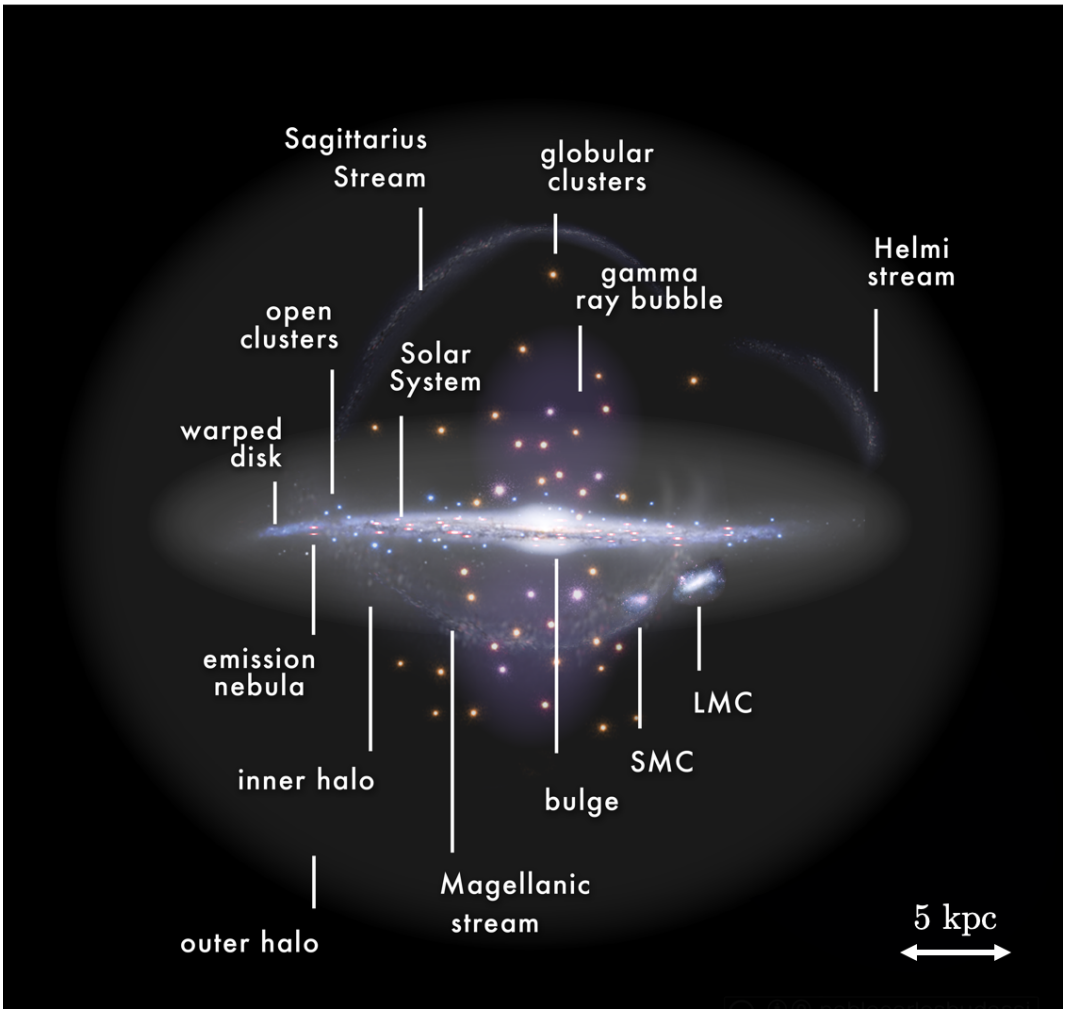


Figure 8.4: **Illustration of the Milky Way.** This artistic representation depicts the various components of the Milky Way, as outlined in Sub-section 8.3.2. Illustration courtesy of Pablo Carlos Budassi.

This subsection provides a topological overview of our galaxy, the MW. The MW has an estimated total mass ranging between 1.2×10^{12} and 1.9×10^{12} $M_\odot$ [Fragione and Loeb, 2017].

Inner component

Structurally, it is categorized as a barred spiral galaxy, featuring a central bar-shaped structure of stars. This central area includes a boxy/peanut-shaped

bulge with a diameter of 6 kpc and thickness of 4 kpc, comprising both young and old stars[5]. At the core of this bulge lies the supermassive black hole Sagittarius A* (Sgr A*), which has a mass of approximately 4.3×10^6 $M_\odot$.

Stellar component

The galaxy also includes a warped disk extending up to 30 kpc in diameter and 0.3 kpc in thickness, populated with young and old stars, gas, dust, and open clusters. These open clusters are loosely bound and tend to dissolve quickly. Our solar system is located within this disk, approximately 8 kpc from the center. Surrounding the disk is a spherical, gasless, and dustless stellar halo with a diameter of about 40 kpc, predominantly containing old stars and globular clusters (GCs) [Mo et al., 2010, Brau, 2001].

Outer components

Moreover, significant infalling structures need to be considered when modeling the MW's potential, such as the Large Magellanic Cloud (LMC), the Small Magellanic Cloud (SMC), and their associated streams, as well as the highly disrupted Sagittarius dwarf galaxy, now mostly in the form of a tidal stream [Brau, 2001]. These features are depicted in Figure 8.4. Lastly, the DM halo, comprising approximately 90% of the galaxy's mass and extending to ~ 100 kpc in radius, remains the most enigmatic component of our galaxy's structure.

8.3.3 Gaia

Motivations

The mapping of DM distribution across the MW necessitates extensive data, particularly detailed kinematics of stellar streams, which were not adequately available prior to the Gaia mission. The complexity of the MW's halo, evidenced by interactions among structures like the Palomar 5 [Küpper et al., 2015] and GD-1 [Grillmair and Dionatos, 2006] streams, alongside the significant disequilibrium induced by interactions such as those with the Large Magellanic Cloud [Gómez et al., 2015], demands more nuanced and time-dependent models which in turn require larger datasets to achieve accurate potential reconstructions.

The European Space Agency's Gaia mission addresses this need by mapping the positions and velocities for nearly two billion stars [Gaia Collaboration, 2016]. This unprecedented data collection which — which saw its latest release through Data Release 3 (DR3) [Gaia Collaboration, 2022] — supports a

[5]Older stars are generally metal-poor as they were formed in an early universe.

wide array of projects, ranging from constraining the MW's potential through observations of its disk [see e.g., Horta et al., 2024] to detailed analyses of its stellar streams [e.g., Malhan and Ibata, 2018, Koposov et al., 2010, Sanders and Binney, 2013, Malhan and Ibata, 2019, Ibata et al., 2021, Ibata et al., 2021, Nibauer et al., 2022].

The Gaia revolution

Observational data

With the Gaia mission, as astronomers we now have access to unprecedented five-dimensional (5D) astrometric data for for hundreds of millions of stars. This 5D data set includes three-dimensional positional information — stellar coordinates on the sky combined with distances calculated using the parallax method — and two-dimensional velocity information derived from proper motions, which simply reflect the apparent motion of stars across the sky corrected for distance.

However, it is important to recognize the inherent uncertainties in these measurements. Stellar distances are subject to significant errors[6]. Additionally, the outer regions of the MW's DM halo-areas critically important for understanding DM properties-tend to have fewer tracers, complicating efforts to derive detailed characteristics of DM from this data [Gaia Collaboration, 2016].

Archaeological and algorithmic revolutions

The wealth of data provided by Gaia has revolutionized our understanding of the MW's structure and history. We have identified several dwarf galaxies that have been accreted and phase-mixed into the MW [Belokurov et al., 2018, Myeong et al., 2019, Helmi et al., 2018], as well as moving groups and dissolved star clusters within the Solar neighborhood [Ramos et al., 2018, Antoja et al., 2018]. Additionally, Gaia's precise measurements have facilitated the detection and confirmation of numerous new stellar streams, leading to an order of magnitude increase-from about ten to over a hundred-in the known number of such streams [Bonaca and Price-Whelan, 2024] mostly thanks to Malhan and Ibata [2018], Malhan et al. [2018], Malhan et al. [2022].

On the computational front, which plays a critical role in this thesis, the era of Gaia has sparked the development of innovative methods for detecting new streams around the MW [Necib et al., 2020, Dodd et al., 2023, Shih et al., 2021, Pettee et al., 2023]. A prime example is the `STREAMFINDER` algorithm, which identifies overdensities of stars near computed orbits based on a model

[6] Approximately $100\,\mu$as, and the velocity components perpendicular to the line of sight can have uncertainties around $1,000\,\mu$as.

of the MW's mass distribution [Malhan and Ibata, 2018]. This algorithm will be briefly described in sub-section 9.3.1.

For a comprehensive atlas of streams identified in the MW up to Gaia DR3, see Ibata et al. [2021a]. For detailed reviews of MW archaeology in the era of Gaia, refer to [Deason and Belokurov, 2024, Bonaca and Price-Whelan, 2024].

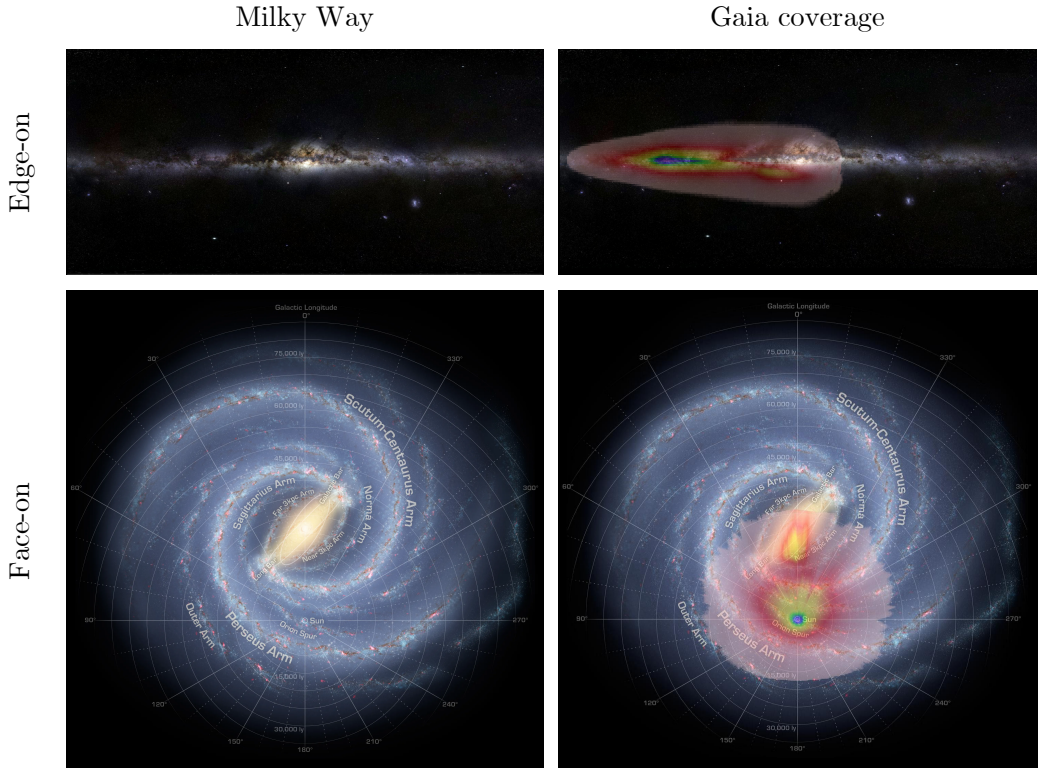


Figure 8.5: **Gaia Coverage.** This figure illustrates the relative density of stars observed by Gaia, with color gradients indicating stellar density-ranging from very high densities (purple-blue) near the Sun to lower densities (pink) farther away. The background provides two perspectives of the Milky Way: an edge-on view of the galaxy (Gigagalaxy Zoom, ESO/S. Brunier/S. Guisard: as seen from ESO, Chile) and an artistic face-on impression (NASA/JPL-Caltech/R. Hurt) [Robin et al., 2012].

The 6D sample

Availability

With the release of Gaia DR3 [Gaia Collaboration, 2022], we now possess comprehensive 6D phase-space information (three-dimensional positions and velocities) for approximately 33 million stars allowing action-based modeling. This is rendered possible by the inclusion of radial velocity measurements derived from stellar spectra. It is important to note, however, that proper motion

measurements rely on accurate distance estimations, which in turn depend on parallax. The further a star, the less precise its parallax measurement due to the limitations of trigonometric methods.

Constraints and coverage

Most stars for which we have complete 6D data, and exhibit a parallax signal-to-noise ratio greater than 10, are therefore located relatively close to our Sun-within about 3 kpc. These observational constraints have limited the depth of Gaia's general astrometric coverage, affecting more than just radial velocity measurements, as illustrated in Figure 8.5. Despite this spatial limitation, the availability of full orbital data from a uniform survey framework presents significant opportunities for advancing our understanding of galactic dynamics, even though the survey depth is somewhat restricted.

It is important to recognize that while some stars from stellar streams near the Sun were previously detected [Helmi et al., 1999], it was not until Gaia Data Release 2 (DR2) that a significant number of stellar stream structures in the inner Galaxy were cataloged [Ibata et al., 2019].

Near-Sun structures

Contrary to the expectation that tidal debris in the vicinity of the Sun would phase-mix rapidly due to short dynamical times, the release of Gaia DR3 has enabled us (this is a pivotal subject of Chapter 9) and others to demonstrate otherwise [Gaia Collaboration, 2022]. Specifically, this data release sparked a significant increase in the discovery of stellar streams and structures by utilizing the comprehensive 6D data set available for a select group of stars [Viswanathan et al., 2023, Tenachi et al., 2022, Oria et al., 2022b, Dodd et al., 2023] very much like the notable Nyx stellar stream, which was identified by focusing on samples near the Sun [Necib et al., 2020].

8.3.4 Probing Dark Matter with Stellar Streams

Stellar streams are crucial for mapping the DM distribution within the MW, having already shown the presence of massive DM halos around both the MW and the LMC [Bonaca and Price-Whelan, 2024].

Mapping the halo shape

Theoretical studies suggest that the extensive phase-space distribution of stellar streams, which can stretch beyond 100 kpc, makes them excellent probes for measuring both the mass and the shape of the DM halo at large scales [Dubinski et al., 1999]. Their great extent offers unique insights into the DM halo

structure, positioning them as one of our most informative tracers [Deason and Belokurov, 2024].

Detecting subhalos

Furthermore, stellar streams could facilitate the detection of DM subhalos at smaller scales [Ibata et al., 2002]. Observations of underdensities or gaps within these streams often indicate interactions with subhalos, providing a direct method to detect otherwise invisible star-free DM halos, a key prediction of the ΛCDM model [Carlberg, 2012]. Notably, Bonaca et al. [2019] have analyzed features in the GD-1 stream, such as gaps and spurs, attributing them to encounters with dark matter substructures.

Challenges and priorities

While the data from stellar streams is invaluable, it is also complex. Each stream studied in detail suggests additional dynamical interactions affecting its density structure. Identifying these subhalos would not only support ΛCDM predictions but also help resolve the long-standing missing satellites puzzle [Klypin et al., 1999]. Current community priorities include discovering new streams in the outer halo and conducting large-scale spectroscopic follow-ups to obtain radial velocities [Bonaca and Price-Whelan, 2024] — a work we conduct in Chapter 9.

Additionally, it is worth mentioning that the acceleration field can be probed using the near Sun 6D sample from Gaia located in the disk, which offers alternative insights into galactic dynamics. This alternative methodology will be explored in Chapter 10.

Chapter 9

Milky Way Archaeology

Portions of the content presented in this Chapter have been previously discussed in the following publications:

2022 *Typhon: A Polar Stream from the Outer Halo Raining through the Solar Neighborhood*
W. Tenachi, P.-A. Oria, R. Ibata, B. Famaey, Z. Yuan, A. Arentsen, N. Martin, A. Viswanathan
ApJL **935** L22, arXiv:2206.10405

2022 *Antaeus: A Retrograde Group of Tidal Debris in the Milky Way's Disk Plane*
P.-A. Oria, **W. Tenachi**, R. Ibata, B. Famaey, Z. Yuan, A. Arentsen, N. Martin, A. Viswanathan
ApJL **936** L3, arXiv:2206.10404

2023 *Charting the Galactic acceleration field II. A global mass model of the Milky Way from the STREAMFINDER Atlas of Stellar Streams detected in Gaia DR3*
R. Ibata, K. Malhan, **W. Tenachi**, et al
ApJ **967** 89, arXiv:2311.17202

Summary.

We discuss our contributions to identifying new probes of dark matter within the Milky Way through the detection of stellar streams. Specifically, we detail the discovery of two new streams, Typhon and Antaeus, identified from Gaia's near-Sun 6D sample. Additionally, I outline my involvement in follow-up observational campaigns that have leveraged 5D data from Gaia DR3 to detect multiple new stellar streams.

As explored in the previous Chapter, understanding the accretion history of the Milky Way — a field known as Galactic Archeology — through the study of stellar streams[1] can assist us in testing the ΛCDM model[2] of galaxy formation. It is also of crucial importance for probing the acceleration field of the Milky Way and its underlying dark matter distribution, thereby providing insights into its nature. However, due to the complex nature of accretion events, these studies demand an increasing amount of observational probes data.

Accordingly, a primary focus within the community is the detection and dynamical characterization of new stellar stream structures [Bonaca and Price-Whelan, 2024]. In this Chapter, we present our contributions to this effort. We explore methodologies for discovering and analyzing stellar streams in the era of the Gaia mission, with a particular emphasis on leveraging the extensive data available from Gaia Data Release 3 (DR3) [Gaia Collaboration, 2022].

First, we introduce two new structures identified by analyzing the near-Sun 6D sample[3] from Gaia, focusing on members passing through the solar neighborhood at high velocities. Specifically, in Section 9.1, we describe a newly discovered stream, which we have named *Typhon*. This polar stream extends to the outer halo of the MW, reaching distances up to $\sim 100\,\mathrm{kpc}$. In Section 9.2, we discuss *Antaeus*, a retrograde moving group located within the disk plane of the MW, and explore its connections to other significant accretion events in the MW's history.

Note that Section 9.2 is the result of collaborative efforts led by Pierre-Antoine Oria with myself, Rodrigo Ibata, Benoit Famaey, Zhen Yuan, Anke Arentsen, Nicols Martin, Akshara Viswanathan.

Finally, Section 9.3 provides an overview of the most comprehensive atlas of stellar streams in the MW up to the release of Gaia DR3, including updated parameters for current MW mass models. This atlas incorporates new detections made possible through an enhanced version of the `STREAMFINDER` algorithm [Malhan and Ibata, 2018][4].

Note that Section 9.3 is the result of collaborative efforts led by Rodrigo Ibata with Khyati Malhan, myself, Anke Ardern-Arentsen, Michele Bellazzini, Paolo Bianchini, Piercarlo Bonifacio, Elisabetta Caffau, Foivos Diakogiannis, Raphael Errani, Benoit Famaey, Salvatore Ferrone, Nicolas F. Martin, Paola di Matteo, Giacomo Monari, Florent Renaud, Else Starkenburg, Guillaume

[1] Thin elongated tidal structures formed during accretion events between a host galaxy and a smaller structure, as detailed in Section 8.1.

[2] The dominant cosmological model which relies on a cold dark matter paradigm as detailed in Section 8.1.

[3] This sample comprises high quality 3D position and 3D velocity information.

[4] This algorithm is capable of utilizing 5D data lacking radial velocity information.

Thomas, Akshara Viswanathan, and Zhen Yuan. Here I mostly focus on my specific contribution to that effort which takes the form of dedicated spectroscopic follow-up observations using VLT/UVES[5] and INT/IDS[6] telescopes/instruments, which have enabled us to acquire radial velocity measurements and confirm the identification of 28 new stellar streams increasing the number of MW long and thin stellar streams identified by the community to a total of 87.

9.1 Typhon: An Outer Halo Stream raining through the Solar Neighberhood

One of the principal goals of the Gaia space mission [Gaia Collaboration et al., 2016a] is to survey the Milky Way, so as to allow us to understand how our home galaxy was built up over cosmic time. Although we only observe the end state of this majestic structure, fortunately the processes of formation and growth have left copious amounts of evidence in the form of debris that is now scattered throughout our Galaxy [Belokurov et al., 2006, Shipp et al., 2018, Ibata et al., 2021b, Malhan et al., 2022a]. Some of these residues are due to the accretion of small galaxies and globular clusters, which disrupted under the action of tidal forces, leaving long stellar streams. In some cases they can still remain as elongated structures, many billion years after the dissolution of their progenitors [Helmi, 2008]. Studying these structures is of great importance since their trajectories probe the galactic acceleration field and the underlying dark matter distribution [e.g., Koposov et al., 2010, Sanders and Binney, 2013, Malhan and Ibata, 2019, Ibata et al., 2021].

As detailed in sub-section 8.2.2, a particularly powerful means to uncover such fossil remnants is by searching for groups of stars with common integrals of motion. Action coordinates are perhaps the best choice for this, as they are adiabatic invariants that will have been preserved along orbits if the Milky Way's potential evolved only slowly through time [Binney and Tremaine, 2011]. However, to transform our stellar measurements into actions (and their conjugate angles), we require the full 6D positions and velocities. With present instrumentation this is only achievable close to the Solar position in the Galaxy.

[5]Very Large Telescope, European Southern Observatory, Cerro Paranal, Chile

[6]Isaac Newton Telesceope, Roque de los Muchachos Observatory, La Palma, Canary Islands, Spain

The Gaia mission has recently made accessible its third data release (DR3) [Gaia Collaboration, 2022] of its all-sky survey. It contains approximately 33 million stars with mean radial velocities down to $G \sim 15$, which, complemented with the excellent proper motions and parallaxes published in the earlier EDR3 release [Gaia Collaboration, 2021], provide the required phase-space constraints. Because the DR3 radial velocity limit is quite shallow, it almost exclusively probes the very nearby regions of the Galaxy (the median distance of the sample with 10σ parallaxes is only 1.26 kpc). In the vicinity of the Sun, dynamical times are short and tidal debris are expected to phase-mix rapidly [Helmi et al., 1999, 2003], erasing any initial stream-like coherence that might have been present.

In this Section we show that, surprisingly and contrary to those expectations, the Solar neighborhood contains a very wide yet kinematically coherent metal poor stellar stream, which we name Typhon[7], whose apocenter reaches out to > 100 kpc – the edge of the Galactic halo.

Sub-section 9.1.1 details our selection process from Gaia data, sub-section 9.1.2 offers a chemo-dynamical characterization of the stream, sub-section 9.1.3 discusses our finding and sub-section 9.1.4 gives our Typhon sample.

9.1.1 Selection

Pre-selection

From the Gaia DR3 catalog, we select the 25,355,580 stars with well-constrained distances (having parallaxes $\varpi/\delta\varpi > 10$), radial velocities measured by Gaia's Radial Velocity Spectrometer (RVS) instrument [Recio-Blanco et al., 2022], having at least a 5-parameter astrometric solution, and with magnitudes in the range $0 \leq G \leq 22$, $0 \leq G_{BP} \leq 30$, $0 \leq G_{RP} \leq 30$. To convert the apparent motions to motions in a frame[8] at rest with respect to the Galaxy, we assume that the Sun is located at $(x, y, z)_\odot = (-8.2240, 0, 0.0028)$ kpc (Solar radius from Bovy 2020 and z-position of the Sun from Widmark et al. 2021), and that it moves with a peculiar velocity $(v_x, v_y, v_z)_\odot = (11.10, 7.20, 7.25)\,\mathrm{km\,s^{-1}}$ (Schönrich et al. 2010, with the ϕ-direction velocity from Bovy 2020), and we take the circular velocity at the Solar radius to be $243\,\mathrm{km\,s^{-1}}$ [Bovy, 2020]. We use the resulting phase space measurements to derive the orbital parameters of the stars, including the pericenter and apocenter distances, as well as action-angle coordinates calculated using the AGAMA package [Vasiliev, 2019a] in a realistic potential model [McMillan, 2017a] for the Milky Way. Since we

[7] The serpent Typhon is the child of Gaia and Tartarus (the deep abyss) in Greek myth.
[8] Throughout this Chapter, we use a right-handed Galactic Cartesian coordinate system.

are particularly interested in finding debris from the outer halo that could be associated to ancient merger events, we impose an apocenter cut at $r_{apo} > 75\,\mathrm{kpc}$, which yields a sub-sample of 870 stars.

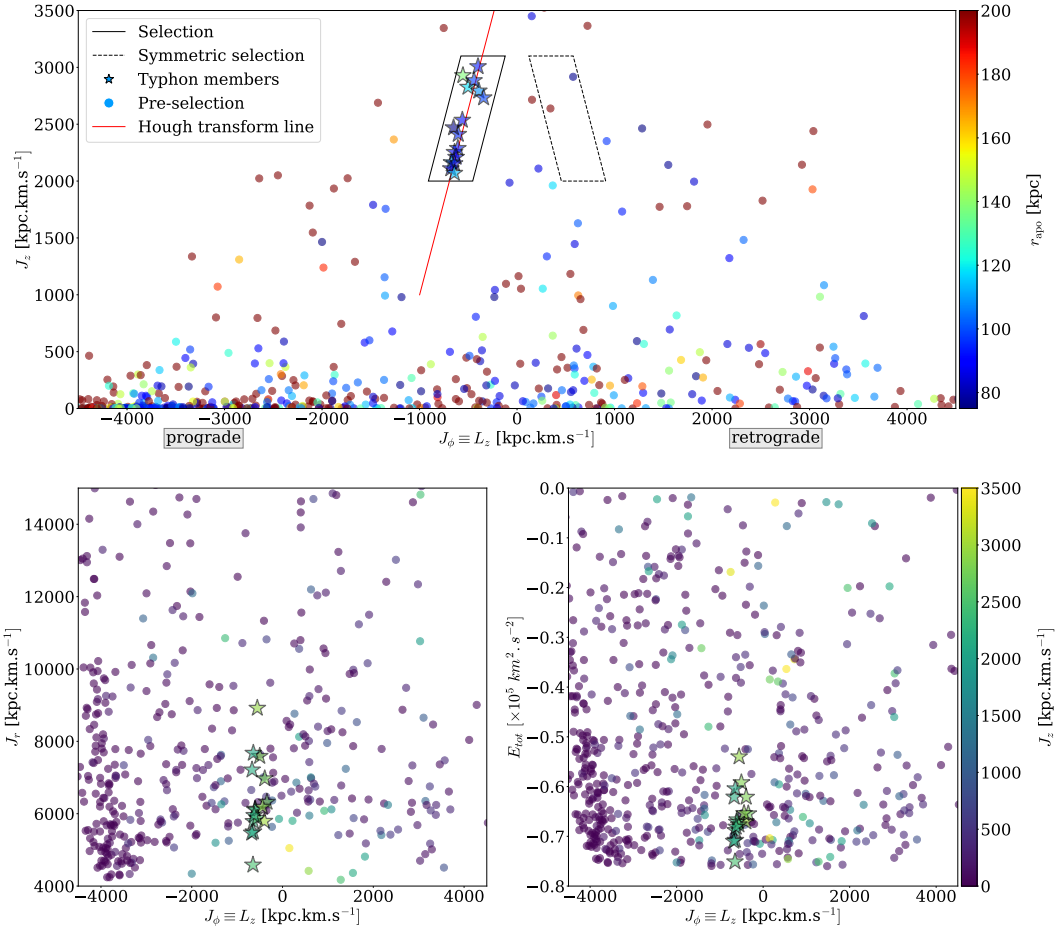


Figure 9.1: **Selection process in the space of actions.** Typhon members are indicated by star-symbols along with the 573 pre-selected stars having $\varpi/\delta\varpi > 10$, $r_{apo} > 75\,\mathrm{kpc}$ and $d_\odot < 4\,\mathrm{kpc}$ (denoted by circles). Stars are colored by their apocenter values in the upper panel and and by their vertical action values in the lower row of panels. Upper panel: (J_ϕ, J_z) plane used for the selection where the overdensity was discovered. The most significant detection obtained using a Hough transform technique [Illingworth and Kittler, 1988] on stars with $J_z > 1000\,\mathrm{kpc\,km\,s^{-1}}$ (i.e. with large departures from the Galactic mid-plane) is shown with a red line. This line runs through the Typhon structure. The parallelogram selection of the structure is depicted in a solid line encompassing 16 stars, and is defined by: $J_z \in [2000, 3100]\,\mathrm{kpc\,km\,s^{-1}}$ and $3.3J_\phi + 3500\,\mathrm{kpc\,km\,s^{-1}} < J_z < 3.3J_\phi + 5000\,\mathrm{kpc\,km\,s^{-1}}$. The symmetric (retrograde) selection with respect to the $J_\phi = 0$ line is shown with a dashed line. Bottom-left and bottom-right panels respectively show the (J_ϕ, J_r) and (J_ϕ, E_{tot}) planes.

Typhon selection

Further analysis is performed in the space of actions (J_r, J_ϕ, J_z), which encode, respectively, the amplitude of orbital motion in the radial, azimuthal, and vertical directions. In particular, we plot the (J_ϕ, J_z) projection colored by r_{apo} in Figure 9.1. There, a polar structure can be spotted as a tight, almost vertical, linear grouping between $(J_\phi \sim -650\,\mathrm{kpc\,km\,s^{-1}}, J_z \sim 2100\,\mathrm{kpc\,km\,s^{-1}})$ and $(J_\phi \sim -400\,\mathrm{kpc\,km\,s^{-1}}, J_z \sim 3000\,\mathrm{kpc\,km\,s^{-1}})$. We find that this feature is most striking when the sample is limited to stars with heliocentric distances $d_\odot < 4\,\mathrm{kpc}$, approximately at the limit of useful 6-D phase-space data in the DR3 catalog. In particular, performing the Hough transformation [Illingworth and Kittler, 1988] line detection technique on the stars in the (J_ϕ, J_z) plane (binning the action data into pixels of size $30\,\mathrm{kpc\,km\,s^{-1}}$ on a side and adopting a 1° discretization for the angle of the fitted lines), we find that the most significant linear grouping of stars with $J_z > 1000\,\mathrm{kpc\,km\,s^{-1}}$ (i.e. that experience large excursions from the Galactic mid-plane) corresponds to this quasi-linear overdensity (red line in Figure 9.1). These 16 stars possess similar apocenter distances ($r_{apo} \approx 100\,\mathrm{kpc}$), and are also highly correlated in the angle coordinates $(\theta_r, \theta_\phi, \theta_z)$ conjugate to the actions.

We then separate this structure from the bulk of the data by applying a simple parallelogram selection in the (J_ϕ, J_z) plane, as follows: $J_z \in [2000, 3100]$ and $3.3J_\phi + 3500 < J_z < 3.3J_\phi + 5000$, which results in a final sample of 16 stars. This selection box is displayed as a solid black parallelogram in Figure 9.1.

Typhon structure

Furthermore, it should be noted that the symmetric control selection around $J_\phi = 0$, shown as a dashed parallelogram, encompasses only two stars and they do not possess homogeneous dynamical properties. Assuming that the halo is symmetric in angular momentum, there is no a priori reason for the prograde selection to contain significantly more stars than the symmetric retrograde selection as is the case here, other than the selection containing a coherent dynamical group. Taking the symmetric selection as a control sample we estimate the significance of the detection to be $\approx 3.5\sigma$. We note in passing that the Gaia Universe Model Snapshot (GUMS, Robin et al. 2012a, updated for Gaia DR3) contains no artificial stars with the selection criteria used to detect Typhon, suggesting that Typhon is a coherent structure that can only be explained by an external body not included in that simulation.

9.1.2 Characteristics

Dynamical characteristics

The positions and velocities of the sample members of the Typhon stream are shown in Figure 9.2. We find that member stars of this polar stream are spread out all around us, passing through the Solar neighborhood with a high vertical velocity, and exiting the disk at an angle of $\sim 50°$ with respect to it.

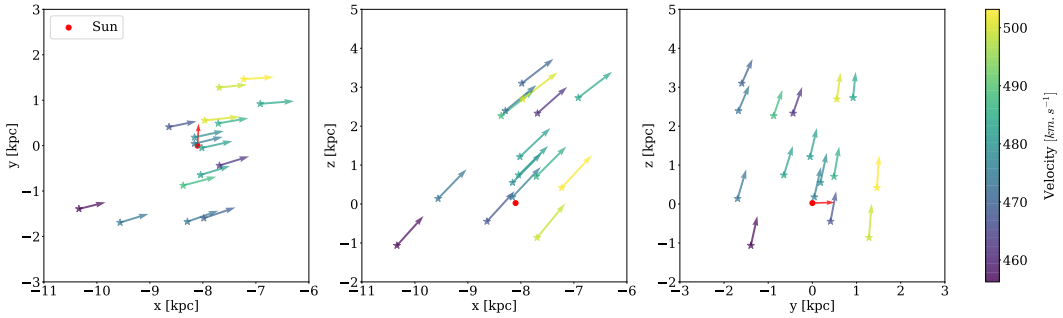


Figure 9.2: **Typhon members.** Positions and velocity vectors in galactic Cartesian coordinates of Typhon sample members. Velocity vectors scale $1 : 3 \times 10^3$. The sample shows very clear streaming motion. For reference, the position and velocity vector of the Sun is also shown in red.

In Figure 9.3 we show the result of integrating Typhon members backwards in time for 5 Gyr in the [McMillan, 2017a] Milky Way potential model. Although the stars were selected from a small region in the (J_ϕ, J_z) plane (but with no constraint on J_r), and so should therefore possess similar orbits, there was no a-priori reason for the sample to be in phase, as is clearly the case from an inspection of Figure 9.3. The sample is dynamically coherent, with very similar orbital parameters: $r_{peri} = 6.0 \pm 0.5\,\mathrm{kpc}$, $r_{apo} = 99 \pm 15\,\mathrm{kpc}$, $J_r = 6400 \pm 1000\,\mathrm{kpc\,km\,s^{-1}}$, $J_\phi = -560 \pm 110\,\mathrm{kpc\,km\,s^{-1}}$, $J_z = 2500 \pm 300\,\mathrm{kpc\,km\,s^{-1}}$ and eccentricity $e = 0.88 \pm 0.02$.

We estimate the 3-dimensional velocity dispersion of the stream to be $\sigma_{v,3D} \approx 13\,\mathrm{km\,s^{-1}}$ by considering the velocity differences of the stars to the computed orbit of the star with *Gaia* ID 3939346894405032576 (whose orbit through the Solar neighborhood appears closest to the middle of the sample). Assuming isotropy, the one-dimensional velocity dispersion is then $\sigma_v \approx 7.5\,\mathrm{km\,s^{-1}}$.

Chemical characteristics

We cross-matched our sample with the LAMOST DR8 [Wang et al., 2022a] catalog, in particular the "FEH_PASTEL" column which covers a wide range

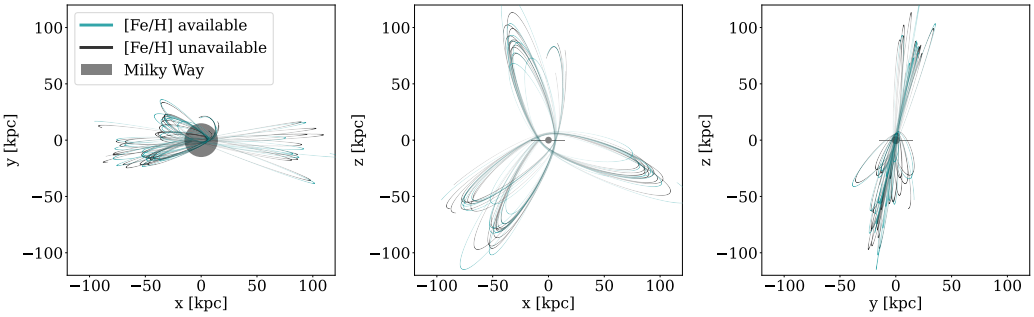


Figure 9.3: **Typhon orbit.** Trajectories of the sample members of Typhon during a 5 Gyr backward integration in the [McMillan, 2017a] potential in galactic Cartesian coordinates. Trajectories of the 7 stars whose metallicity is available through LAMOST DR8 [Wang et al., 2022a] are colored in blue.

of metallicities especially on the very metal-poor regime, enabling us to obtain high quality spectroscopic metallicities for 7 stars of the Typhon stream (the stellar parameters of which lie within the reliable range of the PASTEL catalog). These measurements span between $[\mathrm{Fe/H}] = -2.23 \pm 0.06$ dex and $[\mathrm{Fe/H}] = -1.25 \pm 0.09$ dex.

As shown in Figure 9.3, where we color orbits of stars of known metallicity in yellow, these stars are dynamically representative of the full sample. In Figure 9.4 (left panel), we show the likelihood distribution (black contour lines) for the mean metallicity and for the intrinsic dispersion of the metallicity distribution (correcting for the LAMOST uncertainty estimates, assuming that they are reliable). We find $\langle[\mathrm{Fe/H}]\rangle = -1.60^{+0.15}_{-0.16}$ dex, and $\sigma([\mathrm{Fe/H}]) = 0.32^{+0.17}_{-0.06}$ dex, which indicates that the system has a resolved dispersion in metallicity. We note however, that this result depends on the inclusion of the most metal-poor star in the sample; if it is removed (although we have no a-priori reason to do so) these values become $\langle[\mathrm{Fe/H}]\rangle = -1.41^{+0.05}_{-0.09}$ dex, and $\sigma([\mathrm{Fe/H}]) = 0.06^{+0.17}_{-0.06}$ dex, consistent with no dispersion at the 1σ level.

These metallicities are consistent with the color magnitude diagram shown in Figure 9.4, where we use the 3D extinction estimates by Anders et al. 2022 to deredden the stars. In addition, based on the PARSEC stellar population models [Bressan et al., 2012], and using the canonical two-part-power law initial mass function corrected for unresolved binaries [Kroupa, 2001], and Gaia's detection limit, we compute the order of magnitude of the density of the Typhon stream to be of $\sim 25\,\mathrm{M}_\odot/\,\mathrm{kpc}^{-3}$ in the $d_\odot < 1.5$ kpc solar vicinity fragment. However, without further information we refrain from extrapolating this value out to compute the mass of the full stream structure.

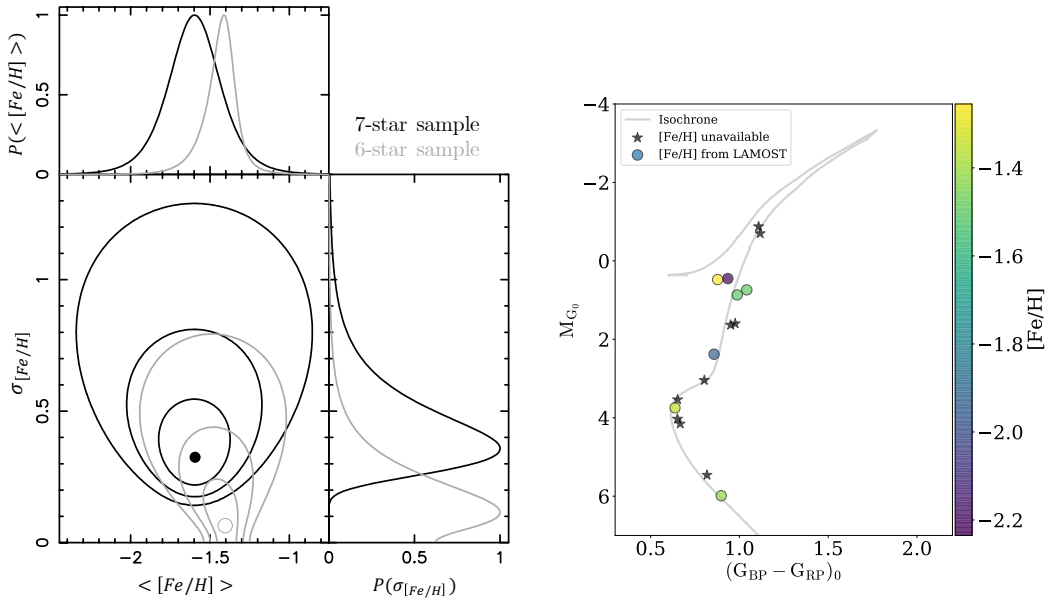


Figure 9.4: **Chemical characteristics of Typhon.** Left: Likelihood contours of the mean metallicity and metallicity dispersion of the spectroscopic sample, shown for the full 7 star sample (black lines), and removing the most metal poor star (grey lines). Right: Color magnitude diagram of the sample members of Typhon. For reference, the grey line shows a PARSEC isochrone model [Bressan et al., 2012] of age 12.5 Gyr and of metallicity [Fe/H] $= -1.60$ dex. The reasonable correspondence of this model shows that the population is predominantly very old.

9.1.3 Discussion and conclusions

Although the search for new stellar streams is currently a very active field, to the best of our knowledge the structure discussed here (Typhon) that we isolated thanks to the new and excellent Gaia DR3 data was never identified before. It should be noted that although Typhon is very close to the DTG-11 stream identified in [Yuan et al., 2020a] in the (J_ϕ, J_z) plane, we verified that Typhon is a distinct structure. In particular, we see that Typhon members have much higher apocenters ($\approx$ 100 kpc vs. $\approx$ 15 kpc for DTG-11) which becomes obvious when comparing their very different J_r values. In addition, we compared our sample to the thorough Malhan et al. [2022a] atlas of stellar streams and found no previously mapped equivalent structure. We note that the discovery of the Typhon structure was confirmed by Dodd et al. [2022a] shortly after the first release of our work using a formal clustering metric.

Chemical considerations

In addition, a follow up observational study focusing on the chemical abundances of Typhon was conducted by Ji et al. [2022] usign the Las Campanas Observatory. That contribution presents high resolution spectra for 7 Typhon members chosen solely based on observability, including 3 members whose metallicities are not available in LAMOST DR8, which nevertheless show consistent metallicities with the LAMOST sub-sample, thereby supporting our conclusions regarding the metallicity distribution of the structure.

The characteristics of Typhon members given in sub-section 9.1.1 lead us to believe that Typhon is likely the tidal remnant of a dwarf galaxy. In particular the metallicity spread, vertical action spread and structure width appear completely incompatible with a globular cluster progenitor. With metallicities reaching [Fe/H] ~ -1.3 dex, and with a mean of [Fe/H] ~ -1.6 dex, the mass-metallicity relation of dwarf galaxies [Kirby et al., 2013a] suggests that the progenitor likely possessed a luminosity of 10^6 – $10^7\,L_\odot$, perhaps similar to the Sculptor dSph. Ji et al. [2022] concur with us on this point. The estimated velocity dispersion value of $\sigma_v \approx 7.5\,\mathrm{km\,s^{-1}}$ lies between that of the Orphan Stream ($\sigma_v \approx 5\,\mathrm{km\,s^{-1}}$, Koposov et al. 2019) and the stream of the Sagittarius dwarf galaxy ($\sigma_v \approx 13\,\mathrm{km\,s^{-1}}$, Gibbons et al. 2017), suggesting that the mass of the Typhon progenitor likely exceeded $10^8\,M_\odot$ (an estimate for the mass of Orphan Stream progenitor, Fardal et al. 2019), but was not as massive as the Sagittarius dwarf.

Dynamical considerations

We noticed that although in the heavy McMillan [2017a] gravitational potential ($M_{\rm vir} = 1.3 \times 10^{12}\,M_\odot$) all stars in the sample are bound, in the lighter MWPotential2014 (Bovy 2015a, $M_{\rm vir} = 8 \times 10^{11}\,M_\odot$), half of the Typhon stream members are unbound[9]. This underlines how having constraints on the trajectories of streams such as Typhon is of great value as the trajectories of these streams are very dependant on the acceleration field of the Milky Way and its underlying dark matter distribution.

We also checked whether the Typhon members could have close encounters with the Large Magellanic Cloud (LMC) or the Sagittarius dwarf galaxy. Taking the trajectories of the two satellites from Vasiliev et al. [2021], we find that the LMC remains always very distant ($\gtrsim 40\,\mathrm{kpc}$). However, the Typhon stars probably did experience a relatively close flyby of Sagittarius ($\sim 20\,\mathrm{kpc}$, 0.10 Gyr ago). We note that Typhon and Sagittarius share very similar orbital planes, although they possess opposite angular momentum vectors (i.e. the direction of motion in the plane is opposite). The interaction

[9]Note that none of the Typhon stars were flagged as hyper-velocity stars in the Marchetti et al. [2019] census.

between Typhon and Sagittarius will be interesting to analyse with N-body simulations, but we defer that investigation to a future contribution.

Perspectives

The identification of this high apocenter polar stream passing so close to the Sun raises many questions. Assuming that the Solar vicinity is not special and is representative of an average location in the disk, the present detection could be used to place constraints on the number of highly radial accretions that took place during the formation of the Milky Way. The picture suggested by Typhon is that there may be a large population of outer halo dwarf galaxies or dwarf galaxy fragments residing near their apocenters, akin to the "Oort Cloud" around the Sun. A more thorough survey of local phase space for other Typhon-like structures and also deeper next-generation sky surveys (with LSST, for instance) that might detect them in place in the outer halo will help quantify this possibility.

This discovery also underlines the relevance of stream research in the Solar vicinity where great quantities of high quality data are available in addition to spatially wider searches. This poses several challenges and may require the development of new algorithmic approaches suited to exploit Gaia era data for nearby structures with incomplete astrometry (e.g. missing line of sight velocities) as sections of streams passing near us are not easily identifiable as streams when projected onto a sky map.

In future work, it will be very useful to attempt to extend the detections along the stream so as to chart it out further in its orbit through the Galaxy. As we alluded to above, such stars may provide very useful dynamical probes for the Milky Way's dark halo, and they will be invaluable to inform follow-up simulation studies attempting to model the N-body evolution of the system. Similarly, having full metallicity information for the member stars would be of great value in order to confirm the present hypothesis regarding the nature of the progenitor.

9.1.4 Data availability

The final Typhon sample is provided at DOI:10.5281/zenodo.6979887, including both Gaia data and other parameters deduced here such as action-angle coordinates.

9.2 Antaeus: A Retrograde Tidal Group in the Milky Way Disk Plane

The complex formation and merging history of the Milky Way (MW) can perhaps be best understood by examining its stellar halo, host to many tidal debris of disrupted galaxies and globular clusters. Dynamical times in the halo are long, so the debris can persist there as coherent phase space structures for billions of years (see e.g. Helmi and de Zeeuw 2000), making them easier for us to detect.

With the advent of the Gaia mission [Gaia Collaboration et al., 2016b] and its superb astrometric data, the task of digging into the stellar halo to uncover the past has been made more accessible. The stellar halo of the MW is now understood to be the product of several important accretion events making up most of its population [Di Matteo et al., 2019], the biggest of which being Gaia-Sausage/Enceladus [Belokurov et al., 2018, Helmi et al., 2018]. Stream finding algorithms [Malhan et al., 2018, Ibata et al., 2021b] have now detected dozens of kinematically coherent structures which will help chart the acceleration field of our Galaxy, providing a wealth of model-agnostic information.

The Gaia data also makes it possible to use action coordinates (J_r, J_ϕ, J_z) to detect stellar structures. Actions keep relevance over very long times if the potential evolves slowly and are thus especially useful to trace past mergers. Recently, Yuan et al. [2020b], Naidu et al. [2020] and Malhan et al. [2022b] used these quantities to detect and construct maps of the MW's dynamical groups and link them to important merger events.

A similar technique was employed by Myeong et al. [2018] to find several retrograde structures in the stellar halo, which were then tentatively associated to the ω Centauri globular cluster, which Majewski et al. [2012] had already suspected of bringing in such material. Retrograde structures have been linked to accretion events for a long time [Carollo et al., 2007], and it has been confirmed by Helmi et al. [2017] that the less bound stars in the halo are typically on retrograde orbits. Sestito et al. [2021] also highlight the importance of the metal poor retrograde halo population for tracing the early building blocks of the galaxy.

Myeong et al. [2019] reexamined the structures from Myeong et al. [2018] and linked them to a substantial merger event they named Sequoia. The Sequoia progenitor galaxy could have brought those retrograde groups and possibly ω Centauri as well. The fact that its stellar population is distinct in metallicity and orbital parameters from the Gaia-Sausage makes the event another important piece of the stellar halo puzzle.

In this Section we present the discovery of Antaeus[10], a retrograde high energy group of tidal debris in the MW's disk plane, made using action-angle coordinates derived from the Gaia DR3 catalog [Gaia Collaboration, 2022] and the Stäckel fudge implemented in Agama [Vasiliev, 2019b]. The new structure has several properties which are similar to those of Sequoia stars, so we discuss its possible affiliation to this event, although both its position in the disk of the MW and its extraordinary low vertical action make it stand out.

Sub-section 9.2.1 details our selection process from Gaia data, sub-section 9.2.2 offers a chemo-dynamical characterization of the structure, sub-section 9.2.3 discusses our finding and sub-section 9.2.4 gives our Antaeus sample.

9.2.1 Selection process

Pre-selection

Throughout this article, we use the right-hand side Galactic Cartesian coordinates for the MW with the Sun located at $(x, y, z)_\odot = (-8.2240, 0, 0.0028)$ kpc (taking the Solar radius from Bovy 2020 and the height above the mid-plane from Widmark et al. 2021) having peculiar velocity $(v_x, v_y, v_z)_\odot = (11.10, 7.20, 7.25)\,\mathrm{km\,s^{-1}}$ (Schönrich et al. 2010, but with the velocity in the direction of Galactic rotation taken from Bovy 2020), and circular velocity $v_c(R = R_\odot) = 243\,\mathrm{km\,s^{-1}}$ [Bovy, 2020]. Our starting point is the Radial Velocity Spectrometer (RVS, Recio-Blanco et al. [2022]) sample of Gaia DR3, for which we derive action-angle coordinates (J_r, J_ϕ, J_z) and orbital parameters using Agama [Vasiliev, 2019b] in the MW gravitational potential of McMillan [2017b]. From this catalog, we take the stars with good parallax measurements ($\varpi/\delta\varpi \geq 10$) and $d \leq 1.5\,\mathrm{kpc}$ so as to retain a good quality Solar neighborhood sample. Since our aim is to investigate the structures that are falling down onto the Milky Way, we choose to select stars with large apocenter distances, $r_{\mathrm{apo}} \geq 25\,\mathrm{kpc}$. These cuts leave us with 3624 stars; we plot the resulting selection in the $J_\phi J_z$ plane, coloured by r_{apo}, in Figure 9.5 (top panel).

Antaeus selection

Among the many interesting structures that stand out from this view, we focus our attention on the low J_z, retrograde moving group of stars delimited by the black rectangle ($2500 \leq J_\phi \leq 3500\,\mathrm{km\,s^{-1}\,kpc}$, $J_z \leq 150\,\mathrm{km\,s^{-1}\,kpc}$), into which we zoom in Figure 9.5 (middle panel). We notice a good agreement in

[10] In Greek mythology, Antaeus is the child of Gaia and Poseidon, a giant whose name comes from "opponent".

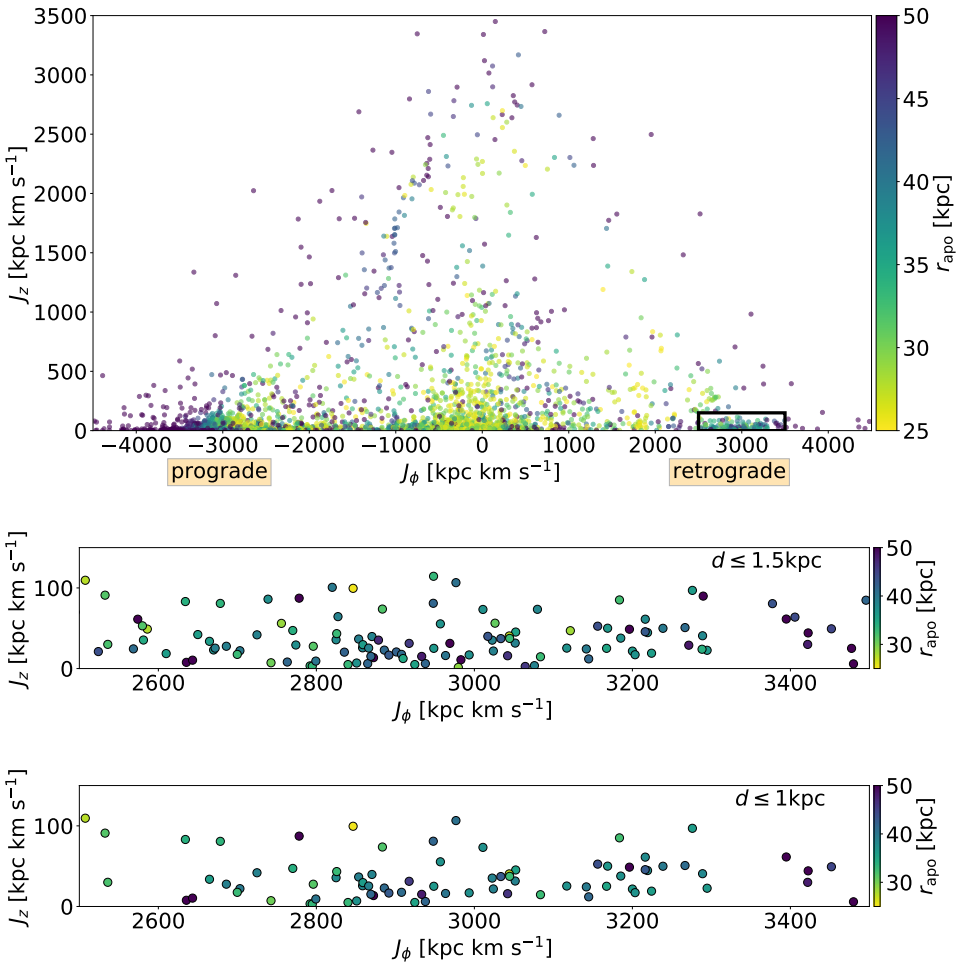


Figure 9.5: **Selection process in the space of actions.** Top panel: Gaia DR3 stars from the selection process described in Section 9.2.1 (i.e. $\varpi/\delta\varpi > 10$, $r_{apo} \geq 25\,\mathrm{kpc}$ and $d \leq 1.5\,\mathrm{kpc}$). Middle panel: zoom on the low J_z region delimited by the rectangle in the top panel ($2500 \leq J_\phi \leq 3500\,\mathrm{km\,s^{-1}\,kpc}$, $J_z \leq 150\,\mathrm{km\,s^{-1}\,kpc}$). Bottom panel: same region as the middle panel, but for our final cut using distances $d \leq 1\,\mathrm{kpc}$ from the Sun.

apocenters for stars in this region, further suggesting the presence of a stellar structure with coherent motion.

Finally, we experimented with the heliocentric distance cut to see how the selection changes. We noticed that by selecting stars within a distance of $d \leq 1\,\mathrm{kpc}$ from the Sun (Figure 9.5, bottom panel) the agreement in apocenters is slightly better, removing in particular some extreme values from the previous cut. This leaves a sample of 80 stars which are given in 9.2.4.

Antaeus structure

In order to establish the statistical significance of this detection, we repeat the same selection on the *Gaia* Universe Model Snapshot (GUMS, Robin et al. [2012b]) simulation updated for DR3. The initial $d \leq 1.5\,\mathrm{kpc}$ cut on GUMS gives 3781 stars, very close to the number of stars in our DR3 selection. Normalizing for this small difference, we find that there is, in the final selection (black rectangle), more than 5 times the number of stars in DR3 than there is in GUMS. Furthermore, the distribution along the J_ϕ axis is bimodal in the GUMS data, with a main peak in the prograde region ($J_\phi \approx -3000$) and a small peak around $J_\phi = 0$, while the same distribution in our DR3 selection is trimodal with an additional peak in the retrograde region ($J_\phi \approx 3000$) corresponding to Antaeus, and the peak around $J_\phi = 0$ being more pronounced. Using the GUMS simulation as an estimate of the expected Galactic populations, the Antaeus feature corresponds to a $\approx 7\sigma$ detection.

9.2.2 Sample characteristics

Dynamical characteristics

We show the positions and velocities of our selection of stars in Figure 9.6 (top panel). It appears clear that the stars belong to a coherent structure dynamically, moving in a retrograde motion in the disk plane of the MW. The structure is rather thick, with a width of at least 1.5 kpc. We identify some outliers from this bulk motion, which all have a distinctive positive velocity in the x direction ($v_x \geq 0$). For the remainder of this study, we will exclude those 15 outliers from our sample, leaving us with 65 stars of the Antaeus stream. In Figure 9.6 (middle panel), we plot velocity planes $v_r v_\phi$, $v_r v_z$, $v_\phi v_z$ with this separation taken into account, showing the compactness of Antaeus stars in those projections.

Chemical characteristics

We crossmatch our selection with the LAMOST DR8 catalog [Wang et al., 2022b] and find 8 stars in common, for which we obtain metallicities from their "FEH_PASTEL" values. These LAMOST stars have a mean $[\mathrm{Fe/H}] = -1.74^{+0.06}_{-0.07}$, with an intrinsic spread of $\sigma = 0.11^{+0.10}_{-0.04}$ (correcting for the LAMOST metallicity uncertainty estimates) and individual values ranging from $[\mathrm{Fe/H}] = -1.33 \pm 0.23$ to $[\mathrm{Fe/H}] = -2.09 \pm 0.30$. The colour magnitude diagram (CMD) of the sample is shown on Figure 9.7, compared to old metal poor isochrones (12 Gyr, $[\mathrm{Fe/H}] = -1.75$ and $[\mathrm{Fe/H}] = -1.50$) from the PARSEC library [Bressan et al., 2012]. The photometry is corrected for interstellar extinction using the 3D extinction estimates calculated by Anders et al. [2022].

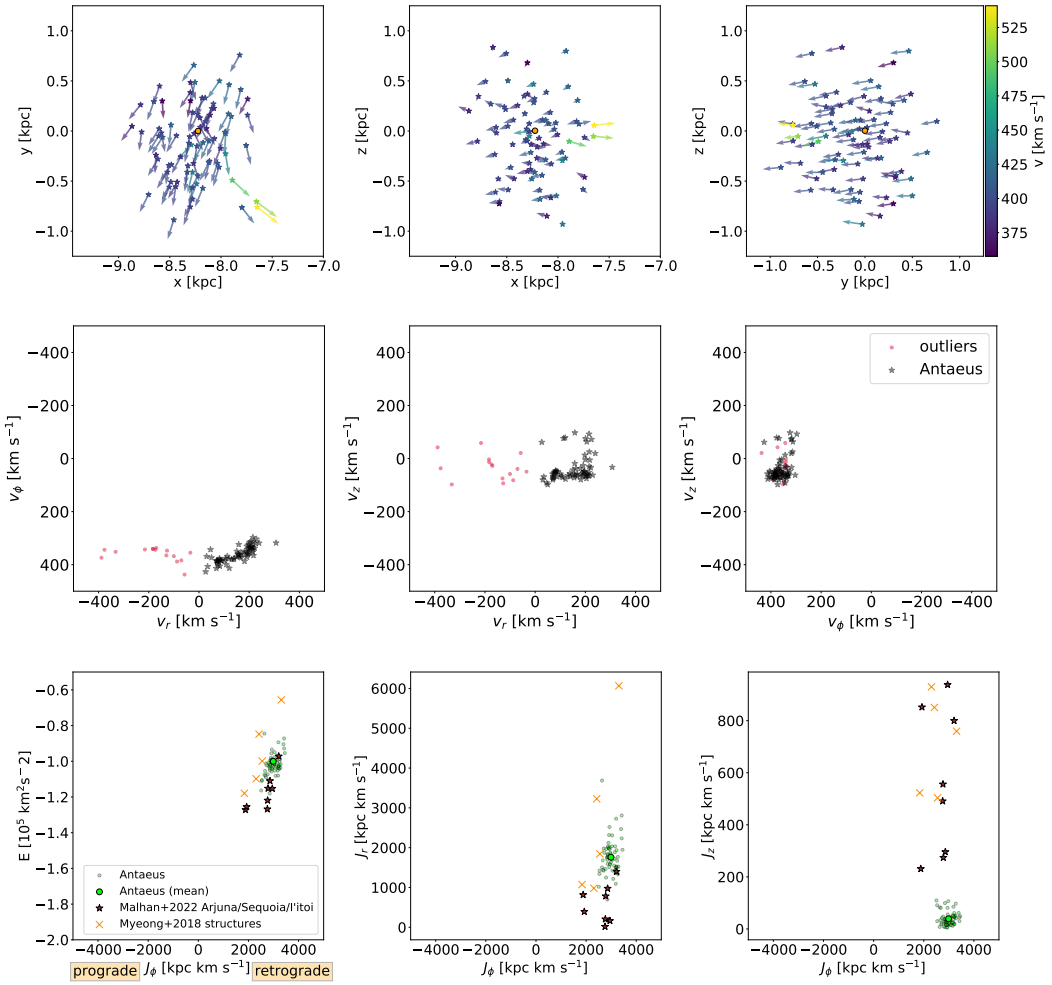


Figure 9.6: **Antaeus members.** Top panel: position and velocity vectors of our selection of stars from Section 9.2.1 colored by total velocity; we plot bulk motion outliers with a slightly transparent line. The orange ball represents the Sun. Antaeus stars are currently passing through our Solar neighbourhood, going in a retrograde motion in the Milky Way's disk plane. Middle panel: velocity planes $v_r v_\phi$, $v_r v_z$, $v_\phi v_z$ with the outliers (red dots) from the top panel bulk motion separated from Antaeus' stars (black). Note that we inverted the v_ϕ axes to be coherent with usual velocity plots. Bottom panel: position of Antaeus (green dots) in energy E and actions J_r, J_ϕ, and J_z, compared to Sequoia-associated retrograde structures from Myeong et al. [2018] (orange crosses) and Arjuna/Sequoia/I'itoi-associated streams and globular clusters from Malhan et al. [2022b] (brown stars).

Orbit integration

Finally, we integrate back in time the orbits of the Antaeus stars in the McMillan MW potential for 1.5 Gyr, and in the MWPotential2014 [Bovy, 2015b]; we show the results in Figure 9.8. Here also the structure appears very coherent dynamically. We find, for the McMillan MW potential

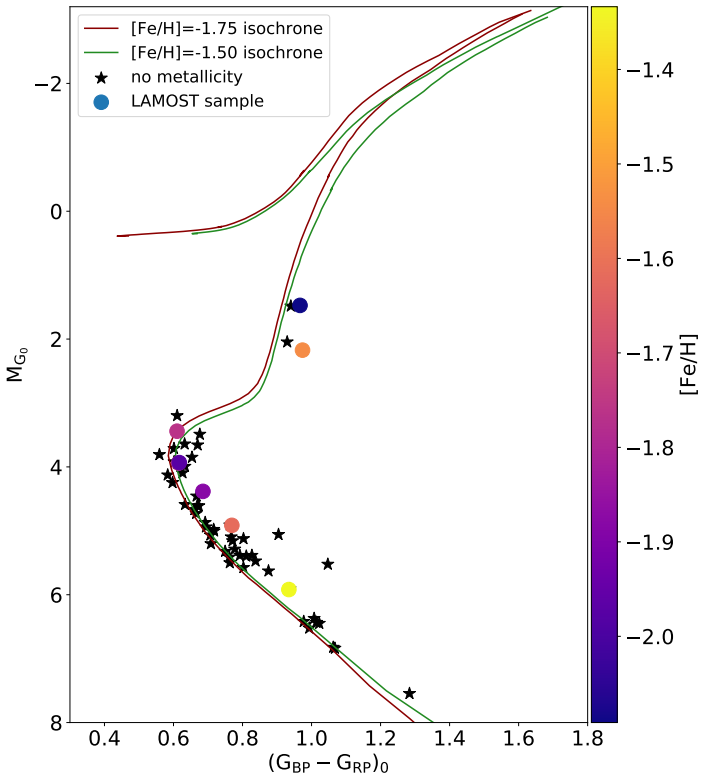


Figure 9.7: **Antaeus CMD.** Colour magnitude diagram for our sample of Antaeus stars, compared to PARSEC model isochrones [Bressan et al., 2012] of age 12 Gyr and metallicities [Fe/H] $= -1.75$ (red) and [Fe/H] $= -1.50$ (green). The colorbar gives the [Fe/H] for the 8 LAMOST stars.

($M_{\rm vir} = 1.3 \times 10^{12}\,{\rm M}_\odot$), a mean pericenter radius of $r_{\rm peri} = 7.3\,{\rm kpc}$, a mean apocenter radius of $r_{\rm apo} = 39.3\,{\rm kpc}$, a mean orbital eccentricity of $e = 0.69$, and a mean orbital time of $t_{\rm orbit} = 1.1\,{\rm Gyr}$. For the lighter MWPotential2014 however ($M_{\rm vir} = 8 \times 10^{11}\,{\rm M}_\odot$), those values become mean $r_{\rm peri} = 7.3\,{\rm kpc}$, mean $r_{\rm apo} = 71.9\,{\rm kpc}$, mean $e = 0.81$, and mean $t_{\rm orbit} = 1.5\,{\rm Gyr}$. The 8 LAMOST stars, whose orbits are plotted in solid black, appear to be good representative members of the stream.

The mean actions of stars in the structure are ($J_r = 1761, J_\phi = 2990, J_z = 39$) kpc km s^{-1}, and their mean energy is $E = -10^5$ km^2 s^{-2} (in the McMillan 2017b potential model); we show this information for individual stars in Figure 9.6 (bottom panel).

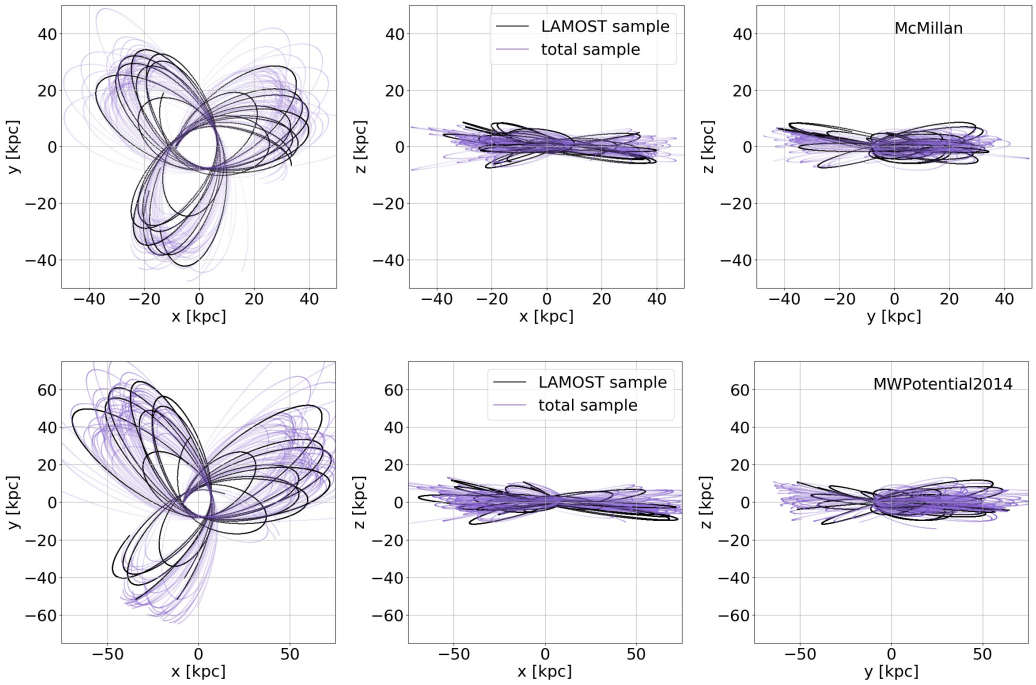


Figure 9.8: **Antaeus orbit.** Orbits of members stars seen in Galactic Cartesian coordinates, integrated backwards in the McMillan [2017b] MW potential for 1.5 Gyr (top panel), and in the MWPotential2014 model for 2.5 Gyr (bottom panel). Notice the change of scales, as stars go farther when integrated in the lighter MWPotential2014. Orbits of the LAMOST sample (8 stars) are in solid black, and orbits of the rest of our sample (57 stars) are in purple.

9.2.3 Discussion and Conclusions

Progenitor

Based on the characteristics derived in sub-section 9.2.2, in particular the thickness of the structure (width $\simeq 1.5$ kpc) and the range of metallicity of its constituent stars, it seems highly likely that this group of stars is the remnant of a tidal stream of a disrupted dwarf galaxy. The CMD (Figure 9.7) seems to indicate that the progenitor is seemingly very old, probably around ~ 12 Gyr in age. The agreement is better with a model metallicity of [Fe/H] $= -1.50$, although we derive a mean value of [Fe/H] $= -1.74^{+0.06}_{-0.07}$. It would thus be very helpful to extend our sample of metallicities to help decide the matter. Such metallicities give an estimated stellar mass of 10^6 to 10^7 $M_\odot$ according to the $z = 0$ mass-metallicity relation of Kirby et al. [2013b]. Taking into account the redshift evolution of such relations (for a given metallicity, higher mass at higher redshift is required), we can consider that those constitute lower bounds and that the progenitor probably has a rather high stellar mass of $\geq 10^7$ $M_\odot$, making it likely that it is linked to an already known accretion event.

Related structures

Indeed, when comparing with known halo structures, we find that the mean J_ϕ, energy, and eccentricities of our sample of Antaeus stars show many similarities with the Arjuna/I'itoi/Sequoia group of mergers [Naidu et al., 2020]. However Antaeus seems more akin to the retrograde structures of Myeong et al. [2018] and to the retrograde tail of the Sequoia event [Myeong et al., 2019] (see the bottom row in Figure 9.6 for a comparison to the previously mentioned groups), especially when factoring in the metallicity of its population. The $\sim 12\,\mathrm{Gyr}$ age derived from the CMD comparison is also consistent with estimates for Sequoia groups [Ruiz-Lara et al., 2022].

Nonetheless, Antaeus' extraordinarily low mean J_z and its position in the disk plane of the MW both make it unique, even when compared to the global atlas of halo structures from Malhan et al. [2022b]. It may be the distinct, low J_z tail of the L-RL64 cluster discovered by Ruiz-Lara et al. [2022] and also detected by Dodd et al. [2022b]. If the structure is indeed related to Sequoia, this difference has to be explained.

Origin

The mere existence of such a streamy, retrograde structure in the disk of the MW is very puzzling. It is not clear how such kinematic coherence could be retained if this population came in with Sequoia $9 \sim 11\,\mathrm{Gyr}$ ago [Myeong et al., 2019]. Of course Antaeus' progenitor could have arrived initially with a small inclination, although this possibility appears somewhat contrived. See however the simulations from Amarante et al. [2022] in which nearly radial mergers could potentially produce such populations. It seems more natural to explain the very low quantity of vertical motion by dissipation due to dynamical friction, which might be consistent with an early arrival in the MW. This scenario would invite the possibility that Antaeus is the debris of the dense core of the Sequoia progenitor, which would have stabilized in the disk through dynamical friction before tidal disruption completely destroyed it.

Perspectives

The discovery of Antaeus opens many exciting possibilities for follow-up studies. A first step would be finding other members of the structure in Gaia with the information we now possess. Creating an N-body model for the infall of the progenitor dwarf galaxy in the potential well of the MW and exploring the possibilities for its survival in the disk would also be highly informative. Finally, it would be very helpful to measure the metallicity of more stars of our selection in order to facilitate discussions regarding the origin of the structure, and links to Sequoia in particular.

9.2.4 Data availability

The final Antaeus sample is provided at DOI:10.5281/zenodo.6912366, including both Gaia data and other parameters deduced here such as action-angle coordinates.

9.3 The Atlas of Milky Way Stellar Streams

This Section provides a brief overview of the comprehensive atlas of Milky Way stellar streams, compiled following the release of Gaia DR3 [Gaia Collaboration, 2022]. For detailed information, see [Ibata et al., 2021a].

Sub-section 9.3.1 briefly discusses the `STREAMFINDER` algorithm used to identify stellar streams from the Gaia dataset. Sub-section 9.3.2 details the observational follow-up campaign to which I have contributed. Sub-section 9.3.3 presents the 87 compiled thin stellar streams, including 28 newly identified streams that emerged from our efforts. Finally, Sub-section 9.3.4 outlines how these streams can be utilized to constrain mass models of the Milky Way.

9.3.1 The `STREAMFINDER` algorithm

This promise of being able to map out the acceleration field of our Galaxy has motivated the development of a dedicated stream-detection algorithm, the `STREAMFINDER`, with the intention to deploy it on the Gaia mission catalogs [Gaia Collaboration et al., 2016b].

Overview

The `STREAMFINDER` is effectively a friend-finding algorithm, with a "distance" in the parameter space of observables defined so as to make objects on similar orbits and with similar stellar populations appear close together. The procedure is presented in detail in Malhan and Ibata 2018, Malhan et al. 2018, Ibata et al. 2021, Ibata et al. 2021a. [Malhan and Ibata, 2018] applied the algorithm to the Gaia DR2 catalog, based on 22 months of astrometric observations, while [Ibata et al., 2021] extended the search to the Gaia EDR3 catalog, with 33 months of observations. In [Malhan et al., 2018] the `STREAMFINDER` sources were cross-matched with the Pristine survey catalog [Starkenburg et al., 2017, Martin et al., 2023], providing metallicity estimates for the stars and hence better discrimination against contamination, which allowed for lower the detection threshold and so find further stream candidates.

Alterations for DR3

Here, the STREAMFINDER algorithm is applied to the full Gaia DR3 [Gaia Collaboration, 2022] catalog, hunting for these structures using an improved version of the STREAMFINDER algorithm, employing a stream template with a spatial width of 100 pc so as to find structures resulting from the expected dissolution of globular clusters or very small dwarf satellite galaxies. It is also modified so as to be able to exploit the 33 millions radial velocity measurements provided in DR3, which are complemented with measurements from other large spectroscopic surveys, specifically, DR3 is cross-matched with the APOGEE-2 survey [Majewski et al., 2017], the GALAH DR3 survey [Buder et al., 2018], the LAMOST DR7 survey [Cui et al., 2012], the Radial Velocity Experiment (RAVE DR5) [Kunder et al., 2017], the SDSS/Segue survey [Yanny et al., 2009], the *Gaia*-ESO survey [Randich et al., 2022], and the S5 survey [Li et al., 2019].

9.3.2 Spectroscopic observations

VLT/UVES follow-up

We used the VLT/UVES spectrograph [D'Odorico et al., 2000] to follow up selected STREAMFINDER sources detected in the *Gaia* EDR3 and DR3 catalogs. These runs comprise runs 105.20AL.001 (2.5 nights in visitor mode), 110.246A.001 (40 hours service in service mode), and 111.2517.001 (3.6 nights in visitor mode). Our instrumental setup uses the DIC2 dichroic beamsplitter in the "437+760" setting, covering the wavelength ranges 3730–4990 Å and 5650–9460 Å. To reduce read noise, we binned the CCD in 2×2 pixel blocks, which in conjunction with a $1''.0$ wide slit yields a spectral resolution of approximately 40,000. Exposure times were selected on a star-by-star basis to reach $S/N \sim 3$–5 for the fainter stars in the sample, so as to measure their radial velocities, but we set a minimum exposure time of 5 min. For the brighter stars, this minimum exposure time also allowed some elemental abundances to be measured. The spectra were reduced with the "esoreflex" pipeline using daytime calibration arc lamps and flat-field images, resulting in extracted wavelength-calibrated one-dimensional spectra.

INT/IDS follow-up

We also secured observations with the IDS long-slit spectrograph on the 2.5m Isaac Newton Telescope in several runs over the course of 2022. Bright northern hemisphere stream stars were targeted with typically ~ 1 hour exposures at $G = 16$ mag. The instrument was configured with the RED+2 detector, the

R1200R grating with a central wavelength set to 8500 Å, a 1″ wide slit and the GG495 order sorting filter.

The radial velocities of the target stars were measured with the IRAF `fxcor` algorithm, using the bright and relatively metal-poor star HD 182572 as a radial velocity standard. The UVES spectra were of sufficient quality and resolution to obtain excellent radial velocity measurements with $< 1\,\mathrm{km\,s^{-1}}$ uncertainty for stars to $G = 18$ mag, while the IDS spectra resulted in velocity uncertainties of $\sim 10\,\mathrm{km\,s^{-1}}$ at $G = 16$ mag.

9.3.3 Atlas of Milk Way streams

Utilizing the `STREAMFINDER` algorithm and subsequent spectroscopic follow-up observations, 87 thin stream-like structures have been identified within the Milky Way. This includes 59 streams previously detected by the community — the bulk by [Malhan and Ibata, 2018, Malhan et al., 2018] — and 28 new discoveries, culminating in the most comprehensive and current atlas of thin streams available as of Gaia DR3 [Bonaca and Price-Whelan, 2024][11] The 87 streams are depicted in Figure 9.9 with a color scheme that facilitates distinction among them.

9.3.4 Mass constraints

Let us now detail how such streams were exploited to refine our current mass model for the MW.

Stream progenitors are modeled as Plummer spheres, a classical spherical model where the potential is given by:

$$\Phi(r) = -\frac{G.M}{\sqrt{r^2 + b^2}} \tag{9.1}$$

where M is the mass and b is a scale length parameter.

For the Milky Way's disk component, the adopted density distribution is described by:

$$\rho_d(R, z) = \frac{\Sigma_d}{2h_z} \exp\left(-\frac{R}{h_R} - \frac{|z|}{h_z}\right) \tag{9.2}$$

where Σ_d represents the central surface density, h_R the scale length, and h_z the scale height.

[11] It is important to note that this compilation does not include the numerous streams which like Typhon, were identified by studying their near-Sun members and for which we only have access to a limited section near the Sun and not their full extended structure to date.

The halo component is modeled as:

$$\rho_s(s) = \rho_0 \left(\frac{s}{r_0}\right)^{-\gamma} \left(1 + \frac{s}{r_0}\right)^{\gamma-\beta} e^{-s^2/r_t^2} \tag{9.3}$$

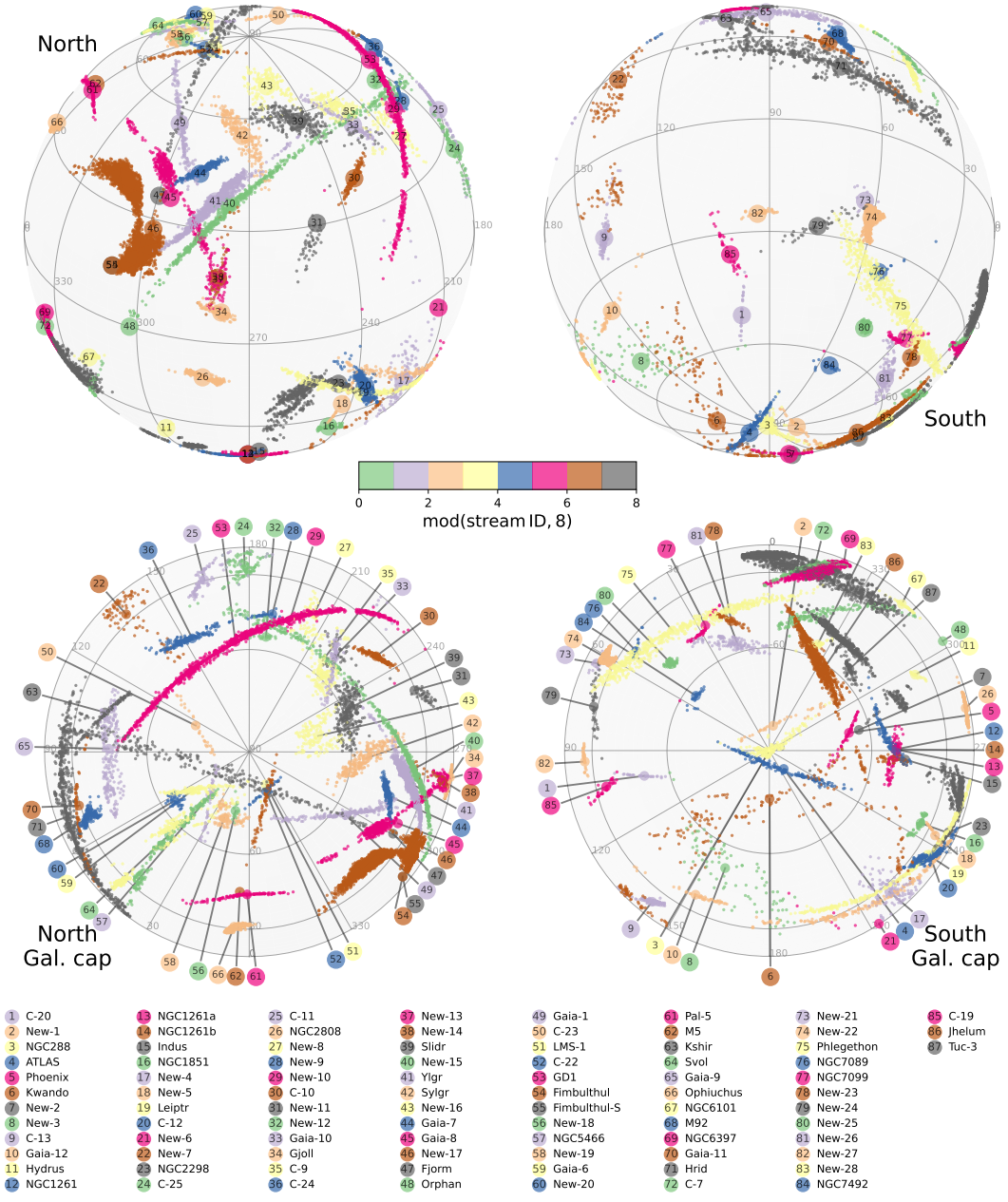


Figure 9.9: **Atlas of stellar streams of the Milky Way.** This figure displays the projected members of the 87 stellar streams identified in the Milky Way as detailed in Ibata et al. [2021a]. Each stream is color-coded using a modulo 8 scheme to differentiate between them.

where ρ_0 is the central density, β and γ are the inner and outer power-law slopes, respectively, r_0 is the scale radius, r_t is the truncation radius, and $s = \sqrt{R^2 + z^2/q_m^2}$ represents an ellipsoidal coordinate, with q_m indicating density flattening.

The Large Magellanic Cloud (LMC) is modeled using an NFW profile. With considerations for reflex motion and dynamical friction all free parameters - including those for the Milky Way, its satellites, and 29 long, thin, velocity-confirmed streams (that offer robust probing capabilities) are optimized using a Markov Chain Monte Carlo methods [Ibata et al., 2011, Goodman and Weare, 2010]. The resulting parameters of this state-of-the-art model of the Milky Way with its satellites and streams, are detailed in Ibata et al. [2021a].

This approach is emblematic of traditional galactic modeling strategies, where assumed shapes, expressed as functional forms, define the components of our Milky Way models. Yet, the true nature of these components may be more nuanced or complex than our current analytical models suggest. What if instead, we allowed for models with free-form structures? Exploring this possibility and its implications will be the focus of the next Chapter.

Chapter 10

Free-Form Potential Recovery from Stellar Coordinates

Portions of the content presented in this Chapter have been previously discussed in the following publication:

2024 *An end-to-end strategy for recovering a free-form potential from a snapshot of stellar coordinates*
W. Tenachi, R. Ibata, F. Diakogiannis
IAU **S379** 147, arXiv:2305.16845

Summary.

We present an observation-driven agnostic unsupervised learning framework we name `MassFinder` for recovering a free-form gravitational potential and its underlying dark matter distribution from a mere snapshot of stellar positions and velocities. Our method leverages a canonical transformation to the space of orbits through a normalizing flow neural network and by making use of auto-differentiation — the automated tracking of derivatives through a computational graph. We then discuss the potential of such approaches to dynamics and its application to Gaia's 6D sample.

As detailed in Chapter 8, the Λ cold dark matter (ΛCDM) paradigm is very successful at reproducing large scale observations. However, it poses several challenges at the galactic scale [Bullock and Boylan-Kolchin, 2017] that could be resolved by having access to a high resolution map of the potential of the Milky Way and its underlying dark matter distribution. This challenging task is being rendered feasible by the European Space Agency's Gaia mission which is measuring the distance and radial velocity of 33 million stars of the Milky Way (MW), enabling us to have access for the first time to a very large dataset of 6D (position and velocity) stellar coordinates [Gaia Collaboration, 2022].

This Chapter introduces a novel framework, which we dub `MassFinder`, designed to compute a detailed map of the acceleration field and its underlying dark matter distribution from observational data.

In Section 10.1, we provide the observational and methodological contexts that motivate this study. Section 10.2 delves into the normalizing flow neural architecture, which is at the heart of our and many other unsupervised learning approaches to dynamics. Section 10.3 details our framework for deriving a free-form neural network model of the galactic potential using auto-differentiation from currently available stellar data snapshots and demonstrates our approach by applying it to a toy textbook synthetic case. In Section 10.4, we demonstrate how the captured neural potential can be distilled into a physically meaningful functional form through symbolic regression, exploring the space of possible equations. Lastly, Section 10.5 discusses the implications of our findings and outlines future research directions.

10.1 Context & Motivations

In sub-section 10.1.1, we detail the observational context of our method and sub-section 10.1.2 discusses the methodological implications of our approach to dynamics.

10.1.1 Observational context

Luminous tracers of dark matter

Given the impossibility of directly observing dark matter, the movement of stars, which are luminous and observable, provides a practical means to infer

the underlying gravitational field predominantly influenced by dark matter (DM). While the spatial distribution of DM can, to a first approximation, be determined from the gravitational field using Newtonian physics, the challenge lies in accurately recovering this gravitational field from stellar observations. It is critical to note that summing the mass of stars, dust, gas, and other visible matter only accounts for about 10% of the mass in our Galaxy, which is insufficient to describe the total gravitational influence [Battaglia et al., 2005, Kafle et al., 2014].

Direct observation of the acceleration field

Directly measuring the acceleration field of the MW by observing changes in stellar velocities over time is not feasible due to the vast galactic time scales involved. Under current observational capabilities, it would take approximately 10^6 years to observe any meaningful change in a star's orbital velocity, with exceptions only in regions of extreme gravitational influence such as near the Galactic center or black holes [Ravi et al., 2019, Quercellini et al., 2008]. Given this impossibility of reading the MW's acceleration field directly, we are limited to snapshots of stars' positions and velocities — essentially, their 6D phase-space coordinates.

Near-Sun Gaia sample

The 6D data available from Gaia, primarily concentrated near the Sun, in the disk, may not be heavily dominated by DM. However, accessing the acceleration field through this data can assist in detecting DM substructures. Furthermore, this sample includes stars that, although currently near the Sun, traverse much larger distances across the galaxy as we saw in Chapter 9 — providing a broader dynamical perspective. As we will see, our methodology mitigates the spatial constraints of the 6D sample by employing canonical transformations to the action space, which allows for a direct exploitation of the orbits themselves.

10.1.2 Exploiting a frozen phase-space snapshot

Intrinsic acceleration field recovery

One could argue that stellar streams are unique in their ability to trace orbits and therefore gauge accelerations by simply reading the difference in velocity along this orbit (Δv)[1]. However, despite the aforementioned observational

[1]With the notable exception of approaches that track stellar populations by their metallicities, which can provide consistent orbits even within the stellar samples from the disk [Horta et al., 2024], offering another perspective on the acceleration field.

challenges, innovative unsupervised learning techniques have been developed to infer the acceleration field from even phase-mixed stellar coordinates.

The foundational *Deep Potential* framework by Green and Ting [2020] has paved the way to multiple studies [An et al., 2021, Buckley et al., 2023, Lim et al., 2023, Kalda et al., 2024, Tenachi et al., 2023b] demonstrating the potential of these methods. In such approaches one essentially computes the acceleration field or potential map — represented by a neural network — that stabilizes the observed stellar distribution, represented using a so called *normalizing flow* architecture.

It is important to emphasize here that these approaches, focused solely on recovering a numerical acceleration or potential, are agnostic with respect to the underlying theoretical framework. Consequently, the results could align with predictions from Cold Dark Matter (ΛCDM), alternative DM theories, or even Modified Newtonian Dynamics (MOND).

Free form models

A contrario, the conventional approach to modeling the Galactic potential (such as the one described in subsection 9.3.3), typically involves simplified analytic models for both the distribution function and gravitational potential. Yet, as pointed out by Green and Ting [2020] recent surveys [Antoja et al., 2018, Trick et al., 2019], have uncovered a richly structured distribution function within the Milky Way — a complexity that might surpass the capabilities of traditional parametric functional forms such as the ones in Binney and Tremaine [2011]. This realization prompts a shift beyond the classical methodologies toward more adaptable and intricate modeling techniques.

In contrast to these traditional methods, new approaches involve representing both the gravitational potential and the stellar probability distribution function through neural networks. These networks, particularly adept at modeling any functional form[2], are free to capture the subtle intricacies observed in galactic structures or even unknown physics.

This novel modeling paradigm typically employs normalizing flow neural networks to represent probability distribution functions, details of which will be elaborated in the next Section.

[2]That is any smooth functions regardless of its complexity as detailed in sub-section 2.1.2.

10.2 Normalizing Flows

In this section, we delve into the inner workings of the *normalizing flow* (NF) architecture used for modeling probability distribution functions. A normalizing flow transforms a given probability space, denoted as $\mathbf{X}$, into a target probability space, $\mathbf{Z}$. The only prerequisite for the initial space $\mathbf{X}$ is the availability of sample data necessary for the transformation process. Conversely, the target space $\mathbf{Z}$ requires a coordinate sampler and a known analytical probability function f_Z to permit the training of the NF transformation. This transformation is designed to be invertible and bijective, ensuring a one-to-one correspondence between elements in $\mathbf{X}$ and $\mathbf{Z}$ [Kobyzev et al., 2020].

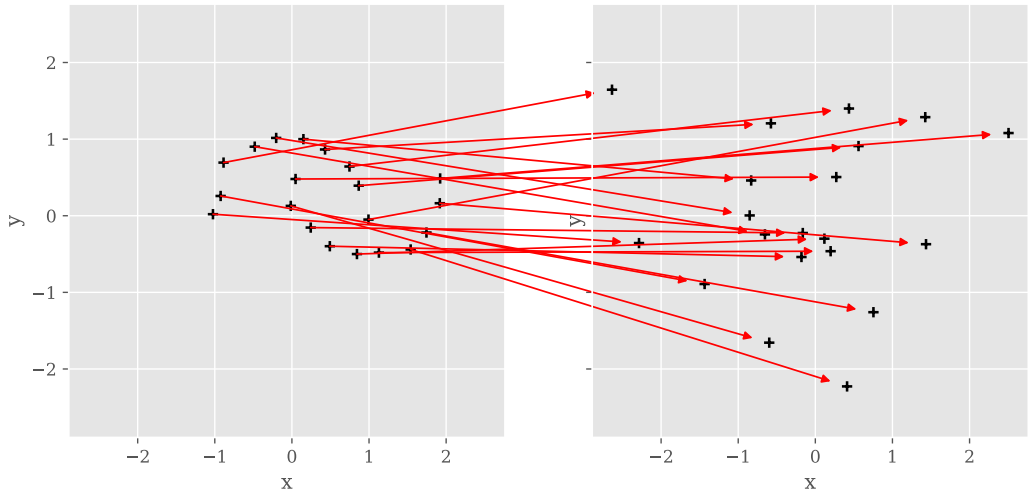


Figure 10.1: **Normalizing Flow transformation.** Here from a toy Moons distribution to a Gaussian

The utility of this technique extends to inferring a smooth probability density function from an arbitrary distribution of points in space $\mathbf{X}$, denoted as $\{(\mathbf{x}, \mathbf{v})_i\}_{i<n_*}$. By mapping this distribution to a Gaussian probability space (where $\mathbf{Z} \equiv$ Gaussian), the transformation leverages the well-understood properties of Gaussian distributions to construct the corresponding density function in the original space $\mathbf{X}$.

Sub-section 10.2.1 gives an intuitive overview of the NF architecture and sub-section 10.2.2 details its formalism.

10.2.1 Overview

This sub-section presents the NF architecture using a 2D "Moons" distribution for illustrative purposes, which simplifies the visualization compared

to a 6D space. Figure 10.1 shows how the NF can transform coordinates from an arbitrary space $\mathbf{X}$ (represented here by the Moons distribution) into coordinates in space $\mathbf{Z}$ (represented here by a Gaussian), where these coordinates correspond to equivalent probabilities. Importantly, the training of an NF does not require prior knowledge of the probability function in space $\mathbf{X}$.

Figure 10.2 depicts the core process of the NF transformation. Beginning with input data points from the space $\mathbf{X}$, specifically the Moons 'phase-space' distribution: $\{X_i\}_{i<n_*} = \{(X_x, X_v)_i\}_{i<n_*} = \{(\mathbf{x}, \mathbf{v})_i\}_{i<n_*}$, $X_i \in \mathbf{X}$, $\forall i < n_*$, these points are fed into the upper left portion of the diagram. Initially, these points are split according to their dimensions (typically in dynamics, samples are separated position component vs. velocity components wise : $\{X_x X_v\}$) and channeled into the initial, or 0^{th}, coupling layer. This initial layer operates as an invertible function, $f_0^{-1}(\beta)$, dictated by a set of tunable parameters β.

Within the first coupling layer, the velocities, X_v, undergo an alteration influenced by a second-order term stemming from a *coupling function* that processes the position data, X_x. This alteration adheres to a defined *coupling law*, here a straightforward summation. The coupling function is executed via a multi-layer perceptron (MLP) (this fundamental architecture was previously detailed in 2.1.2), dependent on the parameters β. As a result, the velocities are modified to: $X_v^{\langle 0 \rangle} = X_v + \text{MLP}_\beta(X_x)$, where $X_v^{\langle 0 \rangle}$ indicates the adjusted velocities at the 0^{th} coupling layer. Concurrently, the positions $X_x^{\langle 0 \rangle}$ retain their initial values X_x through an identity. This sequential modification is repeated through multiple layers, with the positions and velocities undergoing periodic exchanges. After several transformations through these layers, the coordinates initially from space $\mathbf{X}$ transition into $\mathbf{Z}$, ultimately rendering $X_x^{\langle n \rangle} = Z_x$ and $X_v^{\langle n \rangle} = Z_v$ at the n^{th} layer.

About invertibility

Although we rely on an MLP architecture that is not inherently invertible, the entire NF computation remains invertible due to the coupling law utilized. This is because, even in reverse mode, the MLP is computed in a forward manner. To clarify, consider Figure 10.2. Reversing, for example, the n-th coupling layer block (on the right) to retrieve values at the $(n-1)$-th layer, $X_x^{\langle n-1 \rangle}$ and $X_v^{\langle n-1 \rangle}$, from $X_x^{\langle n \rangle}$ and $X_v^{\langle n \rangle}$ can be achieved by recognizing the trivial relationship: $X_v^{\langle n \rangle} = X_v^{\langle n-1 \rangle}$. This can be used to compute the forward MLP of this layer to obtain the residual that can be substracted from $X_x^{\langle n \rangle}$ to recover $X_x^{\langle n-1 \rangle}$, i.e., $X_x^{\langle n-1 \rangle} = X_x^{\langle n \rangle} - \text{MLP}_{\beta^{\langle n \rangle}}\left(X_v^{\langle n-1 \rangle}\right)$. By iterating this process layer by layer, the NF can be traversed in reverse to recover the original input coordinates.

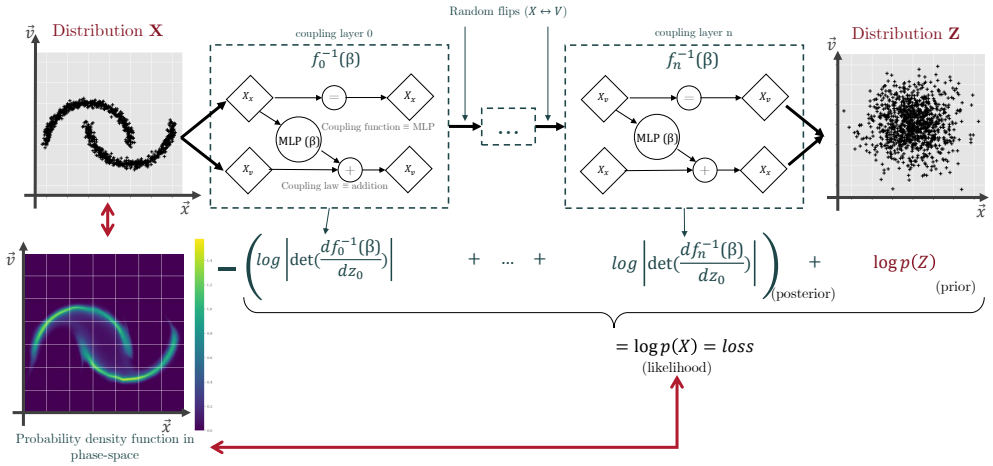


Figure 10.2: **Inner workings diagram of the normalizing flow architecture.** Case of an arbitrary phase-space input distribution. $\{(\mathbf{x}, \mathbf{v})_i\}_{i<n_*}$, mapped to a Gaussian distribution resulting in a smooth, differentiable and normalized density function $f(\mathbf{x}, \mathbf{v})$.

10.2.2 Formalism

The precision of this transformation method in effectively transforming coordinates depends on the adjustment of the parameters β used in the intermediate MLPs and the formulation of the loss function that steers the parameter fitting. To delve deeper into the theoretical roots of the NF architecture, lets us now consider the change of variables formula:

$$p_\beta(X) = p_\beta(f^{-1}(X)) \left| \det \left(\frac{\partial f^{-1}(X)}{\partial X} \right) \right| \tag{10.1}$$

This translates to $p_\beta(X) = p_\beta(Z) \left| \det \left(\frac{\partial Z}{\partial X} \right) \right|$, or equivalently, $p_\beta(X) = p_\beta(Z) \det(J)$. Here, f represents the entire sequence of transformations in the normalizing flow, and J is the Jacobian of the transformation. The transformation function f_β is depends on the parameters β of its MLP components and unfolds through a series of layer operations (coupling layers):

$$X = f_\beta(Z) = f_n \circ ... \circ f_2 \circ f_1(Z) \tag{10.2}$$

$$i.e. \quad Z = f_\beta^{-1}(X) = f_n^{-1} \circ ... \circ f_2^{-1} \circ f_1^{-1}(X) \tag{10.3}$$

Here, $(f_i)_{i<n}$ denotes the individual coupling layers. The dataset $\mathbf{X}$ *flows* through a series of these coupling functions $(f_i)_{i<n}$. Furthermore, during these transformations, the samples from $\mathbf{X}$ are normalized at each stage, ensuring that $\int_\beta p(X)dX = \int_\beta p(Z)dZ = 1$, as indicated in the following equation, justifying the name of the *normalizing flow* architecture.

$$p_\beta(X) = p_\beta(Z) \prod_1^n \left| \det \left(\frac{\partial f_i^{-1}}{\partial Z^{\langle i \rangle}} \right) \right| = p_\beta(Z) \left| \det \left(\frac{\partial f^{-1}}{\partial X} \right) \right| \tag{10.4}$$

Therefore, the log-probability in $\mathbf{X}$ space is given by:

$$\log p_\beta(X) = \log p_\beta(Z) + \sum_{i=1}^n \log \left| \det \left(\frac{\partial f_i^{-1}}{\partial Z^{\langle i \rangle}} \right) \right| \tag{10.5}$$

As depicted in Figure 10.2, our implementation involves calculating the logarithm of the determinant of Jacobians for each coupling layer during processing, which are then integrated into the loss calculation by adding them to the logarithmic probability in space $\mathbf{Z}$,in which there is a defined probability density function.

For the implementation in this work, we predominantly employ the method outlined by Dinh et al. [2016], which introduces a variation from the standard NF framework detailed in Rezende and Mohamed [2015] and illustrated in Figure 10.2[3]. In our approach, the transformations applied in the coupling layers, $(f_i)_{i<n}$, execute the following updates: $X_x^{\langle i \rangle} = X_x^{\langle i-1 \rangle}$, $X_v^{\langle i \rangle} = X_v^{\langle i-1 \rangle} \odot \exp\big(s(X_x^{\langle i-1 \rangle})\big) + t(X_x^{\langle i-1 \rangle})\big)$. Here, s and t represent $\mathbf{R}^3 \to \mathbf{R}^3$ scaling and translation functions respectively, and $\odot$ denotes the Hadamard product (or element-wise product). This results in the Jacobian of the NF being structured as:

$$J = \frac{\partial f^{-1}(X)}{\partial X} = \begin{bmatrix} \mathbf{I}_d & 0 \\ \frac{\partial Z_v}{\partial X_x^T} & \operatorname{diag}\big(\exp[s(X_x)]\big) \end{bmatrix} \tag{10.6}$$

This matrix arrangement makes the NF invertible and computing its inverse computationally inexpensive. A formal demonstration can be found in Kobyzev et al. [2020].

10.3 The `MassFinder` Framework

In this Section we present the `MassFinder` framework for agnostically recovering a free form neural network potential that stabilizes an observed probability

[3]We highlight recent advancements in the field, including continuous normalizing flow methodologies [Grathwohl et al., 2018] and neuro-symbolic regression-inspired frameworks applied to normalizing flows [Tohme et al., 2024].

distribution function. Sub-section 10.3.1 details each step of our framework, sub-section 10.3.2 gives insights about the auto-differentiation aspect of our approach and sub-section 10.3.3 presents a toy experiment that we performed to test our framework.

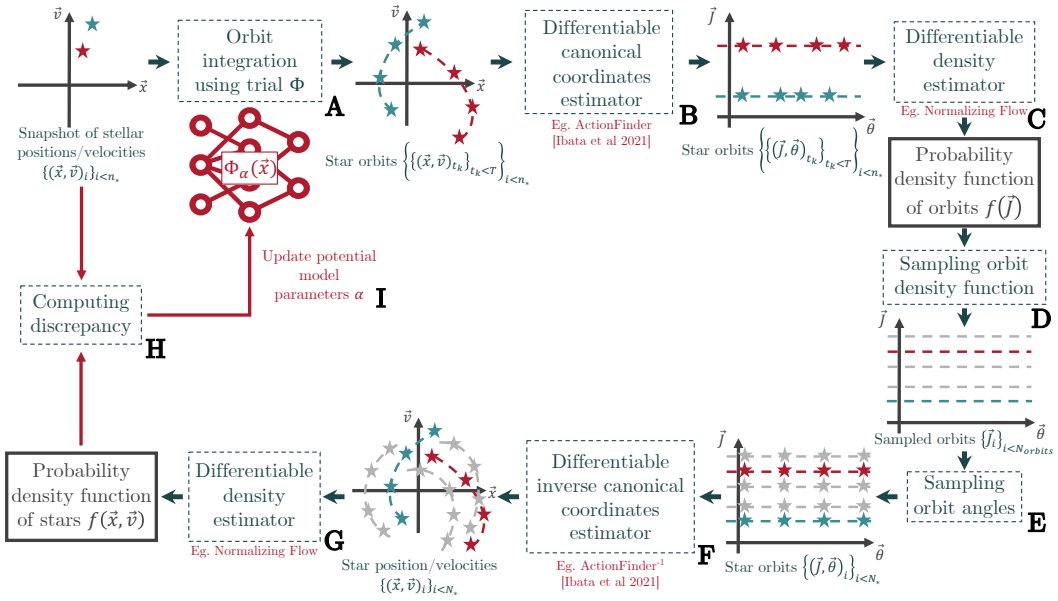


Figure 10.3: **The MassFinder framework.** Proposed strategy for recovering a free form neural network potential that stabilizes a distribution of stars $\{(\mathbf{x}, \mathbf{v})_i\}_{i<n_*}$ using auto-differentiation and gradient descent. See Section10.3 for a full description of the workflow.

10.3.1 Workflow presentation

Our potential recovery framework is shown in Figure 10.3 where we use a normalizing flow as our density estimator. The input data of this framework consists of phase-space stellar coordinates $\{(\mathbf{x}, \mathbf{v})_i\}_{i<n_*}$ obtained from catalogues such as Gaia.

A. These stars are integrated in a trial gravitational potential Φ_α represented by a flexible free form neural network that depends on parameters α.

B. These trajectories are then used to deduce orbits in action space using a differentiable canonical coordinates estimator. For this purpose, the neural network based `ACTIONFINDER` method [Ibata et al., 2021] can be used. This transformation to the space of orbits represented by three integrals of motions i.e. actions : $\mathbf{J} = (J_r, J_\phi, J_z)$ enforces physicality through the Collisionless Boltzmann equation (weak Jeans theorem) and

assumes that the samples are mostly regular (i.e. non chaotic) with non-resonant frequencies (strong Jeans theorem) as detailed in sub-section 8.2.2, which is a reasonable first-order assumption for the Milky Way [Michtchenko et al., 2017].

C. A differentiable density estimator can then be employed to deduce the density function of orbits in action space $f(\mathbf{J})$. For this purpose, a normalizing flow [Papamakarios et al., 2021] or a diffusion[4] model adapted to tabular data such as `TabDDPM` [Kotelnikov et al., 2022] can be used.

D. Sampling this function enables us to obtain new orbits $\{\mathbf{J}_i\}_{i<N_{orbits}}$ in realistic proportions.

E. These can in turn be sampled to deduce stellar coordinates in actions and angles: $\{(\mathbf{J}, \theta)_i\}_{i<N_*}$.

F. By applying an inverse differentiable transformation, one can obtain the Cartesian coordinates of this augmented and phased-mixed stellar population: $\{(\mathbf{x}, \mathbf{v})_i\}_{i<N_*}$.

G. These can be used to infer a smooth density function in phase-space $f(\mathbf{x}, \mathbf{v})$.

H. Finally, this density function can be compared to initial observations, using a Poisson negative log-likelihood loss function : $\sum_{i=1}^{n_*} log(f_\alpha((\mathbf{x}, \mathbf{v})_i))$.

I. Since this final density function depends on the potential neural network's parameters α through all of the steps described above, these can be adjusted to minimize this discrepancy. This process can be repeated iteratively until convergence of Φ_α.

In essence, in our workflow, we are assuming that the system is quasi-stationary (which is arguably verified within a $\sim$ 200 Myr time-scale for the Milky Way [Hou and Han, 2015]) and computing the free form potential that stabilizes the observed stellar distribution.

[4] A deep learning technique [Ho et al., 2020] that involves training a model to gradually denoise data, until it is able to 'de-noise' data from pure Gaussian noise, effectively training it to generate realistic synthetic data from scratch. This method is responsible for the success of current generation image generators such as `DALL-E` [Ramesh et al., 2021] or `StableDiffusion` [Rombach et al., 2022].

10.3.2 Backpropagation

It is worth noting that fitting the large number of parameters that make up neural networks is made possible through backpropagation, which involves computing gradients for each single mathematical operation performed in the workflow [5]. While this approach is powerful, it presents challenges as it necessitates the tracking of all gradients and the utilization of differentiable operations only. We utilize `PyTorch` [Paszke et al., 2019] for this purpose.

In practice, during the training phase, the network evaluates the log probability of observed stellar coordinates under the modeled density function. Ideally, if the model accurately captures the underlying data distribution, these log probabilities should be maximized, resulting in a lower loss, which is computed as the negative sum of these log probabilities. This approach leverages the benefits of the Poisson loss function, notably its convex nature irrespective of the model's complexity. This property ensures that gradient descent methods can be effectively applied to optimize the model's parameters, continually adjusting them to minimize the overall loss and improve the fidelity of the gravitational field estimation derived from the stellar distributions.

10.3.3 Experiment

We demonstrate the efficacy of our scheme using a toy system, substituting Gaia data with synthetic data from an isochrone whose potential is given by:

$$\Phi(r) = -\frac{G.M}{(b + \sqrt{r^2 + b^2})} \tag{10.7}$$

Where M and b are mass and length parameters and using the analytical canonical transformation to actions and angles [see Binney and Tremaine, 2011].

In this toy showcase, we are able to recover the isochrone potential within a mean relative error of 0.1% showing that it is possible to use gradients to backpropagate through all of the steps necessary to recover a gravitational field from observations, including: an orbit integration, a density estimation, a change of coordinates to actions/angle and a data augmentation using actions.

[5] Auto-differentiation was detailed in sub-section 2.1.4.

10.4 Distilling the `MassFinder` Network into an Analytic Function

Symbolic distillation

Although the agnostic recovery of a neural network Φ_α enclosing a potential model for the Milky Way would be of enormous value. We note that such a black box model would contrast with usual empirical laws in that it would be very difficult if not impossible to it connect with theory. Therefore, we suggest the use of symbolic regression which consists in the inference of a free-form symbolic analytical function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ that fits $y = f(\mathbf{x})$ given $(\mathbf{x}, y)$ data for distilling the potential neural network into an intelligible and interpretable analytical function.[6]

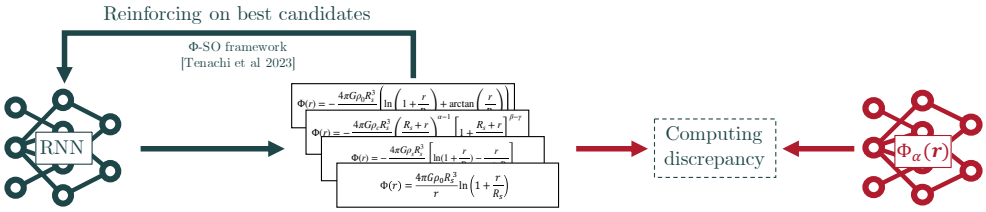


Figure 10.4: **Symbolic distillation of a neural Galactic potential.** Distilling a neural network representing a Galactic gravitational potential into a interpretable analytical expression with Φ-SO (Chapters 4-6). An RNN generates trial expressions, their ability to reproduce the neural network Φ_α is assessed and the best ones are reinforced. This process is repeated iteratively until convergence of the RNN and the extraction a set of high quality expressions.

Here we adopt the Physical Symbolic Optimization (Φ-SO) framework detailed in Chapters 4-6. As illustrated in Figure 10.4 Φ-SO relies on a recurrent neural network (RNN) to generate multiple trial analytical expressions. Fit quality of these expressions can then be assessed against data generated using the Φ_α neural network. Best expressions are then reinforced and the process is repeated until the RNN converges and a set of high quality expressions that reproduce Φ_α predictions is obtained. We note that since this framework relies on reinforcement learning, in addition to fit quality any criteria (even

[6]Symbolic regression is distinct from numerical parameter optimization procedures in that it consists in a search in the space of functional forms themselves by optimizing the arrangement of mathematical symbols (e.g. x, $+$, $-$, $\times$, $/$, sin, exp, log, ...) as well as their parametrizations.

non-differentiable ones) can be used including the condition: $\lim_{r\to\infty} \Phi(r) = 0$. Using the Φ-SO framework, we are able to successfully recover the potential of the toy isochrone system described in Section 10.3.

On the utility of SR

While distilling the resulting numerical potential into an analytical function aligns with our principle of agnosticism — since it avoids imposing any specific functional form — it might conflict with our aim to develop a large, free-form model that can capture the intricate phenomena of the Milky Way, which compact and intelligible analytical functions typically cannot handle.

Nevertheless, we recognize that this additional step could be beneficial for several reasons. First, a substantially more complex analytical function than traditional models that can be found in [e.g., Binney and Tremaine, 2011] — yet compact enough to be intelligible, could potentially provide a sufficiently accurate representation of the Milky Way's potential.

Secondly, even if the Milky Way's potential involves highly intricate structures, symbolic regression could effectively extract lower-order components of this complexity, thereby enhancing the model's interpretability. For instance, the potential could be modeled as $\Phi(\mathbf{x}) = \Phi_0(\mathbf{x}) + \mathrm{MLP}(\mathbf{x})$, combining an analytical foundational function $\Phi_0(\mathbf{x})$ with a lower-order neural network $\mathrm{MLP}(\mathbf{x})$. The MLP component would be able to fine-tune the model to address lower orders complex features of the Milky Way.

10.5 Strategies for Mapping Milky Way Dark Matter

In this section, we explore the methodological implications of our framework as well as similar approaches in the field. We consider enhancements and avenues that could benefit not only our method but also other *Deep Potential*-based models such as : Green and Ting 2020, An et al. 2021, Buckley et al. 2023, Lim et al. 2023, Kalda et al. 2024.

Sub-section 10.5.1 highlights the uniqueness of our approach, particularly our work within the space of actions. Sub-section 10.5.2 discusses the potential of leveraging additional dynamical probes beyond Gaia's 6D near-Sun sample and finally, sub-section 10.5.3 discusses our the current deep learning paradigm in dynamics and how it could be expended.

10.5.1 On the virtues of working in the space of actions

Leveraging canonical transformations

Deep Potential-based methods in the literature often involve fitting a NF to represent a probability distribution function $f(\mathbf{x}, \mathbf{v})$ directly from observational data which is then used to derive a neural representation of the gravitational potential, $\Phi(\mathbf{x})$. Enforcing the potential's compliance with the Collisionless Boltzmann Equation (refer to Equation 8.5), enforcing that f is stabilized by Φ. In contrast, our approach focuses on stabilizing the distribution of orbits themsevles by generating new synthetic samples in realistic proportions within the space of actions. This strategic focus gives us an advantage, particularly when dealing with the spatial limitations of accessible samples, which are predominantly confined to within 3 kpc of the Sun.

Lim et al. [2023] were pioneers in applying a *Deep Potential*-esque framework to actual observational data from Gaia, providing constraints on the MW's potential within 3 kpc of the Sun. We are optimistic that our model, which includes a canonical transformation to action-angle coordinates, will lay the groundwork for spatially extending this approach. Assuming the near-Sun sample of orbits is representative of the broader disk dynamics, our approach could potentially model the entire disk's gravitational potential by working in the space of actions i.e. orbits themselves.

Our current model is but an initial iteration of our framework, with numerous enhancements already planned. However, it is the first of its kind to utilize canonical transformations in this context.

Towards a differentiable action estimator

For this purpose, we currently employ `ActionFinder` [Ibata et al., 2021]. However, integrating `ActionFinder` presents challenges, as it is itself an unsupervised learning method requiring its own convergence through iterative updates. A potential workflow could involve alternating updates — one step for the overall system and one step for `ActionFinder` each epoch — but this integration is inherently complex.

To streamline this process, we might explore developing a traditional deterministic action estimator, such as those based on the *Stäckel fudge* approximation for axi-symmetric systems [Binney, 2012]. The forthcoming versions of the `AGAMA` framework are expected to introduce a canonical coordinates transformation that is differentiable with respect to input coordinates [Vasiliev, 2019]. However, there are no immediate plans to develop an action estimator that is differentiable relative to the potential itself, which presents a unique opportunity for future research initiatives in this area.

10.5.2 Expanding dynamical constraints by exploiting stellar streams and 5D samples

The Milky Way is known to be in significant disequilibrium [Bonaca and Price-Whelan, 2024], a factor that our current framework, like many others, does not account for due to the absence of time-dependence in our models. Ideally, this time-dependence could be modeled by evolving the entire system from an initial state such as to reproduce current observational data accurately by modeling accretion events as time dependent perturbations resulting in new free parameters. One potential elegant solution might involve a model that learns a continuous mapping from a past state — when the Milky Way was spherically symmetric and in equilibrium — to its current state's probability distribution function[7].

However, time-dependent approaches are notoriously data-intensive, which leads us to consider the broader implications of data availability and its constraints on our modeling capabilities.

Exploiting stellar streams

While our framework currently does not specifically incorporate stellar streams, it can potentially be enhanced to utilize the rich data from the numerous stellar streams identified in recent years (as detailed in Section 9.3). Ideally, the potential model would be refined to ensure that stars from a single stream adhere closely to the same orbit, represented by a constant action $\mathbf{J}$. However, the dynamics of most stellar streams are complex, often necessitating more sophisticated modeling approaches going beyond the assumption that all stream members share identical orbits.

A differentiable stream/progenitor model

To address the complexities of modeling stellar streams, we propose the development of a differentiable progenitor/stream model. Traditional stream models often simulate the ejection of stars from a spherical potential representing the progenitor, where stars are randomly ejected to mimic the effects of tidal forces. This method, while effective, relies on discrete operations that are inherently non-differentiable, such as sudden ejections effectively modeled by Heaviside functions.

To make this process differentiable — and thus compatible with gradient-based optimization techniques — we suggest relaxing these discrete operations. For instance, instead of modeling star ejections as binary events (not ejected vs. ejected), we could represent them on a continuous scale using a smooth,

[7] We will mention the `FFJORD` method in sub-section 10.5.3, which is ideally suited for this task.

differentiable function like the arctangent in lieu of Heaviside functions. This approach would allow for partial ejections with varying intensities over time, providing a more flexible and tractable model for the complex dynamics of stellar streams.

De-projecting 5D Data

In addition to exploiting fully characterized 6D data, our framework could leverage the vast majority of Gaia's catalog, which primarily consists of 5D data (missing radial velocity). This would involve a novel approach to de-project these data by inferring probable radial velocities based on the available 5D parameters.

To implement this, we could introduce a radial velocity probability distribution function as an additional free parameter in the model. This distribution would be conditioned on the other five dimensions of data, providing a statistically informed estimate of the missing velocities. The inclusion of this estimated sixth dimension could significantly expand the scope of data usable for modeling the Milky Way's gravitational field, thereby enhancing the robustness and accuracy of the derived potential. The hope being that the degrees of freedom thereby introduced would be compensated by the wealth of additional information exploited.

In practice, this inferred radial velocity distribution could be learned as a secondary correction to an existing reasonable first approximation Milky Way's dynamics in order to facilitate the fitting process [Dropulic et al., 2021].

10.5.3 Deep learning considerations

Improving neural components

Smoothness challenges

The NF architecture we employ here represents an early iteration within the field of neural probability density models. As this area of study grows, numerous advanced alternatives to traditional NF models are being developed. Among these, we previously discussed the `TabDDPM` [Kotelnikov et al., 2022] method, which is a diffusion method tailored for tabular data. However, the leading technique is currently still a NF : the `FFJORD` (Free-Form Jacobian of Reversible Dynamics) approach [Grathwohl et al., 2018]. `FFJORD` generalizes the core principles of NFs, as outlined in Section 10.2, by transforming the discrete steps between probability spaces into a continuous flow, removing the need for multiple discrete transformations. Despite its advances, `FFJORD` and similar approaches encounter significant challenges with ensuring smoothness in the probability density function and in particular its derivatives resulting in

artifacts, a problem highlighted by Kalda et al. [2024] and one that we have also encountered.

`GradNet`

To address these issues, the `GradNet` method, as introduced in sub-section 7.1.1, offers a promising alternative. This approach has the potential to model the probability density function in such a way that its predicted gradients would be as reliable as its integral predictions by construction, which could significantly mitigate the artifact problem encountered in traditional NF implementations.

Symbolic regression

Moreover, an alternative solution could involve employing symbolic regression to model both the probability density function and the associated potential. This method would utilize arbitrary functional forms optimized directly through symbolic regression techniques, offering a smoother and less overfitted model due to the reduced number of degrees of freedom and the inherent differentiability of analytical functions. If the primary objective is to enhance the smoothness of the model, one might even consider using extended and extremely accurate, albeit potentially complex, analytical expressions that may not be immediately intelligible. Implementing this approach with the framework we presented in Chapter 3 would effectively necessitate the use of reinforcement learning techniques.

Towards a reinforcement learning framework

One might question the necessity of ensuring that our framework is fully differentiable. Direct access to gradients via auto-differentiation is certainly advantageous over traditional gradient approximations. Historically, the astrophysics community has relied on methods like Markov Chain Monte Carlo [Gilks et al., 1995] in which gradients are approximated for its optimization needs.

However, the intricacies of neural networks, with their extensive interdependent parameters, make them impossible to optimize using such traditional gradient approximation methods. Nonetheless, the process of approximating gradients to train neural networks effectively without differentiable loss functions is actually an accurate description of the field of deep reinforcement learning (a technique explored in Section 3.3).

Our framework could indeed benefit from employing reinforcement learning, circumventing some of the limitations associated with auto-differentiation. This approach would allow for the integration of non-differentiable elements,

such as action estimators, within a more conventional framework. It is worth noting that reinforcement learning remains unfortunately underutilized within the astrophysics field, yet it offers significant potential for training neural networks where objective functions are derived from non-differentiable systems like simulations.

Chapter 11

Conclusion

Summary.

We synthesize our main findings and discuss the broader implications of this thesis on machine learning approaches to physics and astrophysics.

We then delve into the future prospects for galactic dark matter research offered by our observation-driven, agnostic approach to machine learning. We also discuss, in depth, potential advancements in symbolic learning involving, the extension of this paradigm to incorporate differential equations, the automated formulation of theories, and the impact of symbolic approaches on large language models.

In our final reflections, we emphasize the unique position of astrophysics in addressing the challenges posed by the opaqueness of machine learning in the physical sciences.

This chapter concludes the thesis. In Section 11.1, we synthesize the main findings and contributions of this work. Section 11.2 then explores the broader implications of our findings, with a particular emphasis on the role of inductive bias in local Universe dark matter research and the potential and future prospects of symbolic learning approaches within the fields of physics and astrophysics. Finally, Section 11.3 briefly offers some concluding thoughts and reflections.

11.1 Summary & Overview

We provide a summary of the thesis in sub-section 11.1.1 and highlight the overarching principles that guided our research throughout this work in subsection 11.1.2.

11.1.1 Summary

In Chapter 3, we explored the encoding of formal mathematics into graph structures and discussed methodologies for learning from and generating such graphs. We showed how formal mathematical problems can be viewed as graph optimization problems. Our discussion particularly focused on symbolic regression (SR), which seeks to discover analytical expressions that fit a dataset from scratch. We employed a deep reinforcement learning framework where a neural network sequentially constructs expressions of iteratively increasing fit quality through a trial-and-error process. We described our approach for ensuring the validity of automatically generated expressions using a *prefix notation* and incorporating priors that inherently limit the length of sequences generated by the neural network.

In Chapter 4, we equipped our method, which we dubbed Φ-SO for Physical Symbolic Optimization, with the ability to exploit physical dimensional analysis constraints, significantly narrowing the search space for potential expressions. We achieved this by developing an algorithm capable of conducting highly informative dimensional analyses on partially constructed equations during the expression generation process. Physical units constraints were then applied to ensure the inherent physicality of the expressions using a prior and to teach the neural network, the rules of dimensional analysis, thereby preventing conflicts among the priors.

We demonstrated that Φ-SO is the leading algorithm for exact symbolic recovery by benchmarking it against 17 other SR algorithms using the standard

Feynman benchmark — including 120 equations derived from the Feynman lectures on physics and other textbooks, to be recovered from their associated datasets.

In Chapter 5, we expanded the Φ-SO framework to accommodate the search for a unique functional form that fits multiple realizations of a single class of physical phenomena, allowing each realization to have (possibly) unique free parameter values — an approach particularly relevant to astrophysics. We refer to this new framework as Class SR. Recognizing the novelty of our method, we developed the first benchmark specifically designed to evaluate Class SR systems, demonstrating that Class SR significantly outperforms traditional SR in scenarios where multiple realizations are available. We also showcased the effectiveness of our system using a synthetic dataset of Milky Way streams, successfully deriving an input potential from stellar positions and velocities.

Chapter 6 detailed the `PhySO` software implementation of our Φ-SO framework — demonstrating its unique capabilities and features. We emphasized our commitment to an open-source approach, fostering community interaction and adoption. This commitment is especially vital in the domain of machine learning, where transparency and reproducibility are key. As a result of these efforts, and thanks to the software's clean, robust, and extensible design, `PhySO` has garnered significant adoption within the several physics research communities. Several teams have substantially built upon the `PhySO` framework, enhancing its performances and expanding its capabilities, further solidifying its impact and utility.

In Chapter 7, we explored complementary approaches to our Φ-SO method, able to leverage neural networks to directly capture and embody the graph structure of a dataset, reflecting its underlying analytical representation. Such approaches have the potential to exploit derivatives informing the graph structure with respect to data. Specifically, we introduced a novel method capable of detecting both additive and multiplicative separabilities in a dataset by analyzing the gradients of its neural network representation. Additionally, we discussed an innovative approach where a neural network itself emulates an analytic expression, incorporating basis functions within its structure and promoting sparsity. We then discussed cross-pollination strategies within the field of SR.

We introduced several key improvements to the state-of-the-art in reinforcement learning-based SR. These range from minor refinements to more substantial advancements, including the introduction of an annealing temperature parameter, significant enhancements to the free constants fitting procedure via auto-differentiation, the introduction of Class SR, and the development of new

priors. The most notable of these is the incorporation of dimensional analysis, which informs the neural network of physical units and enables search space reduction while ensuring physical consistency.

Our SR framework has been tested extensively on hundreds of benchmarking synthetic cases, as well as real observational data, such as the study of standard candles to deduce the law of the expansion of the Universe. Additionally, multiple research teams have applied our software to real experimental or observational data across a wide range of physics fields, including astrophysics, aeronautics, mechanics, particle physics, fluid dynamics, telecommunications, and physical geology and even biology. Furthermore, we are currently conducting experiments on data derived from complex simulations, with a focus on the evolution of globular clusters and exploring alternative models for dark matter profiles, as discussed in Section 11.2.

We believe that while developing abstract approaches to physics and astrophysics, it is essential to maintain a strong connection to real-world physics problems. Consequently, we directed our focus to the tangible issue of dark matter at the galactic scale, a topic introduced in Chapter 8.

In Chapter 9, we pursued our observation-driven approach to physical investigations by exploring dark matter probes in the form of structures being accreted by the Milky Way. Studying the near-Sun stellar sample from Gaia, which includes full positional and velocity information, we discovered a new stellar stream which we named Typhon. We expect this polar stream's full sample to extend to the outer halo of the Milky Way at approximately 100 kpc, potentially making it an exceptional probe of dark matter. Following a chemo-dynamical analysis, we identified Typhon as the remnant of a dwarf galaxy — suggesting that many other dwarf galaxy fragments may be lurking in the outer halo. We also contributed to the discovery of another structure — which we named Antaeus — by detecting its members near the Sun, challenging the prevailing view that dynamical times in the disk are short and tidal structures phase-mix quickly, erasing any initial stream-like features.

Chapter 10 focused on methods for recovering the gravitational potential of the Milky Way and its underlying dark matter distribution in a model-agnostic manner. Our approach, dubbed `MassFinder`, essentially requires the potential to stabilize the observed stellar distribution. By employing canonical transformations to work in the space of orbits in the process, we address the spatially limited availability of 6D (position and velocity) samples from Gaia. The chapter concludes by discussing potential advancements in this emerging sub-field of Galactic phenomenology.

11.1.2 Overview

The core principle of this thesis centers around an observation-driven approach, which has steered our integration of machine learning into physics and astrophysics. Contrary to many conventional machine learning methods that learn from examples derived from existing models or simulations, we designed our approaches to generate solutions in an unsupervised manner — without relying on prior models — by inherently requiring adherence to observational data. This model-agnostic strategy is crucial in physics, as it represents the sole pathway to discovering new models that more accurately reflect natural phenomena

Throughout this thesis, we have consistently applied an observation-driven philosophy to our research. For instance, in our work on SR, we did not train neural networks on a pre-established library of mathematical expressions. Instead, we designed models capable of devising *a priori* unknown expressions by requiring them to fit the data but also comply with the rules of dimensional analysis and class constraints. Similarly, in addressing the complexities of dark matter, we avoided relying on predetermined simulations incorporating inductive biases. Our approach has been to develop an unsupervised learning framework that inherently constrains dark matter properties by enforcing adherence to observational constraints, again, aligning our model development with empirical evidence.

On a technical level, we aspire for the methodologies explored in this thesis, particularly auto-differentiation and reinforcement learning techniques, to gain wider adoption within the astrophysical community. We believe that these approaches, though currently underutilized in astrophysics, possess tremendous potential for advancing model-agnostic analyses.

On a deeper level, we aspire for this thesis to mark the beginning of a paradigm shift towards interpretable and meaningful methods in physics and astrophysics, spearheading the integration of symbolic deep learning techniques. Our symbolic learning framework, Φ-SO, represents a pioneering achievement as it is the first and currently only method developed within the physics and astrophysics communities where a neural network directly manipulates mathematical symbols. This work underscores the belief that the prevailing trend of relying solely on supervised learning and black-box models merely scratches the surface of what machine learning can offer to the scientific investigation of natural phenomena.

11.2 Perspectives

This Section explores broad and challenging future directions opened by the research presented in this thesis. While we have discussed numerous potential developments throughout this manuscript, here we focus on the most transformative prospects in sub-sections 11.2.1, 11.2.2, 11.2.3 and 11.2.4, the details of which we will introduce in their respective sub-sections.

Before we delve into these major avenues, let us first recapitulate the most significant perspectives highlighted throughout the thesis. To aid in navigating these discussions, we will reference the sections of the manuscript where these perspectives were previously discussed. Additionally, we will present some new, more straightforward perspectives that, while less central to the thesis's core contributions, we believe still offer valuable pathways for future research.

Advancing symbolic learning

As highlighted in earlier discussions, particularly in sub-section 3.3.3, a critical priority is to enhance the effectiveness of symbolic learning systems. Current methodologies, whether based on neural networks like ours or genetic programming, typically refine symbolic expressions using a scalar, non-differentiable (with respect to symbolic arrangement) metric of fit quality. This approach can inadvertently lead to what we term the *curse of accuracy-guided SR*, where optimization of the metric does not necessarily converge towards the most accurate functional form.

To address this, a significant research direction involves improving the self-correction mechanisms of trial-and-error systems by enabling them to utilize gradients relative symbolic arrangement with respect to symbolic arrangement. This enhancement could involve integrating a supervised learning component that actively learns the local geometry of the functional form search space, as outlined in sub-section 3.3.3. Additionally, fostering synergies with other symbolic regression methodologies could enrich the breadth and depth of symbolic learning, as discussed in sub-section sub-section 3.3.3 and Section 7.3. Further specific prospects related to our approaches incorporating dimensional analysis and class-based symbolic regression have been detailed in Section 4.6 and sub-section 5.4, respectively.

Applying SR to physics & astrophysics

The increasing application of SR methods within the fields of physics and astrophysics represents a significant development, as illustrated by the use of Φ-SO in research as listed in Table 6.1.

Inductive biases

As previously discussed in paragraphs 3.3.2 and 6.1.2, one effective way to enhance the application of SR beyond our current use of class constraints and dimensional analysis is to incorporate domain-specific prior knowledge into the search process. For instance, if the search is for a functional form that is expected to exhibit certain symmetries, limit behaviors, or specific characteristics within a differential equation, these aspects should be integrated into the reward function of Φ-SO. Importantly, because these constraints do not need to be differentiable, Φ-SO can accommodate a wide range of scientifically meaningful restrictions to refine and guide the search process effectively.

Learning analytic approximations to expensive physics

While this thesis primarily focused on observation-driven approaches, symbolic learning also offers significant potential for addressing computationally expensive aspects of simulations. A flourishing area involves neural emulators designed to mimic the complex computations found in cosmological simulations, such as those associated with feedback [Dai and Seljak, 2021]. SR might be particularly well-suited for this role due to its capabilities for generalization and interpretability, potentially offering advantages over other methods.

In galactic dynamics, SR could be utilized to derive approximate analytic expressions for phenomena like dynamical friction [François et al., 2024] or the computationally demanding aspects of stellar physics [Bianchini et al., 2016]. These efficient and comprehensible equations could substantially reduce the need for extensive N-body simulation grids. For instance, a mean field / test particle approach paired with an analytical emulator might be used during parameter searches to bypass frequent simulation runs, with a final validation step involving the full simulation to verify the accuracy of the parameters identified by the emulator.

Learning optimal N-body approximations

Continuing with the theme of enhancing simulation efficiency, N-body simulations, which traditionally approximate the N^2 interactions among all bodies, could benefit significantly from advanced learning methods. These simulations often simplify interactions through techniques that can be conceptualized as graph-based approximations.

There exists a promising opportunity to employ SR or its underlying graph optimization techniques directly to learn these optimal approximations. Inspired by the reinforcement learning methods discussed in Chapter 3, which exploit graph structures, we could develop a system to automatically identify and implement the most effective simplifications for N-body simulations.

This approach has the potential to not only refine the accuracy of N-body simulations but also reduce their computational demands.

Mapping the Milky Way

Investigating Near-Sun structures

In previous discussions, specifically in paragraphs 9.1.3 and 9.2.3, we explored perspectives related to the Typhon and Antaeus stellar streams. One particularly promising avenue is to extend the sampling of the Typhon stream. Although initially detected near the Sun, this stream is projected to extend up to the outer halo. Thus, its complete sampling could provide an exceptional probe for studying dark matter. Additionally, the discovery of such coherent dynamical structures near the Sun prompts further investigation into how these structures can remain non-phased mixed, given the short expected dynamical times near the Sun. This inquiry could lead to valuable insights, potentially through targeted simulations of such structures.

Exploiting 5D Samples

As noted in Section 10.5, another intriguing perspective involves the learning of de-projections for 5D samples that lack radial velocity information. By effectively learning the distribution function of the missing dimension, the hope being that the enhanced input of information will offset the additional degrees of freedom introduced by needing to learn this distribution.

Differentiable dynamical methods

As explored in Section 10.5, we discussed the potential to develop a differentiable model for stellar streams that could be integrated into mean field approaches. Additionally, we considered the feasibility of constructing a differentiable and deterministic action estimator that operates in canonical coordinates. This would involve utilizing classical approximations, which would be adapted to allow for differentiability.

11.2.1 Constraining dark matter

Constraining the dark matter particle mass

An interesting avenue would be to compute tractable and intelligible equations that encapsulate key properties of stellar streams along their length as a function of the dark matter particle mass. To achieve this, we could utilize multiple streams obtained from simulations conducted at varying particle mass levels [Carlberg et al., 2024]. By fitting these functional expressions

to actual stream observations, on can then extract the parameter values that encode the dark matter particle mass using the Class SR framework we designed (Chapter 5). The equations generated in this process would be designed by construction to capture the particle mass-dependent behavior, allowing us to harness the presence of gaps resulting from potential sub-halos in a statistically meaningful way. This statistical approach may enable us to establish connections between all available observations and multiple models, using an intelligible bottleneck in the form of an equation. Such an approach offers a plethora of opportunities for constraining dark matter properties as a function of observational constraints.

Learning dynamical profiles

Beyond the NFW Profile

We are currently applying the Φ-SO framework to develop alternatives to the empirical [Navarro, Frenk, and White, 1996] — NFW profile given in Eqn. 8.1, traditionally used to describe dark matter distributions within galaxies. Preliminary results on the NIHAO set of simulations [Wang et al., 2015] indicate that we have identified several profiles that are both simpler and more accurate than the NFW profile: "Pareto-dominating" it in terms of predictive power and complexity in both hydro-dynamical and dark matter only scenarios. Additionally, we are working at addressing a significant drawback of the NFW profile — its non-converging enclosed-mass at infinity — by incorporating this constraint in our reward function, ensuring physical validity in our newly discovered profiles.

Modeling Globular Cluster Rotation Curves

The standard empirical model currently used for globular cluster (GC) rotation curves fails to capture post-peak velocity trends and lacks temporal dynamics, which are crucial given the evolutionary nature of these clusters due to mass loss [Bianchini et al., 2018]. We are developing a new profile based on state-of-the-art simulations data of GCs[1] that not only models the velocity as a function of radius but also incorporates temporal changes. Initial successes include modeling the time-dependent evolution of peak velocities, offering a robust tool potentially useful for determining the ages of globular clusters from their dynamic profiles in the Milky Way.

[1]These N-body simulations, conducted by Paolo Bianchini at the Observatoire Astronomique de Strasbourg, represent the first to model globular clusters with a one-to-one correspondence of stars over a 13 Gyr timeframe while accounting for stellar evolution, tidal fields, and initial rotational dynamics, providing a comprehensive and realistic representation of globular cluster evolution.

Recovering a general distribution from extra-galactic streams.

Another exciting possibility is to build upon the *unsupervised* learning framework described in Chapter 10 to recover a "universal" dark matter distribution fitting multiple observational constraints related to low surface brightness structures surrounding distant galaxies [Nibauer et al., 2023, Sola et al., 2022]. Varghese et al. [2011] pioneered the exploitation of such features to constrain mass distributions. These constraints can be sourced from observational data collected through CFHT, Euclid [Laureijs et al., 2011], or the Roman Space Telescope. To accomplish this, we consider developing a differentiable framework designed to de-project the numerous structures and compute the free-form potential (only parameterized by a few galaxy-specific scale parameters) that can reproduce these structures in an *unsupervised* manner, fitting de-projection parameters in the process[2]. Despite the number of degrees of introduced, one can be hopeful that the extensive amount of observational constraints available will render this project both viable and informative. Note that although the velocity of extra-galactic tidal features is typically not measured, recent studies suggest if could be traced by their globular clusters for which 6D information is available [Ferrone et al., 2023].

We will see in the next-subsection, that one can also envision exploiting extra-galactic velocity maps for similar purposes.

11.2.2 Uncovering differential equations from data

In the the future, we plan to push the boundaries of symbolic learning methods by delving into the realm of automatic analytical differential equation generation as illustrated on Figure 11.1. The aim is to extend Φ-SO with the ability to generate differential equations whose solutions fit a given dataset or meet specific criteria, (e.g., additional physical principles, symmetries, or asymptotic conditions) extending operators to e.g., $\{\partial/\partial t, \partial/\partial x, \vec{\nabla}.\square, \vec{\nabla} \times \square ...\}$. The inclusion of differential equations in symbolic learning marks a transition away from empirical laws and toward more abstract yet interpretable constructs. Working in the space of differential equations is particularly relevant since they can express simple counterparts to solutions that can be very complex or might not even exist explicitly.

Generating differential equations inherently requires a sequential process, which is currently only possible within a few symbolic learning frameworks[3]

[2] This de-projection approach would be akin to the one we suggested for Milky Way stars, enabling cross-pollination in methodological approaches to this problem or even neural cross-pollinations in the form of transfer learning.

[3] A requirement for frameworks in which a neural network generates mathematical symbols directly such as ours.

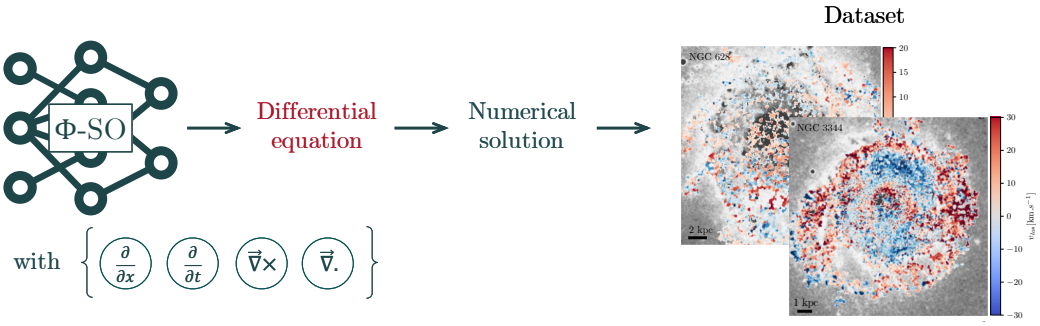


Figure 11.1: **Uncovering differential equations governing observations.** A planned extension of our Φ-SO framework involves the search for symbolic differential equations whose numerical solution fit a dataset. In the context of dark matter research, this could be used to learn alternatives or extensions to the Poisson equation governing dark matter and dynamics in galaxies based on e.g., their velocity maps. Here we illustrate our point by showing line-of-sight (los) velocity fields from two galaxies adapted from [Urrejola-Mora et al., 2022].

like Φ-SO. This sequential generation is crucial as it enables the integration of the necessary *in situ* priors required to handle multi-dimensional variables effectively and enforce a maximum nesting of differential operators for numerical stability.

Existing methods in this area often simply involve a regular SR framework in which one includes derivatives by treating them as additional variables, such as $\{x_1, \frac{\partial x_1}{\partial t}, x_2, \frac{\partial x_2}{\partial t}, t\}$ in lieu of $\{x_1, x_2, t\}$, rather than integrating them dynamically into the learning process. While our current implementation of Φ-SO already supports this straightforward possibility, the approach we propose goes beyond by allowing for the nesting of differential operators and the incorporation of vector operations such as scalar and cross products alongside a ∇ operator.

This methodological expansion is somewhat uncharted in machine learning, primarily because it aims to produce interpretable models in the form of differential equations which would have to be solved to be used, making them much more computationally intensive than typical machine learning objectives — which often prioritize predictive accuracy and computational efficiency over interpretability.

Learning alternatives to the Poisson equation governing dark matter

Our plans include utilizing this more advanced symbolic learning framework to delve into alternative formulations or higher-order expansions of the Poisson equation (given in Eqn. 8.4), which is central to understanding the dynamics

of dark matter within galaxies. This exploration will be based on observational velocity maps of galaxies.

This approach could be used as a systematic method to assess whether observational data from galaxies support the ΛCDM (Λ Cold Dark Matter) or MOND (Modified Newtonian Dynamics) frameworks as discussed in Section 8.1, or perhaps indicate the presence of entirely new physical phenomena. By combining this approach with our Class SR framework, detailed in Chapter 5, we aim to enable the learning of a 'universal' differential equation that encapsulates the dynamics of multiple galaxies at the same time as illustrated in Figure 11.1.

Learning alternative cosmologies

Other interesting applications include galaxy evolution, especially with the aid of JWST data [Gardner et al., 2006] as well as cosmological investigations. Specifically, in light of the Planck mission [Aghanim et al., 2020], we plan on exploring extensions to the Friedman equation by requiring the resulting analytical model to predict both the observed cosmic expansion through standard candles and the cosmic microwave background (CMB), which encodes crucial information about matter distribution in the early Universe.

11.2.3 Toward the automatic formulation of theories

What is a theory?

While the discovery of new extensions or alternatives to laws governing specific scales, such as the cosmological or galactic scales, would undoubtedly be of immense value, it would not equate to developing a comprehensive theory. Defining what constitutes a "theory" is a delicate epistemological question. Considering an observation-driven perspective, here we define a theory as a cohesive set of interrelated equations capable of accurately predicting natural phenomena across vastly different scales. For instance, Newton's laws were historically regarded as a first 'theory of everything' because they could predict both the trajectory of an apple falling from a tree and the movements of celestial bodies — phenomena that occur at drastically different scales. This concept underpins the challenges faced by quantum gravity theories [Rovelli, 2004] and, albeit on a less grand scale, the challenges we encounter in reconciling phenomena across galactic and cosmological scales.

Physical theory optimization

For this purpose, we aim to push the boundaries of the symbolic learning paradigm by developing a framework capable of formulating comprehensive

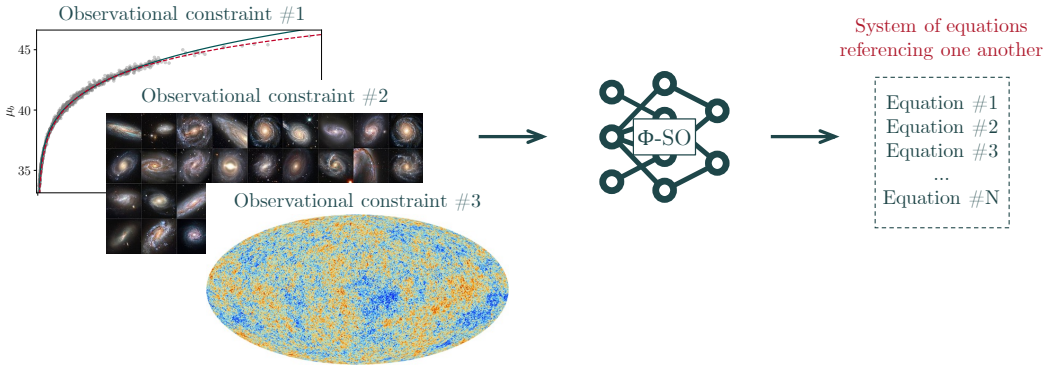


Figure 11.2: **Toward the automatic formulation of theories** Illustration of the planned extension of the Φ-SO framework allowing it to autonomously learn "theories" i.e. multiple analytical (possibly differential) equations referencing one another satisfying multiple observational constraints (each constraint possibly consisting of multiple realizations of a single phenomena). We illustrate our point by suggesting the automated exploration of "theories" fitting standard candles, a key galactic property across multiple realizations and the cosmological microwave background (CMB).

"theories". This involves creating an algorithm that can autonomously generate and refine systems of interrelated (possibly differential) equations to precisely match multiple observational constraints (each constraint possibly consisting of multiple realizations of a single phenomenon). This concept is visually represented in Figure 11.2.

In practice, we envisage the development of a framework that refines $N+n$ equations, where the first N equations are tailored to fit N distinct observational datasets-leveraging our Class framework optimized for multi-realization datasets (such as observations of distant galaxies). The additional n equations would function as auxiliary equations, potentially encoding redundancies that capture the core underlying principles of the theory. The objective would be to derive the simplest possible set of equations that collectively describe all observed phenomena effectively.

Applications

Given the complexity and the abstract nature of such a system, along with the anticipated challenges of ensuring its robust performance in real scientific applications, we propose an initial focus on the recovery of a "simple theory", such as Maxwell's laws of electromagnetism. Subsequently, the aim is to transition toward addressing substantial real-world astrophysical scenarios to maintain practical relevance. The specific applications may vary. However, one promising avenue involves applying this system to develop a predictive model able to predict behavior at both the galactic and cosmological scales as

illustrated in Figure 11.2.

11.2.4 Making large language models data and mathematics-literate

Current-generation LLMs

Throughout the duration of this PhD, we have observed the rapid development and adoption of Large Language Models (LLMs) that utilize Generative Pre-trained Transformer architectures (similar to those employed in pre-trained approaches to symbolic regression discussed in Section 2.2.2). While these models have achieved human-level or near-expert performance in various tasks, including language translation and programming language tasks, their proficiency in scientific domains remains relatively underdeveloped [Saxena et al., 2023].

Tokenization and learning challenges

One potential pathway to enhance the scientific capabilities of LLMs lies in optimizing how they process and learn from scientific content. Science fundamentally involves observing natural phenomena and formulating predictive models. Current LLMs struggle with scientific tasks partly due to their training approach, which heavily relies on supervised learning from vast corpora of text that are tokenized in a specific manner, as detailed in Section 3.1.2.

The tokenization process for mathematical equations it typically apllied directly to `Latex` strings. This method poses a significant limitation: if an LLM proposes the equation `b + a` but the correct format in the training data is `a + b`, the model is penalized despite the mathematical equivalence, due to its token-by-token learning approach. Additionally, LLMs must independently learn the syntax and structure of valid mathematical expressions, including the rules of parentheses and expression formatting, without the aid of specialized embeddings for mathematical constructs. This contrasts sharply with their language processing capabilities, where embeddings for words or sub-words significantly simplify the learning process by abstracting away the need to construct words from individual letters.

In essence, current LLMs lack direct access to the underlying graph structure of mathematical expressions; they only interact with their `Latex` representations. This limitation underscores a significant gap in their training: without embeddings analogous to those used for textual data, LLMs are tasked with a far more complex learning challenge when dealing with analytical expressions.

Multi-modality

Emerging advancements in multi-modality are now allowing LLMs to engage directly with diverse data forms, including images and audio, in addition to text. This development introduces a transformative capability where LLMs can process data through modality-specific inputs — such as an audio neural *head* capable of directly analyzing voice recordings[4].

Symbolic & data modalities

Given recent advances in deep learning SR, we suggest the introduction of specialized science-focused modalities into LLMs, as illustrated in Figure 11.3 This adaptation aims to enhance LLMs' ability to handle scientific tasks by incorporating domain-specific knowledge directly into their framework.

Potential platforms for incorporating these new modalities include the `AstroLlama` LLM [Nguyen et al., 2023], fined tuned for astrophysics applications, and `nanoGPT` [Karpathy, 2023], which offers a simplified architecture[5] built for prototyping. Additionally, the current state-of-the-art `Llama 3.1` [Llama Team, 2024] could provide a robust environment for deploying these complex, multi-modal learning strategies.

Symbolic modality

We suggest a specific "symbolic" modality that would enable LLMs to interpret and learn from the underlying graph structure of mathematical expressions, proofs, or even computer programs[6]. This specialized symbolic head could adopt a trial-and-error approach: instead of generating an expression in a single attempt, it could iteratively refine the expression, starting from an initial latent space representation — a method akin to diffusion processes [Ho et al., 2020] or Kamienny et al. [2023]'s approach in the context of SR. Implementing such a system would necessitate a robust symbolic expression graph management tool, which we have developed as part of the Φ-SO framework. Our system is uniquely capable in this regard, offering full graph representation and vectorization across both batch and equation length dimensions, positioning Φ-SO at the forefront of this exciting endeavor.

[4]This integration allows the models to respond to nuances in audio data, such as emotional intonations, tunes or contextual sounds which would be lost in translation through a traditional speech-to-text transcription process.

[5]Though offering performances on par with `GPT2` [Radford et al., 2019]

[6]Automated computer program generation, holds significant industrial interest and is likely to drive substantial innovation (see, e.g., , Lin et al. [2024]).

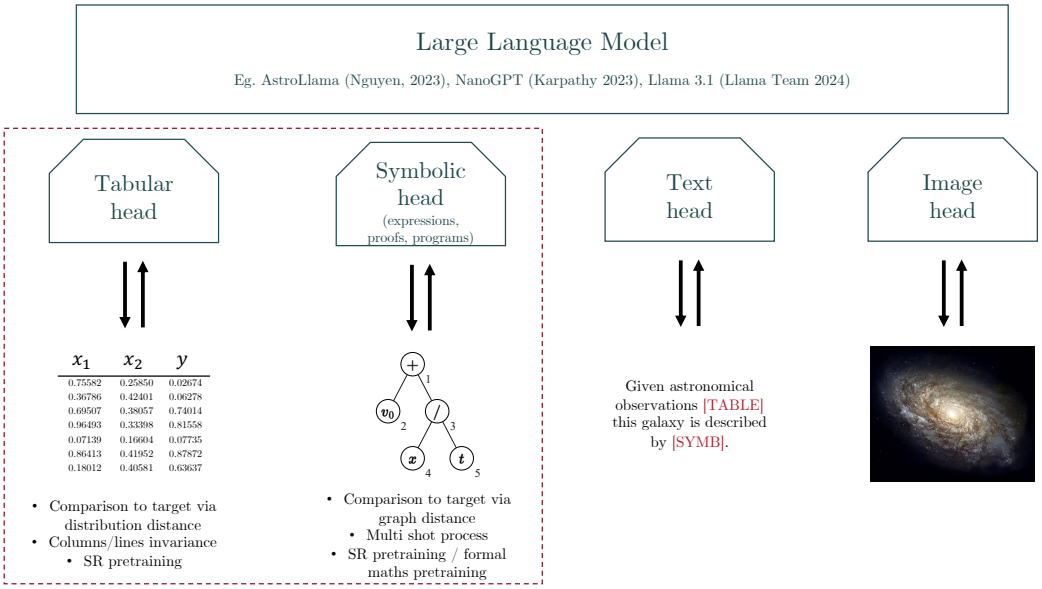


Figure 11.3: **Towards symbolic & data modalities for LLMs.** This diagram illustrates proposed enhancements (highlighted by the red contour) to Large Language Models (LLMs) that would enable them to process and learn on scientific data and mathematical expressions in addition to their existing capabilities with text and images. The extensions include specialized *heads* for ingesting and outputting symbolic expressions `[SYMB]` or tabular data `[TABLE]` in lieu of tokens, enhancing their literacy in data and formal mathematics. For a detailed discussion, refer to sub-section 11.2.4.

Graph distance metric

An additional advancement could involve developing a differentiable metric for measuring distances between equations. This metric would account for properties like commutativity (e.g., assigning zero distance between expressions like $a + b$ and $b + a$) and incorporating more complex algebraic identities to evaluate similarity. Implementing this would require a neural network, which, while not infallible, offers the speed and differentiability necessary for such a task. This approach aims to create a rapid and universally applicable symbolic expression distance metric, harnessing the capabilities of neural networks to achieve efficiency and scalability.

Tabular data modality

Current-generation LLMs often encode numerical values as text strings, such as e.g., $s.aaa.10^{bb}$, where s represents the sign (positive or negative), aaa are digits, and bb are the exponent digits, with each digit treated as a distinct class. Consequently, the LLM must empirically learn numerical closeness-for instance, that 42.1 is closer to 42.2 than to 92.1, given that all digits are being

treated as separate classes. Given pioneering efforts by Lalande et al. [2023] to integrate actual numerical values into transformer models rather than treating them as discrete tokens, there is potential to develop a robust method for incorporating tabular data directly into LLMs. This approach would handle data in a column and line invariant manner, differentiating it from image modalities as suggested by e.g., Kotelnikov et al. [2022].

SR pre-training

To enhance the LLM's capacity for integrating complex relationships between symbolic and data modalities, we propose pre-training the LLM's symbolic and data heads on SR tasks. This initial training phase would focus solely on modeling the relationship between symbolic expressions and data before the LLM is trained on an ensemble of multimodal datasets such as research papers, which often include text, numerical data, and mathematical expressions. Additionally, pre-training the symbolic head on formal mathematical problems could further refine its ability to handle complex symbolic information. This foundational training is expected to significantly boost the LLM's proficiency in scientific tasks where precise data interpretation and symbolic manipulation are crucial.

Scientific LLMs prospects

Throughout this thesis, and particularly within this section, we have incorporated a rich tapestry of prior knowledge into our symbolic frameworks, such as the inherent graph structure of mathematical expressions, prefix notation, and the integration of dimensional analysis. These elements-combined with the requirement to fit multiple realizations, and the application of Occam's razor to favor concise expressions have framed our current approach. Yet, the rapid evolution of LLMs suggests that many of the constraints we have meticulously encoded might soon become redundant, learned implicitly by more advanced models.[7]

Modern LLMs, for instance, no longer require explicit rules like prefix notation to generate balanced mathematical expressions. Errors like producing $a + +b$ instead of $a + b$ or producing $(a+)b.c$ instead of $(a + b).c$ are exceedingly rare, indicating a significant leap in their understanding of syntactic rules without direct programming. This natural proficiency raises a compelling question about the future capabilities of multi-purpose models: Could they, one

[7] This situation parallels the early resistance encountered by proponents of black box neural networks, who argued against manually coding rules with the help of experts. Instead, they advocated for systems that learn rules directly from data without explicit human intervention, a shift that marked a pivotal moment in the history of artificial intelligence [Schmidhuber, 2015].

day, perform tasks like SR directly out of the box? Imagine a scenario where an LLM, merely by processing a dataset, could autonomously generate an analytic expression reflecting on all previously encountered data/expressions occurrences, without specific training for SR tasks.

This prospect mirrors the unexpected abilities seen in early versions of `GPT` [Radford et al., 2019], where the model demonstrated a capacity for translating English to French — despite not being explicitly trained for translation — by applying its broad learning from English to the negligible French examples encountered during its training. This type of cross-application of learned knowledge hints at a future where advanced models not only meet but exceed their training directives, tackling complex and unanticipated tasks [8].

11.3 Concluding Remarks

In conclusion, unlike some other domains like computer vision, control, or computer science, I firmly believe that, physics and astrophysics necessitate not only traditional machine learning approaches but also an additional symbolic learning paradigm to advance effectively in the era of Big Data. Through this thesis we propose an ambitious framework and set of methodologies for extending the symbolic machine learning paradigm into the domain of physics. Our strategies draw from our experiences being confronted to concrete astrophysical challenges. The overarching statement of the present thesis being the establishment of a mutually beneficial relationship between the development of such approaches and the maximization of science returns from observational missions — and in particular the investigation of the dark matter problem, one of the most prominent challenge of physics.

The current landscape of machine learning is dominated by industrial applications that offer remarkable predictive capabilities but often fall short in terms of intelligibility and interpretability, aspects that hold paramount importance in the natural sciences. Given the current technological context, particularly regarding language processing, and the data abundance era we are entering in astrophysics, I firmly believe that the time is ripe to develop symbolic machine learning tools capable of producing comprehensible models in the form of analytical expressions.

[8] Beyond simple translations, `GPT4` has demonstrated proficiency in various complex tasks that it was not explicitly trained to perform, showcasing its generalization capabilities [de Wynter, 2024, Bubeck et al., 2023, Fan et al., 2022].

Obviously these methods hold the potential to benefit various branches of physics, but their particular relevance shines brightest in astrophysics, given the unprecedented influx of data in our field. Historically, astrophysics has frequently pioneered new numerical methods that later benefited the broader natural sciences. I believe that now more than ever, astrophysics has the responsibility to remain at the forefront of physical sciences by addressing these critical new challenges.

Formal Acknowledgments

Founding

We gratefully acknowledge the support from the École Doctorale Physique et Chimie-Physique (ED182) of the Université de Strasbourg, which selected the author and this PhD project for funding.

We acknowledge funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 834148).

HPC of the Université de Strasbourg

We acknowledge the High Performance Computing Center of the Université de Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

European Space Agency

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

European Southern Observatory

Part of the work presented here is based on observations collected at the European Southern Observatory under ESO programmes 105.20AL.001, 110.246A.001, and 111.2517.001.

Isaac Newton Telescope

We wish to thank Pablo Galán, Rosa Hoogenboom, Sara Vitali, Clár-Bríd Tohill, Sara Vitali and Paige Yarker for their invaluable help in supporting the INT observations reported here.

Detailed Summary (fr.)

Ce chapitre offre un résumé détaillé de la thèse, rédigé en français. Il débute par l'introduction et la contextualisation des méthodologies d'apprentissage automatique interprétables pour la physique et l'astrophysique. Ce résumé articule ensuite de manière succincte les contributions scientifiques réalisées au cours de cette thèse, accompagnées des perspectives futures détaillées.

Introduction

Les théories physiques, notamment en astrophysique, sont historiquement issues de lois empiriques. Les physiciens, observant les phénomènes naturels, élaborent des lois empiriques pour les décrire, puis construisent des théories englobantes qui intègrent ces lois. À titre d'exemple, la loi de la gravitation universelle de Newton rend compte avec élégance tant du mouvement des objets terrestres que des lois du mouvement planétaire de Kepler. Toutefois, l'avènement de l'apprentissage profond a transformé de nombreuses lois empiriques en modèles complexes basés sur des réseaux de neurones, rendant leur intégration dans des théories plus vastes nettement plus complexe.

Avec le lancement de nouvelles missions observationnelles telles que Gaia, Euclid, LSST et SKA, l'astrophysique se trouve à l'aube d'une ère prolifique en données, avec des volumes approchant le pétaoctet. Cette profusion de données suscite un vif intérêt pour l'identification de nouvelles lois empiriques qui pourraient, potentiellement, ouvrir la voie à des découvertes physiques inédites. Cependant, cette abondance de données impose également d'importants défis conceptuels. Si l'apprentissage profond permet d'extraire des informations précieuses de ces vastes corpus, il est néanmoins doublement influencé par les réseaux neuronaux qui constituent l'un de ses composants les plus efficaces, mais aussi les plus problématiques.

Le problème de l'opacité en apprentissage automatique

Les réseaux de neurones, tout en étant extrêmement flexibles et puissants, capables de modéliser presque tous les systèmes physiques et de fonctionner dans des espaces de haute dimension, demeurent pour la plupart des boîtes noires opaques. L'interprétabilité et la compréhensibilité, essentielles en physique, interrogent ainsi sur notre capacité à tirer parti des vastes ensembles de données tout en préservant la faculté d'interpréter et de relier ces informations à des théories substantielles. Peut-on, après avoir entraîné un réseau neuronal profond, dévoiler les mécanismes internes de cette boîte noire ? Peut-on en extraire et comprendre la physique sous-jacente ?

Un paradigme d'apprentissage symbolique

Face à ces défis uniques à la physique et à l'astrophysique, cette thèse propose le développement d'un nouveau paradigme en apprentissage automatique : un paradigme qui manipule les symboles mathématiques de manière non supervisée, un paradigme capable de transformer des réseaux neuronaux ou des ensembles de données en modèles physiques exprimés par des lois ana-

lytiques et symboliques concises. Cette approche novatrice, en enrichissant les méthodes conventionnelles d'une dimension d'interprétabilité précieuse, se présente comme une solution prometteuse pour relever le défi croissant de connecter de manière agnostique les observations à la théorie.

La nécessité des approches symboliques

Prédiction vs. explication : le dilemme épistémologique posé par les réseaux de neurones

Puissance prédictive par l'unification. Cette question épistémologique, soulevée par l'avènement des réseaux neuronaux, se situe au cœur même de la recherche physique. Devons-nous nous satisfaire de modèles qui, bien que précis dans leurs prédictions, ne nous offrent aucune lumière sur les processus sous-jacents ? Imaginez, à titre hypothétique, un réseau neuronal entièrement opaque, certes omniscient, capable de prédire avec une exactitude irréprochable tout phénomène physique. Un tel outil pourrait-il vraiment assouvir notre soif de connaissance scientifique ? Probablement pas, car le désir intrinsèque de comprendre, cette quête d'explications qui anime chaque physicien, demeurerait inassouvi. Ce scénario nous amène à une interrogation fondamentale : quelle est la vocation ultime de la physique ? Est-elle simplement de prédire ou aussi d'expliquer ? Et, le cas échéant, quel aspect définit le plus profondément la discipline ?

Puissance prédictive à travers la complexité. Historiquement, les avancées significatives en physique ont souvent résulté de l'unification de théories, à la fois simples et puissantes, aptes à expliquer et à prédire des phénomènes à différentes échelles. Newton et ses lois en sont un exemple éloquent. Cette propension pour la simplicité et l'élégance, fréquemment résumée par le principe du rasoir d'Occam, suggère une préférence pour des théories moins paramétrées mais dotées d'une capacité explicative exhaustive.

Les mathématiques : le langage de l'unification. À l'opposé, les modèles basés sur les réseaux de neurones marquent une rupture de paradigme. Ils excèlent dans la prédiction au sein de leur domaine d'entraînement mais se caractérisent souvent par une densité paramétrique élevée et une absence de la simplicité explicative propre aux modèles analytiques. On pourrait arguer que la puissance prédictive, à elle seule, justifie un écart par rapport au principe d'Occam — si un modèle peut élucider des phénomènes jusqu'alors inexpliqués, ne peut-on pas alors excuser sa complexité ?

Néanmoins, nous soutenons que cette question ne se pose pas encore, car aucun modèle actuel d'apprentissage profond n'est capable d'apprendre et

de prédire universellement tous les phénomènes physiques. Nous assistons plutôt à une fragmentation des modèles à travers diverses sous-disciplines de la physique, chaque modèle étant spécifiquement adapté à certains ensembles de données ou phénomènes. Il est concevable que des indices de nouvelles physiques soient déjà présents, dissimulés au sein de l'un ou plusieurs de ces réseaux spécialisés, formés sur d'immenses corpus de données observationnelles ou expérimentales.

La méthode traditionnelle de synthèse des observations empiriques en théories globales s'est toujours effectuée à travers le langage universel des mathématiques. Cette tradition nous enseigne que, malgré l'exploitation des capacités des réseaux de neurones, il demeure un impératif critique pour des modèles mathématiques interprétables. Ces derniers sont indispensables pour faciliter la communication des concepts physiques entre les divers domaines de la physique.

Constructions mathématiques en physique

Galilée, dans son œuvre Opere Il Saggiatore, a perceptivement observé que le livre de l'Univers est "écrit en langue mathématique". Depuis lors, l'une des préoccupations centrales de la physique a été de tenter d'expliquer les propriétés de la nature en termes mathématiques, en proposant ou en dérivant des expressions mathématiques qui encapsulent nos observations et expériences. Cette démarche s'est avérée d'une efficacité remarquable. Au fil des siècles, grâce à une méthode d'essai et d'erreur, les grands maîtres de la physique ont développé et légué une panoplie de techniques nous permettant de déchiffrer le monde et de bâtir notre civilisation technologique moderne. Aujourd'hui, avec l'évolution des réseaux neuronaux, se profile l'espoir d'une accélération de cette entreprise, exploitant le fait que les machines peuvent explorer un espace de solutions bien plus vaste que ne le pourrait un être humain seul.

Régression Symbolique

Cette perspective souligne le rôle crucial de la "Régression Symbolique" dans cette thèse. Au-delà des méthodes traditionnelles émergeant depuis l'avènement de l'informatique, qui impliquent généralement l'ajustement de coefficients à des fonctions linéaires ou non linéaires prédéfinies, la régression symbolique ambitionne davantage. Elle aspire non seulement à optimiser les coefficients au sein d'une fonction mathématique donnée mais aussi à découvrir les formes fonctionnelles elles-mêmes. Plus précisément, elle vise à déduire une fonction analytique symbolique libre $f : \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_2}$ qui ajuste $\mathbf{y} = f(\mathbf{x})$ à partir de données $(\mathbf{x}, \mathbf{y})$.

Une philosophie de modélisation axée sur l'observation

Le problème de biais en apprentissage automatique

Les méthodologies traditionnelles d'apprentissage automatique en physique et en astrophysique reposent fréquemment sur l'entraînement supervisé de réseaux de neurones, où d'importantes hypothèses physiques sont intégrées de manière prépondérante dans les exemples d'entraînement, notamment par le biais de simulations fondées sur des modèles physiques établis. Si de telles démarches peuvent s'avérer utiles dans certains contextes, elles confinent inéluctablement la découverte de nouvelles physiques en alignant rigidement les modèles résultants sur les paradigmes théoriques préexistants.

L'agnosticisme nécessaire à la découverte scientifique

Cette thèse aspire à inaugurer des méthodologies novatrices pour la découverte scientifique en physique et astrophysique, en prônant une philosophie résolument guidée par l'observation. Cette philosophie, qui imprègne l'ensemble du travail présenté, postule que les véritables découvertes physiques ne sauraient émerger de la simple adhésion à des modèles pré-établis, mais requièrent plutôt une exploitation éclairée et agnostique des données observationnelles.

Nous déployons ainsi des paradigmes novateurs qui s'affranchissent de l'apprentissage dépendant des modèles au bénéfice de stratégies d'apprentissage non supervisées. Ces dernières ne s'appuient pas sur des modèles physiques prédéfinis, mais aspirent plutôt à édifier des modèles physiques intrinsèquement fidèles aux contraintes imposées par les observations. Le paradigme Φ-SO Optimisation Symbolique Physique, établi au travers de cette thèse, illustre parfaitement cette approche. Φ-SO s'attache à formuler des expressions analytiques symboliques ex nihilo, par le biais d'un processus itératif d'essais et erreurs, strictement encadré par l'impératif de conformité aux données empiriques, sans aucune exposition préalable à des expressions symboliques. Cette méthodologie incarne notre engagement à révéler des modèles physiques par le prisme de contraintes comportementales strictes, affranchies de toute présupposition modélisatrice : une démarche véritablement libre de tout biais inductif.

En épousant la philosophie formulée par Donald Lynden-Bell, nous nous engageons résolument à "suivre les données", permettant ainsi aux motifs intrinsèques et aux vérités contenues dans les observations de guider nos avancées théoriques. Cette approche ne stimule pas seulement le potentiel de découvertes fondamentales mais s'harmonise également avec la mission centrale de la physique : élucider les principes régissant l'univers à travers le prisme de la preuve empirique.

Le problème de la matière noire

La quête de progrès méthodologiques, bien que porteuse d'innovations, peut aisément nous entraîner vers l'abstraction. Il est donc essentiel d'ancrer ces avancées dans des défis scientifiques tangibles. Le mystère de la matière noire illustre parfaitement ce type de défi, son comportement énigmatique à l'échelle galactique laissant présager des manques dans notre compréhension et peut-être des pistes vers de nouvelles physiques, offrant ainsi un terreau fertile pour l'éprouve de nouvelles théories et méthodologies.

L'ambition ultime de cette thèse est de cultiver une relation symbiotique entre le développement de stratégies d'apprentissage automatique innovantes et d'approches d'apprentissage symbolique, et leur application en astrophysique, notamment dans la démystification de la matière noire, l'un des défis les plus ardus de la physique contemporaine.

Plan & Objectifs

Nos objectifs se déclinent en deux axes : primo, étendre les frontières de l'apprentissage automatique symbolique au-delà de ses domaines traditionnels, qui privilégient souvent les communautés informatiques et de contrôle, pour explorer de nouveaux terrains prometteurs qui confèrent à la physique une valeur ajoutée par l'interprétabilité. Secundo, mettre en œuvre ces méthodes avant-gardistes pour relever le défi actuel posé par la matière noire, en veillant scrupuleusement à ce que l'évolution de ces méthodologies demeure solidement ancrée dans des études de cas scientifiques concrets.

Ici, nous proposons un résumé de la thèse. À titre indicatif et pour faciliter la navigation, voici une esquisse de sa structure, détaillant chaque chapitre.

Le chapitre 2 examine une diversité d'approches interprétables en apprentissage automatique adaptées à la physique et à l'astrophysique, et expose des stratégies fondamentales issues de la littérature susceptibles de catalyser des avancées dans ces disciplines. Il met en exergue l'importance cruciale de l'apprentissage symbolique dans ce cadre et fournit une contextualisation approfondie de ces techniques.

Le chapitre 3 s'attarde sur la conceptualisation des problématiques mathématiques comme des problèmes numériques d'optimisation de graphes et aborde la représentation des mathématiques formelles en tant que données numériques pouvant être traitées numériquement. Ce segment présente notre méthode destinée à entraîner des réseaux de neurones à élaborer des expressions mathématiques répondant à des contraintes spécifiques, telles que l'ajustement à un jeu de données, il s'agit de la problématique de la régression symbolique, par une méthode d'essais et d'erreurs utilisant l'apprentissage pro-

fond par renforcement.

Dans le chapitre 4, nous proposons une technique d'intégration des contraintes d'analyse dimensionnelle physique à l'optimisation symbolique, que nous combinons par la suite à notre stratégie d'apprentissage par renforcement, aboutissant à notre paradigme Φ-SO qui affiche des performances de premier plan évalué sur un benchmark de régression symbolique standardisé.

Le chapitre 5 développe le paradigme Φ-SO afin de permettre la recherche d'une forme fonctionnelle unique qui s'adapte à plusieurs réalisations d'une classe spécifique de phénomènes, chaque réalisation pouvant comporter des valeurs paramétriques potentiellement distinctes. Nous baptisons cette approche Régression Symbolique de Classe (Class SR). L'efficacité de cette nouvelle méthode est démontrée par la mise en place et l'exécution d'un premier benchmark pour Class SR, ainsi que par l'optimisation d'un potentiel gravitationnel galactique synthétique analytique à partir de données de courants stellaires correspondantes.

Le chapitre 6 détaille le logiciel `PhySO`, qui constitue notre implémentation du paradigme Φ-SO.

Le chapitre 7 expose des méthodologies complémentaires à Φ-SO. Ces techniques utilisent des réseaux de neurones pour représenter directement la structure hiérarchique sous-jacente des expressions analytiques, augmentant ainsi la portée et l'efficacité de nos approches d'apprentissage symbolique.

Le chapitre 8 aborde les défis liés à la compréhension de la matière noire à l'échelle galactique, en se concentrant spécifiquement sur la Voie Lactée. Ce chapitre introduis ces problématiques pour une étude plus poussée du rôle et des caractéristiques de la matière noire dans notre galaxie.

Le chapitre 9 précise nos contributions à l'investigation de sondes observationnelles de la matière noire dans la Voie Lactée, soulignant la découverte et l'analyse de nouveaux courants stellaires. Nous présentons notamment un courant nouvellement identifié, que nous baptisons Typhon.

Le chapitre 10 présente une méthode innovante pour cartographier la distribution de la matière noire dans la Voie Lactée à partir de coordonnées stellaires. Cette méthode s'inscrit dans notre philosophie agnostique et guidée par l'observation, faisant appel à des techniques d'apprentissage non supervisé.

Le chapitre 11 clôt cette thèse en synthétisant nos découvertes et en présentant nos perspectives. Il accentue l'importance des méthodologies prospectives destinées à dévoiler de nouvelles contraintes sur la matière noire et envisage des avancées futures dans le domaine de l'apprentissage symbolique.

Approches Interprétables en Physique et en Astrophysique

Les Découvertes Scientifiques à l'Ère du Machine Learning

Le paradigme de l'apprentissage supervisé

L'apprentissage profond, ou le processus de calibrage d'un réseau neuronal profond, a considérablement évolué depuis sa revitalisation par LeCun et al. [1998]. Réanimé initialement dans le secteur de la recherche en ingénierie, il s'est rapidement diffusé dans les applications industrielles[9], grâce à sa capacité à modéliser ou simuler presque tous les systèmes, supplantant effectivement des champs entiers comme le traitement des signaux et de l'image [Schmidhuber, 2015]. L'apprentissage profond marque un tournant non seulement en termes de capacités mais aussi de méthodologie, passant de règles établies par des experts à un apprentissage empirique directement à partir des données. Cela soulève une question fondamentale : un jeu de données peut-il lui-même être considéré comme un modèle direct ?[10]

Un domaine axé sur l'ingénierie ?

Le besoin d'une inférence rapide et précise. L'orientation du machine learning centrée sur l'ingénierie reflète un ensemble de priorités distinctes : bien que l'interprétabilité soit souvent reléguée au second plan, l'importance d'une inférence rapide et précise est cruciale. Cette orientation contraste fortement avec les exigences des sciences naturelles, où la compréhension et l'interprétabilité sont essentielles, et bien que l'inférence rapide soit avantageuse, elle n'est pas toujours primordiale. Étant donné que chaque découverte scientifique est unique, une fois une percée réalisée, le besoin d'inférences répétées s'amoindrit.

Paradigme d'apprentissage supervisé. L'accent mis par l'ingénierie sur la précision et la rapidité a favorisé le paradigme de l'apprentissage su-

[9] Ce qui est souligné par le fait que, bien que open-source, des plateformes majeures d'apprentissage profond comme `TensorFlow` [Abadi et al., 2016] et `JAX` [Bradbury et al., 2018] sont développées par de grandes corporations telles que Google, tandis que `PyTorch` [Paszke et al., 2019] est géré par Meta.

[10] Cette idée fait écho à la définition des langues humaines, qui sont souvent appréhendées à travers des corpus plutôt que des règles préétablies [Hunston, 2006].

pervisé, où les réseaux de neurones sont formés sur des exemples d'entrée-sortie appariés. Pendant l'inférence, les paramètres du modèle sont "figés" (i.e. fixés), permettant au réseau de prédire les résultats reflétant son processus d'apprentissage. Cette méthode est généralement efficace dans la gamme de ses données d'entraînement en raison de la flexibilité inhérente aux réseaux de neurones. Cependant, cette approche restreint intrinsèquement la portée des découvertes en physique et en astrophysique, où l'objectif va au-delà de la simple prédiction pour inclure la compréhension de nouveaux processus fondamentaux.

Les limitations de l'apprentissage supervisé pour la découverte de nouvelles physiques

Nous n'avons accès qu'à un seul Univers.. La contrainte fondamentale de l'application de l'apprentissage supervisé pour découvrir de nouvelles physiques réside dans notre unique jeu de données d'observation : l'Univers lui-même. Contrairement à d'autres domaines où des données provenant de sources variées peuvent servir à former et valider des modèles, la physique doit faire face à la tâche de dériver des lois universelles à partir d'observations limitées à une seule instance. Cette situation unique limite l'utilité de l'apprentissage supervisé, qui repose traditionnellement sur des ensembles de données divers pour généraliser et prédire des résultats dans des contextes non familiers. Si nous avions accès à plusieurs univers, chacun régi par des lois physiques différentes, l'apprentissage supervisé pourrait potentiellement "trianguler" des lois physiques applicables à un univers jusque-là inconnu.[11]

Approches fallacieuses. Bien qu'il soit possible d'entraîner des réseaux de neurones sur des simulations intégrant certaines hypothèses physiques, le véritable test se présente lorsque ces modèles sont appliqués à de véritables données observationnelles. Idéalement, les données simulées devraient imiter étroitement les données observationnelles pour assurer que le modèle fonctionne dans les paramètres pour lesquels il a été entraîné. Cependant, cette méthode présuppose que les lois physiques intégrées dans la simulation reflètent fidèlement la réalité. Il y a un risque que des chercheurs employant ce type d'approches puissent involontairement confirmer les hypothèses intégrées dans la simulation lors de l'application de ces modèles aux données réelles, confondant l'écho de leurs suppositions pour une découverte. Ceci met en lumière un piège critique de l'utilisation de l'apprentissage supervisé, où le modèle n'est aussi bon que les hypothèses de ses données d'entraînement et pourrait ne pas véritablement s'étendre à la découverte de nouveaux principes dans les données

[11] C'est pourquoi les approches bayésiennes de la probabilité sont souvent privilégiées par rapport aux approches fréquentistes en physique et en astrophysique, où la répétition expérimentale à l'échelle universelle est impossible.

observationnelles.

Approches raisonnables d'apprentissage supervisé en (astro)-physique

Traitement de vastes ensembles de données Malgré ses limitations pour découvrir de nouvelles lois physiques, l'apprentissage supervisé reste un outil précieux dans l'analyse préliminaire de vastes ensembles de données en astrophysique.

Prenons pour exemple l'analyse des spectres stellaires[12]. Dans ce contexte, les modèles d'apprentissage supervisé sont aptes à déduire des caractéristiques stellaires essentielles, telles que la métallicité[13] ou la gravité de surface, à partir des données spectrales. Ces modèles exploitent de larges ensembles de données bien caractérisés où les propriétés des étoiles sont bien comprises et cohérentes. En entraînant des réseaux de neurones sur ces ensembles de données, les chercheurs peuvent automatiser l'analyse des spectres stellaires, standardisant ainsi efficacement cet aspect de la recherche astrophysique, comme démontré par l'approche utilisée dans le catalogue `APOGEE` [Holtzman et al., 2018].

Approches utilitaristes : Quand la fin justifie les moyens

Dans certains contextes, le processus de génération du modèle devient secondaire face à l'utilité et la précision de celui-ci. Cela est particulièrement pertinent dans des cas où le modèle final peut être indépendamment vérifié et testé, sans égard pour son origine. Dans de telles circonstances, la méthode de découverte, qu'elle soit conventionnelle ou via une boîte noire produisant la solution, est subordonnée à la validité et à l'applicabilité du modèle.

La régression symbolique (SR) incarne parfaitement cette approche. Utilisant des techniques d'apprentissage automatique, la SR génère un modèle physique sous forme d'expression analytique qui s'adapte aux données observationnelles. L'atout majeur ici est que le résultat, les expressions analytiques, est intrinsèquement interprétable et vérifiable, se distinguant nettement de la méthode computationnelle employée pour le trouver.

Un autre domaine d'application est la résolution de conjectures mathématiques par la génération de preuves mathématiques formelles. L'accent est ici mis sur l'efficacité de la solution apportée, plutôt que sur les mécanismes du réseau neuronal qui l'a générée. Si un réseau neuronal, même un modèle de type boîte noire, peut proposer une preuve valide pour

[12] C'est notamment l'analyse des spectres stellaires et la détection des premiers motifs indiquant l'évolution stellaire qui ont marqué la naissance de l'astrophysique à partir de l'astronomie.

[13] En astrophysique, la métallicité désigne la proportion de masse d'une étoile qui n'est ni hydrogène ni hélium, souvent mesurée relativement au contenu métallique du Soleil.

un problème mathématique, cette preuve peut être examinée et validée indépendamment de la méthode de découverte.

L'optimisation des paramètres en astrophysique implique fréquemment le réglage des paramètres de simulation afin que les résultats de celle-ci concordent avec les données observées. Alors que les approches traditionnelles telles que la Chaîne de Markov Monte Carlo (MCMC)[14] prévalent en raison de leur robustesse pour l'estimation des incertitudes, les réseaux de neurones présentent une alternative directe et potentiellement plus rapide. En entraînant des réseaux sur des paires de sorties de simulation et de paramètres, il est possible de prédire les paramètres qui engendreront un résultat désiré, lesquels peuvent ensuite être vérifiés par un unique cycle de simulation. Cette méthode offre un accès direct à la validation de la solution, bien qu'elle ne fournisse souvent pas les estimations d'incertitude caractéristiques des méthodes telles que MCMC. Nous aborderons ultérieurement dans cette section des stratégies palliant cette limitation.

De manière plus globale, l'usage de réseaux de neurones pour émuler des simulations complexes devient une pratique de plus en plus courante. En capturant efficacement la dynamique des simulations, les réseaux de neurones peuvent proposer des alternatives plus rapides à l'exécution de modèles coûteux en calcul.

Émulation neuronale de simulations

Recherches de paramètres

Optimisation de paramètres via l'inférence basée sur la simulation. L'émulation neuronale de simulations, souvent désignée sous le terme d'"inférence basée sur la simulation" (SBI), constitue un outil puissant pour l'accélération de la recherche de paramètres [Cranmer et al., 2020a]. Les réseaux neuronaux, même de grande taille, étant nettement plus rapides à évaluer que les simulations complexes en astrophysique ou en physique, ils s'avèrent particulièrement bénéfiques pour les problèmes de recherche paramétrique de haute dimension où les approches en grille traditionnelles se révèlent immensément coûteuses. Ces méthodologies permettent des contrôles ponctuels avec des paramètres choisis aléatoirement, garantissant ainsi la précision de l'émulateur sans nécessiter des simulations exhaustives sur une vaste étendue de paramètres.

Application de SBI en cosmologie. Par exemple, des études en cosmologie ont prouvé que ces approches peuvent surpasser les techniques de

[14] Cette méthode consiste à construire une chaîne de Markov dans l'espace de recherche des paramètres, la distribution de cette chaîne représentant la distribution sous-jacente étudiée.

Monte Carlo par chaînes de Markov (MCMC) pour récupérer avec précision les distributions postérieures[15] des paramètres [Zhao et al., 2022].

Cartographie des distributions postérieures avec les modèles de flux normalisés (NF). De plus, en menant une série de simulations et en utilisant des estimateurs de distribution de probabilité neuronaux (tels que les modèles de flux normalisants (NF) qui sont examinés dans la Section 10.2), les chercheurs peuvent estimer approximativement la distribution postérieure complète. Cette méthode élimine le besoin de sonder une pléthore de paramètres au sein des simulations, simplifiant considérablement le processus de recherche.

Les vertus de l'émulation neurale

Au-delà de l'accélération. Les émulateurs neuronaux, tout en accélérant considérablement les processus de calcul, offrent des avantages qui dépassent la simple rapidité. Ils présentent des atouts distinctifs qui peuvent les rendre essentiels, même dans des contextes où les simulations pourraient être réalisées instantanément.

Aborder des problèmes inverses L'un des principaux atouts des émulateurs neuronaux est leur aptitude à résoudre des problèmes inverses. En inversant le processus d'entraînement du réseau de neurones, le conditionnant à prédire les paramètres d'entrée θ à partir d'un résultat de simulation et non l'inverse, nous simplifions la recherche des conditions initiales ou des paramètres expliquant les phénomènes observés. Cette méthodologie est illustrée dans la Figure 12.1.

Un émulateur différentiable Un autre avantage crucial est la différentiabilité des émulateurs neuronaux. Grâce à l'auto-différenciation (approfondie dans 2.1.4), il est possible de dériver des gradients à travers les réseaux de neurones, facilitant l'utilisation de la descente de gradient pour optimiser les paramètres de simulation directement en fonction des résultats souhaités, typiquement pour correspondre aux observations.

Bien que les techniques d'apprentissage supervisé fournissent des outils précieux pour l'investigation scientifique, elles ne sont pas conçues pour découvrir de nouvelles lois physiques. Ces approches, centrées sur l'ingénierie, sont naturellement limitées dans le contexte de la physique et

[15] Dans le cadre de la statistique bayésienne, la probabilité d'une valeur de paramètre est évaluée par sa probabilité postérieure, qui intègre des connaissances antérieures via la distribution a priori et inclut un terme de marginalisation qui prend en compte toutes les autres variables [Bayes, 1763].

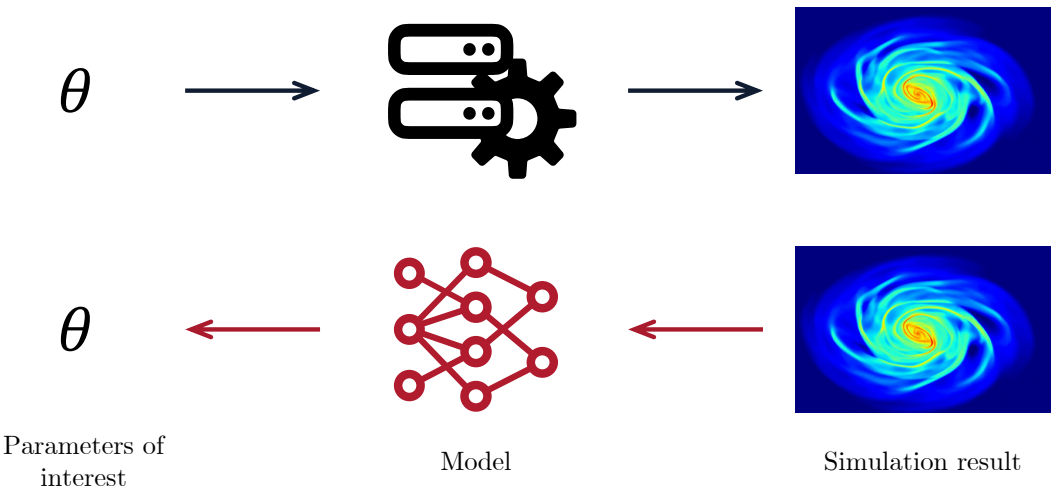


Figure 12.1: **Émulateur neuronal pour aborder les problèmes inverses.** Un émulateur neuronal formé peut renverser le processus de simulation habituel. Au lieu de générer des résultats à partir de paramètres, il infère les paramètres θ qui mèneraient à un résultat de simulation spécifique. Cette capacité facilite la résolution de problèmes inverses en prédisant les conditions initiales ou les paramètres qui correspondent à des résultats arbitraires.

de l'astrophysique. En tant que physiciens, il est essentiel de mettre en œuvre ces technologies de manière judicieuse, en transcendant les paradigmes traditionnels pour exploiter au maximum le potentiel de l'apprentissage profond.

Approches agnostiques

Cadre et exemples

Cadre Un éloignement des paradigmes d'entraînement conventionnels nous incite à envisager des méthodes où les réseaux de neurones ne se limitent pas à apprendres à partir d'exemples d'entraînement aux issues connues. Dans ce cadre, ils sont sollicités pour effectuer des prédictions tout en respectant un ensemble de contraintes, généralement d'ordre physique ou observationnel. Ce procédé, désigné sous le terme d'"apprentissage non supervisé", engage l'entraînement des réseaux au moyen d'essais et d'erreurs sans issues prédéterminées.

Exemples Un exemple éminent de cette méthodologie est notre démarche de régression symbolique (SR), où le réseau génère une expression analytique. L'impératif ici est que l'expression corresponde avec exactitude aux données observationnelles, sans que le réseau n'ait préalablement été exposé à des exemples d'expressions symboliques.

L'apprentissage non supervisé trouve aussi son application dans le clustering, qui consiste à identifier des groupes d'éléments similaires au sein d'un jeu de données. À titre d'illustration, Dodd et al. [2023] a mis en œuvre des techniques de clustering pour distinguer les structures de la Voie Lactée à partir d'un vaste ensemble de données sur les positions et vitesses stellaires à proximité du Soleil.

En astrophysique, l'algorithme `ActionFinder` établi par Ibata et al. [2021] illustre également cette approche. Cet algorithme apprend une transformation canonique (et son Hamiltonien sous-jacent) vers l'espace des actions, essentiellement les orbites, de manière non supervisée, en assurant que les étoiles d'un même courant stellaire[16] présentent des valeurs similaires dans l'espace latent. Cette réalisation se fait sans aucun exemple préalable ni dépendance à un modèle dynamique physique, l'unique postulat étant que les étoiles d'un même courant stellaire suivent approximativement la même orbite.

Agnosticité. Dans le domaine des sciences naturelles, surtout lorsque de nouveaux modèles physiques sont à l'étude, l'agnosticité se révèle essentielle. Les démarches non supervisées assurent que l'apprentissage est exempt de biais induits par des théories ou des simulations préexistantes, pavant ainsi la voie à de réelles avancées physiques. L'apprentissage non supervisé constitue la seule approche viable exempte des biais habituellement introduits par un entraînement sur des résultats connus.

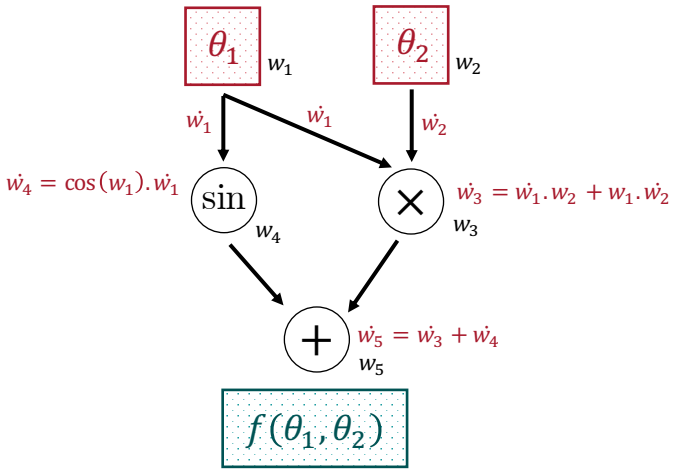
Sur la puissance de l'auto-différenciation

L'auto-différenciation est un outil précieux, quoique souvent sous-estimé, introduit par l'apprentissage profond. Examinons son principe fondamental et son utilité.

Approximation par surparamétrisation. Il pourrait être tentant de penser que la performance de l'apprentissage profond découle simplement de l'abondance de paramètres au sein des réseaux neuronaux, leur permettant de modéliser finement une vaste gamme de fonctions, semblablement aux séries de Taylor. Cette faculté est formellement validée par le théorème d'approximation universelle, selon lequel un perceptron multicouche (MLP) de taille adéquate peut approximer n'importe quelle fonction intégrable au sens de Lebesgue [Hornik et al., 1989].

Pour exemple, comparons des modèles physiques de complexité croissante : un modèle de chute libre simpliste supposant une absence de résistance

[16] Ces structures étendues et fines se forment lorsque des corps célestes sont accretés par la Voie Lactée. Nous approfondissons ce concept dans les discussions contextuelles du Chapitre 9.



(a) Computational graph (b) Pytorch syntax

Figure 12.2: **Illustration de l'auto-différenciation.** Pour un ensemble de paramètres $\theta = \theta_1, ..., \theta_n$, les dérivées de chaque étape de calcul sont mémorisées, permettant l'application du théorème de la dérivation des fonctions composées pour calculer les dérivées $\frac{\partial x}{\partial \theta_1}, ..., \frac{\partial x}{\partial \theta_n}$ relativement à θ pour toute variable x. Cette figure illustre un graphe de calcul (a) et une syntaxe typique dans le cadre de `PyTorch` [Paszke et al., 2019] pour une opération simple : $\sin(\theta_1) + \theta_1\theta_2$.

atmosphérique, $z(t) = -\frac{1}{2}gt^2 + v_0t + z_0$, avec trois paramètres, est moins précis qu'un modèle incluant une pression atmosphérique uniforme, $z(t) = H \ln \frac{1+e^{-2t/T}}{2} + v_0t + z_0$, avec six paramètres[17], qui est à son tour moins précis qu'un réseau de neurones pouvant impliquer des milliers de paramètres.

Une observation notable dans la recherche actuelle sur l'apprentissage profond est que les réseaux de neurones surparamétrés fonctionnent souvent exceptionnellement bien sans surajustement, à condition d'être correctement entraînés, y compris l'utilisation d'un ensemble de test distinct que le réseau n'a jamais rencontré durant l'entraînement [Li and Liang, 2018].

L'auto-différenciation : le secret de l'apprentissage profond

Bien que le nombre élevé de paramètres contribue sans aucun doute au succès de l'apprentissage profond, un autre facteur déterminant est la rétropropagation. Ce processus consiste à suivre chaque opération mathématique durant l'inférence, potentiellement des millions, et à enregistrer sa dérivée dans un graphe de calcul. Cela permet la différentiation

[17] Où z, t, g, v_0, z_0 représentent respectivement l'altitude, le temps, la gravité terrestre, la vitesse initiale, l'altitude initiale et H et T désignent une hauteur d'échelle et un temps caractéristique, respectivement.

automatique et analytique de la fonction de coût par rapport aux paramètres ajustables via le théorème des dérivées composées, facilitant grandement la convergence. Sans cette capacité, bien qu'il serait théoriquement possible de régler les paramètres permettant aux réseaux de neurones d'émuler n'importe quelle fonction, il serait pratiquement irréalisable de les trouver, c'est-à-dire de former des réseaux à travers des couches profondes et complexes. L'auto-différenciation rend possible la propagation des dérivées de la fonction de coût à travers même des réseaux neuronaux très profonds. Cette technique fondamentale est illustrée dans la Figure 12.2.

Simulations différentiables

L'implémentation de simulations entières dans un cadre auto-différentiable, où chaque opération est différentiable ou peut être approximée comme telle, ouvre des horizons extraordinaires. Par exemple, Li et al. [2022] a développé une simulation cosmologique entièrement dans ce cadre, leur permettant d'optimiser les conditions initiales pour répondre à des critères observationnels spécifiques. Cette capacité à "rétropropager" à travers une simulation pour ajuster les conditions initiales ou toute autre variable illustre profondément l'impact de l'auto-différenciation, initialement promue par l'apprentissage profond mais fondamentalement indépendante de celui-ci. Une telle flexibilité signifie que l'on pourrait théoriquement optimiser les conditions initiales d'un univers simulé pour correspondre à tout résultat observationnel souhaité, démontrant l'utilité puissante de cette approche.[18]

Techniques d'apprentissage profond

Apprentissage non supervisé : Nous avons exploré des configurations d'apprentissage non supervisé, qui engagent la formation de réseaux de neurones basés sur toute contrainte différentiable. Ces contraintes peuvent être complexes, s'étendant aux calculs dans des simulations physiques (comme nous le faisons nous-mêmes au Chapitre 10) ou à tout processus permettant la différentiation.

Auto-différenciation : L'auto-différenciation se distingue comme un outil puissant pour l'apprentissage des paramètres au sein des systèmes physiques, offrant une méthode directe pour optimiser les valeurs en fonction des données observationnelles à travers un modèle physique.

[18] Li et al. [2022] démontrent les capacités de cette approche en optimisant les conditions initiales d'un univers simulé de manière à ce que les observations actuelles révèlent un motif épelant le nom de leur logiciel, `pmwd`, à travers des structures cosmiques à grande échelle.

Apprentissage par renforcement : Face à des fonctions objectives non différentiables, l'apprentissage par renforcement profond devient essentiel. Dans ce cadre, les réseaux de neurones, souvent qualifiés de politiques (*policies*), apprennent à maximiser une récompense en adaptant leurs stratégies basées sur les résultats de leurs actions, guidés par une fonction de récompense, approximant efficacement les gradients. Cette méthode est cruciale pour les paradigmes développés dans cette thèse et est amplement discutée dans la Section 3.3. Sa capacité à gérer des objectifs non différentiables la rend particulièrement précieuse pour des applications en robotique et interactions humaines, puisque nous ne pouvons évidemment pas auto-différencier la réalité, ainsi que pour des simulations non différentiables, i.e. la plupart des simulations actuelles[19]. Malgré son utilité, elle reste l'une des rares méthodes d'apprentissage non supervisé largement développées dans les domaines de l'ingénierie en raison de ses applications pratiques [Schmidhuber, 2015].

Auto-encodeurs. Un point d'honneur doit être accordé aux auto-encodeurs utilisés de manière non supervisée. Dans ces configurations, l'objectif est de reproduire les données d'entrée après les avoir traitées à travers une couche de goulot d'étranglement très compressée et de faible dimension. Ce processus réduit non seulement la dimensionnalité des données mais révèle également des éclairages profonds sur la structure des données dans l'espace latent.

Les Auto-Encodeurs Variationnels (VAEs) étendent cette idée en modélisant la distribution des données dans l'espace latent, apprenant des paramètres tels que la moyenne et la variance. Ces modèles sont inestimables pour leur capacité à simplifier des données complexes en formes plus gérables sans perdre des informations essentielles.

Un exemple remarquable de cette approche est illustré par le travail de Laroche and Speagle [2024], qui a démontré que des spectres stellaires entiers pourraient être efficacement codés en utilisant seulement six valeurs scalaires grâce à cette méthode. Cet exemple souligne le potentiel des VAEs à condenser considérablement de vastes quantités de données tout en conservant des informations cruciales, une technique qui présente des implications profondes étant donné la manière dont elle se rapproche de la recherche en physique.

Gestion des incertitudes. Dans le domaine des sciences naturelles, la prise en compte des incertitudes est primordiale. La technique connue sous le nom d'abandon (*dropout*) [Srivastava et al., 2014], initialement conçue pour prévenir le surajustement en désactivant aléatoirement une fraction des neurones durant l'entraînement, facilite également l'estimation des incertitudes

[19] Cela inclut également, par exemple, les jeux vidéo.

[Gal and Ghahramani, 2016]. Cette approche forme efficacement plusieurs variantes du modèle simultanément, chacune fonctionnant avec un sous-ensemble différent de neurones. En conséquence, la variabilité des prédictions du réseau peut être interprétée comme une mesure d'incertitude, offrant une gamme de résultats possibles au lieu d'une prédiction fixe.

En s'appuyant sur ce concept, on peut envisager chaque neurone non seulement comme une unité déterministe mais aussi comme une mini-distribution régulée par sa moyenne et sa variance. Cette notion est à la base des Réseaux Neuronaux Bayésiens [Goan and Fookes, 2020], où l'exploitation de ces distributions permet une quantification de l'incertitude dans les prédictions, offrant une vision plus approfondie de la fiabilité des sorties neuronales.

Vers l'apprentissage symbolique

Tout au long de cette discussion, nous avons exploré diverses manières par lesquelles l'apprentissage profond peut contribuer aux entreprises scientifiques, offrant parfois un degré d'interprétabilité. Les réseaux de neurones sont inestimables dans des domaines comme le traitement d'images ou la modélisation de systèmes complexes, où de tels modèles "souples" sont capables de saisir des nuances subtiles. Nous examinons comment cela peut se rapporter aux efforts pour cartographier le potentiel de la Voie Lactée au Chapitre 10.

Cependant, lorsque notre objectif se tourne vers la découverte de lois physiques fondamentales, l'exigence d'une interprétabilité symbolique se manifeste clairement. Le langage des mathématiques offre une description plus nette et plus définitive des phénomènes naturels. La section suivante introduit la régression symbolique, préparant le terrain pour le Chapitre 3 où nous plongerons plus profondément dans la manière dont les approches symboliques représentent et manipulent les mathématiques formelles.

Vue d'ensemble

La leçon fondamentale de cette section réside dans la constatation que sous la surface habituelle des applications de l'apprentissage machine en physique et astrophysique, des applications souvent héritées directement du domaine de l'ingénierie, se cache un immense "iceberg" de méthodes novatrices. Ces stratégies, ancrées profondément dans l'interprétabilité, possèdent un réel potentiel pour catalyser des découvertes scientifiques authentiques. Cette réalité est illustrée visuellement dans la Figure 12.3, qui esquisse cet iceberg des approches d'apprentissage machine, mettant en lumière les vastes possibilités encore peu exploitées pour la recherche avant-gardiste en physique et astrophysique.

Figure 12.3: Un iceberg des approches d'apprentissage machine pour la physique et l'astrophysique

Approches Symboliques

Depuis le début de la révolution scientifique, les chercheurs ont tenté de trouver des régularités répétables dans les expériences et les observations. Des structures mathématiques ont été utilisées dans cette exploration, et de nombreuses nouvelles, y compris les fonctions et les équations différentielles, ont été développées pour répondre à ce besoin de modéliser la nature. Peut-être en raison des symétries partagées entre la nature et les mathématiques, ces structures abstraites se sont souvent avérées exceptionnellement efficaces pour reproduire et prédire les propriétés du monde, au point que certains ont même envisagé que l'univers soit en réalité mathématique dans son essence [Tegmark, 2008].

La Régression Symbolique (SR), qui est centrale dans cette thèse, a un long pedigree. Peut-être son application la plus célèbre a été celle de Kepler aux éphémérides planétaires, lui permettant ainsi de trouver la loi de régression qui porte son nom [Kepler, 1609]. Cette loi empirique a fourni la base observationnelle sur laquelle Newton a pu construire les théories physiques développées dans ses *Principia Mathematica* [Newton, 1687].

Dans cette Section, nous introduisons la SR moderne qui vise à utiliser les immenses ressources informatiques à notre disposition pour rechercher parmi les descriptions analytiques possibles en termes d'un ensemble de fonctions et d'opérateurs (par exemple x, $+$, $-$, $\times$, $/$, sin, cos, exp log, ...) pour mieux ajuster un ensemble de données numériques $(\mathbf{x}, y)$ que nous souhaitons

modéliser. Concrètement, on cherche une fonction analytique $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ qui ajuste $y = f(\mathbf{x})$ étant donné ces données.

Régression symbolique

La régression symbolique (SR) pallie l'opacité des méthodes conventionnelles d'apprentissage automatique en élaborant des modèles qui allient compacité, interprétabilité et capacité de généralisation. L'ambition est de formuler des principes aussi élémentaires que la loi de la gravitation universelle de Newton, qui expliquent de manière exhaustive un large éventail d'expériences et d'observations. Les avantages de formuler les lois physiques sous forme d'expressions mathématiques succinctes, plutôt que comme de vastes modèles numériques, sont multiples.

Compacité

La SR est capable de générer des modèles remarquablement compacts, par exemple, des expressions contenant environ $\sim 10^1$ symboles [La Cava et al., 2021], un ordre de grandeur comparable à la longueur typique des expressions trouvées dans les *Feynman Lectures on Physics* [Feynman et al., 1971], qui est de 16. Les techniques les plus performantes de SR peuvent même produire des expressions bien en deçà d'une longueur de 10^3. À l'inverse, les modèles numériques, tels que les réseaux de neurones, s'appuient sur un nombre considérablement plus élevé de paramètres. Cette économie de moyens rend l'exécution des modèles SR moins coûteuse et, théoriquement, permet à la SR de redécouvrir l'expression mathématique exacte sous-jacente à un jeu de données avec beaucoup moins de données que les méthodes d'apprentissage machine conventionnelles [Wilstrup and Kasak, 2021], tout en offrant une résilience notable au bruit même dans le cadre d'une récupération parfaite du modèle [Reinbold et al., 2021, La Cava et al., 2021].

Généralisation

En outre, sauf dans le cas où les équations cibles sont constituées de polynômes de longueur arbitraire, les expressions concises générées par la SR tendent à être moins sujettes au surajustement sur les erreurs de mesure. Elles se montrent également beaucoup plus robustes et fiables hors de la gamme des données initiales, exhibant de meilleures capacités de généralisation, comme l'ont démontré plusieurs études [Sahoo et al., 2018, Kamienny et al., 2022, Kamienny and Lamprier, 2022, Wilstrup and Kasak, 2021]. Cela positionne la SR comme un outil potentiellement puissant pour élucider les représentations les plus concises et universelles des données mesurées.

Intelligibilité et interprétabilité

Les modèles élaborés par la SR, constitués d'expressions mathématiques, sont intrinsèquement compréhensibles pour les scientifiques, contrairement aux vastes modèles numériques. Cette caractéristique est d'une importance capitale en physique [Wu and Tegmark, 2019], où les modèles de SR peuvent faciliter l'intégration de nouvelles lois physiques découvertes dans un cadre théorique existant, ouvrant la voie à d'autres développements théoriques. Plus généralement, cette démarche s'inscrit dans un mouvement plus large favorisant des modèles d'apprentissage machine intelligibles [Sabbatini and Calegari, 2022], explicables [Arrieta et al., 2020] et interprétables [Murdoch et al., 2019], ce qui est d'autant plus crucial dans les domaines où ces modèles peuvent impacter directement les vies humaines [European Commission, 2021, 117th US Congress, 2022].

Un bref aperçu de la régression symbolique moderne

Approches traditionnelles de la SR

Programmation génétique. Historiquement, la SR a été principalement traitée via la programmation génétique, où un ensemble d'expressions mathématiques potentielles est raffiné de manière itérative grâce à des techniques inspirées de l'évolution naturelle, telles que la sélection naturelle, le croisement et la mutation. Cette méthode comprend le célèbre logiciel `Eureqa` [Schmidt and Lipson, 2009, 2011], reconnu pour ses performances (voir Graham et al. 2013 pour une évaluation de `Eureqa` sur des tests astrophysiques), et est complétée par des recherches plus récentes [Cranmer, 2023, de Franca and Aldeia, 2021, La Cava et al., 2019, Cava et al., 2019, Virgolin et al., 2019, Cranmer et al., 2020b, Virgolin et al., 2021, Stephens, 2015, Kommenda et al., 2020].

Autres approches conventionnelles. La SR a également été appliquée via diverses méthodes, allant de la force brute aux approches Monte-Carlo (guidées ou non), jusqu'à des recherches probabilistes [McConaghy, 2011, Kammerer et al., 2020, Bartlett et al., 2023a, Brence et al., 2021, Jin et al., 2019], et même par le biais d'algorithmes de simplification de problèmes [Luo et al., 2022, Tohme et al., 2023].

Deep learning

Approches principales. Les avancées substantielles des techniques de deep learning dans de nombreux domaines ont naturellement conduit à leur application en régression symbolique, remettant en question la prédominance des méthodes traditionnelles comme `Eureqa` [La Cava et al., 2021, Matsubara

et al., 2022]. Une gamme variée de méthodes intégrant des réseaux de neurones à la SR a été développée, depuis des stratégies sophistiquées de simplification de problèmes [Udrescu and Tegmark, 2020, Udrescu et al., 2020, Cranmer et al., 2020b], jusqu'à des méthodes de régression symbolique de bout en bout où un réseau de neurones est entraîné de manière supervisée à associer des jeux de données à leurs fonctions symboliques correspondantes [Kamienny et al., 2022, Lalande et al., 2023, Biggio et al., 2020, 2021, Vastl et al., 2022, d'Ascoli et al., 2022, Kamienny et al., 2023, Bendinelli et al., 2023, Holt et al., 2023, Li et al., 2024a,b, Chen et al., 2024a, Meidani et al., 2024, Becker et al., 2022, Shojaee et al., 2024, Alnuqaydan et al., 2022, Aréchiga et al., 2021], en passant par l'intégration de fonctions symboliques dans des réseaux neuronaux et leur ajustement par sparsité pour renforcer l'interprétabilité ou pour récupérer des expressions mathématiques [Scholl et al., 2023, Martius and Lampert, 2017, Brunton et al., 2016, Zheng et al., 2022, Sahoo et al., 2018, Valle and Haddadin, 2021, Kim et al., 2020, Panju and Ghodsi, 2020, Ouyang et al., 2018]. Pour des revues récentes des algorithmes de régression symbolique, voir [La Cava et al., 2021, Makke and Chawla, 2022, Angelis et al., 2023].

Apprentissage profond par renforcement. Bien que certaines méthodes précédemment évoquées se soient distinguées dans la production d'approximations symboliques extrêmement précises, le cadre de régression symbolique basé sur l'apprentissage profond par renforcement, proposé par Petersen et al. [2021a], est désormais considéré comme la référence pour la récupération exacte de fonctions symboliques, particulièrement en présence de bruit [La Cava et al., 2021, Matsubara et al., 2022]. Cette approche a engendré plusieurs études significatives [Landajuela et al., 2021a,b, Kim et al., 2021, Petersen et al., 2021b, Landajuela et al., 2022, Faris et al., 2024, He et al., 2024a, Du et al., 2022, Tian et al., 2024, Michishita, 2024, DiPietro and Zhu, 2022, Zheng et al., 2022, Landajuela et al., 2021b, Usama and Lee, 2022].

Vue d'ensemble

Pour conclure, nous mettons en lumière `PySR` [Cranmer, 2023], une initiative open source visant à recréer le logiciel `Eureqa` [Schmidt and Lipson, 2009, 2011], offrant des performances comparables. Bien que `PySR` n'emploie pas de techniques de deep learning, il a rapidement gagné en popularité au sein de la communauté astrophysique.

Une analyse comparative des principales méthodologies de régression symbolique, y compris notre propre méthode d'apprentissage profond par renforcement, qui est l'objet des Chapitres 3-6 sera exposée via le benchmark de Feynman standardisé [La Cava et al., 2021] dans la Figure 4.3 démontrant sa supériorité nette en terme de performances. Notre méthode est partic-

ulièrement notable, car elle est la seule jusqu'à présent où un réseau de neurones manipule des symboles mathématiques développée au sein d'une quelconque communauté physique ou astrophysique.

Nous explorerons également une méthode de simplification de problèmes dans la Section 7.1 et une approche neuro-symbolique dans la Section 7.2.

Résumé du corps

Dans le chapitre 3, nous avons investigué la transcription des mathématiques formelles en structures de graphes et avons examiné des méthodologies d'apprentissage à partir et de génération de ces graphes. Nous avons démontré que les problèmes mathématiques formels peuvent être abordés comme des problèmes d'optimisation de graphes. Notre discussion s'est spécialement focalisée sur la régression symbolique (SR), qui aspire à découvrir des expressions analytiques correspondant parfaitement à un ensemble de données initial. Nous avons mis en œuvre un paradigme d'apprentissage profond par renforcement où un réseau de neurones construit de manière séquentielle des expressions d'une qualité d'ajustement croissante au fil d'un processus d'essais et d'erreurs. Nous avons exposé notre méthode visant à garantir la validité des expressions générées automatiquement en utilisant une *notation préfixe* et en intégrant des principes qui limitent intrinsèquement la longueur des séquences produites par le réseau de neurones.

Dans le chapitre 4, nous avons enrichi notre méthode, que nous avons baptisée Φ-SO pour *Physical Symbolic Optimization*, avec la capacité d'exploiter les contraintes de l'analyse dimensionnelle physique, restreignant ainsi significativement l'espace de recherche pour les expressions potentielles. Cela a été réalisé en développant un algorithme capable de réaliser des analyses dimensionnelles extrêmement informatives sur des équations en cours de construction durant le processus de génération des expressions. Des contraintes sur les unités physiques ont ensuite été appliquées pour assurer la consistance physique des expressions grâce à un principe a priori et pour inculquer au réseau de neurones les règles de l'analyse dimensionnelle, prévenant ainsi les conflits entre les principes établis.

Nous avons montré que Φ-SO est l'algorithme prééminent pour la récupération symbolique précise en le comparant à 17 autres algorithmes de SR via le benchmark standard de Feynman, incluant 120 équations issues des conférences de Feynman sur la physique et d'autres manuels, à extraire de

leurs ensembles de données associés.

Dans le chapitre 5, nous avons élargi le paradigme Φ-SO pour permettre la recherche d'une forme fonctionnelle unique adaptée à plusieurs réalisations d'une même classe de phénomènes physiques, autorisant chaque réalisation à posséder des valeurs de paramètres libres (potentiellement) uniques, une approche particulièrement pertinente pour l'astrophysique. Nous avons baptisé ce nouveau cadre de travail *Class SR*. Étant donné l'originalité de notre méthode, nous avons développé le premier benchmark spécifiquement conçu pour évaluer les systèmes de Class SR, démontrant que Class SR surpasse significativement la SR traditionnelle dans les scénarios où plusieurs réalisations sont disponibles. Nous avons aussi démontré l'efficacité de notre système en utilisant un jeu de données synthétique de courants de la Voie Lactée, dérivant avec succès un potentiel d'entrée à partir des positions et vitesses stellaires.

Le chapitre 6 a détaillé l'implémentation logicielle `PhySO` de notre paradigme Φ-SO, en démontrant ses capacités et fonctionnalités uniques. Nous avons mis en avant notre engagement envers une démarche open-source, encourageant l'interaction et l'adoption par la communauté. Cet engagement est particulièrement essentiel dans le domaine de l'apprentissage automatique, où la transparence et la reproductibilité sont cruciales. Grâce à ces efforts, et à la conception du logiciel qui est à la fois concise, robuste et extensible, `PhySO` a connu une adoption significative au sein de plusieurs communautés de recherche en physique. Plusieurs équipes ont considérablement enrichi l'algorithme de `PhySO`, améliorant ses performances et élargissant ses capacités, renforçant ainsi son impact et son utilité.

Dans le chapitre 7, nous avons exploré des approches complémentaires à notre méthode Φ-SO, capables d'exploiter les réseaux de neurones pour capturer directement et incarner la structure de graphe d'un ensemble de données, reflétant sa représentation analytique sous-jacente. Ces approches ont le potentiel d'exploiter les dérivées informant la structure du graphe par rapport aux données. Nous avons notamment introduit une méthode novatrice capable de détecter les séparabilités additives et multiplicatives dans un ensemble de données en analysant les gradients de sa représentation par réseau de neurones. De plus, nous avons discuté d'une approche innovante où un réseau de neurones lui-même émule une expression analytique, intégrant des fonctions de base dans sa structure et favorisant la sparsité. Nous avons ensuite abordé des stratégies de fertilisation croisée dans le domaine de la SR.

Nous avons introduit plusieurs améliorations clés à l'état de l'art de la régression symbolique basée sur l'apprentissage par renforcement. Celles-ci vont des raffinements mineurs aux avancées plus substantielles, incluant l'introduction d'un paramètre de température de recuit, des améliorations significatives du processus d'ajustement des constantes libres via l'auto-

différenciation, l'introduction de la Class SR, ainsi que le développement de nouvelles contraintes. La plus notable de ces améliorations est l'incorporation de l'analyse dimensionnelle, qui informe le réseau de neurones des unités physiques et permet une réduction de l'espace de recherche tout en garantissant la cohérence physique.

Notre algorithme de SR a été testé de manière approfondie sur des centaines de cas synthétiques, ainsi que sur des données réelles observées, comme l'étude des chandelles standards pour déduire la loi de l'expansion de l'Univers. De plus, plusieurs équipes de recherche ont appliqué notre logiciel à des données expérimentales ou observationnelles réelles dans un large éventail de domaines de la physique, tels que l'astrophysique, l'aéronautique, la mécanique, la physique des particules, la dynamique des fluides, les télécommunications, la géologie physique, et même la biologie. En outre, nous menons actuellement des expériences sur des données issues de simulations complexes, avec un accent sur l'évolution des amas globulaires et l'exploration de modèles alternatifs pour les profils de matière noire, comme discuté dans la Section 11.2.

Nous souscrivons à l'idée selon laquelle en développant des approches abstraites pour la physique et l'astrophysique, il est essentiel de maintenir un lien solide avec les problèmes physiques concrets. Par conséquent, nous avons orienté notre attention vers la question tangible de la matière noire à l'échelle galactique, un sujet introduit dans le chapitre 8.

Dans le chapitre 9, nous avons poursuivi notre démarche guidée par l'observation des enquêtes physiques en explorant des sondes de matière noire sous forme de structures étant accrétées par la Voie Lactée. En étudiant l'échantillon stellaire proche du Soleil de Gaia, qui comprend des informations complètes sur la position et la vitesse, nous avons découvert un nouveau courant stellaire que nous avons baptisé Typhon. Nous prévoyons que l'échantillon complet de ce courant polaire s'étende jusqu'à l'halo externe de la Voie Lactée à environ 100 kpc, potentiellement en faisant une sonde exceptionnelle de la matière noire. Après une analyse chémo-dynamique, nous avons identifié Typhon comme le reliquat d'une galaxie naine, suggérant que de nombreux autres fragments de galaxies naines pourraient se cacher dans l'halo externe. Nous avons également contribué à la découverte d'une autre structure, que nous avons nommée Antaeus, en détectant ses membres près du Soleil, remettant en question la vue prédominante selon laquelle les temps dynamiques dans le disque sont courts et que les structures de marée se mélangent rapidement, effaçant toutes les caractéristiques initiales de type courant. Par ailleurs, nous avons partcipé à une large campagne observationnelle à l'Isaac Newton Telescope ainsi qu'au Very Large Telescope ayant permis de mettre

en évidence 28 nouveaux courants stellaires résultant au catalogue le plus exhaustif des courants de la Voie Lactée à ce jour.

Le chapitre 10 s'est concentré sur les méthodes de récupération du potentiel gravitationnel de la Voie Lactée et de sa distribution sous-jacente de matière noire de manière agnostique au modèle. Notre approche, baptisée `MassFinder`, nécessite essentiellement que le potentiel stabilise la distribution stellaire observée. En employant des transformations canoniques pour travailler dans l'espace des orbites dans le processus, nous abordons la disponibilité spatialement limitée des échantillons 6D (position et vitesse) de Gaia. Le chapitre conclut en discutant des avancées potentielles dans ce domaine émergent de la phénoménologie galactique.

Vue d'ensemble

La pierre angulaire de cette thèse est ancrée dans une démarche résolument guidée par l'observation, qui a profondément influencé notre intégration de l'apprentissage automatique au cœur de la physique et de l'astrophysique. À rebours des méthodes conventionnelles d'apprentissage automatique, qui s'appuient généralement sur des exemples issus de modèles ou simulations préexistants, nous avons élaboré nos méthodes dans l'optique de générer des solutions de manière non supervisée, sans dépendre de modèles antérieurs, en requérant de manière intrinsèque une conformité aux données observationnelles. Cette orientation agnostique au modèle est indispensable en physique, car elle constitue l'unique voie pour l'émergence de nouveaux modèles qui reflètent avec une plus grande fidélité les phénomènes naturels.

Durant l'ensemble de ce travail doctoral, nous avons scrupuleusement adhéré à une philosophie de recherche guidée par l'observation. À titre d'exemple, dans notre étude sur la régression symbolique, nous avons évité d'entraîner les réseaux de neurones sur une collection prédéfinie d'expressions mathématiques. Nous avons plutôt élaboré des modèles aptes à formuler des expressions *a priori* inédites, en exigeant non seulement qu'elles s'ajustent aux données mais également qu'elles respectent les principes de l'analyse dimensionnelle et les contraintes de classes. De même, dans notre approche des complexités liées à la matière noire, nous avons délibérément délaissé les simulations prédéfinies qui incluent des biais inductifs. Nous avons choisi de développer un cadre d'apprentissage non supervisé qui contraint de façon inhérente les propriétés de la matière noire à se conformer aux contraintes observationnelles, harmonisant ainsi le développement de nos modèles avec les preuves empiriques.

Sur le plan technique, nous ambitionnons que les méthodologies évoquées dans cette thèse, notamment l'auto-différenciation et les techniques d'apprentissage par renforcement, soient davantage adoptées au sein de la communauté astrophysique. Nous sommes convaincus que ces approches, bien que

peu exploitées actuellement en astrophysique, détiennent un potentiel considérable pour révolutionner les analyses indépendantes des modèles.

À une échelle plus profonde, nous espérons que ce travail doctoral inaugure une ère nouvelle vers des méthodologies interprétables et porteuses de sens en physique et astrophysique, menant à l'intégration des techniques d'apprentissage profond symbolique. Notre cadre d'apprentissage symbolique, Φ-SO, est une innovation majeure : il représente la première et, à ce jour, la seule méthode développée dans les domaines de la physique et de l'astrophysique où un réseau de neurones manipule directement des symboles mathématiques. Ce travail renforce la conviction que la tendance dominante à se reposer exclusivement sur l'apprentissage supervisé et les modèles de type boîte noire ne fait qu'effleurer les possibilités offertes par l'apprentissage automatique dans l'exploration scientifique des phénomènes naturels.

Perspectives

Progrès en apprentissage symbolique

Comme cela a été souligné dans les discussions précédentes, notamment dans la sous-section 3.3.3, il est primordial d'améliorer l'efficacité des systèmes d'apprentissage symbolique. Les méthodologies actuelles, qu'elles s'appuient sur des réseaux neuronaux comme les nôtres ou sur la programmation génétique, raffinent généralement les expressions symboliques en utilisant une métrique scalaire non différentiable (par rapport à l'arrangement symbolique) de la qualité de l'ajustement. Cette approche peut conduire involontairement à ce que nous désignons par la *malédiction de la SR guidée par la précision*, où l'optimisation de la métrique ne converge pas nécessairement vers la forme fonctionnelle la plus exacte.

Pour pallier ce problème, une orientation de recherche importante consiste à améliorer les mécanismes d'auto-correction des systèmes d'essais et d'erreurs en leur permettant d'utiliser des gradients relatifs à l'arrangement symbolique. Cette amélioration pourrait nécessiter l'intégration d'un composant d'apprentissage supervisé qui apprendrait activement la géométrie locale de l'espace de recherche de la forme fonctionnelle, comme décrit dans la sous-section 3.3.3. Par ailleurs, encourager les synergies avec d'autres méthodologies de régression symbolique pourrait enrichir la portée et la profondeur de l'apprentissage symbolique, comme discuté dans la sous-section 3.3.3 et dans la Section 7.3. Des perspectives spécifiques liées à nos approches

intégrant l'analyse dimensionnelle et la régression symbolique basée sur les classes ont été détaillées dans la Section 4.6 et la sous-section 5.4, respectivement.

Application de la SR en physique et en astrophysique

L'adoption croissante des méthodes de SR dans les domaines de la physique et de l'astrophysique représente un développement notable, comme illustré par l'utilisation de Φ-SO dans la recherche énumérée dans la Table 6.1.

Biais inductifs. Comme déjà discuté dans les paragraphes 3.3.2 et 6.1.2, une méthode efficace pour améliorer l'application de la SR au-delà de notre utilisation actuelle des contraintes de classe et de l'analyse dimensionnelle consiste à intégrer des connaissances préalables spécifiques au domaine dans le processus de recherche. Par exemple, si la recherche concerne une forme fonctionnelle censée présenter certaines symétries, des comportements limites ou des caractéristiques spécifiques dans une équation différentielle, ces aspects devraient être intégrés dans la fonction de récompense de Φ-SO. En particulier, puisque ces contraintes ne nécessitent pas d'être différentiables, Φ-SO peut accommoder un large éventail de restrictions scientifiquement significatives pour affiner et orienter efficacement le processus de recherche.

Apprentissage des approximations analytiques pour des physiques coûteuses. Bien que cette thèse se soit principalement concentrée sur des approches guidées par l'observation, l'apprentissage symbolique offre aussi un potentiel significatif pour aborder les aspects computationnellement exigeants des simulations. Un domaine en expansion concerne les émulateurs neuronaux conçus pour simuler les calculs complexes trouvés dans les simulations cosmologiques, tels que ceux liés aux rétroactions [Dai and Seljak, 2021]. La SR pourrait être particulièrement bien adaptée à ce rôle en raison de ses capacités de généralisation et d'interprétabilité, offrant potentiellement des avantages sur d'autres méthodes.

En dynamique galactique, la SR pourrait être utilisée pour dériver des expressions analytiques approximatives pour des phénomènes tels que la friction dynamique [François et al., 2024] ou les aspects de la physique stellaire exigeants en calcul [Bianchini et al., 2016]. Ces équations efficaces et compréhensibles pourraient grandement réduire le besoin de vastes grilles de simulation N-corps. Par exemple, une approche de champ moyen / particule test couplée à un émulateur analytique pourrait être utilisée lors de recherches de paramètres pour éviter des simulations fréquentes, avec une étape de validation finale utilisant la simulation complète pour vérifier l'exactitude des paramètres identifiés par l'émulateur.

Apprentissage des approximations optimales N-corps. En continuant sur le thème de l'amélioration de l'efficacité des simulations, les simulations N-corps, qui approximent traditionnellement les interactions N^2 entre tous les corps, pourraient bénéficier grandement de méthodes d'apprentissage avancées. Ces simulations simplifient souvent les interactions à travers des techniques pouvant être conceptualisées comme des approximations basées sur des graphes.

Il existe une opportunité prometteuse d'utiliser la SR ou ses techniques d'optimisation de graphes sous-jacentes directement pour apprendre ces approximations optimales. Inspirés par les méthodes d'apprentissage par renforcement discutées dans le Chapitre 3, qui exploitent les structures de graphes, nous pourrions développer un système pour identifier et mettre en œuvre automatiquement les simplifications les plus efficaces pour les simulations N-corps. Cette approche a le potentiel non seulement d'affiner l'exactitude des simulations N-corps, mais aussi de réduire leurs exigences computationnelles.

Cartographie de la Voie Lactée

Exploration des structures proches du Soleil. Dans nos discussions précédentes, notamment dans les paragraphes 9.1.3 et 9.2.3, nous avons examiné les perspectives concernant les courants stellaires Typhon et Antaeus. Une piste particulièrement prometteuse réside dans l'extension de l'échantillonnage du courant Typhon. Bien qu'initialement détecté à proximité du Soleil, ce courant est prévu pour s'étendre jusqu'à l'halo externe. Par conséquent, un échantillonnage complet pourrait constituer une sonde exceptionnelle pour l'étude de la matière noire. De surcroît, la découverte de telles structures dynamiques et cohérentes près du Soleil encourage une investigation approfondie sur la façon dont ces structures peuvent demeurer non mélangées en phase, compte tenu des courts temps dynamiques prévus près du Soleil. Cette exploration pourrait générer des connaissances précieuses, potentiellement à travers des simulations spécifiquement ciblées de ces structures.

Exploitation des échantillons 5D. Comme mentionné dans la Section 10.5, une autre perspective fascinante est celle d'apprendre à déprojeter les échantillons 5D dépourvus d'informations sur la vitesse radiale. En maîtrisant de manière efficace la fonction de distribution de la dimension manquante, l'espoir est que le surplus d'informations compensera les degrés de liberté additionnels induits par la nécessité d'apprendre cette distribution.

Méthodes dynamiques différentiables. Comme abordé dans la Section 10.5, nous avons envisagé le développement d'un modèle différentiable pour les courants stellaires qui pourrait être intégré à des approches de champ moyen. En outre, nous avons examiné la possibilité de mettre au point un es-

timateur d'action différentiable et déterministe opérant en coordonnées canoniques. Cela nécessiterait l'adoption d'approximations classiques, qui seraient adaptées pour permettre une différentiabilité.

Contraindre la matière noire

Contrainte sur la masse de la particule de matière noire

Il serait fructueux de développer des équations claires et maniables qui récapitulent les propriétés essentielles des courants stellaires en fonction de la masse des particules de matière noire. Pour réaliser cela, nous pourrions exploiter divers courants issus de simulations menées à différents niveaux de masse particulaire [Carlberg et al., 2024]. En ajustant ces modèles fonctionnels aux observations actuelles des courants, il serait alors possible d'extraire les valeurs paramétriques décrivant la masse de ces particules en utilisant le paradigme de Régression Symbolique de Classe que nous avons élaboré (Chapitre 5). Les formules ainsi générées seraient spécifiquement conçues pour refléter les comportements dépendant de la masse des particules, permettant d'exploiter la présence de lacunes dues à des sous-halos potentiels de manière statistiquement significative. Cette méthode statistique pourrait nous permettre de lier toutes les observations disponibles à plusieurs modèles, via une simplification intelligible sous forme d'équation, offrant ainsi de vastes possibilités pour contraindre les propriétés de la matière noire basées sur des contraintes observationnelles.

Apprentissage de profils dynamiques

Au-delà du profil NFW. Nous nous employons actuellement à mettre en œuvre le paradigme Φ-SO pour développer des alternatives au profil empirique NFW, tel qu'il est formulé dans l'Eqn. 8.1. Traditionnellement utilisé pour décrire les distributions de matière noire au sein des galaxies, ce profil est remis en question par nos résultats préliminaires obtenus sur l'ensemble de simulations NIHAO [Wang et al., 2015]. Nous avons identifié plusieurs profils qui, en termes de capacité prédictive et de simplicité, surpassent nettement le profil NFW. De plus, nous nous attaquons à un problème majeur du profil NFW, sa masse intégrée non convergente à l'infini, en intégrant cette limitation à notre fonction de récompense, garantissant ainsi la validité physique de nos nouveaux profils.

Modélisation des courbes de rotation des amas globulaires. Le modèle empirique couramment utilisé pour les courbes de rotation des amas globulaires ne parvient pas à saisir les tendances de vitesse au-delà du pic ni à intégrer la dynamique temporelle, essentielle compte tenu de la nature

évolutive de ces amas liée à la perte de masse [Bianchini et al., 2018]. Nous élaborons un nouveau profil fondé sur les données des simulations les plus récentes concernant ces amas globulaires[20], qui modélise la vitesse en fonction du rayon tout en intégrant des modifications temporelles. Nos premiers succès incluent la modélisation de l'évolution temporelle des vitesses maximales, offrant un outil robuste potentiellement utile pour estimer l'âge des amas globulaires à partir de leurs profils dynamiques observés dans la Voie Lactée.

Extraction d'une distribution générale à partir de courants extragalactiques

Une autre perspective stimulante consisterait à s'appuyer sur le cadre d'apprentissage non supervisé décrit dans le Chapitre 10 pour élaborer une distribution de matière noire "universelle" répondant à diverses contraintes observationnelles relatives aux structures de faible luminosité de surface entourant des galaxies lointaines [Nibauer et al., 2023, Sola et al., 2022]. Varghese et al. [2011] a ouvert la voie à l'exploitation de ces caractéristiques pour contraindre les distributions de masse. Ces données pourraient être tirées d'observations réalisées par CFHT, Euclid [Laureijs et al., 2011], ou le Télescope Spatial Roman. Nous envisageons de développer un cadre différentiable destiné à déprojeter ces nombreuses structures et à calculer le potentiel libre (paramétré uniquement par quelques paramètres d'échelle spécifiques à chaque galaxie) qui reproduirait ces structures de manière non supervisée, tout en ajustant les paramètres de déprojection en cours de processus[21]. Malgré le nombre élevé de degrés de liberté introduits, on peut espérer que l'abondance des contraintes observationnelles rendra ce projet viable et riche en informations. Il est à noter que, bien que la vitesse des caractéristiques de marée extragalactiques ne soit généralement pas mesurée, des études récentes indiquent qu'elle pourrait être déduite de leurs amas globulaires pour lesquels des informations en 6D sont disponibles [Ferrone et al., 2023].

Nous aborderons dans la sous-section suivante la possibilité d'utiliser également des cartes de vitesse extragalactiques à des fins similaires.

[20] Ces simulations N-corps, réalisées par Paolo Bianchini à l'Observatoire Astronomique de Strasbourg, sont les premières à modéliser les amas globulaires avec une correspondance un-à-un des étoiles sur une période de 13 Gyr tout en prenant en compte l'évolution stellaire, les champs de marée, et la dynamique rotationnelle initiale, offrant ainsi une représentation exhaustive et réaliste de l'évolution des amas globulaires.

[21] Cette méthode de déprojection serait semblable à celle que nous avons proposée pour les étoiles de la Vo

ie Lactée, permettant ainsi une pollinisation croisée des approches méthodologiques face à ce défi ou même une pollinisation croisée neuronale sous la forme de l'apprentissage par transfert.

Génération d'équations différentielles à partir des données

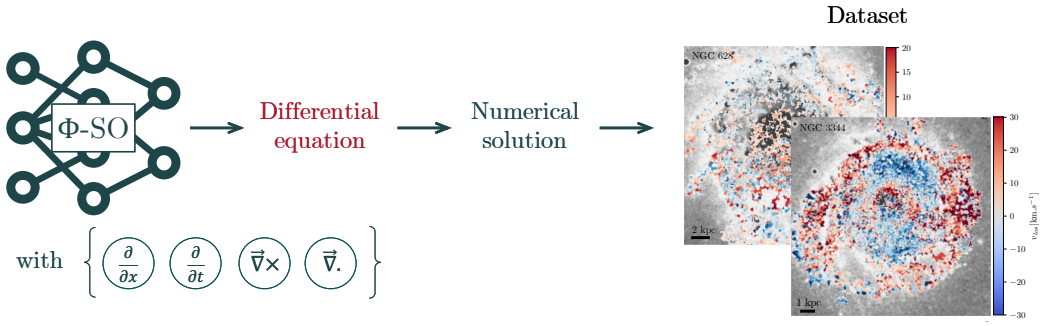


Figure 12.4: **Génération d'équations différentielles régissant des observations.** Une extension envisagée de notre paradigme Φ-SO vise la découverte d'équations différentielles symboliques dont les solutions numériques ajustent un ensemble de données. Dans le cadre de la recherche sur la matière noire, cela pourrait être employé à apprendre des alternatives ou des extensions à l'équation de Poisson qui régit la matière noire et les dynamiques dans les galaxies, basées sur, par exemple, leurs cartes de vitesse. Nous illustrons ce point en montrant les champs de vitesse en ligne de mire (los) de deux galaxies, adaptés de [Urrejola-Mora et al., 2022].

À l'avenir, nous projetons de repousser les limites des méthodes d'apprentissage symbolique en explorant le domaine de la génération automatique d'équations différentielles analytiques, comme illustré sur la Figure 12.4. L'objectif est d'étendre le paradigme Φ-SO afin de lui permettre de générer des équations différentielles dont les solutions correspondent à un jeu de données donné ou répondent à des critères spécifiques, tels que des principes physiques supplémentaires, des symétries ou des conditions asymptotiques, étendant les opérateurs à, par exemple: $\partial/\partial t, \partial/\partial x, \vec{\nabla}.\square, \vec{\nabla} \times \square$.... L'intégration des équations différentielles dans l'apprentissage symbolique marque un éloignement des lois empiriques vers des constructions plus abstraites mais interprétables. Travailler dans le domaine des équations différentielles est particulièrement pertinent car elles permettent d'exprimer des solutions simples à des problèmes qui peuvent être très complexes ou dont les solutions explicites peuvent même ne pas exister.

La génération d'équations différentielles requiert naturellement un processus séquentiel, actuellement réalisable uniquement dans quelques paradigmes d'apprentissage symbolique[22] tels que Φ-SO. Cette génération séquentielle est cruciale car elle permet l'intégration des *priors in situ* nécessaires pour

[22]Dû à la nécessité pour les paradigmes dans lesquels un réseau de neurones génère directement des symboles mathématiques comme le nôtre.

gérer efficacement les variables multidimensionnelles et pour imposer un enchevêtrement maximal des opérateurs différentiels assurant la stabilité numérique.

Les méthodes existantes dans ce domaine se contentent souvent d'un cadre de régression symbolique régulier dans lequel on intègre des dérivées en les traitant comme des variables supplémentaires, telles que $x_1, \frac{\partial x_1}{\partial t}, x_2, \frac{\partial x_2}{\partial t}, t$ au lieu de x_1, x_2, t, plutôt que de les intégrer dynamiquement dans le processus d'apprentissage. Bien que notre implémentation actuelle de Φ-SO prenne déjà en charge cette possibilité de manière directe, l'approche que nous proposons va plus loin en permettant l'enchevêtrement des opérateurs différentiels et l'incorporation d'opérations vectorielles telles que les produits scalaires et vectoriels aux côtés d'un opérateur ∇.

Cette extension méthodologique est quelque peu inexplorée en apprentissage automatique, principalement parce qu'elle vise à produire des modèles interprétables sous la forme d'équations différentielles qui doivent être résolues pour être utilisées, les rendant bien plus exigeantes en calcul que les objectifs typiques de l'apprentissage automatique, qui privilégient souvent la précision prédictive et l'efficacité computationnelle plutôt que l'interprétabilité.

Apprentissage d'alternatives à l'équation de Poisson régissant la matière noire

Nos projets incluent l'utilisation de ce cadre d'apprentissage symbolique avancé pour explorer des formulations alternatives ou des expansions d'ordre supérieur de l'équation de Poisson (donnée dans l'Eqn. 8.4), qui est essentielle pour comprendre la dynamique de la matière noire dans les galaxies. Cette exploration se basera sur les cartes de vitesse observationnelles des galaxies.

Cette démarche pourrait servir de méthode systématique pour évaluer si les données observationnelles des galaxies appuient les paradigmes ΛCDM (Λ Cold Dark Matter) ou MOND (Dynamique Newtonienne Modifiée), comme discuté dans la Section 8.1, ou pourraient même révéler la présence de phénomènes physiques entièrement nouveaux. En combinant cette méthode avec notre cadre de Régression Symbolique de Classe, détaillé dans le Chapitre 5, nous aspirons à permettre l'apprentissage d'une équation différentielle "universelle" qui encapsule la dynamique de plusieurs galaxies simultanément, comme illustré sur la Figure 12.4.

Apprentissage de cosmologies alternatives

Parmi les autres applications intéressantes figurent l'évolution des galaxies, notamment avec le soutien des données JWST [Gardner et al., 2006], ainsi que la recherche cosmologique. Spécifiquement, à la lumière de la mission Planck [Aghanim et al., 2020], nous envisageons d'explorer des extensions à l'équation de Friedman, en exigeant que le modèle analytique résultant prédise

à la fois l'expansion cosmique observée à travers des chandelles standard et le fond diffus cosmique (CMB), qui contient des informations cruciales sur la distribution de la matière dans l'Univers primitif.

Vers la génération automatique de théories

Qu'est-ce qu'une théorie ?

Bien que la découverte de nouvelles extensions ou alternatives aux lois qui régissent des échelles spécifiques, telles que cosmologiques ou galactiques, représenterait un apport indéniable, cela ne constituerait pas en soi l'élaboration d'une théorie exhaustive. Définir ce qui constitue une "théorie" relève d'une interrogation épistémologique complexe. Adoptant une perspective orientée par l'observation, nous concevons une théorie comme un ensemble cohérent de lois interconnectées, capable de prédire avec précision des phénomènes naturels à des échelles extrêmement variées. Historiquement, les lois de Newton ont été perçues comme une première "théorie de tout", capables de prédire aussi bien la trajectoire d'une pomme tombant d'un arbre que les mouvements des corps célestes, des phénomènes se manifestant à des échelles diamétralement opposées. Ce principe illustre les défis posés par les théories de la gravité quantique [Rovelli, 2004] et, bien que de manière moins formidables, les défis que nous rencontrons en tentant de concilier les phénomènes aux échelles galactiques et cosmologiques.

Optimisation de théories physiques

Dans cet objectif, nous aspirons à repousser les frontières du paradigme de l'apprentissage symbolique en développant un cadre capable de formuler des "théories" complètes. Cela implique la création d'un algorithme capable de générer et de raffiner de manière autonome des systèmes d'équations interconnectées (éventuellement différentielles) pour se conformer précisément à de multiples contraintes observationnelles (chaque contrainte pouvant comprendre plusieurs manifestations d'un seul phénomène). Ce concept est illustré visuellement dans la Figure 12.5.

Dans la pratique, nous envisageons le développement d'un cadre de travail où sont optimisées $N+n$ équations, où les N premières équations sont conçues pour ajuster N jeux de données observationnelles distincts, avec la possibilité pour chaque équation d'ajuster des jeux de données à multiples réalisations (comme les observations de galaxies lointaines) en exploitant le paradigme Class SR. Les n équations supplémentaires agiraient comme des équations auxiliaires, encodant potentiellement des redondances qui saisissent les principes fondamentaux de la théorie. L'objectif serait de dériver l'ensemble d'équations le plus simple possible qui décrive collectivement tous les phénomènes observés

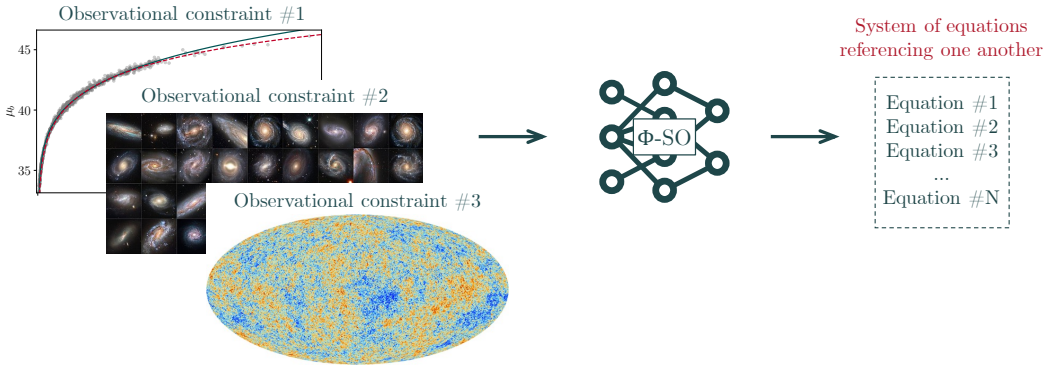


Figure 12.5: **Vers l'élaboration automatique de théories.** Cette illustration illustre l'extension prévue de notre paradigme Φ-SO qui lui permettrait d'apprendre de manière autonome des "théories", soit plusieurs équations analytiques (éventuellement différentielles) se référant mutuellement et satisfaisant à diverses contraintes observationnelles (chaque contrainte pouvant impliquer plusieurs réalisations d'un même phénomène). Nous soulignons notre propos en proposant une exploration automatisée de "théories" ajustant des chandelles standard, une caractéristique galactique clée à travers de multiples réalisations, et le fond diffus cosmologique (CMB).

de manière efficace.

Applications

Compte tenu de la complexité et de la nature abstraite de ce système, ainsi que des défis anticipés pour garantir sa performance robuste dans des applications scientifiques réelles, nous proposons de concentrer initialement nos efforts sur la récupération d'une "théorie simple", telle que les lois de l'électromagnétisme de Maxwell. Par la suite, l'objectif serait de s'attaquer à des scénarios astrophysiques concrets significatifs pour maintenir une pertinence pratique. Les applications spécifiques peuvent varier. Toutefois, une voie prometteuse consisterait à appliquer ce système pour développer un modèle prédictif capable de prévoir les comportements tant à l'échelle galactique que cosmologique, comme illustré dans la Figure 12.5.

Améliorer la compétence des grands modèles de langage en matière de données et de mathématiques

Modèles LLM de génération actuelle

Ce projet doctoral, a été témoins du développement rapide et de l'adoption généralisée des grands modèles de langage (LLMs) exploitant des architectures de transformateurs génératifs pré-entraînés (similaires à ceux utilisés dans les approches pré-entraînées de régression symbolique discutées dans la Section

2.2.2). Ces modèles ont atteint des performances de niveau humain ou quasi-expert dans diverses tâches, y compris la traduction linguistique et la programmation, mais leur maîtrise dans les domaines scientifiques demeure relativement embryonnaire [Saxena et al., 2023].

Tokenisation et défis d'apprentissage. L'amélioration des capacités scientifiques des LLMs pourrait résider dans l'optimisation de leur traitement et apprentissage du contenu scientifique. La science implique essentiellement l'observation de phénomènes naturels et la formulation de modèles prédictifs. Les LLMs actuels rencontrent des difficultés avec les tâches scientifiques, en partie à cause de leur méthode d'entraînement qui repose sur l'apprentissage supervisé à partir de vastes corpus de textes tokenisés de manière spécifique, comme expliqué dans le paragraphe 3.1.2.

Le processus de tokenisation pour les équations mathématiques est souvent directement appliqué aux chaînes `Latex`. Cette approche présente une contrainte majeure : si un LLM génère l'équation `b + a` alors que le format correct dans les données d'entraînement est `a + b`, le modèle est sanctionné malgré l'équivalence mathématique, du fait de son approche d'apprentissage basée sur les tokens. En outre, les LLMs doivent apprendre de manière autonome la syntaxe et la structure des expressions mathématiques valides, y compris les règles de parenthésage et de formatage des expressions, sans l'aide de plongements (*embeddings*) spécialisés pour les constructions mathématiques. Cette situation contraste nettement avec leurs capacités de traitement linguistique, où les plongements pour mots ou sous-mots simplifient grandement le processus d'apprentissage.

En définitive, les LLMs actuels n'ont pas accès directement à la structure de graphe sous-jacente des expressions mathématiques ; ils interagissent seulement avec leurs représentations en `Latex`. Cette lacune souligne une importante limitation dans leur entraînement : sans plongements analogues à ceux utilisés pour les données textuelles, les LLMs sont confrontés à un défi d'apprentissage bien plus complexe lorsqu'ils traitent des expressions analytiques.

Multi-modalité. Les avancées récentes en multi-modalité permettent désormais aux LLM d'interagir directement avec divers types de données, incluant images et audios, en plus des textes. Cette évolution introduit une capacité transformative où les LLMs peuvent traiter des données via des entrées spécifiques à chaque modalité, comme une unité neuronale audio capable d'analyser directement les enregistrements vocaux[23].

[23] Cette intégration permet aux modèles de répondre aux subtilités des données audio, telles que les intonations émotionnelles, les mélodies ou les sons contextuels qui seraient perdus lors d'une transcription traditionnelle de parole en texte.

Modalités symboliques et de données

Au vu des récentes avancées en régression symbolique profonde, nous proposons d'intégrer des modalités spécialisées orientées vers la science dans les LLM, comme le montre la Figure 12.6. Cette adaptation vise à renforcer la capacité des LLM à accomplir des tâches scientifiques en intégrant directement des connaissances spécifiques au domaine dans leur structure.

Les plateformes potentielles pour intégrer ces nouvelles modalités incluent le LLM `AstroLlama` [Nguyen et al., 2023], spécifiquement créé pour des applications en astrophysique, et `nanoGPT` [Karpathy, 2023], qui propose une architecture simplifiée[24] conçue pour le prototypage. De plus, le modèle de pointe `Llama 3.1` [Llama Team, 2024] pourrait offrir un environnement robuste pour déployer ces stratégies d'apprentissage complexes et multi-modales.

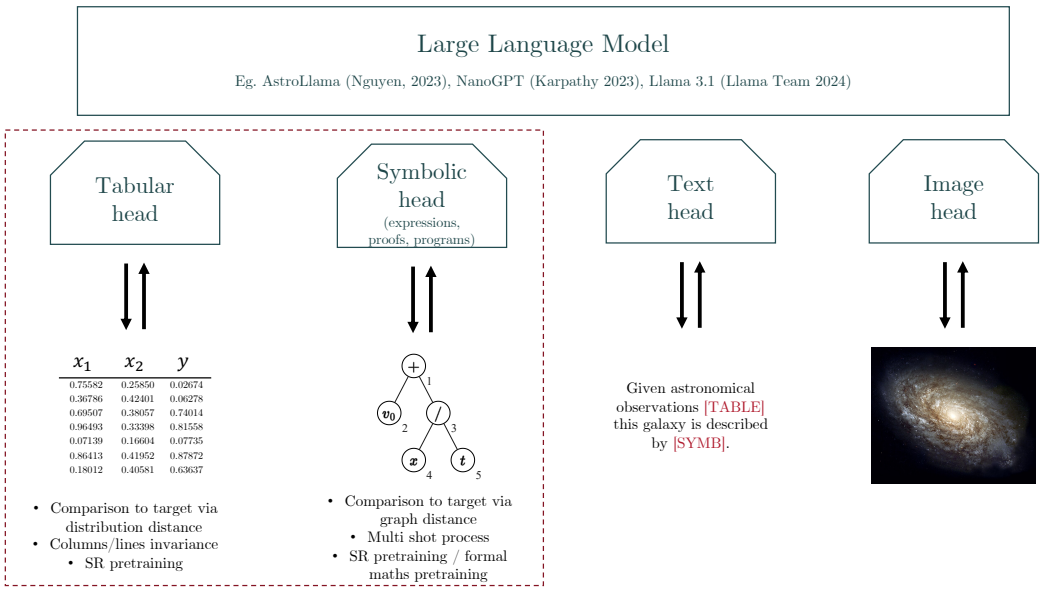


Figure 12.6: **Vers des modalités symboliques et de données pour les LLMs.** Ce schéma illustre les améliorations proposées (mises en évidence par le contour rouge) aux Grands Modèles de Langage (LLMs) qui leur permettraient de traiter et d'apprendre à partir de données scientifiques et d'expressions mathématiques, en plus de leurs capacités existantes avec le texte et les images. Les extensions comprennent des unités spécialisées pour ingérer et produire des expressions symboliques `[SYMB]` ou des données tabulaires `[TABLE]` au lieu de tokens, améliorant leur compétence en données et en mathématiques formelles.

Modalité symbolique. Nous suggérons une modalité "symbolique" spécifique qui permettrait aux LLM d'interpréter et d'apprendre à partir

[24] Offrant des performances comparables à `GPT2` [Radford et al., 2019]

de la structure de graphe sous-jacente des expressions mathématiques, des preuves ou même des programmes informatiques[25]. Cette unité symbolique spécialisée pourrait adopter une approche d'essais et d'erreurs : plutôt que de générer une expression en une seule tentative, elle pourrait affiner itérativement l'expression, en partant d'une représentation initiale dans l'espace latent, une méthode similaire aux processus de diffusion [Ho et al., 2020] ou à l'approche de Kamienny et al. [2023] dans le contexte de la SR. La mise en œuvre d'un tel système nécessiterait un outil robuste de gestion de graphes d'expressions symboliques, que nous avons développé dans le cadre du Φ-SO. Notre système est unique en son genre, offrant une représentation et vectorisation complètes des graphes à la fois en lot et en longueur d'équation, plaçant Φ-SO à l'avant-garde de cet effort passionnant.

Métrique de distance de graphe. Un progrès supplémentaire pourrait consister à développer une métrique différentiable pour mesurer les distances entre les équations. Cette métrique tiendrait compte de propriétés telles que la commutativité (par exemple, attribuant une distance nulle entre des expressions telles que $a+b$ et $b+a$) et intégrant des identités algébriques plus complexes pour évaluer la similitude. La mise en œuvre de cela nécessiterait un réseau neuronal, qui, bien que non infaillible, offre la rapidité et la différentiabilité nécessaires pour une telle tâche. Cette approche vise à créer une métrique de distance d'expression symbolique rapide et universellement applicable, exploitant les capacités des réseaux neuronaux pour atteindre l'efficacité et la scalabilité.

Modalité de données tabulaires. Les LLM de génération actuelle codent souvent les valeurs numériques sous forme de chaînes de texte, par exemple $s.aaa.10^{bb}$, où s représente le signe (positif ou négatif), aaa sont les chiffres, et bb sont les chiffres de l'exposant, chaque chiffre étant traité comme une classe distincte. En conséquence, le LLM doit apprendre empiriquement la proximité numérique, par exemple, que 42.1 est plus proche de 42.2 que de 92.1, étant donné que tous les chiffres sont traités comme des classes séparées. Étant donné les efforts pionniers de Lalande et al. [2023] pour intégrer des valeurs numériques réelles dans des architectures *transformers* plutôt que de les traiter comme des tokens distincts, il existe un potentiel pour développer une méthode robuste pour incorporer des données tabulaires directement dans les LLM. Cette approche traiterait les données de manière invariante par colonne et ligne, la différenciant des modalités d'image comme suggéré par exemple par Kotelnikov et al. [2022].

[25] La génération automatisée de programmes informatiques suscite un intérêt industriel considérable et est susceptible de stimuler une innovation importante (voir, par exemple, Lin et al. [2024]).

Pré-entraînement SR. Pour améliorer la capacité du LLM à intégrer des relations complexes entre les modalités symboliques et de données, nous proposons un pré-entraînement des têtes symboliques et de données du LLM sur des tâches SR. Cette phase initiale de formation se concentrerait uniquement sur la modélisation de la relation entre les expressions symboliques et les données avant que le LLM ne soit entraîné sur un ensemble de données multimodales telles que des articles de recherche, qui incluent souvent du texte, des données numériques et des expressions mathématiques. De plus, un pré-entraînement de la tête neurale symbolique sur des problèmes mathématiques formels pourrait affiner davantage sa capacité à gérer des informations symboliques complexes. Cette formation de base devrait considérablement améliorer la compétence du LLM dans les tâches scientifiques où l'interprétation précise des données et la manipulation symbolique sont cruciales.

Perspectives pour les LLMs scientifiques

Tout au long de cette thèse, et particulièrement dans cette section, nous avons suggéré l'intégration d'un riche ensemble de connaissances préalables dans nos paradigmes symboliques, tels que la structure de graphe inhérente aux expressions mathématiques, la notation préfixe, et l'intégration de l'analyse dimensionnelle. Ces éléments, combinés à la nécessité d'ajuster plusieurs réalisations et à l'application du rasoir d'Occam pour favoriser des expressions concises, ont façonné notre approche actuelle. Cependant, l'évolution rapide des LLMs suggère que de nombreuses contraintes que nous avons méticuleusement encodées pourraient bientôt devenir redondantes, apprises implicitement par des modèles plus avancés.[26]

Les LLMs modernes, par exemple, n'exigent plus de règles explicites comme la notation préfixe pour générer des expressions mathématiques sensées. Les erreurs consistant à produire par exemple $a + +b$ au lieu de $a + b$ ou produire $(a+)b.c$ au lieu de $(a + b).c$ sont extrêmement rares, indiquant un bond significatif dans leur compréhension des règles syntaxiques sans programmation directe. Cette compétence naturelle soulève une question intrigante sur les capacités futures des modèles polyvalents : pourraient-ils, un jour, effectuer des tâches comme la régression symbolique directement clé en mains ? Imaginons un scénario où un LLM, simplement en traitant un jeu de données, pourrait générer de manière autonome une

[26] Cette situation fait écho à la résistance initiale rencontrée par les partisans des réseaux neuronaux boîte noire, qui contestaient l'encodage manuel des règles avec l'aide d'experts. Au lieu de cela, ils préconisaient des systèmes qui apprennent les règles directement à partir des données sans intervention humaine explicite, un changement qui a marqué un moment pivot dans l'histoire de l'intelligence artificielle [Schmidhuber, 2015].

expression analytique reflétant toutes les occurrences de données/expressions précédemment rencontrées, sans entraînement spécifique pour les tâches de SR.

Cette perspective reflète les capacités inattendues observées dans les premières versions de `GPT` [Radford et al., 2019], où le modèle a démontré une capacité à traduire de l'anglais au français, bien qu'il n'ait pas été explicitement entraîné dans cet objectif, en appliquant son apprentissage étendu de l'anglais aux quelques exemples français rencontrés pendant son entraînement. Ce type de transfert de connaissances apprises laisse entrevoir un avenir où les modèles avancés ne se contentent pas de répondre à leurs directives d'entraînement, mais les dépassent, en abordant des tâches complexes et imprévues [27].

Considérations Finales

En guise de conclusion, il est manifeste que, contrairement à d'autres domaines comme la vision par ordinateur, le contrôle ou l'informatique, la physique et l'astrophysique requièrent non seulement l'application des approches classiques d'apprentissage automatique mais exigent également l'adoption d'un paradigme supplémentaire d'apprentissage symbolique pour progresser efficacement à l'ère de l'abondance des données. Cette thèse propose ainsi un cadre de travail novateur et un ensemble de méthodologies destinées à étendre le paradigme de l'apprentissage machine symbolique au domaine de la physique. Nos stratégies tirent parti de nos confrontations à des problèmes astrophysiques tangibles. L'objectif principal de cette thèse est de forger une relation symbiotique entre le développement de ces nouvelles approches et la maximisation des retombées scientifiques des missions d'observation, notamment en ce qui concerne l'énigme de la matière noire, l'un des plus grands défis de la physique actuelle.

Le paysage actuel de l'apprentissage automatique est dominé par des applications industrielles offrant des capacités prédictives exceptionnelles mais souvent dépourvues de compréhensibilité et d'interprétabilité, des qualités essentielles dans les sciences naturelles. Compte tenu du contexte technologique actuel, notamment en ce qui concerne le traitement du langage, et de l'entrée

[27] Au-delà de simples traductions, `GPT4` a fait preuve de compétence dans diverses tâches complexes pour lesquelles il n'avait pas été explicitement formé, démontrant ainsi ses capacités de généralisation [de Wynter, 2024, Bubeck et al., 2023, Fan et al., 2022].

dans une ère caractérisée par une profusion de données en astrophysique, il est impératif de développer des outils d'apprentissage machine symbolique capables de générer des modèles analytiques intelligibles.

Ces méthodes présentent un potentiel indéniable pour enrichir diverses branches de la physique, mais leur applicabilité est particulièrement pertinente en astrophysique, compte tenu de l'abondance de données sans précédent dans ce domaine. Historiquement, l'astrophysique a souvent été à l'origine de nouvelles méthodes numériques qui ont ensuite bénéficié aux sciences naturelles dans leur ensemble. Plus que jamais, l'astrophysique est appelée à jouer un rôle de pionnier dans les sciences physiques, en relevant ces nouveaux défis cruciaux.

References

Isaac Newton. *Philosophiae Naturalis Principia Mathematica.* 1687. doi: 10.3931/e-rara-440.

Johann Kepler. *Astronomia Nova.* 1609.

Timo Gaia Collaboration, Prusti, JHJ De Bruijne, Anthony GA Brown, Antonella Vallenari, C Babusiaux, CAL Bailer-Jones, U Bastian, M Biermann, Dafydd Wyn Evans, L Eyer, et al. The gaia mission. *Astronomy & astrophysics*, 595:A1, 2016a.

R. Laureijs, J. Amiaux, S. Arduini, J. L. Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, B. Garilli, P. Gondoin, L. Guzzo, J. Hoar, H. Hoekstra, R. Holmes, T. Kitching, T. Maciaszek, Y. Mellier, F. Pasian, W. Percival, J. Rhodes, G. Saavedra Criado, M. Sauvage, R. Scaramella, L. Valenziano, S. Warren, R. Bender, F. Castander, A. Cimatti, O. Le Fèvre, H. Kurki-Suonio, M. Levi, P. Lilje, G. Meylan, R. Nichol, K. Pedersen, V. Popa, R. Rebolo Lopez, H. W. Rix, H. Rottgering, W. Zeilinger, F. Grupp, P. Hudelot, R. Massey, M. Meneghetti, L. Miller, S. Paltani, S. Paulin-Henriksson, S. Pires, C. Saxton, T. Schrabback, G. Seidel, J. Walsh, N. Aghanim, L. Amendola, J. Bartlett, C. Baccigalupi, J. P. Beaulieu, K. Benabed, J. G. Cuby, D. Elbaz, P. Fosalba, G. Gavazzi, A. Helmi, I. Hook, M. Irwin, J. P. Kneib, M. Kunz, F. Mannucci, L. Moscardini, C. Tao, R. Teyssier, J. Weller, G. Zamorani, M. R. Zapatero Osorio, O. Boulade, J. J. Foumond, A. Di Giorgio, P. Guttridge, A. James, M. Kemp, J. Martignac, A. Spencer, D. Walton, T. Blümchen, C. Bonoli, F. Bortoletto, C. Cerna, L. Corcione, C. Fabron, K. Jahnke, S. Ligori, F. Madrid, L. Martin, G. Morgante, T. Pamplona, E. Prieto, M. Riva, R. Toledo, M. Trifoglio, F. Zerbi, F. Abdalla, M. Douspis, C. Grenet, S. Borgani, R. Bouwens, F. Courbin, J. M. Delouis, P. Dubath, A. Fontana, M. Frailis, A. Grazian, J. Koppenhöfer, O. Mansutti, M. Melchior, M. Mignoli, J. Mohr, C. Neissner, K. Noddle, M. Poncet, M. Scodeggio, S. Serrano, N. Shane, J. L. Starck, C. Surace, A. Taylor, G. Verdoes-Kleijn, C. Vuerli, O. R. Williams, A. Zacchei, B. Altieri, I. Escudero Sanz, R. Kohley, T. Oosterbroek, P. Astier, D. Bacon, S. Bardelli, C. Baugh, F. Bellagamba, C. Benoist, D. Bianchi, A. Biviano, E. Branchini, C. Carbone, V. Cardone, D. Clements, S. Colombi, C. Conselice, G. Cresci, N. Deacon, J. Dunlop, C. Fedeli, F. Fontanot, P. Franzetti, C. Giocoli, J. Garcia-Bellido, J. Gow, A. Heavens, P. Hewett, C. Heymans, A. Holland, Z. Huang, O. Ilbert, B. Joachimi, E. Jennins, E. Kerins, A. Kiessling, D. Kirk, R. Kotak, O. Krause, O. Lahav, F. van Leeuwen, J. Lesgourgues, M. Lombardi, M. Magliocchetti, K. Maguire, E. Majerotto, R. Maoli, F. Marulli, S. Maurogordato, H. McCracken, R. McLure, A. Melchiorri, A. Merson, M. Moresco, M. Nonino, P. Norberg, J. Peacock, R. Pello, M. Penny, V. Pettorino, C. Di Porto, L. Pozzetti, C. Quercellini, M. Radovich, A. Rassat, N. Roche, S. Ronayette, E. Rossetti, B. Sartoris, P. Schneider, E. Semboloni, S. Serjeant, F. Simpson, C. Skordis, G. Smadja, S. Smartt, P. Spano, S. Spiro, M. Sullivan, A. Tilquin, R. Trotta, L. Verde, Y. Wang, G. Williger, G. Zhao, J. Zoubian, and E. Zucca. Euclid Definition Study Report. *arXiv e-prints*, art. arXiv:1110.3193, October 2011.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L.

Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, David J. Mills, Connor Miraval, Joachim Moeyens, Fred E. Moolekamp, David G. Monet, Marc Moniez, Serge Monkewitz, Christopher Montgomery, Christopher B. Morrison, Fritz Mueller, Gary P. Muller, Freddy Muñoz Arancibia, Douglas R. Neill, Scott P. Newbry, Jean-Yves Nief, Andrei Nomerotski, Martin Nordby, Paul O'Connor, John Oliver, Scot S. Olivier, Knut Olsen, William O'Mullane, Sandra Ortiz, Shawn Osier, Russell E. Owen, Reynald Pain, Paul E. Palecek, John K. Parejko, James B. Parsons, Nathan M. Pease, J. Matt Peterson, John R. Peterson, Donald L. Petravick, M. E. Libby Petrick, Cathy E. Petry, Francesco Pierfederici, Stephen Pietrowicz, Rob Pike, Philip A. Pinto, Raymond Plante, Stephen Plate, Joel P. Plutchak, Paul A. Price, Michael Prouza, Veljko Radeka, Jayadev Rajagopal, Andrew P. Rasmussen, Nicolas Regnault, Kevin A. Reil, David J. Reiss, Michael A. Reuter, Stephen T. Ridgway, Vincent J. Riot, Steve Ritz, Sean Robinson, William Roby, Aaron Roodman, Wayne Rosing, Cecille Roucelle, Matthew R. Rumore, Stefano Russo, Abhijit Saha, Benoit Sassolas, Terry L. Schalk, Pim Schellart, Rafe H. Schindler, Samuel Schmidt, Donald P. Schneider, Michael D. Schneider, William Schoening, German Schumacher, Megan E. Schwamb, Jacques Sebag, Brian Selvy, Glenn H. Sembroski, Lynn G. Seppala, Andrew Serio, Eduardo Serrano, Richard A. Shaw, Ian Shipsey, Jonathan Sick, Nicole Silvestri, Colin T. Slater, J. Allyn Smith, R. Chris Smith, Shahram Sobhani, Christine Soldahl, Lisa Storrie-Lombardi, Edward Stover, Michael A. Strauss, Rachel A. Street, Christopher W. Stubbs, Ian S. Sullivan, Donald Sweeney, John D. Swinbank, Alexander Szalay, Peter Takacs, Stephen A. Tether, Jon J. Thaler, John Gregg Thayer, Sandrine Thomas, Adam J. Thornton, Vaikunth Thukral, Jeffrey Tice, David E. Trilling, Max Turri, Richard Van Berg, Daniel Vanden Berk, Kurt Vetter, Francoise Virieux, Tomislav Vucina, William Wahl, Lucianne Walkowicz, Brian Walsh, Christopher W. Walter, Daniel L. Wang, Shin-Yawn Wang, Michael Warner, Oliver Wiecha, Beth Willman, Scott E. Winters, David Wittman, Sidney C. Wolff, W. Michael Wood-Vasey, Xiuqin Wu, Bo Xin, Peter Yoachim, and Hu Zhan. Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, mar 2019. doi: 10.3847/1538-4357/ab042c. URL https://dx.doi.org/10.3847/1538-4357/ab042c.

LSST Science Collaboration. Lsst science book, version 2.0, 2009.

C.L. Carilli and S. Rawlings. Motivation, key science projects, standards and assumptions. *New Astronomy Reviews*, 48(11):979–984, 2004. ISSN 1387-6473. doi: https://doi.org/10.1016/j.newar.2004.09.001. URL https://www.sciencedirect.com/science/article/pii/S1387647304000880. Science with the Square Kilometre Array.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

Galileo Galilei. *Il saggiatore*. 1623.

William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

Ana Bonaca and Adrian M. Price-Whelan. Stellar Streams in the Gaia Era. *arXiv e-prints*, art. arXiv:2405.19410, May 2024. doi: 10.48550/arXiv.2405.19410.

Ian Goodfellow. *Deep learning*. The MIT Press, Cambridge, Massachusetts, 2016. ISBN 9780262035613.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Papers With Code. Papers with code trends, 2023. URL https://paperswithcode.com/trends.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

S. Hunston. Corpus linguistics. In Keith Brown, editor, *Encyclopedia of Language and Linguistics (Second Edition)*, pages 234–248. Elsevier, Oxford, second edition edition, 2006. ISBN 978-0-08-044854-1. doi: https://doi.org/10.1016/B0-08-044854-2/00944-5.

Jon A Holtzman, Sten Hasselquist, Matthew Shetrone, Katia Cunha, Carlos Allende Prieto, Borja Anguiano, Dmitry Bizyaev, Jo Bovy, Andrew Casey, Bengt Edvardsson, et al. Apogee data releases 13 and 14: data and analysis. *The Astronomical Journal*, 156(3): 125, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020a.

Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

Xiaosheng Zhao, Yi Mao, Cheng Cheng, and Benjamin D Wandelt. Simulation-based inference of reionization parameters from 3d tomographic 21 cm light-cone images. *The Astrophysical Journal*, 926(2):151, 2022.

Jonathan Chardin, Grégoire Uhlrich, Dominique Aubert, Nicolas Deparis, Nicolas Gillet, Pierre Ocvirk, and Joseph Lewis. A deep learning model to emulate simulations of cosmic reionization. *Monthly Notices of the Royal Astronomical Society*, 490(1):1055–1065, 2019.

Emma Dodd, Thomas M. Callingham, Amina Helmi, Tadafumi Matsuno, Tomás Ruiz-Lara, Eduardo Balbinot, and Sofie Lövdal. Gaia DR3 view of dynamical substructure in the stellar halo near the Sun. *A&A*, 670:L2, February 2023. doi: 10.1051/0004-6361/202244546.

Rodrigo Ibata, Foivos I. Diakogiannis, Benoit Famaey, and Giacomo Monari. The ACTIONFINDER: An Unsupervised Deep Learning Algorithm for Calculating Actions and the Acceleration Field from a Set of Orbit Segments. *ApJ*, 915(1):5, July 2021. doi: 10.3847/1538-4357/abfda9.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf.

Yin Li, Libin Lu, Chirag Modi, Drew Jamieson, Yucheng Zhang, Yu Feng, Wenda Zhou, Ngai Pok Kwan, François Lanusse, and Leslie Greengard. pmwd: A differentiable cosmological particle-mesh n-body library. *arXiv preprint arXiv:2211.09958*, 2022.

Alexander Laroche and Joshua S Speagle. Closing the stellar labels gap: Stellar label independent evidence for [*alpha*/m] information in gaia bp/rp spectra. *arXiv preprint arXiv:2404.07316*, 2024.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87, 2020.

Max Tegmark. The Mathematical Universe. *Foundations of Physics*, 38(2):101–150, February 2008. doi: 10.1007/s10701-007-9186-9.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabricio de Franca, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason Moore. Contemporary symbolic regression methods and their relative performance. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c0c7c76d30bd3dcaefc96f40275bdc0a-Paper-round1.pdf.

Richard Phillips Feynman, Robert B Leighton, Matthew Sands, et al. *The Feynman lectures on physics*, volume 1-3. Addison-Wesley Reading, MA, 1971.

Casper Wilstrup and Jaan Kasak. Symbolic regression outperforms other models for small data sets. *arXiv preprint arXiv:2103.15147*, 2021.

Patrick AK Reinbold, Logan M Kageorge, Michael F Schatz, and Roman O Grigoriev. Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression. *Nature communications*, 12(1):1–8, 2021.

Subham Sahoo, Christoph Lampert, and Georg Martius. Learning equations for extrapolation and control. In *International Conference on Machine Learning*, pages 4442–4450. PMLR, 2018.

Pierre-Alexandre Kamienny, Stéphane d'Ascoli, Guillaume Lample, and Francois Charton. End-to-end symbolic regression with transformers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=GoOuIrDHG_Y.

Pierre-Alexandre Kamienny and Sylvain Lamprier. Symbolic-model-based reinforcement learning. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022. URL https://openreview.net/forum?id=yeF6cyYU7W.

Tailin Wu and Max Tegmark. Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100(3):033311, 2019.

Federico Sabbatini and Roberta Calegari. Evaluation metrics for symbolic knowledge extracted from machine learning black boxes: A discussion paper. *arXiv preprint arXiv:2211.00238*, 2022.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

European Commission. The artificial intelligence act, 2021. URL https://artificialintelligenceact.eu/.

117th US Congress. Algorithmic accountability act, 2022. URL https://www.congress.gov/bill/117th-congress/house-bill/6580/.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

Michael Schmidt and Hod Lipson. *Age-Fitness Pareto Optimization*, pages 129–146. Springer New York, New York, NY, 2011. ISBN 978-1-4419-7747-2. doi: 10.1007/978-1-4419-7747-2_8. URL https://doi.org/10.1007/978-1-4419-7747-2_8.

Matthew J Graham, SG Djorgovski, Ashish A Mahabal, Ciro Donalek, and Andrew J Drake. Machine-assisted discovery of relationships in astronomy. *Monthly Notices of the Royal Astronomical Society*, 431(3):2371–2384, 2013.

ME Thing and SM Koksbang. cp3-bench: a tool for benchmarking symbolic regression algorithms demonstrated with cosmology. *Journal of Cosmology and Astroparticle Physics*, 2025(01):040, 2025.

Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Fabricio Olivetti de Franca and Guilherme Seidyo Imai Aldeia. Interaction–transformation evolutionary algorithm for symbolic regression. *Evolutionary computation*, 29(3):367–390, 2021.

William La Cava, Thomas Helmuth, Lee Spector, and Jason H Moore. A probabilistic and multi-objective analysis of lexicase selection and ε-lexicase selection. *Evolutionary Computation*, 27(3):377–402, 2019.

William La Cava, Tilak Raj Singh, James Taggart, Srinivas Suri, and Jason Moore. Learning concise representations for regression by evolving networks of trees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hke-JhA9Y7.

Marco Virgolin, Tanja Alderliesten, and Peter AN Bosman. Linear scaling with and within semantic backpropagation-based genetic programming for symbolic regression. In *Proceedings of the genetic and evolutionary computation conference*, pages 1084–1092, 2019.

Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive

biases. *Advances in Neural Information Processing Systems*, 33:17429–17442, 2020b.

M. Virgolin, T. Alderliesten, C. Witteveen, and P. A. N. Bosman. Improving Model-Based Genetic Programming for Symbolic Regression of Small Expressions. *Evolutionary Computation*, 29(2):211–237, 06 2021. ISSN 1063-6560. doi: 10.1162/evco_a_00278. URL https://doi.org/10.1162/evco_a_00278.

Trevor Stephens. Gplearn, 2015. URL https://gplearn.readthedocs.io/en/stable/index.html.

Michael Kommenda, Bogdan Burlacu, Gabriel Kronberger, and Michael Affenzeller. Parameter identification for symbolic regression using nonlinear least squares. *Genetic Programming and Evolvable Machines*, 21(3):471–501, 2020.

Trent McConaghy. Ffx: Fast, scalable, deterministic symbolic regression technology. In *Genetic Programming Theory and Practice IX*, pages 235–260. Springer, 2011.

Lukas Kammerer, Gabriel Kronberger, Bogdan Burlacu, Stephan M Winkler, Michael Kommenda, and Michael Affenzeller. Symbolic regression by exhaustive search: reducing the search space using syntactical constraints and efficient semantic structure deduplication. In *Genetic Programming Theory and Practice XVII*, pages 79–99. Springer, 2020.

Deaglan J. Bartlett, Harry Desmond, and Pedro G. Ferreira. Exhaustive symbolic regression. *IEEE Transactions on Evolutionary Computation*, pages 1–1, 2023a. doi: 10.1109/TEVC.2023.3280250.

Jure Brence, Ljupčo Todorovski, and Sašo Džeroski. Probabilistic grammars for equation discovery. *Knowledge-Based Systems*, 224:107077, 2021.

Ying Jin, Weilin Fu, Jian Kang, Jiadong Guo, and Jian Guo. Bayesian symbolic regression. *arXiv preprint arXiv:1910.08892*, 2019.

Changtong Luo, Chen Chen, and Zonglin Jiang. Divide and conquer: A quick scheme for symbolic regression. *International Journal of Computational Methods*, 19(08):2142002, 2022.

Tony Tohme, Dehong Liu, and KAMAL YOUCEF-TOUMI. GSR: A generalized symbolic regression approach. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=lheUXtDNvP.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. SRSD: Rethinking datasets of symbolic regression for scientific discovery. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022. URL https://openreview.net/forum?id=oKwyEqClqkb.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. Ai feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems*, 33:4860–4871, 2020.

Florian Lalande, Yoshitomo Matsubara, Naoya Chiba, Tatsunori Taniai, Ryo Igarashi, and Yoshitala Ushiku. A transformer model for symbolic regression towards scientific discovery. *arXiv preprint arXiv:2312.04070*, 2023.

Luca Biggio, Tommaso Bendinelli, Aurelien Lucchi, and Giambattista Parascandolo. A seq2seq approach to symbolic regression. In *Learning Meets Combinatorial Algorithms at NeurIPS2020*, 2020. URL https://openreview.net/forum?id=W7jCKuyPn1.

Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, and Giambattista Parascandolo. Neural symbolic regression that scales. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 936–945. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/biggio21a.html.

Martin Vastl, Jonáš Kulhánek, Jirí Kubalík, Erik Derner, and Robert Babuška. Symformer: End-to-end symbolic regression using transformer-based architecture. *arXiv preprint arXiv:2205.15764*, 2022.

Stéphane d'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, and François Charton. Deep symbolic regression for recurrent sequences. *arXiv preprint arXiv:2201.04600*, 2022.

Pierre-Alexandre Kamienny, Guillaume Lample, Sylvain Lamprier, and Marco Virgolin. Deep generative symbolic regression with monte-carlo-tree-search. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15655–15668. PMLR, 2023. URL https://proceedings.mlr.press/v202/kamienny23a.html.

Tommaso Bendinelli, Luca Biggio, and Pierre-Alexandre Kamienny. Controllable neural symbolic regression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2063–2077. PMLR, 2023. URL https://proceedings.mlr.press/v202/bendinelli23a.html.

Samuel Holt, Zhaozhi Qian, and Mihaela van der Schaar. Deep generative symbolic regression. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=o7koEEMA1bR.

Y. Li, J. Liu, W. Li, L. Yu, M. Wu, W. Li, M. Hao, S. Wei, and Y. Deng. MMSR: Symbolic Regression is a Multimodal Task, 2024a.

Y. Li, W. Li, L. Yu, M. Wu, J. Liu, W. Li, M. Hao, S. Wei, and Y. Deng. Discovering Mathematical Formulas from Data via GPT-guided Monte Carlo Tree Search, 2024b.

T. Chen, P. Xu, and H. Zheng. Bootstrapping OTS-Funcimg Pre-training Model (Botfip) – A Comprehensive Symbolic Regression Framework, 2024a.

K. Meidani, P. Shojaee, C. K. Reddy, and A. B. Farimani. SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KZSEgJGPxu.

Sören Becker, Michal Klein, Alexander Neitz, Giambattista Parascandolo, and Niki Kilbertus. Discovering ordinary differential equations that govern time-series. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022. URL https://openreview.net/forum?id=vhrtZYgxLzV.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*, 2024.

Abdulhakim Alnuqaydan, Sergei Gleyzer, and Harrison Prosper. Symba: Symbolic computation of squared amplitudes in high energy physics with machine learning. *Machine Learning: Science and Technology*, 2022.

Nikos Aréchiga, Francine Chen, Yan-Ying Chen, Yanxia Zhang, Rumen Iliev, Heishiro Toyoda, and Kent Lyons. Accelerating understanding of scientific experiments with end to end symbolic regression. *arXiv preprint arXiv:2112.04023*, 2021.

Camilla Fiorini, Clément Flint, Louis Fostier, Emmanuel Franck, Reyhaneh Hashemi, Victor Michel-Dansac, and Wassim Tenachi. Generalizing the sindy approach with nested neural networks. *arXiv preprint arXiv:2404.15742*, 2024.

Philipp Scholl, Katharina Bieker, Hillary Hauger, and Gitta Kutyniok. Parfam–symbolic regression based on continuous global optimization. *arXiv preprint arXiv:2310.05537*, 2023.

Georg Martius and Christoph H Lampert. Extrapolation and learning equations, 2017. URL https://openreview.net/forum?id=BkgRp0FYe.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, mar 2016. doi: 10.1073/pnas.1517384113.

Wenqing Zheng, SP Sharan, Zhiwen Fan, Kevin Wang, Yihan Xi, and Zhangyang Wang. Symbolic visual reinforcement learning: A scalable framework with object-level abstraction and differentiable expression search. *arXiv preprint arXiv:2212.14849*, 2022.

Carlos Magno CO Valle and Sami Haddadin. Syrenets: Symbolic residual neural networks. *arXiv preprint arXiv:2105.14396*, 2021.

Samuel Kim, Peter Y Lu, Srijon Mukherjee, Michael Gilbert, Li Jing, Vladimir Čeperić, and Marin Soljačić. Integration of neural network-based symbolic regression in deep learning for scientific discovery. *IEEE transactions on neural networks and learning systems*, 32 (9):4166–4177, 2020.

Maysum Panju and Ali Ghodsi. A neuro-symbolic method for solving differential and functional equations. *arXiv preprint arXiv:2011.02415*, 2020.

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.*, 2:083802, Aug 2018. doi: 10.1103/PhysRevMaterials.2.083802. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083802.

Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A review. *arXiv preprint arXiv:2211.10873*, 2022.

Dimitrios Angelis, Filippos Sofos, and Theodoros E. Karakasidis. Artificial intelligence in physical sciences: Symbolic regression trends and perspectives. *Archives of Computational Methods in Engineering*, 30(6):3845–3865, April 2023. ISSN 1886-1784. doi: 10.1007/s11831-023-09922-z. URL http://dx.doi.org/10.1007/s11831-023-09922-z.

Brenden K Petersen, Mikel Landajuela Larma, Terrell N. Mundhenk, Claudio Prata Santiago, Soo Kyung Kim, and Joanne Taery Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=m5Qsh0kBQG.

Wassim Tenachi, Rodrigo Ibata, and Foivos I. Diakogiannis. Deep Symbolic Regression for Physics Guided by Units Constraints: Toward the Automated Discovery of Physical Laws. *ApJ*, 959(2):99, December 2023a. doi: 10.3847/1538-4357/ad014c.

W. Tenachi, R. Ibata, and F. I. Diakogiannis. Physical Symbolic Optimization. *arXiv e-prints*, art. arXiv:2312.03612, 2023b. doi: 10.48550/arXiv.2312.03612.

Wassim Tenachi, Rodrigo Ibata, Thibaut L. François, and Foivos I. Diakogiannis. Class Symbolic Regression: Gotta Fit 'Em All. *ApJL*, 969(2):L26, July 2024. doi: 10.3847/2041-8213/ad5970.

Mikel Landajuela, Brenden K Petersen, Soo K Kim, Claudio P Santiago, Ruben Glatt, T Nathan Mundhenk, Jacob F Pettit, and Daniel M Faissol. Improving exploration in policy gradient search: Application to symbolic optimization. In *1st Mathematical Reasoning in General Artificial Intelligence, International Conference on Learning Representations (ICLR)*, 2021a.

Mikel Landajuela, Brenden K Petersen, Sookyung Kim, Claudio P Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F Pettit, and Daniel Faissol. Discovering symbolic policies with deep reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5979–5989. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/landajuela21a.html.

Joanne T Kim, Mikel Landajuela, and Brenden K Petersen. Distilling wikipedia mathematical knowledge into neural network models. In *1st Mathematical Reasoning in General Artificial Intelligence, International Conference on Learning Representations (ICLR)*, 2021.

Brenden K Petersen, Claudio Santiago, and Mikel Landajuela. Incorporating domain knowledge into neural-guided search via in situ priors and constraints. In *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021b. URL https://openreview.net/

forum?id=yAis5yB9MQ.

Mikel Landajuela, Chak Shing Lee, Jiachen Yang, Ruben Glatt, Claudio P Santiago, Ignacio Aravena, Terrell Mundhenk, Garrett Mulcahy, and Brenden K Petersen. A unified framework for deep symbolic regression. *Advances in Neural Information Processing Systems*, 35:33985–33998, 2022.

J. G. Faris, C. F. Hayes, A. R. Goncalves, K. G. Sprenger, D. Faissol, B. K. Petersen, M. Landajuela, and F. Leno da Silva. Pareto Front Training For Multi-Objective Symbolic Optimization. In *The Sixteenth Workshop on Adaptive and Learning Agents*, 2024. URL https://openreview.net/forum?id=e0gswuNjcb.

Y. He, B. Sheng, and Z. Li. Channel Modeling Based on Transformer Symbolic Regression for Inter-Satellite Terahertz Communication. *Applied Sciences*, 14(7), 2024a. ISSN 2076-3417. doi: 10.3390/app14072929. URL https://www.mdpi.com/2076-3417/14/7/2929.

Zachary Bastiani, Robert M Kirby, Jacob Hochhalter, and Shandian Zhe. Complexity-aware deep symbolic regression with robust risk-seeking policy gradients. *arXiv preprint arXiv:2406.06751*, 2024.

Mengge Du, Yuntian Chen, and Dongxiao Zhang. Discover: Deep identification of symbolic open-form pdes via enhanced reinforcement-learning. *arXiv preprint arXiv:2210.02181*, 2022.

Y. Tian, W. Zhou, H. Dong, D. S. Kammer, and Olga Fink. Sym-Q: Adaptive Symbolic Regression via Sequential Decision-Making, 2024.

Y. Michishita. Alpha Zero for Physics: Application of Symbolic Regression with Alpha Zero to find the analytical methods in physics, 2024.

Daniel M DiPietro and Bo Zhu. Symplectically integrated symbolic regression of hamiltonian dynamical systems. *arXiv preprint arXiv:2209.01521*, 2022.

Muhammad Usama and In-Young Lee. Data-driven non-linear current controller based on deep symbolic regression for spmsm. *Sensors*, 22(21):8240, 2022.

OpenAI. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

Andrej Bauer, Matej Petković, and Ljupco Todorovski. MLFMF: Data sets for machine learning for mathematical formalization. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a. URL https://openreview.net/forum?id=KZjSvE2mJz.

Kaiyu Yang and Jia Deng. Learning to prove theorems via interacting with proof assistants. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6984–6994. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yang19a.html.

Rodrigo Ochigame. Automated mathematics and the reconfiguration of proof and labor. *Bulletin of the American Mathematical Society*, 61(3):423–437, May 2024. ISSN 1088-9485. doi: 10.1090/bull/1821. URL http://dx.doi.org/10.1090/bull/1821.

A. B. Kempe. On the geographical problem of the four colours. *American Journal of Mathematics*, 2(3):193–200, 1879. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2369235.

The Coq Development Team. The Coq reference manual – release 8.19.0. https://coq.inria.fr/doc/V8.19.0/refman, 2024.

Leonardo De Moura, Soonho Kong, Jeremy Avigad, Floris Van Doorn, and Jakob von Raumer. The lean theorem prover (system description). In *Automated Deduction-CADE-25: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25*, pages 378–388. Springer, 2015.

Gabriel Poesia, David Broman, Nick Haber, and Noah D Goodman. Learning formal mathematics from intrinsic motivation. *arXiv preprint arXiv:2407.00695*, 2024.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. A survey on deep learning for theorem proving, 2024c. URL https://arxiv.org/abs/2404.09939.

Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Een, Francois Chollet, and Josef Urban. Deepmath - deep sequence models for premise selection. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/f197002b9a0853eca5e046d9ca4663d5-Paper.pdf.

Miroslav Olšák, Cezary Kaliszyk, and Josef Urban. Property invariant embedding for automated reasoning. In *ECAI 2020*, pages 1395–1402. IOS Press, 2020.

Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving, 2020. URL https://arxiv.org/abs/2009.03393.

Jesse Michael Han, Jason Rute, Yuhuai Wu, Edward Ayers, and Stanislas Polu. Proof artifact co-training for theorem proving with language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=rpxJc9j04U.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. Formal mathematics statement curriculum learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-P7G-8dmSh4.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Guillaume Lample, Timothee Lacroix, Marie anne Lachaux, Aurelien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. Hypertree proof search for neural theorem proving. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=J4pX8Q8cxHH.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

James S Bullock and Michael Boylan-Kolchin. Small-scale challenges to the λ cdm paradigm. *Annual Review of Astronomy and Astrophysics*, 55:343–387, 2017.

J.H. Davenport, Y. Siret, and E. Tournier. *Computer Algebra: Systems and Algorithms for Algebraic Computation*. Academic Press, 1993. ISBN 9780122092329. URL https://books.google.fr/books?id=h9tQAAAAMAAJ.

Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

OpenAI. tiktoken is a fast bpe tokeniser for use with openai's models., 2023. URL https://github.com/openai/tiktoken.

Marco Virgolin and Solon P Pissis. Symbolic regression is NP-hard. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=LTiaPxqe2e.

Bezalel Gavish and Stephen C Graves. The travelling salesman problem and related problems. 1978.

Ahmed Stohy, Heba-Tullah Abdelhakam, Sayed Ali, Mohammed Elhenawy, Abdallah A Hassan, Mahmoud Masoud, Sebastien Glaser, and Andry Rakotonirainy. Hybrid pointer networks for traveling salesman problems optimization. *Plos one*, 16(12):e0260995, 2021.

Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel

Miranda, Jordi Pallarès, and Marta Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5):eaav6971, 2020. doi: 10.1126/sciadv.aav6971. URL https://www.science.org/doi/abs/10.1126/sciadv.aav6971.

Deaglan Bartlett, Harry Desmond, and Pedro Ferreira. Priors for symbolic regression. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, page 2402–2411, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701207. doi: 10.1145/3583133.3596327. URL https://doi.org/10.1145/3583133.3596327.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Jakob Bauer, Kate Baumli, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Satinder Singh, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei M Zhang. Human-timescale adaptation in an open-ended task space. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1887–1935. PMLR, 23–29 Jul 2023b. doi: 10.5555/3618408.3618488. URL https://proceedings.mlr.press/v202/bauer23a.html.

Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. EPOpt: Learning robust neural network policies using model ensembles. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SyWvgP5el.

Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

Stephen Wolfram. *The mathematica book*, volume 1. Wolfram Research, Inc., 2003.

Aaron Meurer, Christopher P Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K Moore, Sartaj Singh, et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017.

Ekaterina J. Vladislavleva, Guido F. Smits, and Dick den Hertog. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*, 13(2):333–349, 2009. doi: 10.1109/TEVC.2008.926486.

Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. Machine learning in the search for new fundamental physics. *Nature Reviews Physics*, 4 (6):399–412, 2022.

Mojtaba Valipour, Bowen You, Maysum Panju, and Ali Ghodsi. Symbolicgpt: A generative transformer model for symbolic regression. *arXiv preprint arXiv:2106.14131*, 2021.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL

https://openreview.net/forum?id=rc8o_j8I8PX.

Ryan Grindle. *Perils and pitfalls of symbolic regression.* The University of Vermont and State Agricultural College, 2021.

Konstantin T Matchev, Katia Matcheva, and Alexander Roman. Analytical modeling of exoplanet transit spectroscopy with dimensional analysis and symbolic regression. *The Astrophysical Journal*, 930(1):33, 2022.

Liron Simon Keren, Alex Liberzon, and Teddy Lazebnik. A computational framework for physics-informed symbolic regression with straightforward integration of domain knowledge. *Scientific Reports*, 13:1249, January 2023. doi: 10.1038/s41598-023-28328-2.

Muhammad Sarmad Ali, Meghana Kshirsagar, Enrique Naredo, and Conor Ryan. Automated grammar-based feature selection in symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '22, page 902–910, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392372. doi: 10.1145/3512290.3528852. URL https://doi.org/10.1145/3512290.3528852.

Laure Crochepierre, Lydia Boudjeloud-Assala, and Vincent Barbesant. A reinforcement learning approach to domain-knowledge inclusion using grammar guided symbolic regression. *arXiv preprint arXiv:2202.04367*, 2022.

Michael F Korns. Abstract expression grammar symbolic regression. *Genetic programming theory and practice VIII*, pages 109–128, 2011.

N.X. Hoai, R.I. McKay, D. Essam, and R. Chau. Solving the symbolic regression problem with tree-adjunct grammar guided genetic programming: the comparative results. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, volume 2, pages 1326–1331 vol.2, 2002. doi: 10.1109/CEC.2002.1004435.

Daniel Manrique, Juan Ríos, and Alfonso Rodríguez-Patón. Grammar-guided genetic programming. *Encyclopedia of Artificial Intelligence*, pages 767–773, 2009.

Tony Worm and Kenneth Chiu. Prioritized grammar enumeration: Symbolic regression by dynamic programming. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, GECCO '13, page 1021–1028, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319638. doi: 10.1145/2463372.2463486. URL https://doi.org/10.1145/2463372.2463486.

Edgar Buckingham. On physically similar systems; illustrations of the use of dimensional equations. *Physical review*, 4(4):345, 1914.

Thomas AR Purcell, Matthias Scheffler, and Luca M Ghiringhelli. Recent advances in the sisso method and their implementation in the sisso++ code. *arXiv preprint arXiv:2305.01242*, 2023.

Jure Brence, Sašo Džeroski, and Ljupčo Todorovski. Dimensionally-consistent equation discovery through probabilistic attribute grammars. *Information Sciences*, 632:742–756, 2023. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2023.03.073. URL https://www.sciencedirect.com/science/article/pii/S0020025523003705.

Maximilian Reissmann, Yuan Fang, Andrew Ooi, and Richard Sandberg. Constraining genetic symbolic regression via semantic backpropagation. *arXiv preprint arXiv:2409.07369*, 2024.

H. Goldstein, C.P. Poole, and J.L. Safko. *Classical Mechanics.* Addison Wesley, 2002. ISBN 9780201657029. URL https://books.google.fr/books?id=tJCuQgAACAAJ.

J.D. Jackson. *Classical Electrodynamics.* Wiley, 2012. ISBN 9788126510948. URL https://books.google.fr/books?id=8qHCZjJHRUgC.

S. Weinberg. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity.* Wiley, 1972. ISBN 9780471925675. URL https://books.google.fr/books?id=XLbvAAAAMAAJ.

M.D. Schwartz. *Quantum Field Theory and the Standard Model.* Quantum Field Theory and the Standard Model. Cambridge University Press, 2014. ISBN 9781107034730. URL

https://books.google.fr/books?id=HbdEAgAAQBAJ.

Ignacio Arnaldo, Krzysztof Krawiec, and Una-May O'Reilly. Multiple regression genetic programming. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, GECCO '14, page 879–886, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326629. doi: 10.1145/2576768.2598291. URL https://doi.org/10.1145/2576768.2598291.

William La Cava, Kourosh Danai, and Lee Spector. Inference of compact nonlinear dynamic models by epigenetic local search. *Engineering Applications of Artificial Intelligence*, 55: 292–306, 2016.

Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *arXiv preprint arXiv: 2409.09359*, 2024.

Bogdan Burlacu. Gecco'2022 symbolic regression competition: Post-analysis of the operon framework. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, page 2412–2419, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701207. doi: 10.1145/3583133.3596390. URL https://doi.org/10.1145/3583133.3596390.

D. M. Scolnic, D. O. Jones, A. Rest, Y. C. Pan, R. Chornock, R. J. Foley, M. E. Huber, R. Kessler, G. Narayan, A. G. Riess, S. Rodney, E. Berger, D. J. Brout, P. J. Challis, M. Drout, D. Finkbeiner, R. Lunnan, R. P. Kirshner, N. E. Sanders, E. Schlafly, S. Smartt, C. W. Stubbs, J. Tonry, W. M. Wood-Vasey, M. Foley, J. Hand, E. Johnson, W. S. Burgett, K. C. Chambers, P. W. Draper, K. W. Hodapp, N. Kaiser, R. P. Kudritzki, E. A. Magnier, N. Metcalfe, F. Bresolin, E. Gall, R. Kotak, M. McCrum, and K. W. Smith. The Complete Light-curve Sample of Spectroscopically Confirmed SNe Ia from Pan-STARRS1 and Cosmological Constraints from the Combined Pantheon Sample. *ApJ*, 859(2):101, June 2018. doi: 10.3847/1538-4357/aab9bb.

James Binney and Scott Tremaine. *Galactic dynamics*, volume 13. Princeton university press, 2011.

Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. The Structure of Cold Dark Matter Halos. *ApJ*, 462:563, May 1996. doi: 10.1086/177173.

I. Marinescu, Y. Strittmatter, C. Williams, and S. Musslick. Expression Sampler as a Dynamic Benchmark for Symbolic Regression. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL https://openreview.net/forum?id=i3PecpoiPG.

Rodrigo Ibata, Khyati Malhan, Nicolas Martin, Dominique Aubert, Benoit Famaey, Paolo Bianchini, Giacomo Monari, Arnaud Siebert, Guillaume F. Thomas, Michele Bellazzini, Piercarlo Bonifacio, Elisabetta Caffau, and Florent Renaud. Charting the galactic acceleration field. i. a search for stellar streams with gaia dr2 and edr3 with follow-up from espadons and uves. *ApJ*, 914(2):123, jun 2021. doi: 10.3847/1538-4357/abfcc2. URL https://dx.doi.org/10.3847/1538-4357/abfcc2.

Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. A Universal Density Profile from Hierarchical Clustering. *ApJ*, 490(2):493–508, December 1997. doi: 10.1086/304888.

Ewa L. Łokas and Gary A. Mamon. Properties of spherical galaxies and clusters with an NFW density profile. *MNRAS*, 321(1):155–166, February 2001. doi: 10.1046/j.1365-8711.2001.04007.x.

Etienne Russeil, Fabrício Olivetti de França, Konstantin Malanchev, Bogdan Burlacu, Emille E. O. Ishida, Marion Leroux, Clément Michelin, Guillaume Moinard, and Emmanuel Gangler. Multi-view symbolic regression, 2024.

Wassim Tenachi, Rodrigo Ibata, and Foivos I Diakogiannis. Symbolic regression driven by dimensional analysis for the automated discovery of physical laws and constants of nature. In *Journées 2023 de la Société Française d'Astronomie & d'Astrophysique (SF2A).*, pages 107–108, Strasbourg, France, June 2023a. URL https://hal.science/hal-04325284.

Robert C Martin. *Clean code: a handbook of agile software craftsmanship*. Pearson Education, 2009.

David Melching, Florian Paysan, Tobias Strohmann, and Eric Breitbarth. A universal crack tip correction algorithm discovered by physical deep symbolic regression, 2024. URL https://arxiv.org/abs/2403.10320.

Qingliang Li, Cheng Zhang, Zhongwang Wei, Xiaochun Jin, Wei Shangguan, Hua Yuan, Jinlong Zhu, Lu Li, Pingping Liu, Xiao Chen, et al. Advancing symbolic regression for earth science with a focus on evapotranspiration modeling. *npj Climate and Atmospheric Science*, 7(1):321, 2024d.

Yi Xie, Tianyu Qiu, Yun Xiong, Xiuqi Huang, Xiaofeng Gao, and Chao Chen. An efficient and generalizable symbolic regression method for time series analysis. *arXiv preprint arXiv:2409.03986*, 2024.

Yuanzhi He, Biao Sheng, and Zhiqiang Li. Channel modeling based on transformer symbolic regression for inter-satellite terahertz communication. *Applied Sciences*, 14(7), 2024b. ISSN 2076-3417. doi: 10.3390/app14072929. URL https://www.mdpi.com/2076-3417/14/7/2929.

Taimur Rahman, Shamima Sultana, Tanjir Ahmed, Md Farhad Momin, and Afra Anam Provasha. Deep symbolic regression for numerical formulation of fundamental period in concentrically steel-braced rc frames. *Asian Journal of Civil Engineering*, pages 1–20, 2024.

Shijun Cheng and Tariq Alkhalifah. Discovery of physically interpretable wave equations. *arXiv preprint arXiv:2404.17971*, 2024.

Lena Podina, Diba Darooneh, Joshveer Grewal, and Mohammad Kohandel. Enhancing symbolic regression and universal physics-informed neural networks with dimensional analysis. *arXiv preprint arXiv: 2411.15919*, 2024.

Clément Stahl, Nicolas Mai, Benoit Famaey, Yohan Dubois, and Rodrigo Ibata. From inflation to dark matter halo profiles: the impact of primordial non-gaussianities on the central density cusp. *Journal of Cosmology and Astroparticle Physics*, 2024(05):021, 2024.

EMC Sijben, JC Jansen, Peter AN Bosman, and Tanja Alderliesten. Function class learning with genetic programming: Towards explainable meta learning for tumor growth functionals. *arXiv preprint arXiv:2402.12510*, 2024.

Xia Chen, Alexander Rex, Janis Woelke, Christoph Eckert, Boris Bensmann, Richard Hanke-Rauschenbach, and Philipp Geyer. Machine learning in proton exchange membrane water electrolysis—a knowledge-integrated framework. *Applied Energy*, 371:123550, 2024b.

Xiao-Yun Wang, Chen Dong, and Xiang Liu. Analysis of strong coupling constant with machine learning and its application. *Chinese Physics Letters*, 41(3):031201, 2024. doi: 10.1088/0256-307X/41/3/031201. URL https://cpl.iphy.ac.cn/EN/abstract/article_116517.shtml.

Mengge Du, Yuntian Chen, Longfeng Nie, Siyu Lou, and Dongxiao Zhang. Physics-constrained robust learning of open-form partial differential equations from limited and noisy data. *Physics of Fluids*, 36(5):057123, 2024.

Eunhye Shin, Jinseop Jang, and Junghyo Jo. Physics education and symbolic regression. *New Physics: Sae Mulli*, 74(7):678–687, 07 2024. doi: 10.3938/NPSM.74.678. URL http://www.npsm-kps.org/journal/view.html?doi=10.3938/NPSM.74.678.

Pei Li, Joo-Ho Choi, Dingyang Zhang, Shuyou Zhang, and Yiming Zhang. Reinforced symbolic learning with logical constraints for predicting turbine blade fatigue life. *arXiv preprint arXiv:2412.03580*, 2024e.

Teppei Kurita, Yuhi Kondo, Legong Sun, Takayuki Sasaki, Sho Nitta, Yasuhiro Hashimoto, Yoshinori Muramatsu, and Yusuke Moriuchi. Revisiting disparity from dual-pixel images: Physics-informed lightweight depth estimation. *arXiv preprint arXiv:2411.04714*, 2024.

Boqian Zhang and Juanmian Lei. Unit-constrained data-driven turbulence modeling for separated flows using symbolic regression. *arXiv preprint arXiv:2405.08656*, 2024.

Changtong Luo, Chen Chen, and Zonglin Jiang. A divide and conquer method for symbolic regression, 2017. URL https://arxiv.org/abs/1705.08061.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Ananth Ranganathan. The levenberg-marquardt algorithm. *Tutoral on LM algorithm*, 11 (1):101–110, 2004.

John Taylor, Wenyi Wang, Biswajit Bala, and Tomasz Bednarz. Optimizing the optimizer for data driven deep neural networks and physics informed neural networks, 2022. URL https://arxiv.org/abs/2205.07430.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

Planck Collaboration. Planck 2015 results. *Astronomy & Astrophysics*, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830. URL https://doi.org/10.1051/0004-6361/201525830.

Marius Cautun, Alejandro Benitez-Llambay, Alis J Deason, Carlos S Frenk, Azadeh Fattahi, Facundo A Gómez, Robert JJ Grand, Kyle A Oman, Julio F Navarro, and Christine M Simpson. The milky way total mass profile as inferred from gaia dr2. *Monthly Notices of the Royal Astronomical Society*, 494(3):4291–4313, 2020.

Thomas M Callingham, Marius Cautun, Alis J Deason, Carlos S Frenk, Wenting Wang, Facundo A Gómez, Robert JJ Grand, Federico Marinacci, and Ruediger Pakmor. The mass of the milky way from satellite dynamics. *Monthly Notices of the Royal Astronomical Society*, 484(4):5453–5467, 2019.

Vera C Rubin and W Kent Ford Jr. Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *Astrophysical Journal, vol. 159, p. 379*, 159:379, 1970.

Matthias Bartelmann. Gravitational lensing. *Classical and Quantum Gravity*, 27(23):233001, 2010.

Douglas Clowe, Anthony Gonzalez, and Maxim Markevitch. Weak-lensing mass reconstruction of the interacting cluster 1e 0657–558: Direct evidence for the existence of dark matter. *The Astrophysical Journal*, 604(2):596, 2004.

George F Smoot, Charles L Bennett, A Kogut, Edward L Wright, Jon Aymon, Nancy W Boggess, Edward S Cheng, G De Amici, S Gulkis, MG Hauser, et al. Structure in the cobe differential microwave radiometer first-year maps. *Astrophysical Journal, Part 2-Letters (ISSN 0004-637X), vol. 396, no. 1, Sept. 1, 1992, p. L1-L5. Research supported by NASA.*, 396:L1–L5, 1992.

R Agnese, Alan J Anderson, M Asai, D Balakishiyeva, R Basu Thakur, DA Bauer, J Beaty, J Billard, A Borgland, MA Bowles, et al. Search for low-mass weakly interacting massive particles with supercdms. *Physical review letters*, 112(24):241302, 2014.

J Kopecky, J-Ch Sublet, JA Simpson, RA Forrest, and D Nierop. Atlas of neutron capture cross sections. Technical report, International Atomic Energy Agency, 1997.

Jules Gascon. Direct dark matter searches review. In *EPJ Web of Conferences*, volume 95, page 02004. EDP Sciences, 2015.

Eugene Oks. Brief review of recent advances in understanding dark matter and dark energy. *New Astronomy Reviews*, 93:101632, 2021.

Elena Aprile, K Arisaka, F Arneodo, A Askin, L Baudis, A Behrens, E Brown, JMR Cardoso, B Choi, D Cline, et al. The xenon100 dark matter experiment. *Astroparticle Physics*, 35 (9):573–590, 2012.

Martin F Sohnius. Introducing supersymmetry. *Physics reports*, 128(2-3):39–204, 1985.

Martin Schmaltz and David Tucker-Smith. Little higgs theories. *Annu. Rev. Nucl. Part. Sci.*, 55:229–270, 2005.

Jihn E Kim and Gianpaolo Carosi. Axions and the strong c p problem. *Reviews of Modern Physics*, 82(1):557, 2010.

Jonathan L Feng. Dark matter candidates from particle physics and methods of detection. *Annual Review of Astronomy and Astrophysics*, 48:495–545, 2010.

Mark Vogelsberger, Shy Genel, Volker Springel, Paul Torrey, Debora Sijacki, Dandan Xu, G Snyder, Simeon Bird, Dylan Nelson, and Lars Hernquist. Properties of galaxies reproduced by a hydrodynamic simulation. *Nature*, 509(7499):177–182, 2014.

Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, and et al. The eagle project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, Nov 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2058. URL http://dx.doi.org/10.1093/mnras/stu2058.

Ruth Durrer. *The cosmic microwave background*. Cambridge University Press, 2020.

Amandine Doliva-Dolinsky, Nicolas F. Martin, Guillaume F. Thomas, Annette M. N. Ferguson, Rodrigo A. Ibata, Geraint F. Lewis, Dougal Mackey, Alan W. McConnachie, and Zhen Yuan. The pandas view of the andromeda satellite system. iii. dwarf galaxy detection limits. *The Astrophysical Journal*, 933(2):135, jul 2022. doi: 10.3847/1538-4357/ac6fd5. URL https://dx.doi.org/10.3847/1538-4357/ac6fd5.

Giuseppina Battaglia and Carlo Nipoti. Stellar dynamics and dark matter in local group dwarf galaxies. *Nature Astronomy*, 6(6):659–672, 2022.

Michael Boylan-Kolchin, James S Bullock, and Manoj Kaplinghat. Too big to fail? the puzzling darkness of massive milky way subhaloes. *Monthly Notices of the Royal Astronomical Society: Letters*, 415(1):L40–L44, 2011.

Go Ogiya and Andreas Burkert. Re-examining the too-big-to-fail problem for dark matter haloes with central density cores. *Monthly Notices of the Royal Astronomical Society*, 446 (3):2363–2369, 2015.

Fangzhou Jiang and Frank C van den Bosch. Comprehensive assessment of the too big to fail problem. *Monthly Notices of the Royal Astronomical Society*, 453(4):3575–3592, 2015.

Isabel ME Santos-Santos, Julio F Navarro, Andrew Robertson, Alejandro Benítez-Llambay, Kyle A Oman, Mark R Lovell, Carlos S Frenk, Aaron D Ludlow, Azadeh Fattahi, and Adam Ritz. Baryonic clues to the puzzling diversity of dwarf galaxy rotation curves. *Monthly Notices of the Royal Astronomical Society*, 495(1):58–77, 2020.

Benoit Famaey and Stacy S McGaugh. Modified newtonian dynamics (mond): observational phenomenology and relativistic extensions. *Living reviews in relativity*, 15:1–159, 2012.

Pavel Kroupa, Christian Theis, and Christian M Boily. The great disk of milky-way satellites and cosmological sub-structures. *Astronomy & Astrophysics*, 431(2):517–521, 2005.

Rodrigo A Ibata, Geraint F Lewis, Anthony R Conn, Michael J Irwin, Alan W McConnachie, Scott C Chapman, Michelle L Collins, Mark Fardal, Annette MN Ferguson, Neil G Ibata, et al. A vast, thin plane of corotating dwarf galaxies orbiting the andromeda galaxy. *Nature*, 493(7430):62–65, 2013.

Jenna Samuel, Andrew Wetzel, Sierra Chapman, Erik Tollerud, Philip F Hopkins, Michael Boylan-Kolchin, Jeremy Bailin, and Claude-André Faucher-Giguère. Planes of satellites around milky way/m31-mass galaxies in the fire simulations and comparisons with the

local group. *Monthly Notices of the Royal Astronomical Society*, 504(1):1379–1397, 2021.

Yang-Shyang Li and Amina Helmi. Infall of substructures on to a milky way-like dark halo. *Monthly Notices of the Royal Astronomical Society*, 385(3):1365–1373, 2008.

Nicolas Garavito-Camargo, Ekta Patel, Gurtina Besla, Adrian M Price-Whelan, Facundo A Gomez, Chervin FP Laporte, and Kathryn V Johnston. The clustering of orbital poles induced by the lmc: hints for the origin of planes of satellites. *The Astrophysical Journal*, 923(2):140, 2021.

Marcel S Pawlowski and Stacy S McGaugh. Co-orbiting planes of sub-halos are similarly unlikely around paired and isolated hosts. *The Astrophysical Journal Letters*, 789(1):L24, 2014.

Marla Geha, Risa H Wechsler, Yao-Yuan Mao, Erik J Tollerud, Benjamin Weiner, Rebecca Bernstein, Ben Hoyle, Sebastian Marchi, Phil J Marshall, Ricardo Muñoz, et al. The saga survey. i. satellite galaxy populations around eight milky way analogs. *The Astrophysical Journal*, 847(1):4, 2017.

Yao-Yuan Mao, Marla Geha, Risa H Wechsler, Benjamin Weiner, Erik J Tollerud, Ethan O Nadler, and Nitya Kallivayalil. The saga survey. ii. building a statistical sample of satellite systems around milky way–like galaxies. *The Astrophysical Journal*, 907(2):85, 2021.

Till Sawala, Marius Cautun, Carlos Frenk, John Helly, Jens Jasche, Adrian Jenkins, Peter H Johansson, Guilhem Lavaux, Stuart McAlpine, and Matthieu Schaller. The milky way's plane of satellites is consistent with λ cdm. *Nature Astronomy*, 7(4):481–491, 2023.

Laura V Sales, Andrew Wetzel, and Azadeh Fattahi. Baryonic solutions and challenges for cosmological models of dwarf galaxies. *Nature Astronomy*, 6(8):897–910, 2022.

Mark R. Lovell, Carlos S. Frenk, Vincent R. Eke, Adrian Jenkins, Liang Gao, and Tom Theuns. The properties of warm dark matter haloes. *Monthly Notices of the Royal Astronomical Society*, 439(1):300–317, 02 2014. ISSN 0035-8711. doi: 10.1093/mnras/stt2431. URL https://doi.org/10.1093/mnras/stt2431.

Annika HG Peter. Dark matter: a brief review. *arXiv preprint arXiv:1201.3942*, 2012.

David N Spergel and Paul J Steinhardt. Observational evidence for self-interacting cold dark matter. *Physical review letters*, 84(17):3760, 2000.

M. Milgrom. A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *ApJ*, 270:365–370, July 1983. doi: 10.1086/161130.

J. Bekenstein and M. Milgrom. Does the missing mass problem signal the breakdown of Newtonian gravity? *ApJ*, 286:7–14, November 1984. doi: 10.1086/162570.

P. A. Oria, B. Famaey, G. F. Thomas, R. Ibata, J. Freundlich, L. Posti, M. Korsaga, G. Monari, O. Müller, N. I. Libeskind, and M. S. Pawlowski. The Phantom Dark Matter Halos of the Local Volume in the Context of Modified Newtonian Dynamics. *ApJ*, 923 (1):68, December 2021. doi: 10.3847/1538-4357/ac273d.

Garry W Angus, Benoit Famaey, and HongSheng Zhao. Can mond take a bullet? analytical comparisons of three versions of mond beyond spherical symmetry. *Monthly Notices of the Royal Astronomical Society*, 371(1):138–146, 2006.

Sebastiano Ghigna, Ben Moore, Fabio Governato, George Lake, Thomas Quinn, and Joachim Stadel. Dark matter haloes within clusters. *Monthly Notices of the Royal Astronomical Society*, 300(1):146–162, 1998.

Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. A universal density profile from hierarchical clustering. *The Astrophysical Journal*, 490(2):493–508, Dec 1997. ISSN 1538-4357. doi: 10.1086/304888. URL http://dx.doi.org/10.1086/304888.

J. Einasto. On the Construction of a Composite Model for the Galaxy and on the Determination of the System of Galactic Parameters. *Trudy Astrofizicheskogo Instituta Alma-Ata*, 5:87–100, January 1965.

A. Burkert. The Structure of Dark Matter Halos in Dwarf Galaxies. *ApJL*, 447:L25–L28, July 1995. doi: 10.1086/309560.

Paul J. McMillan. The mass distribution and gravitational potential of the Milky Way. *MNRAS*, 465(1):76–94, February 2017a. doi: 10.1093/mnras/stw2759.

Eugene Vasiliev. Agama: action-based galaxy modelling architecture. *Monthly Notices of the Royal Astronomical Society*, 482(2):1525–1544, 2019.

Houjun Mo, Frank C. van den Bosch, and Simon White. *Galaxy Formation and Evolution*. 2010.

Leonard Searle and Robert Zinn. Compositions of halo clusters and the formation of the galactic halo. *Astrophysical Journal, Part 1, vol. 225, Oct. 15, 1978, p. 357-379.*, 225: 357–379, 1978.

Alis J. Deason and Vasily Belokurov. Galactic archaeology with gaia. *New Astronomy Reviews*, 99:101706, 2024. ISSN 1387-6473. doi: https://doi.org/10.1016/j.newar.2024.101706. URL https://www.sciencedirect.com/science/article/pii/S1387647324000137.

Joss Bland-Hawthorn and Ortwin Gerhard. The Galaxy in Context: Structural, Kinematic, and Integrated Properties. *ARAA*, 54:529–596, September 2016. doi: 10.1146/annurev-astro-081915-023441.

Françoise Combes, Stéphane Leon, and Georges Meylan. N-body simulations of globular cluster tides. *A&A*, 352:149–162, December 1999. doi: 10.48550/arXiv.astro-ph/9910148.

D Lynden-Bell and RM Lynden-Bell. Ghostly streams from the formation of the galaxy's halo. *Monthly Notices of the Royal Astronomical Society*, 275(2):429–442, 1995.

Joshua D. Simon and Marla Geha. The kinematics of the ultra-faint milky way satellites: Solving the missing satellite problem. *The Astrophysical Journal*, 670(1):313, nov 2007. doi: 10.1086/521816. URL https://dx.doi.org/10.1086/521816.

Benjamin P Moster, Rachel S Somerville, Christian Maulbetsch, Frank C Van Den Bosch, Andrea V Macciò, Thorsten Naab, and Ludwig Oser. Constraints on the relationship between stellar mass and halo mass at low and high redshift. *The Astrophysical Journal*, 710(2):903, 2010.

Jacob Nibauer, Ana Bonaca, and Kathryn V Johnston. Constraining the gravitational potential from the projected morphology of extragalactic tidal streams. *The Astrophysical Journal*, 954(2):195, 2023.

Michal Bílek, Pierre-Alain Duc, Jean-Charles Cuillandre, Stephen Gwyn, Michele Cappellari, David V Bekaert, Paolo Bonfini, Theodoros Bitsakis, Sanjaya Paudel, Davor Krajnović, et al. Census and classification of low-surface-brightness structures in nearby early-type galaxies from the matlas survey. *Monthly Notices of the Royal Astronomical Society*, 498(2):2138–2166, 2020.

Elisabeth Sola, Pierre-Alain Duc, Felix Richards, Adeline Paiement, Mathias Urbano, Julie Klehammer, Michal Bílek, Jean-Charles Cuillandre, Stephen Gwyn, and Alan McConnachie. Characterization of low surface brightness structures in annotated deep images. *A&A*, 662:A124, June 2022. doi: 10.1051/0004-6361/202142675.

Amandine Doliva-Dolinsky, Nicolas F. Martin, Zhen Yuan, Alessandro Savino, Daniel R. Weisz, Annette M. N. Ferguson, Rodrigo A. Ibata, Stacy Y. Kim, Geraint F. Lewis, Alan W. McConnachie, and Guillaume F. Thomas. The pandas view of the andromeda satellite system. iv. global properties. *The Astrophysical Journal*, 952(1):72, jul 2023. doi: 10.3847/1538-4357/acdcf6. URL https://dx.doi.org/10.3847/1538-4357/acdcf6.

Rodrigo A Ibata, Gerry Gilmore, and MJ Irwin. A dwarf satellite galaxy in sagittarius. *Nature*, 370(6486):194–196, 1994.

Rodrigo Ibata, Michael Irwin, Geraint F Lewis, and Andrea Stolte. Galactic halo substructure in the sloan digital sky survey: The ancient tidal stream from the sagittarius dwarf galaxy. *The Astrophysical Journal*, 547(2):L133, 2001.

Amina Helmi. Velocity trends in the debris of sagittarius and the shape of the dark matter halo of our galaxy. *The Astrophysical Journal*, 610(2):L97, 2004.

Kathryn V Johnston, David R Law, and Steven R Majewski. A two micron all sky survey view of the sagittarius dwarf galaxy. iii. constraints on the flattening of the galactic halo. *The Astrophysical Journal*, 619(2):800, 2005.

David R Law and Steven R Majewski. The sagittarius dwarf galaxy: a model for evolution in a triaxial milky way halo. *The Astrophysical Journal*, 714(1):229, 2010.

Pierre-Antoine Oria, Rodrigo Ibata, Pau Ramos, Benoit Famaey, and Raphaël Errani. Revisiting a disky origin for the faint branch of the sagittarius stellar stream. *The Astrophysical Journal Letters*, 932(2):L14, 2022a.

Steven R Majewski, William E Kunkel, David R Law, Richard J Patterson, Allyson A Polak, Helio J Rocha-Pinto, Jeffrey D Crane, Peter M Frinchaboy, Cameron B Hummels, Kathryn V Johnston, et al. A two micron all sky survey view of the sagittarius dwarf galaxy. ii. swope telescope spectroscopy of m giant stars in the dynamically cold sagittarius tidal stream. *The Astronomical Journal*, 128(1):245, 2004.

Ken Freeman and Joss Bland-Hawthorn. The New Galaxy: Signatures of Its Formation. *ARAA*, 40:487–537, January 2002. doi: 10.1146/annurev.astro.40.060401.093840.

Giacomo Fragione and Abraham Loeb. Constraining the milky way mass with hypervelocity stars. *New Astronomy*, 55:32–38, 2017.

James E. Brau. The Milky Way Galaxy, 2001. URL https://pages.uoregon.edu/jimbrau/astr123-2001/Notes/Chapter23.html.

Andreas HW Küpper, Eduardo Balbinot, Ana Bonaca, Kathryn V Johnston, David W Hogg, Pavel Kroupa, and Basilio X Santiago. Globular cluster streams as galactic high-precision scales—the poster child palomar 5. *The Astrophysical Journal*, 803(2):80, 2015.

Carl J Grillmair and Odysseas Dionatos. Detection of a 63 cold stellar stream in the sloan digital sky survey. *The Astrophysical Journal*, 643(1):L17, 2006.

Facundo A Gómez, Gurtina Besla, Daniel D Carpintero, Álvaro Villalobos, Brian W O'Shea, and Eric F Bell. And yet it moves: the dangers of artificially fixing the milky way center of mass in the presence of a massive large magellanic cloud. *The Astrophysical Journal*, 802(2):128, 2015.

Gaia Collaboration. The gaia mission. *arXiv preprint arXiv:1609.04153*, 2016.

Gaia Collaboration. Gaia data release 3. summary of the content and survey properties. *Astronomy & Astrophysics*, June 2022. doi: 10.1051/0004-6361/202243940. URL https://doi.org/10.1051/0004-6361/202243940.

Danny Horta, Adrian M. Price-Whelan, David W. Hogg, Kathryn V. Johnston, Lawrence Widrow, Julianne J. Dalcanton, Melissa K. Ness, and Jason A. S. Hunt. Orbital torus imaging: Acceleration, density, and dark matter in the galactic disk measured with element abundance gradients. *The Astrophysical Journal*, 962(2):165, feb 2024. doi: 10.3847/1538-4357/ad16e8. URL https://dx.doi.org/10.3847/1538-4357/ad16e8.

Khyati Malhan and Rodrigo A Ibata. Streamfinder–i. a new algorithm for detecting stellar streams. *Monthly Notices of the Royal Astronomical Society*, 477(3):4063–4076, 2018.

Sergey E. Koposov, Hans-Walter Rix, and David W. Hogg. Constraining the Milky Way Potential with a Six-Dimensional Phase-Space Map of the GD-1 Stellar Stream. *ApJ*, 712 (1):260–273, March 2010. doi: 10.1088/0004-637X/712/1/260.

Jason L. Sanders and James Binney. Stream-orbit misalignment - II. A new algorithm to constrain the Galactic potential. *MNRAS*, 433(3):1826–1836, August 2013. doi: 10.1093/mnras/stt816.

Khyati Malhan and Rodrigo A. Ibata. Constraining the Milky Way halo potential with the GD-1 stellar stream. *MNRAS*, 486(3):2995–3005, July 2019. doi: 10.1093/mnras/stz1035.

Jacob Nibauer, Vasily Belokurov, Miles Cranmer, Jeremy Goodman, and Shirley Ho. Charting Galactic Accelerations with Stellar Streams and Machine Learning. *ApJ*, 940(1):22, November 2022. doi: 10.3847/1538-4357/ac93ee.

Vasily Belokurov, Denis Erkal, NW Evans, SE Koposov, and Alis J Deason. Co-formation of the disc and the stellar halo. *Monthly Notices of the Royal Astronomical Society*, 478 (1):611–619, 2018.

GC Myeong, E Vasiliev, G Iorio, NW Evans, and Vasily Belokurov. Evidence for two early accretion events that built the milky way stellar halo. *Monthly Notices of the Royal Astronomical Society*, 488(1):1235–1247, 2019.

Amina Helmi, Carine Babusiaux, Helmer H. Koppelman, Davide Massari, Jovan Veljanoski, and Anthony G. A. Brown. The merger that led to the formation of the Milky Way's inner stellar halo and thick disk. *Nat*, 563(7729):85–88, October 2018. doi: 10.1038/s41586-018-0625-x.

P Ramos, T Antoja, and F Figueras. Riding the kinematic waves in the milky way disk with gaia. *Astronomy & Astrophysics*, 619:A72, 2018.

T Antoja, A Helmi, M Romero-Gómez, D Katz, C Babusiaux, Ronald Drimmel, DW Evans, F Figueras, Eloisa Poggio, C Reylé, et al. A dynamically young and perturbed milky way disk. *Nature*, 561(7723):360–362, 2018.

Khyati Malhan, Rodrigo A. Ibata, and Nicolas F. Martin. Ghostly tributaries to the Milky Way: charting the halo's stellar streams with the Gaia DR2 catalogue. *MNRAS*, 481(3): 3442–3455, December 2018. doi: 10.1093/mnras/sty2474.

Khyati Malhan, Rodrigo A. Ibata, Sanjib Sharma, Benoit Famaey, Michele Bellazzini, Raymond G. Carlberg, Richard D'Souza, Zhen Yuan, Nicolas F. Martin, and Guillaume F. Thomas. The global dynamical atlas of the milky way mergers: Constraints from gaia edr3–based orbits of globular clusters, stellar streams, and satellite galaxies. *ApJ*, 926 (2):107, feb 2022. doi: 10.3847/1538-4357/ac4d2a. URL https://dx.doi.org/10.3847/1538-4357/ac4d2a.

Lina Necib, Bryan Ostdiek, Mariangela Lisanti, Timothy Cohen, Marat Freytsis, and Shea Garrison-Kimmel. Chasing accreted structures within gaia dr2 using deep learning. *The Astrophysical Journal*, 903(1):25, oct 2020. doi: 10.3847/1538-4357/abb814. URL https://dx.doi.org/10.3847/1538-4357/abb814.

David Shih, Matthew R Buckley, Lina Necib, and John Tamanas. via machinae: Searching for stellar streams using unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*, 509(4):5992–6007, 11 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab3372. URL https://doi.org/10.1093/mnras/stab3372.

Mariel Pettee, Sowmya Thanvantri, Benjamin Nachman, David Shih, Matthew R Buckley, and Jack H Collins. Weakly supervised anomaly detection in the Milky Way. *Monthly Notices of the Royal Astronomical Society*, 527(3):8459–8474, 11 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad3663. URL https://doi.org/10.1093/mnras/stad3663.

Rodrigo Ibata, Khyati Malhan, Nicolas Martin, Dominique Aubert, Benoit Famaey, Paolo Bianchini, Giacomo Monari, Arnaud Siebert, Guillaume F. Thomas, Michele Bellazzini, Piercarlo Bonifacio, Elisabetta Caffau, and Florent Renaud. Charting the Galactic Acceleration Field. I. A Search for Stellar Streams with Gaia DR2 and EDR3 with Follow-up from ESPaDOnS and UVES. *ApJ*, 914(2):123, June 2021a. doi: 10.3847/1538-4357/abfcc2.

AC Robin, X Luri, C Reylé, Y Isasi, E Grux, S Blanco-Cuaresma, F Arenou, C Babusiaux, M Belcheva, R Drimmel, et al. Gaia universe model snapshot-a statistical analysis of the expected contents of the gaia catalogue. *Astronomy & Astrophysics*, 543:A100, 2012.

Amina Helmi, Simon D. M. White, P. Tim de Zeeuw, and Hongsheng Zhao. Debris streams in the solar neighbourhood as relicts from the formation of the Milky Way. *Nat*, 402 (6757):53–55, November 1999. doi: 10.1038/46980.

Rodrigo A. Ibata, Khyati Malhan, and Nicolas F. Martin. The streams of the gaping abyss:

A population of entangled stellar streams surrounding the inner galaxy. *The Astrophysical Journal*, 872(2):152, feb 2019. doi: 10.3847/1538-4357/ab0080. URL https://dx.doi.org/10.3847/1538-4357/ab0080.

Akshara Viswanathan, Else Starkenburg, Helmer H Koppelman, Amina Helmi, Eduardo Balbinot, and Anna F Esselink. Hidden deep in the halo: selection of a reduced proper motion halo catalogue and mining retrograde streams in the velocity space. *Monthly Notices of the Royal Astronomical Society*, 521(2):2087–2102, 03 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad380. URL https://doi.org/10.1093/mnras/stad380.

Wassim Tenachi, Pierre-Antoine Oria, Rodrigo Ibata, Benoit Famaey, Zhen Yuan, Anke Arentsen, Nicolas Martin, and Akshara Viswanathan. Typhon: A polar stream from the outer halo raining through the solar neighborhood. *ApJL*, 935(2):L22, aug 2022. doi: 10.3847/2041-8213/ac874f. URL https://dx.doi.org/10.3847/2041-8213/ac874f.

Pierre-Antoine Oria, Wassim Tenachi, Rodrigo Ibata, Benoit Famaey, Zhen Yuan, Anke Arentsen, Nicolas Martin, and Akshara Viswanathan. Antaeus: A retrograde group of tidal debris in the milky way's disk plane. *ApJL*, 936(1):L3, aug 2022b. doi: 10.3847/2041-8213/ac86d3. URL https://dx.doi.org/10.3847/2041-8213/ac86d3.

Lina Necib, Bryan Ostdiek, Mariangela Lisanti, Timothy Cohen, Marat Freytsis, Shea Garrison-Kimmel, Philip F. Hopkins, Andrew Wetzel, and Robyn Sanderson. Evidence for a vast prograde stellar stream in the solar vicinity. *Nature Astronomy*, 4:1078–1083, July 2020. doi: 10.1038/s41550-020-1131-2.

John Dubinski, J Christopher Mihos, and Lars Hernquist. Constraining dark halo potentials with tidal tails. *The Astrophysical Journal*, 526(2):607, 1999.

RA Ibata, GF Lewis, MJ Irwin, and T Quinn. Uncovering cold dark matter halo substructure with tidal streams. *Monthly Notices of the Royal Astronomical Society*, 332(4):915–920, 2002.

Raymond G Carlberg. Dark matter sub-halo counts via star stream crossings. *The Astrophysical Journal*, 748(1):20, 2012.

Ana Bonaca, David W Hogg, Adrian M Price-Whelan, and Charlie Conroy. The spur and the gap in gd-1: Dynamical evidence for a dark substructure in the milky way halo. *The Astrophysical Journal*, 880(1):38, 2019.

Anatoly Klypin, Andrey V. Kravtsov, Octavio Valenzuela, and Francisco Prada. Where are the missing galactic satellites? *The Astrophysical Journal*, 522(1):82, sep 1999. doi: 10.1086/307643. URL https://dx.doi.org/10.1086/307643.

V. Belokurov, D. B. Zucker, N. W. Evans, G. Gilmore, S. Vidrih, D. M. Bramich, H. J. Newberg, R. F. G. Wyse, M. J. Irwin, M. Fellhauer, P. C. Hewett, N. A. Walton, M. I. Wilkinson, N. Cole, B. Yanny, C. M. Rockosi, T. C. Beers, E. F. Bell, J. Brinkmann, Ž. Ivezić, and R. Lupton. The Field of Streams: Sagittarius and Its Siblings. *ApJL*, 642 (2):L137–L140, May 2006. doi: 10.1086/504797.

N. Shipp, A. Drlica-Wagner, E. Balbinot, P. Ferguson, D. Erkal, T. S. Li, K. Bechtol, V. Belokurov, B. Buncher, D. Carollo, M. Carrasco Kind, K. Kuehn, J. L. Marshall, A. B. Pace, E. S. Rykoff, I. Sevilla-Noarbe, E. Sheldon, L. Strigari, A. K. Vivas, B. Yanny, A. Zenteno, T. M. C. Abbott, F. B. Abdalla, S. Allam, S. Avila, E. Bertin, D. Brooks, D. L. Burke, J. Carretero, F. J. Castander, R. Cawthon, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, C. Davis, J. De Vicente, S. Desai, H. T. Diehl, P. Doel, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. Hartley, K. Honscheid, B. Hoyle, D. J. James, M. D. Johnson, E. Krause, N. Kuropatkin, O. Lahav, H. Lin, M. A. G. Maia, M. March, P. Martini, F. Menanteau, C. J. Miller, R. Miquel, R. C. Nichol, A. A. Plazas, A. K. Romer, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, R. Schindler, M. Schubnell, M. Smith, R. C. Smith, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, A. R. Walker, R. H. Wechsler, and DES Collaboration. Stellar Streams Discovered in the Dark Energy Survey. *ApJ*, 862(2):114, August 2018. doi: 10.3847/1538-4357/aacdab.

Rodrigo Ibata, Khyati Malhan, Nicolas Martin, Dominique Aubert, Benoit Famaey, Paolo Bianchini, Giacomo Monari, Arnaud Siebert, Guillaume F. Thomas, Michele Bellazzini, Piercarlo Bonifacio, Elisabetta Caffau, and Florent Renaud. Charting the Galactic Acceleration Field. I. A Search for Stellar Streams with Gaia DR2 and EDR3 with Follow-up from ESPaDOnS and UVES. *ApJ*, 914(2):123, June 2021b. doi: 10.3847/1538-4357/abfcc2.

Khyati Malhan, Rodrigo A. Ibata, Sanjib Sharma, Benoit Famaey, Michele Bellazzini, Raymond G. Carlberg, Richard D'Souza, Zhen Yuan, Nicolas F. Martin, and Guillaume F. Thomas. The Global Dynamical Atlas of the Milky Way Mergers: Constraints from Gaia EDR3-based Orbits of Globular Clusters, Stellar Streams, and Satellite Galaxies. *ApJ*, 926(2):107, February 2022a. doi: 10.3847/1538-4357/ac4d2a.

Amina Helmi. The stellar halo of the Galaxy. *A&AR*, 15(3):145–188, June 2008. doi: 10.1007/s00159-008-0009-6.

Gaia Collaboration. Gaia Early Data Release 3. Summary of the contents and survey properties. *A&A*, 649:A1, May 2021. doi: 10.1051/0004-6361/202039657.

Amina Helmi, Simon D. M. White, and Volker Springel. The phase-space structure of cold dark matter haloes: insights into the Galactic halo. *MNRAS*, 339(3):834–848, March 2003. doi: 10.1046/j.1365-8711.2003.06227.x.

A. Recio-Blanco, P. de Laverny, P. A. Palicio, G. Kordopatis, M. A. Álvarez, M. Schultheis, G. Contursi, H. Zhao, G. Torralba Elipe, C. Ordenovic, M. Manteiga, C. Dafonte, I. Oreshina-Slezak, A. Bijaoui, Y. Fremat, G. Seabroke, F. Pailler, E. Spitoni, E. Poggio, O. L. Creevey, A. Abreu Aramburu, S. Accart, R. Andrae, C. A. L. Bailer-Jones, I. Bellas-Velidis, N. Brouillet, E. Brugaletta, A. Burlacu, R. Carballo, L. Casamiquela, A. Chiavassa, W. J. Cooper, A. Dapergolas, L. Delchambre, T. E. Dharmawardena, R. Drimmel, B. Edvardsson, M. Fouesneau, D. Garabato, P. Garcia-Lario, M. Garcia-Torres, A. Gavel, A. Gomez, I. Gonzalez-Santamaria, D. Hatzidimitriou, U. Heiter, A. Jean-Antoine Piccolo, M. Kontizas, A. J. Korn, A. C. Lanzafame, Y. Lebreton, Y. Le Fustec, E. L. Licata, H. E. P. Lindstrom, E. Livanou, A. Lobel, A. Lorca, A. Magdaleno Romeo, F. Marocco, D. J. Marshall, N. Mary, C. Nicolas, L. Pallas-Quintela, C. Panem, B. Pichon, F. Riclet, C. Robin, J. Rybizki, R. Santovena, A. Silvelo, R. L. Smart, L. M. Sarro, R. Sordo, C. Soubiran, M. Suvege, A. Ulla, A. Vallenari, J. Zorec, E. Utrilla, and J. Bakker. Gaia Data Release 3: Analysis of RVS spectra using the General Stellar Parametriser from spectroscopy. *arXiv e-prints*, art. arXiv:2206.05541, June 2022.

Jo Bovy. A purely acceleration-based measurement of the fundamental Galactic parameters. *arXiv e-prints*, art. arXiv:2012.02169, December 2020.

A. Widmark, P. F. de Salas, and G. Monari. Weighing the Galactic disk in sub-regions of the solar neighbourhood using Gaia DR2. *A&A*, 646:A67, February 2021. doi: 10.1051/0004-6361/202039852.

Ralph Schönrich, James Binney, and Walter Dehnen. Local kinematics and the local standard of rest. *MNRAS*, 403(4):1829–1833, April 2010. doi: 10.1111/j.1365-2966.2010.16253.x.

Eugene Vasiliev. AGAMA: action-based galaxy modelling architecture. *MNRAS*, 482(2):1525–1544, January 2019a. doi: 10.1093/mnras/sty2672.

John Illingworth and Josef Kittler. A survey of the hough transform. *Computer vision, graphics, and image processing*, 44(1):87–116, 1988.

A. C. Robin, X. Luri, C. Reylé, Y. Isasi, E. Grux, S. Blanco-Cuaresma, F. Arenou, C. Babusiaux, M. Belcheva, R. Drimmel, C. Jordi, A. Krone-Martins, E. Masana, J. C. Mauduit, F. Mignard, N. Mowlavi, B. Rocca-Volmerange, P. Sartoretti, E. Slezak, and A. Sozzetti. Gaia Universe model snapshot. A statistical analysis of the expected contents of the Gaia catalogue. *A&A*, 543:A100, July 2012a. doi: 10.1051/0004-6361/201118646.

Chun Wang, Yang Huang, Haibo Yuan, Huawei Zhang, Maosheng Xiang, and Xiaowei Liu. The Value-added Catalog for LAMOST DR8 Low-resolution Spectra. *ApJS*, 259(2):51, April 2022a. doi: 10.3847/1538-4365/ac4df7.

Alessandro Bressan, Paola Marigo, Léo. Girardi, Bernardo Salasnich, Claudia Dal Cero, Stefano Rubele, and Ambra Nanni. PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code. *MNRAS*, 427(1):127–145, Nov 2012. doi: 10.1111/j.1365-2966.2012.21948.x.

F. Anders, A. Khalatyan, A. B. A. Queiroz, C. Chiappini, J. Ardèvol, L. Casamiquela, F. Figueras, Ó. Jiménez-Arranz, C. Jordi, M. Monguió, M. Romero-Gómez, D. Altamirano, T. Antoja, R. Assaad, T. Cantat-Gaudin, A. Castro-Ginard, H. Enke, L. Girardi, G. Guiglion, S. Khan, X. Luri, A. Miglio, I. Minchev, P. Ramos, B. X. Santiago, and M. Steinmetz. Photo-astrometric distances, extinctions, and astrophysical parameters for Gaia EDR3 stars brighter than G = 18.5. *A&A*, 658:A91, February 2022. doi: 10.1051/0004-6361/202142369.

Pavel Kroupa. On the variation of the initial mass function. *MNRAS*, 322(2):231–246, April 2001. doi: 10.1046/j.1365-8711.2001.04022.x.

Zhen Yuan, G. C. Myeong, Timothy C. Beers, N. W. Evans, Young Sun Lee, Projjwal Banerjee, Dmitrii Gudin, Kohei Hattori, Haining Li, Tadafumi Matsuno, Vinicius M. Placco, M. C. Smith, Devin D. Whitten, and Gang Zhao. Dynamical Relics of the Ancient Galactic Halo. *ApJ*, 891(1):39, March 2020a. doi: 10.3847/1538-4357/ab6ef7.

Emma Dodd, Thomas M. Callingham, Amina Helmi, Tadafumi Matsuno, Tomás Ruiz-Lara, Eduardo Balbinot, and Sofie Lövdal. The Gaia DR3 view of dynamical substructure in the stellar halo near the Sun. *arXiv e-prints*, art. arXiv:2206.11248, June 2022a.

Alexander P. Ji, Rohan P. Naidu, Kaley Brauer, Yuan-Sen Ting, and Joshua D. Simon. Chemical Abundances of the Typhon Stellar Stream. *arXiv e-prints*, art. arXiv:2207.04016, July 2022.

Evan N. Kirby, Judith G. Cohen, Puragra Guhathakurta, Lucy Cheng, James S. Bullock, and Anna Gallazzi. The Universal Stellar Mass-Stellar Metallicity Relation for Dwarf Galaxies. *ApJ*, 779(2):102, December 2013a. doi: 10.1088/0004-637X/779/2/102.

S. E. Koposov, V. Belokurov, T. S. Li, C. Mateu, D. Erkal, C. J. Grillmair, D. Hendel, A. M. Price-Whelan, C. F. P. Laporte, K. Hawkins, S. T. Sohn, A. del Pino, N. W. Evans, C. T. Slater, N. Kallivayalil, J. F. Navarro, and Orphan Aspen Treasury Collaboration. Piercing the Milky Way: an all-sky view of the Orphan Stream. *MNRAS*, 485(4):4726–4742, June 2019. doi: 10.1093/mnras/stz457.

S. L. J. Gibbons, V. Belokurov, and N. W. Evans. A tail of two populations: chemo-dynamics of the Sagittarius stream and implications for its original mass. *MNRAS*, 464(1):794–809, January 2017. doi: 10.1093/mnras/stw2328.

Mark A. Fardal, Roeland P. van der Marel, Sangmo Tony Sohn, and Andres del Pino Molina. The course of the Orphan Stream in the Northern Galactic hemisphere traced with Gaia DR2. *MNRAS*, 486(1):936–949, June 2019. doi: 10.1093/mnras/stz749.

Jo Bovy. galpy: A python Library for Galactic Dynamics. *ApJS*, 216(2):29, February 2015a. doi: 10.1088/0067-0049/216/2/29.

T. Marchetti, E. M. Rossi, and A. G. A. Brown. Gaia DR2 in 6D: searching for the fastest stars in the Galaxy. *MNRAS*, 490(1):157–171, November 2019. doi: 10.1093/mnras/sty2592.

Eugene Vasiliev, Vasily Belokurov, and Denis Erkal. Tango for three: Sagittarius, LMC, and the Milky Way. *MNRAS*, 501(2):2279–2304, February 2021. doi: 10.1093/mnras/staa3673.

Amina Helmi and P. Tim de Zeeuw. Mapping the substructure in the Galactic halo with the next generation of astrometric satellites. *MNRAS*, 319(3):657–665, December 2000. doi: 10.1046/j.1365-8711.2000.03895.x.

Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, L. Eyer, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, D. J. Milli-

gan, C. Panem, V. Poinsignon, D. Pourbaix, S. Randich, G. Sarri, P. Sartoretti, H. I. Siddiqui, C. Soubiran, V. Valette, F. van Leeuwen, N. A. Walton, C. Aerts, F. Arenou, M. Cropper, R. Drimmel, E. Høg, D. Katz, M. G. Lattanzi, W. O'Mullane, E. K. Grebel, A. D. Holland, C. Huc, X. Passot, L. Bramante, C. Cacciari, J. Castañeda, L. Chaoul, N. Cheek, F. De Angeli, C. Fabricius, R. Guerra, J. Hernández, A. Jean-Antoine-Piccolo, E. Masana, R. Messineo, N. Mowlavi, K. Nienartowicz, D. Ordóñez-Blanco, P. Panuzzo, J. Portell, P. J. Richards, M. Riello, G. M. Seabroke, P. Tanga, F. Thévenin, J. Torra, S. G. Els, G. Gracia-Abril, G. Comoretto, M. Garcia-Reinaldos, T. Lock, E. Mercier, M. Altmann, R. Andrae, T. L. Astraatmadja, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, G. Busso, B. Carry, A. Cellino, G. Clementini, S. Cowell, O. Creevey, J. Cuypers, M. Davidson, J. De Ridder, A. de Torres, L. Delchambre, A. Dell'Oro, C. Ducourant, Y. Frémat, M. García-Torres, E. Gosset, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, M. Hauser, D. Hestroffer, S. T. Hodgkin, H. E. Huckle, A. Hutton, G. Jasniewicz, S. Jordan, M. Kontizas, A. J. Korn, A. C. Lanzafame, M. Manteiga, A. Moitinho, K. Muinonen, J. Osinde, E. Pancino, T. Pauwels, J. M. Petit, A. Recio-Blanco, A. C. Robin, L. M. Sarro, C. Siopis, M. Smith, K. W. Smith, A. Sozzetti, W. Thuillot, W. van Reeven, Y. Viala, U. Abbas, A. Abreu Aramburu, S. Accart, J. J. Aguado, P. M. Allan, W. Allasia, G. Altavilla, M. A. Álvarez, J. Alves, R. I. Anderson, A. H. Andrei, E. Anglada Varela, E. Antiche, T. Antoja, S. Antón, B. Arcay, A. Atzei, L. Ayache, N. Bach, S. G. Baker, L. Balaguer-Núñez, C. Barache, C. Barata, A. Barbier, F. Barblan, M. Baroni, D. Barrado y Navascués, M. Barros, M. A. Barstow, U. Becciani, M. Bellazzini, G. Bellei, A. Bello García, V. Belokurov, P. Bendjoya, A. Berihuete, L. Bianchi, O. Bienaymé, F. Billebaud, N. Blagorodnova, S. Blanco-Cuaresma, T. Boch, A. Bombrun, R. Borrachero, S. Bouquillon, G. Bourda, H. Bouy, A. Bragaglia, M. A. Breddels, N. Brouillet, T. Brüsemeister, B. Bucciarelli, F. Budnik, P. Burgess, R. Burgon, A. Burlacu, D. Busonero, R. Buzzi, E. Caffau, J. Cambras, H. Campbell, R. Cancelliere, T. Cantat-Gaudin, T. Carlucci, J. M. Carrasco, M. Castellani, P. Charlot, J. Charnas, P. Charvet, F. Chassat, A. Chiavassa, M. Clotet, G. Cocozza, R. S. Collins, P. Collins, G. Costigan, F. Crifo, N. J. G. Cross, M. Crosta, C. Crowley, C. Dafonte, Y. Damerdji, A. Dapergolas, P. David, M. David, P. De Cat, F. de Felice, P. de Laverny, F. De Luise, R. De March, D. de Martino, R. de Souza, J. Debosscher, E. del Pozo, M. Delbo, A. Delgado, H. E. Delgado, F. di Marco, P. Di Matteo, S. Diakite, E. Distefano, C. Dolding, S. Dos Anjos, P. Drazinos, J. Durán, Y. Dzigan, E. Ecale, B. Edvardsson, H. Enke, M. Erdmann, D. Escolar, M. Espina, N. W. Evans, G. Eynard Bontemps, C. Fabre, M. Fabrizio, S. Faigler, A. J. Falcão, M. Farràs Casas, F. Faye, L. Federici, G. Fedorets, J. Fernández-Hernández, P. Fernique, A. Fienga, F. Figueras, F. Filippi, K. Findeisen, A. Fonti, M. Fouesneau, E. Fraile, M. Fraser, J. Fuchs, R. Furnell, M. Gai, S. Galleti, L. Galluccio, D. Garabato, F. García-Sedano, P. Garé, A. Garofalo, N. Garralda, P. Gavras, J. Gerssen, R. Geyer, G. Gilmore, S. Girona, G. Giuffrida, M. Gomes, A. González-Marcos, J. González-Núñez, J. J. González-Vidal, M. Granvik, A. Guerrier, P. Guillout, J. Guiraud, A. Gúrpide, R. Gutiérrez-Sánchez, L. P. Guy, R. Haigron, D. Hatzidimitriou, M. Haywood, U. Heiter, A. Helmi, D. Hobbs, W. Hofmann, B. Holl, G. Holland, J. A. S. Hunt, A. Hypki, V. Icardi, M. Irwin, G. Jevardat de Fombelle, P. Jofré, P. G. Jonker, A. Jorissen, F. Julbe, A. Karampelas, A. Kochoska, R. Kohley, K. Kolenberg, E. Kontizas, S. E. Koposov, G. Kordopatis, P. Koubsky, A. Kowalczyk, A. Krone-Martins, M. Kudryashova, I. Kull, R. K. Bachchan, F. Lacoste-Seris, A. F. Lanza, J. B. Lavigne, C. Le Poncin-Lafitte, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, V. Lemaitre, H. Lenhardt, F. Leroux, S. Liao, E. Licata, H. E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, W. Löffler, M. López, A. Lopez-Lozano, D. Lorenz, T. Loureiro, I. MacDonald, T. Magalhães Fernandes, S. Managau, R. G. Mann, G. Mantelet, O. Marchal, J. M. Marchant, M. Marconi, J. Marie, S. Marinoni, P. M. Marrese, G. Marschalkó, D. J. Marshall, J. M. Martín-Fleitas, M. Martino, N. Mary, G. Matijevič, T. Mazeh, P. J. McMillan, S. Messina, A. Mestre, D. Michalik, N. R. Millar, B. M. H. Miranda, D. Molina, R. Molinaro, M. Molinaro, L. Molnár, M. Moniez, P. Montegriffo, D. Monteiro, R. Mor, A. Mora, R. Morbidelli, T. Morel, S. Morgenthaler, T. Morley, D. Morris, A. F. Mulone, T. Muraveva, I. Musella, J. Narbonne, G. Nelemans, L. Nicastro, L. Noval, C. Ordénovic, J. Ordieres-Meré, P. Osborne, C. Pagani, I. Pagano, F. Pailler, H. Palacin, L. Palaversa, P. Parsons, T. Paulsen, M. Pecoraro, R. Pedrosa, H. Pentikäinen, J. Pereira, B. Pichon, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poujoulet, A. Prša, L. Pulone, S. Ragaini, S. Rago, N. Rambaux, M. Ramos-Lerate, P. Ranalli, G. Rauw, A. Read, S. Regibo, F. Renk, C. Reylé, R. A. Ribeiro, L. Rimoldini, V. Ripepi, A. Riva, G. Rixon,

M. Roelens, M. Romero-Gómez, N. Rowell, F. Royer, A. Rudolph, L. Ruiz-Dern, G. Sadowski, T. Sagristà Sellés, J. Sahlmann, J. Salgado, E. Salguero, M. Sarasso, H. Savietto, A. Schnorhk, M. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Segransan, E. Serpell, I. C. Shih, R. Smareglia, R. L. Smart, C. Smith, E. Solano, F. Solitro, R. Sordo, S. Soria Nieto, J. Souchay, A. Spagna, F. Spoto, U. Stampa, I. A. Steele, H. Steidelmüller, C. A. Stephenson, H. Stoev, F. F. Suess, M. Süveges, J. Surdej, L. Szabados, E. Szegedi-Elek, D. Tapiador, F. Taris, G. Tauran, M. B. Taylor, R. Teixeira, D. Terrett, B. Tingley, S. C. Trager, C. Turon, A. Ulla, E. Utrilla, G. Valentini, A. van Elteren, E. Van Hemelryck, M. van Leeuwen, M. Varadi, A. Vecchiato, J. Veljanoski, T. Via, D. Vicente, S. Vogt, H. Voss, V. Votruba, S. Voutsinas, G. Walmsley, M. Weiler, K. Weingrill, D. Werner, T. Wevers, G. Whitehead, Ł. Wyrzykowski, A. Yoldas, M. Žerjal, S. Zucker, C. Zurbach, T. Zwitter, A. Alecu, M. Allen, C. Allende Prieto, A. Amorim, G. Anglada-Escudé, V. Arsenijevic, S. Azaz, P. Balm, M. Beck, H. H. Bernstein, L. Bigot, A. Bijaoui, C. Blasco, M. Bonfigli, G. Bono, S. Boudreault, A. Bressan, S. Brown, P. M. Brunet, P. Bunclark, R. Buonanno, A. G. Butkevich, C. Carret, C. Carrion, L. Chemin, F. Chéreau, L. Corcione, E. Darmigny, K. S. de Boer, P. de Teodoro, P. T. de Zeeuw, C. Delle Luche, C. D. Domingues, P. Dubath, F. Fodor, B. Frézouls, A. Fries, D. Fustes, D. Fyfe, E. Gallardo, J. Gallegos, D. Gardiol, M. Gebran, A. Gomboc, A. Gómez, E. Grux, A. Gueguen, A. Heyrovsky, J. Hoar, G. Iannicola, Y. Isasi Parache, A. M. Janotto, E. Joliet, A. Jonckheere, R. Keil, D. W. Kim, P. Klagyivik, J. Klar, J. Knude, O. Kochukhov, I. Kolka, J. Kos, A. Kutka, V. Lainey, D. LeBouquin, C. Liu, D. Loreggia, V. V. Makarov, M. G. Marseille, C. Martayan, O. Martinez-Rubi, B. Massart, F. Meynadier, S. Mignot, U. Munari, A. T. Nguyen, T. Nordlander, P. Ocvirk, K. S. O'Flaherty, A. Olias Sanz, P. Ortiz, J. Osorio, D. Oszkiewicz, A. Ouzounis, M. Palmer, P. Park, E. Pasquato, C. Peltzer, J. Peralta, F. Péturaud, T. Pieniluoma, E. Pigozzi, J. Poels, G. Prat, T. Prod'homme, F. Raison, J. M. Rebordao, D. Risquez, B. Rocca-Volmerange, S. Rosen, M. I. Ruiz-Fuertes, F. Russo, S. Sembay, I. Serraller Vizcaino, A. Short, A. Siebert, H. Silva, D. Sinachopoulos, E. Slezak, M. Soffel, D. Sosnowska, V. Straižys, M. ter Linden, D. Terrell, S. Theil, C. Tiede, L. Troisi, P. Tsalmantza, D. Tur, M. Vaccari, F. Vachier, P. Valles, W. Van Hamme, L. Veltz, J. Virtanen, J. M. Wallut, R. Wichmann, M. I. Wilkinson, H. Ziaeepour, and S. Zschocke. The Gaia mission. *A&A*, 595:A1, November 2016b. doi: 10.1051/0004-6361/201629272.

P. Di Matteo, M. Haywood, M. D. Lehnert, D. Katz, S. Khoperskov, O. N. Snaith, A. Gómez, and N. Robichon. The Milky Way has no in-situ halo other than the heated thick disc. Composition of the stellar halo and age-dating the last significant merger with Gaia DR2 and APOGEE. *A&A*, 632:A4, December 2019. doi: 10.1051/0004-6361/201834929.

V. Belokurov, D. Erkal, N. W. Evans, S. E. Koposov, and A. J. Deason. Co-formation of the disc and the stellar halo. *MNRAS*, 478(1):611–619, July 2018. doi: 10.1093/mnras/sty982.

Zhen Yuan, G. C. Myeong, Timothy C. Beers, N. W. Evans, Young Sun Lee, Projjwal Banerjee, Dmitrii Gudin, Kohei Hattori, Haining Li, Tadafumi Matsuno, Vinicius M. Placco, M. C. Smith, Devin D. Whitten, and Gang Zhao. Dynamical Relics of the Ancient Galactic Halo. *ApJ*, 891(1):39, March 2020b. doi: 10.3847/1538-4357/ab6ef7.

Rohan P. Naidu, Charlie Conroy, Ana Bonaca, Benjamin D. Johnson, Yuan-Sen Ting, Nelson Caldwell, Dennis Zaritsky, and Phillip A. Cargile. Evidence from the H3 Survey That the Stellar Halo Is Entirely Comprised of Substructure. *ApJ*, 901(1):48, September 2020. doi: 10.3847/1538-4357/abaef4.

Khyati Malhan, Rodrigo A. Ibata, Sanjib Sharma, Benoit Famaey, Michele Bellazzini, Raymond G. Carlberg, Richard D'Souza, Zhen Yuan, Nicolas F. Martin, and Guillaume F. Thomas. The Global Dynamical Atlas of the Milky Way Mergers: Constraints from Gaia EDR3-based Orbits of Globular Clusters, Stellar Streams, and Satellite Galaxies. *ApJ*, 926(2):107, February 2022b. doi: 10.3847/1538-4357/ac4d2a.

G. C. Myeong, N. W. Evans, V. Belokurov, J. L. Sanders, and S. E. Koposov. Discovery of new retrograde substructures: the shards of ω Centauri? *MNRAS*, 478(4):5449–5459, August 2018. doi: 10.1093/mnras/sty1403.

Steven R. Majewski, David L. Nidever, Verne V. Smith, Guillermo J. Damke, William E. Kunkel, Richard J. Patterson, Dmitry Bizyaev, and Ana E. García Pérez. Exploring Halo

Substructure with Giant Stars: Substructure in the Local Halo as Seen in the Grid Giant Star Survey Including Extended Tidal Debris from ωCentauri. *ApJL*, 747(2):L37, March 2012. doi: 10.1088/2041-8205/747/2/L37.

Daniela Carollo, Timothy C. Beers, Young Sun Lee, Masashi Chiba, John E. Norris, Ronald Wilhelm, Thirupathi Sivarani, Brian Marsteller, Jeffrey A. Munn, Coryn A. L. Bailer-Jones, Paola Re Fiorentin, and Donald G. York. Two stellar components in the halo of the Milky Way. *Nat*, 450(7172):1020–1025, December 2007. doi: 10.1038/nature06460.

Amina Helmi, Jovan Veljanoski, Maarten A. Breddels, Hao Tian, and Laura V. Sales. A box full of chocolates: The rich structure of the nearby stellar halo revealed by Gaia and RAVE. *A&A*, 598:A58, February 2017. doi: 10.1051/0004-6361/201629990.

Federico Sestito, Tobias Buck, Else Starkenburg, Nicolas F. Martin, Julio F. Navarro, Kim A. Venn, Aura Obreja, Pascale Jablonka, and Andrea V. Macciò. Exploring the origin of low-metallicity stars in Milky-Way-like galaxies with the NIHAO-UHD simulations. *MNRAS*, 500(3):3750–3762, January 2021. doi: 10.1093/mnras/staa3479.

G. C. Myeong, E. Vasiliev, G. Iorio, N. W. Evans, and V. Belokurov. Evidence for two early accretion events that built the Milky Way stellar halo. *MNRAS*, 488(1):1235–1247, September 2019. doi: 10.1093/mnras/stz1770.

Eugene Vasiliev. AGAMA: action-based galaxy modelling architecture. *MNRAS*, 482(2): 1525–1544, January 2019b. doi: 10.1093/mnras/sty2672.

Paul J. McMillan. The mass distribution and gravitational potential of the Milky Way. *MNRAS*, 465(1):76–94, February 2017b. doi: 10.1093/mnras/stw2759.

A. C. Robin, X. Luri, C. Reylé, Y. Isasi, E. Grux, S. Blanco-Cuaresma, F. Arenou, C. Babusiaux, M. Belcheva, R. Drimmel, C. Jordi, A. Krone-Martins, E. Masana, J. C. Mauduit, F. Mignard, N. Mowlavi, B. Rocca-Volmerange, P. Sartoretti, E. Slezak, and A. Sozzetti. Gaia Universe model snapshot. A statistical analysis of the expected contents of the Gaia catalogue. *A&A*, 543:A100, July 2012b. doi: 10.1051/0004-6361/201118646.

Chun Wang, Yang Huang, Haibo Yuan, Huawei Zhang, Maosheng Xiang, and Xiaowei Liu. The Value-added Catalog for LAMOST DR8 Low-resolution Spectra. *ApJS*, 259(2):51, April 2022b. doi: 10.3847/1538-4365/ac4df7.

Jo Bovy. galpy: A python Library for Galactic Dynamics. *ApJS*, 216(2):29, February 2015b. doi: 10.1088/0067-0049/216/2/29.

Evan N. Kirby, Judith G. Cohen, Puragra Guhathakurta, Lucy Cheng, James S. Bullock, and Anna Gallazzi. The Universal Stellar Mass-Stellar Metallicity Relation for Dwarf Galaxies. *ApJ*, 779(2):102, December 2013b. doi: 10.1088/0004-637X/779/2/102.

Tomás Ruiz-Lara, Tadafumi Matsuno, S. Sofie Lövdal, Amina Helmi, Emma Dodd, and Helmer H. Koppelman. Substructure in the stellar halo near the Sun. II. Characterisation of independent structures. *arXiv e-prints*, art. arXiv:2201.02405, January 2022.

Emma Dodd, Thomas M. Callingham, Amina Helmi, Tadafumi Matsuno, Tomás Ruiz-Lara, Eduardo Balbinot, and Sofie Lövdal. The Gaia DR3 view of dynamical substructure in the stellar halo near the Sun. *arXiv e-prints*, art. arXiv:2206.11248, June 2022b.

João A. S. Amarante, Victor P. Debattista, Leandro Beraldo e Silva, Chervin F. P. Laporte, and Nathan Deg. GASTRO library I: the simulated chemodynamical properties of several GSE-like stellar halos. *arXiv e-prints*, art. arXiv:2204.12187, April 2022.

Else Starkenburg, Nicolas Martin, Kris Youakim, David S. Aguado, Carlos Allende Prieto, Anke Arentsen, Edouard J. Bernard, Piercarlo Bonifacio, Elisabetta Caffau, Raymond G. Carlberg, Patrick Côté, Morgan Fouesneau, Patrick François, Oliver Franke, Jonay I. González Hernández, Stephen D. J. Gwyn, Vanessa Hill, Rodrigo A. Ibata, Pascale Jablonka, Nicolas Longeard, Alan W. McConnachie, Julio F. Navarro, Rubén Sánchez-Janssen, Eline Tolstoy, and Kim A. Venn. The Pristine survey - I. Mining the Galaxy for the most metal-poor stars. *MNRAS*, 471(3):2587–2604, November 2017. doi: 10.1093/mnras/stx1068.

Nicolas F. Martin, Else Starkenburg, Zhen Yuan, Morgan Fouesneau, Anke Arentsen, Francesca De Angeli, Felipe Gran, Martin Montelius, René Andrae, Michele Bellazzini, Paolo Montegriffo, Anna F. Esselink, Hanyuan Zhang, Kim A. Venn, Akshara Viswanathan, David S. Aguado, Giuseppina Battaglia, Manuel Bayer, Piercarlo Bonifacio, Elisabetta Caffau, Patrick Côté, Raymond Carlberg, Sébastien Fabbro, Emma Fernández Alvar, Jonay I. González Hernández, Isaure González Rivera de La Vernhe, Vanessa Hill, Rodrigo A. Ibata, Pascale Jablonka, Georges Kordopatis, Carmela Lardo, Alan W. McConnachie, Camila Navarrete, Julio Navarro, Alejandra Recio-Blanco, Rubén Sánchez Janssen, Federico Sestito, Guillaume F. Thomas, Sara Vitali, and Kristopher Youakim. The Pristine survey – XXIII. Data Release 1 and an all-sky metallicity catalogue based on Gaia DR3 BP/RP spectro-photometry. *arXiv e-prints*, art. arXiv:2308.01344, August 2023. doi: 10.48550/arXiv.2308.01344.

Steven R. Majewski, Ricardo P. Schiavon, Peter M. Frinchaboy, Carlos Allende Prieto, Robert Barkhouser, Dmitry Bizyaev, Basil Blank, Sophia Brunner, Adam Burton, Ricardo Carrera, S. Drew Chojnowski, Kátia Cunha, Courtney Epstein, Greg Fitzgerald, Ana E. García Pérez, Fred R. Hearty, Chuck Henderson, Jon A. Holtzman, Jennifer A. Johnson, Charles R. Lam, James E. Lawler, Paul Maseman, Szabolcs Mészáros, Matthew Nelson, Duy Coung Nguyen, David L. Nidever, Marc Pinsonneault, Matthew Shetrone, Stephen Smee, Verne V. Smith, Todd Stolberg, Michael F. Skrutskie, Eric Walker, John C. Wilson, Gail Zasowski, Friedrich Anders, Sarbani Basu, Stephane Beland, Michael R. Blanton, Jo Bovy, Joel R. Brownstein, Joleen Carlberg, William Chaplin, Cristina Chiappini, Daniel J. Eisenstein, Yvonne Elsworth, Diane Feuillet, Scott W. Fleming, Jessica Galbraith-Frew, Rafael A. García, D. Aníbal García-Hernández, Bruce A. Gillespie, Léo Girardi, James E. Gunn, Sten Hasselquist, Michael R. Hayden, Saskia Hekker, Inese Ivans, Karen Kinemuchi, Mark Klaene, Suvrath Mahadevan, Savita Mathur, Benoît Mosser, Demitri Muna, Jeffrey A. Munn, Robert C. Nichol, Robert W. O'Connell, John K. Parejko, A. C. Robin, Helio Rocha-Pinto, Matthias Schultheis, Aldo M. Serenelli, Neville Shane, Victor Silva Aguirre, Jennifer S. Sobeck, Benjamin Thompson, Nicholas W. Troup, David H. Weinberg, and Olga Zamora. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *AJ*, 154(3):94, Sep 2017. doi: 10.3847/1538-3881/aa784d.

Sven Buder, Martin Asplund, Ly Duong, Janez Kos, Karin Lind, Melissa K. Ness, Sanjib Sharma, Joss Bland -Hawthorn, Andrew R. Casey, Gayandhi M. de Silva, Valentina D'Orazi, Ken C. Freeman, Geraint F. Lewis, Jane Lin, Sarah L. Martell, Katharine J. Schlesinger, Jeffrey D. Simpson, Daniel B. Zucker, Tomaž Zwitter, Anish M. Amarsi, Borja Anguiano, Daniela Carollo, Luca Casagrande, Klemen Čotar, Peter L. Cottrell, Gary da Costa, Xudong D. Gao, Michael R. Hayden, Jonathan Horner, Michael J. Ireland, Prajwal R. Kafle, Ulisse Munari, David M. Nataf, Thomas Nordlander, Dennis Stello, Yuan-Sen Ting, Gregor Traven, Fred Watson, Robert A. Wittenmyer, Rosemary F. G. Wyse, David Yong, Joel C. Zinn, Maruša Žerjal, and Galah Collaboration. The GALAH Survey: second data release. *MNRAS*, 478(4):4513–4552, August 2018. doi: 10.1093/mnras/sty1281.

Xiang-Qun Cui, Yong-Heng Zhao, Yao-Quan Chu, Guo-Ping Li, Qi Li, Li-Ping Zhang, Hong-Jun Su, Zheng-Qiu Yao, Ya-Nan Wang, Xiao-Zheng Xing, Xin-Nan Li, Yong-Tian Zhu, Gang Wang, Bo-Zhong Gu, A. Li Luo, Xin-Qi Xu, Zhen-Chao Zhang, Gen-Rong Liu, Hao-Tong Zhang, De-Hua Yang, Shu-Yun Cao, Hai-Yuan Chen, Jian-Jun Chen, Kun-Xin Chen, Ying Chen, Jia-Ru Chu, Lei Feng, Xue-Fei Gong, Yong-Hui Hou, Hong-Zhuan Hu, Ning-Sheng Hu, Zhong-Wen Hu, Lei Jia, Fang-Hua Jiang, Xiang Jiang, Zi-Bo Jiang, Ge Jin, Ai-Hua Li, Yan Li, Ye-Ping Li, Guan-Qun Liu, Zhi-Gang Liu, Wen-Zhi Lu, Yin-Dun Mao, Li Men, Yong-Jun Qi, Zhao-Xiang Qi, Huo-Ming Shi, Zheng-Hong Tang, Qing-Sheng Tao, Da-Qi Wang, Dan Wang, Guo-Min Wang, Hai Wang, Jia-Ning Wang, Jian Wang, Jian-Ling Wang, Jian-Ping Wang, Lei Wang, Shu-Qing Wang, You Wang, Yue-Fei Wang, Ling-Zhe Xu, Yan Xu, Shi-Hai Yang, Yong Yu, Hui Yuan, Xiang-Yan Yuan, Chao Zhai, Jing Zhang, Yan-Xia Zhang, Yong Zhang, Ming Zhao, Fang Zhou, Guo-Hua Zhou, Jie Zhu, and Si-Cheng Zou. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Research in Astronomy and Astrophysics*, 12(9):1197–1242, Sep 2012. doi: 10.1088/1674-4527/12/9/003.

Andrea Kunder, Georges Kordopatis, Matthias Steinmetz, Tomaž Zwitter, Paul J. McMillan, Luca Casagrande, Harry Enke, Jennifer Wojno, Marica Valentini, Cristina Chiap-

pini, Gal Matijevič, Alessand ro Siviero, Patrick de Laverny, Alejandra Recio-Blanco, Albert Bijaoui, Rosemary F. G. Wyse, James Binney, E. K. Grebel, Amina Helmi, Paula Jofre, Teresa Antoja, Gerard Gilmore, Arnaud Siebert, Benoit Famaey, Olivier Bienaymé, Brad K. Gibson, Kenneth C. Freeman, Julio F. Navarro, Ulisse Munari, George Seabroke, Borja Anguiano, Maruša Žerjal, Ivan Minchev, Warren Reid, Joss Bland-Hawthorn, Janez Kos, Sanjib Sharma, Fred Watson, Quentin A. Parker, Ralf-Dieter Scholz, Donna Burton, Paul Cass, Malcolm Hartley, Kristin Fiegert, Milorad Stupar, Andreas Ritter, Keith Hawkins, Ortwin Gerhard, W. J. Chaplin, G. R. Davies, Y. P. Elsworth, M. N. Lund, A. Miglio, and B. Mosser. The Radial Velocity Experiment (RAVE): Fifth Data Release. *AJ*, 153(2):75, February 2017. doi: 10.3847/1538-3881/153/2/75.

Brian Yanny, Constance Rockosi, Heidi Jo Newberg, Gillian R. Knapp, Jennifer K. Adelman-McCarthy, Bonnie Alcorn, Sahar Allam, Carlos Allende Prieto, Deokkeun An, Kurt S. J. Anderson, Scott Anderson, Coryn A. L. Bailer-Jones, Steve Bastian, Timothy C. Beers, Eric Bell, Vasily Belokurov, Dmitry Bizyaev, Norm Blythe, John J. Bochanski, William N. Boroski, Jarle Brinchmann, J. Brinkmann, Howard Brewington, Larry Carey, Kyle M. Cudworth, Michael Evans, N. W. Evans, Evalyn Gates, B. T. Gänsicke, Bruce Gillespie, Gerald Gilmore, Ada Nebot Gomez-Moran, Eva K. Grebel, Jim Greenwell, James E. Gunn, Cathy Jordan, Wendell Jordan, Paul Harding, Hugh Harris, John S. Hendry, Diana Holder, Inese I. Ivans, Željko Ivezič, Sebastian Jester, Jennifer A. Johnson, Stephen M. Kent, Scot Kleinman, Alexei Kniazev, Jurek Krzesinski, Richard Kron, Nikolay Kuropatkin, Svetlana Lebedeva, Young Sun Lee, R. French Leger, Sébastien Lépine, Steve Levine, Huan Lin, Daniel C. Long, Craig Loomis, Robert Lupton, Olena Malanushenko, Viktor Malanushenko, Bruce Margon, David Martinez-Delgado, Peregrine McGehee, Dave Monet, Heather L. Morrison, Jeffrey A. Munn, Jr. Neilsen, Eric H., Atsuko Nitta, John E. Norris, Dan Oravetz, Russell Owen, Nikhil Padmanabhan, Kaike Pan, R. S. Peterson, Jeffrey R. Pier, Jared Platson, Paola Re Fiorentin, Gordon T. Richards, Hans-Walter Rix, David J. Schlegel, Donald P. Schneider, Matthias R. Schreiber, Axel Schwope, Valena Sibley, Audrey Simmons, Stephanie A. Snedden, J. Allyn Smith, Larry Stark, Fritz Stauffer, M. Steinmetz, C. Stoughton, Mark SubbaRao, Alex Szalay, Paula Szkody, Aniruddha R. Thakar, Thirupathi Sivarani, Douglas Tucker, Alan Uomoto, Dan Vanden Berk, Simon Vidrih, Yogesh Wadadekar, Shannon Watters, Ron Wilhelm, Rosemary F. G. Wyse, Jean Yarger, and Dan Zucker. SEGUE: A Spectroscopic Survey of 240,000 Stars with g = 14-20. *AJ*, 137(5):4377–4399, May 2009. doi: 10.1088/0004-6256/137/5/4377.

S. Randich, G. Gilmore, L. Magrini, G. G. Sacco, R. J. Jackson, R. D. Jeffries, C. C. Worley, A. Hourihane, A. Gonneau, C. Viscasillas Vazquez, E. Franciosini, J. R. Lewis, E. J. Alfaro, C. Allende Prieto, T. Bensby, R. Blomme, A. Bragaglia, E. Flaccomio, P. François, M. J. Irwin, S. E. Koposov, A. J. Korn, A. C. Lanzafame, E. Pancino, A. Recio-Blanco, R. Smiljanic, S. Van Eck, T. Zwitter, M. Asplund, P. Bonifacio, S. Feltzing, J. Binney, J. Drew, A. M. N. Ferguson, G. Micela, I. Negueruela, T. Prusti, H. W. Rix, A. Vallenari, A. Bayo, M. Bergemann, K. Biazzo, G. Carraro, A. R. Casey, F. Damiani, A. Frasca, U. Heiter, V. Hill, P. Jofré, P. de Laverny, K. Lind, G. Marconi, C. Martayan, T. Masseron, L. Monaco, L. Morbidelli, L. Prisinzano, L. Sbordone, S. G. Sousa, S. Zaggia, V. Adibekyan, R. Bonito, E. Caffau, S. Daflon, D. K. Feuillet, M. Gebran, J. I. Gonzalez Hernandez, G. Guiglion, A. Herrero, A. Lobel, J. Maiz Apellaniz, T. Merle, Š. Mikolaitis, D. Montes, T. Morel, C. Soubiran, L. Spina, H. M. Tabernero, G. Tautvaišiene, G. Traven, M. Valentini, M. Van der Swaelmen, S. Villanova, N. J. Wright, U. Abbas, V. Aguirre Børsen-Koch, J. Alves, L. Balaguer-Nunez, P. S. Barklem, D. Barrado, S. R. Berlanas, A. S. Binks, A. Bressan, R. Capuzzo-Dolcetta, L. Casagrande, L. Casamiquela, R. S. Collins, V. D'Orazi, M. L. L. Dantas, V. P. Debattista, E. Delgado-Mena, P. Di Marcantonio, A. Drazdauskas, N. W. Evans, B. Famaey, M. Franchini, Y. Frémat, E. D. Friel, X. Fu, D. Geisler, O. Gerhard, E. A. Gonzalez Solares, E. K. Grebel, M. L. Gutierrez Albarran, D. Hatzidimitriou, E. V. Held, F. Jiménez-Esteban, H. Jönsson, C. Jordi, T. Khachaturyants, G. Kordopatis, J. Kos, N. Lagarde, L. Mahy, M. Mapelli, E. Marfil, S. L. Martell, S. Messina, A. Miglio, I. Minchev, A. Moitinho, J. Montalban, M. J. P. F. G. Monteiro, C. Morossi, N. Mowlavi, A. Mucciarelli, D. N. A. Murphy, N. Nardetto, S. Ortolani, F. Paletou, J. Palouš, E. Paunzen, J. C. Pickering, A. Quirrenbach, P. Re Fiorentin, J. I. Read, D. Romano, N. Ryde, N. Sanna, W. Santos, G. M. Seabroke, A. Spagna, M. Steinmetz, E. Stonkuté, E. Sutorius, F. Thévenin, M. Tosi, M. Tsantaki, J. S. Vink,

N. Wright, R. F. G. Wyse, M. Zoccali, J. Zorec, D. B. Zucker, and N. A. Walton. The Gaia-ESO Public Spectroscopic Survey: Implementation, data products, open cluster survey, science, and legacy. *Astronomy and Astrophysics*, 666:A121, October 2022. doi: 10.1051/0004-6361/202243141.

T. S. Li, S. E. Koposov, D. B. Zucker, G. F. Lewis, K. Kuehn, J. D. Simpson, A. P. Ji, N. Shipp, Y. Y. Mao, M. Geha, A. B. Pace, A. D. Mackey, S. Allam, D. L. Tucker, G. S. Da Costa, D. Erkal, J. D. Simon, J. R. Mould, S. L. Martell, Z. Wan, G. M. De Silva, K. Bechtol, E. Balbinot, V. Belokurov, J. Bland-Hawthorn, A. R. Casey, L. Cullinane, A. Drlica-Wagner, S. Sharma, A. K. Vivas, R. H. Wechsler, B. Yanny, and S5 Collaboration. The southern stellar stream spectroscopic survey (S^5): Overview, target selection, data reduction, validation, and early science. *MNRAS*, 490(3):3508–3531, December 2019. doi: 10.1093/mnras/stz2731.

Sandro D'Odorico, Stefano Cristiani, Hans Dekker, Vanessa Hill, Andreas Kaufer, Taesun Kim, and Francesca Primas. Performance of UVES, the echelle spectrograph for the ESO VLT and highlights of the first observations of stars and quasars. In Jacqueline Bergeron, editor, *Discoveries and Research Prospects from 8- to 10-Meter-Class Telescopes*, volume 4005 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 121–130, June 2000. doi: 10.1117/12.390133.

R. Ibata, A. Sollima, C. Nipoti, M. Bellazzini, S. C. Chapman, and E. Dalessandro. The Globular Cluster NGC 2419: A Crucible for Theories of Gravity. *ApJ*, 738(2):186, September 2011. doi: 10.1088/0004-637X/738/2/186.

Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, January 2010. doi: 10.2140/camcos.2010.5.65.

Giuseppina Battaglia, Amina Helmi, Heather Morrison, Paul Harding, Edward W. Olszewski, Mario Mateo, Kenneth C. Freeman, John Norris, and Stephen A. Shectman. The radial velocity dispersion profile of the Galactic halo: constraining the density profile of the dark halo of the Milky Way. *Monthly Notices of the Royal Astronomical Society*, 364(2):433–442, 12 2005. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2005.09367.x. URL https://doi.org/10.1111/j.1365-2966.2005.09367.x.

Prajwal Raj Kafle, Sanjib Sharma, Geraint F Lewis, and Joss Bland-Hawthorn. On the shoulders of giants: properties of the stellar halo and the milky way mass distribution. *The Astrophysical Journal*, 794(1):59, 2014.

Aakash Ravi, Nicholas Langellier, David F. Phillips, Malte Buschmann, Benjamin R. Safdi, and Ronald L. Walsworth. Probing dark matter using precision measurements of stellar accelerations. *Phys. Rev. Lett.*, 123:091101, Aug 2019. doi: 10.1103/PhysRevLett.123.091101. URL https://link.aps.org/doi/10.1103/PhysRevLett.123.091101.

C. Quercellini, L. Amendola, and A. Balbi. Mapping the galactic gravitational potential with peculiar acceleration. *Monthly Notices of the Royal Astronomical Society*, 391(3):1308–1314, 11 2008. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2008.13968.x. URL https://doi.org/10.1111/j.1365-2966.2008.13968.x.

Gregory M. Green and Yuan-Sen Ting. Deep potential: Recovering the gravitational potential from a snapshot of phase space, 2020.

J. An, A. P. Naik, N. W. Evans, and C. Burrage. Charting galactic accelerations: when and how to extract a unique potential from the distribution function. *MNRAS*, 506(4):5721–5730, October 2021. doi: 10.1093/mnras/stab2049.

Matthew R. Buckley, Sung Hak Lim, Eric Putney, and David Shih. Measuring Galactic dark matter through unsupervised machine learning. *MNRAS*, 521(4):5100–5119, June 2023. doi: 10.1093/mnras/stad843.

Sung Hak Lim, Eric Putney, Matthew R. Buckley, and David Shih. Mapping Dark Matter in the Milky Way using Normalizing Flows and Gaia DR3. *arXiv e-prints*, art. arXiv:2305.13358, May 2023. doi: 10.48550/arXiv.2305.13358.

Taavet Kalda, Gregory M Green, and Soumavo Ghosh. Recovering the gravitational potential in a rotating frame: Deep Potential applied to a simulated barred galaxy. *Monthly Notices of the Royal Astronomical Society*, 527(4):12284–12297, 01 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae011. URL https://doi.org/10.1093/mnras/stae011.

Wassim Tenachi, Rodrigo Ibata, and Foivos I Diakogiannis. An end-to-end strategy for recovering a free-form potential from a snapshot of stellar coordinates. *arXiv preprint arXiv:2305.16845*, 2023b.

T. Antoja, A. Helmi, M. Romero-Gómez, D. Katz, C. Babusiaux, R. Drimmel, D. W. Evans, F. Figueras, E. Poggio, C. Reylé, A. C. Robin, G. Seabroke, and C. Soubiran. A dynamically young and perturbed Milky Way disk. *Nat*, 561(7723):360–362, September 2018. doi: 10.1038/s41586-018-0510-7.

Wilma H. Trick, Johanna Coronado, and Hans-Walter Rix. The Galactic disc in action space as seen by Gaia DR2. *MNRAS*, 484(3):3291–3306, April 2019. doi: 10.1093/mnras/stz209.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Tony Tohme, Mohammad Javad Khojasteh, Mohsen Sadr, Florian Meyer, and Kamal Youcef-Toumi. Isr: Invertible symbolic regression. *arXiv preprint arXiv:2405.06848*, 2024.

Tatiana A Michtchenko, Ronaldo SS Vieira, Douglas A Barros, and Jacques RD Lépine. Modelling resonances and orbital chaos in disk galaxies-application to a milky way spiral model. *Astronomy & Astrophysics*, 597:A39, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *JMLR*, 22(1), jan 2021. ISSN 1532-4435. doi: 10.48550/arXiv.1912.02762.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling Tabular Data with Diffusion Models. *arXiv e-prints*, art. arXiv:2209.15421, September 2022. doi: 10.48550/arXiv.2209.15421.

L. G. Hou and J. L. Han. Offset between stellar spiral arms and gas arms of the Milky Way. *Monthly Notices of the Royal Astronomical Society*, 454(1):626–636, 09 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1904. URL https://doi.org/10.1093/mnras/stv1904.

James Binney. Actions for axisymmetric potentials. *Monthly Notices of the Royal Astronomical Society*, 426(2):1324–1327, 2012.

Adriana Dropulic, Bryan Ostdiek, Laura J Chang, Hongwan Liu, Timothy Cohen, and Mariangela Lisanti. Machine learning the sixth dimension: Stellar radial velocities from 5d phase-space correlations. *The Astrophysical Journal Letters*, 915(1):L14, 2021.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Biwei Dai and Uroš Seljak. Learning effective physical laws for generating cosmological hydrodynamics with lagrangian deep learning. *Proceedings of the National Academy of*

Sciences, 118(16):e2020324118, 2021.

Thibaut L. François, Christian M. Boily, Jonathan Freundlich, Simon Rozier, and Karina Voggel. Forming off-center massive black hole binaries in dwarf galaxies through Jacobi capture. *A&A*, 687:A203, July 2024. doi: 10.1051/0004-6361/202348591.

P. Bianchini, M. A. Norris, G. van de Ven, E. Schinnerer, A. Bellini, R. P. van der Marel, L. L. Watkins, and J. Anderson. The Effect of Unresolved Binaries on Globular Cluster Proper-motion Dispersion Profiles. *ApJL*, 820(1):L22, March 2016. doi: 10.3847/2041-8205/820/1/L22.

Raymond G. Carlberg, Adrian Jenkins, Carlos S. Frenk, and Andrew P. Cooper. Star Stream Velocity Distributions in CDM and WDM Galactic Halos. *arXiv e-prints*, art. arXiv:2405.18522, May 2024. doi: 10.48550/arXiv.2405.18522.

Liang Wang, Aaron A Dutton, Gregory S Stinson, Andrea V Macciò, Camilla Penzo, Xi Kang, Ben W Keller, and James Wadsley. Nihao project–i. reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations. *Monthly Notices of the Royal Astronomical Society*, 454(1):83–94, 2015.

P. Bianchini, R. P. van der Marel, A. del Pino, L. L. Watkins, A. Bellini, M. A. Fardal, M. Libralato, and A. Sills. The internal rotation of globular clusters revealed by Gaia DR2. *MNRAS*, 481(2):2125–2139, December 2018. doi: 10.1093/mnras/sty2365.

Anjali Varghese, R Ibata, and Geraint F Lewis. Stellar streams as probes of dark halo mass and morphology: a bayesian reconstruction. *Monthly Notices of the Royal Astronomical Society*, 417(1):198–215, 2011.

Salvatore Ferrone, Paola Di Matteo, Alessandra Mastrobuono-Battisti, Misha Haywood, Owain N Snaith, Marco Montuori, Sergey Khoperskov, and David Valls-Gabaud. The e-tidalgcs project-modeling the extra-tidal features generated by galactic globular clusters. *Astronomy & Astrophysics*, 673:A44, 2023.

Catalina Urrejola-Mora, Facundo A. Gómez, Sergio Torres-Flores, Philippe Amram, Benoît Epinat, Antonela Monachesi, Federico Marinacci, and Claudia Mendes de Oliveira. WiNDS: An H_α Kinematics Survey of Nearby Spiral Galaxies-Vertical Perturbations in Nearby Disk-type Galaxies. *ApJ*, 935(1):20, August 2022. doi: 10.3847/1538-4357/ac78ec.

Jonathan P Gardner, John C Mather, Mark Clampin, Rene Doyon, Matthew A Greenhouse, Heidi B Hammel, John B Hutchings, Peter Jakobsen, Simon J Lilly, Knox S Long, et al. The james webb space telescope. *Space Science Reviews*, 123:485–606, 2006.

Nabila Aghanim, Yashar Akrami, Frederico Arroja, Mark Ashdown, J Aumont, Carlo Baccigalupi, M Ballardini, Anthony J Banday, RB Barreiro, Nicola Bartolo, et al. Planck 2018 results-i. overview and the cosmological legacy of planck. *Astronomy & Astrophysics*, 641: A1, 2020.

Carlo Rovelli. *Quantum gravity*. Cambridge university press, 2004.

D Saxena, S Khandare, and S Chaudhary. An overview of chatgpt: Impact on academic learning. *FMDB Transactions on Sustainable Techno Learning*, 1(1):11–20, 2023.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, et al. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*, 2023.

Andrej Karpathy. nanogpt. GitHub repository, 2023. URL https://github.com/karpathy/nanoGPT.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Llama Team. The llama 3 herd of models, July 2024. URL https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

Chenxi Lin, Jiayu Ren, Guoxiu He, Zhuoren Jiang, Haiyan Yu, and Xiaomin Zhu. Tree-

based hard attention with self-motivation for large language models. *arXiv preprint arXiv:2402.08874*, 2024.

Adrian de Wynter. Will gpt-4 run doom? *arXiv preprint arXiv:2403.05468*, 2024.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Appendix A

Press Release for the ApJ 959, 99 Publication

Our paper, *Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws*[1], authored by W. Tenachi, R. Ibata, F. Diakogiannis was accompanied by a press release, which is included here in full.

[1] ApJ **959** 99, arXiv:2303.03192

A.1 ObAS Press Release

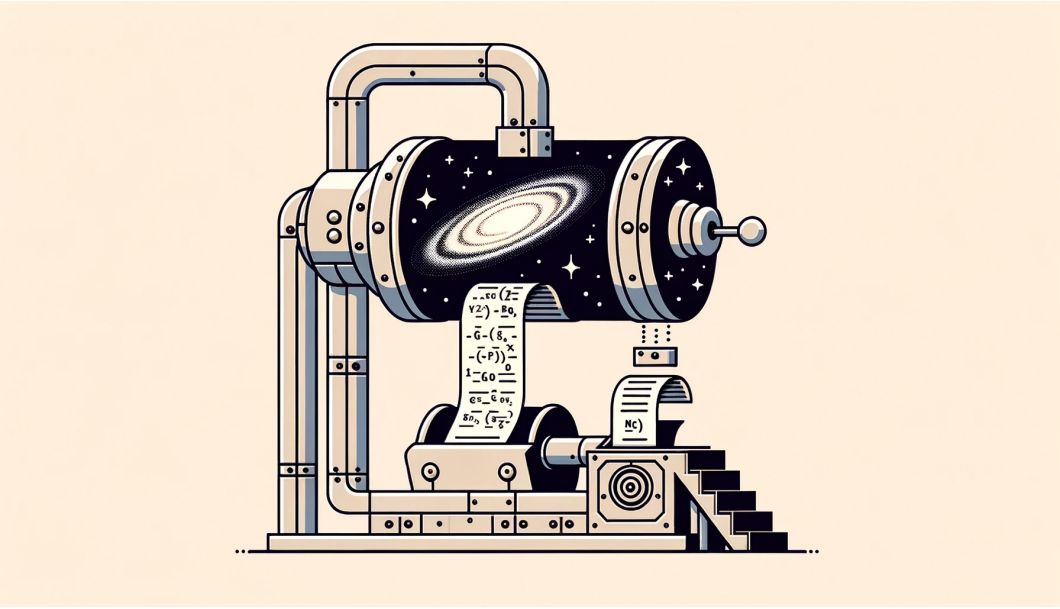
Press release[2] from the Observatoire Astronomique de Strasbourg (ObAS):

An AI rediscovers 74 physical laws

Dec. 11 2023

Is deep learning inevitably synonymous with "black boxes"? These methods are often criticized for their lack of transparency, resulting in unintelligible models. This issue is particularly relevant in physics, where the goal is to model the laws governing our universe as comprehensible equations, rather than opaque neural networks composed of millions of numbers.

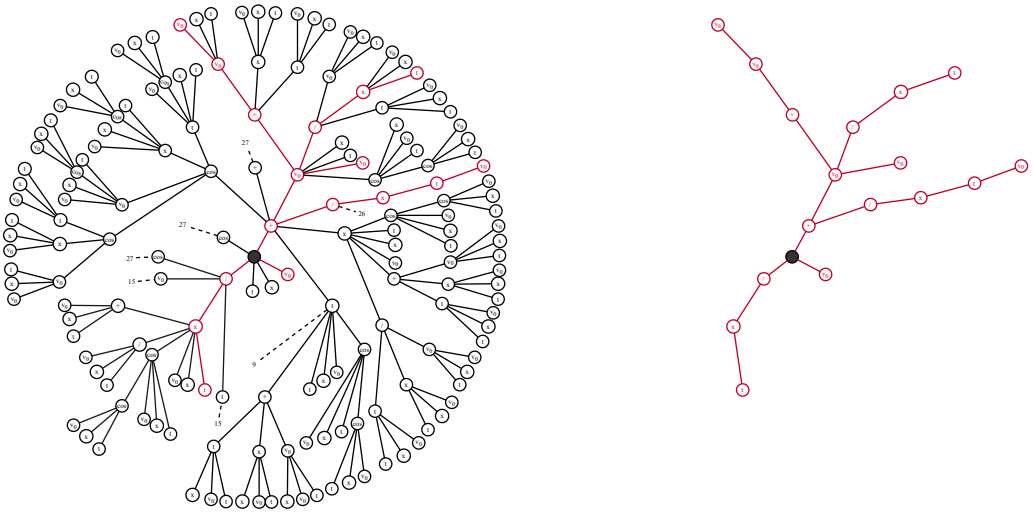
Wassim Tenachi, Rodrigo Ibata, two astrophysicists based in France and Foivos Diakogiannis, a researcher at Australia's national science agency, CSIRO tackled this problem by creating an artificial intelligence algorithm that produces analytical physical models from raw scientific data. Their work was published in the American journal, The Astrophysical Journal, on December 11.



[2] https://astro.unistra.fr/en/2023/12/11/an-ai-rediscovers-74-physical-laws/

Manipulating even elementary mathematical symbols like addition or division can be a complex challenge for neural networks. However, thanks to advances in artificial intelligence techniques related to natural language processing and drawing from methods used in symbolic computation, it is now possible to create neural networks that generate equations.

Nevertheless, the quest for the ideal equation that perfectly models a dataset while having the freedom to combine a plethora of mathematical symbols can quickly become a combinatorial nightmare. As you may have been told many times in school, in physics, you can't "add potatoes and carrots together", for example, you can't add a length and a velocity because it doesn't make physical sense. These rules, known as dimensional analysis, prohibit certain combinations of mathematical symbols when writing a physical equation, greatly reducing the combinatorial space.



Search space

Search space with our *in situ* physical units prior

Figure A.1: Illustration of the combinatorial space reduction provided by dimensional analysis.

The artificial intelligence method, known as "PhySO" (Physical Symbolic Optimization), designed by these researchers French and Australian scientists formulates thousands of equations per second and autonomously learns to formulate increasingly high-quality equations through trial and error while capitalizing on these dimensional analysis rules.

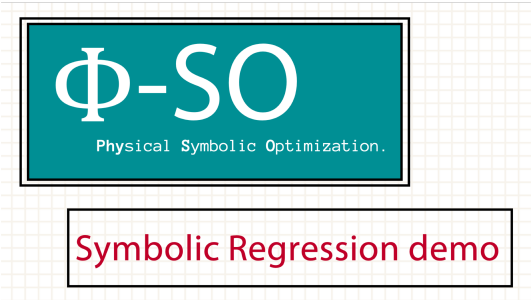
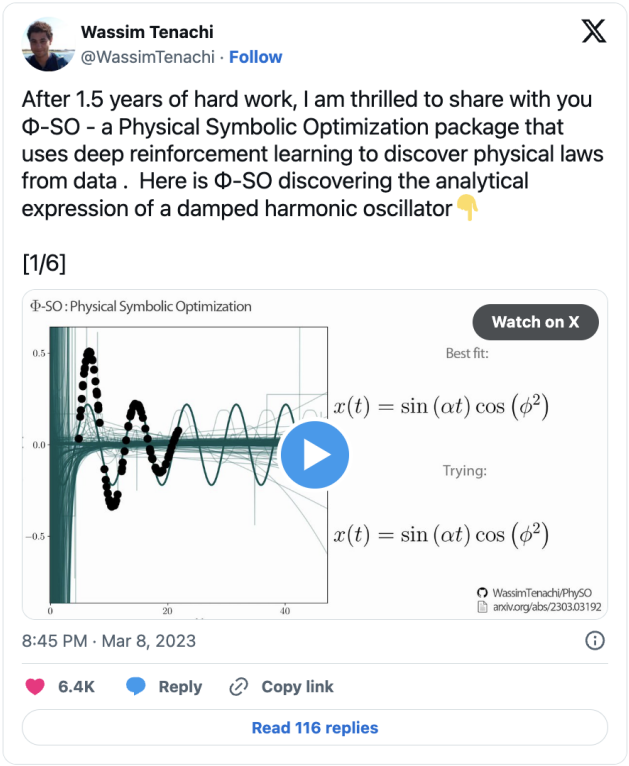


Figure A.2: [SR demo] (https://youtu.be/wubzZMkoTUY) PhySO, the artificial intelligence developed by the Franco-Australian collaboration, rediscovering the equation of the textbook harmonic oscillator from a dataset.

This study made a lot of noise on Twitter[3], quickly becoming the most-discussed scientific article of the week on the social network. It was even shared by Professor Yann Lecun, the scientist often considered the father of modern artificial intelligence and the head of the artificial intelligence department at Meta (formerly Facebook).



[3] https://x.com/WassimTenachi/status/1633645134934949888

This kind of approach raises many questions about the role of humans in the scientific process. "The goal is not to replace the physicist but simply to equip us with a powerful tool to explore the space of equations that empirically meet experimental or observational constraints," emphasize the authors.

In this first study, the Franco-Australian collaboration focused on the automated formulation of empirical equations, aligning more with observational and experimental needs than the theoretical aspects of physics.

It is worth noting the impartiality of this type of unsupervised method regarding the precise configuration of the equations sought. Could this intrinsic impartiality one day lead to a more agnostic scientific research?

Scientific contacts:

- Wassim Tenachi (PhD Student) wassim.tenachi@astro.unistra.fr
- Rodrigo Ibata (DR CNRS) rodrigo.ibata@astro.unistra.fr

Article:

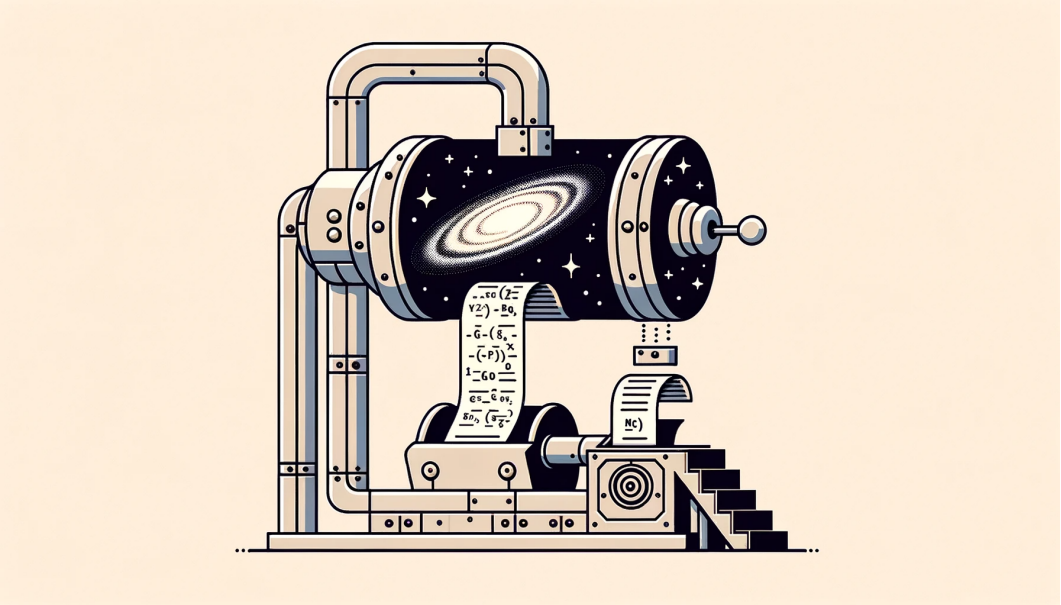
Wassim Tenachi, Rodrigo Ibata, Foivos Diakogiannis, Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws, ApJ (DOI: 10.3847/1538-4357/ad014c)

A.2 CNRS Press Release (fr.)

Press release[4] from the Centre national de la recherche scientifique (CNRS):

Une intelligence artificielle retrouve des lois physiques à partir de données scientifiques

Dec. 11 2023



L'apprentissage profond est-il inévitablement synonyme de "boîtes noires"? On reproche souvent à ces méthodes leur absence de transparence résultant en des modèles inintelligibles. C'est un problème qui se pose tout particulièrement en physique, domaine dans lequel on cherche à modéliser les lois régissant notre Univers sous la forme d'équations compréhensibles et non pas de réseaux de neurones opaques constitués de millions de nombres. Une équipe de recherche internationale comprenant des scientifiques du CNRS-INSU (voir encadré), s'est attaquée à ce problème en créant un algorithme d'intelligence artificielle produisant des modèles physiques analytiques à partir de données scientifiques brutes.

[4] https://www.insu.cnrs.fr/fr/cnrsinfo/une-intelligence-artificielle-retrouve-des-lois-physiques-partir-de-donnees-scientifiques

Manipuler des symboles mathématiques même élémentaires tels que l'addition ou la division peut s'avérer un défi complexe pour les réseaux de neurones. Toutefois, grâce aux progrès réalisés dans les techniques d'intelligence artificielle liées au traitement du langage et en s'appuyant sur les approches utilisées en calcul symbolique, il est désormais possible de créer des réseaux de neurones générant des équations. Néanmoins, la quête de l'équation idéale modélisant parfaitement un jeu de données en ayant la liberté conjuguer pléthore de symboles mathématiques peut rapidement devenir un enfer combinatoire. Comme on vous l'a peut être répété mainte fois à l'école, en physique on ne peut pas "additionner des patates et des carottes", par exemple on ne peut pas additionner une longueur et une vitesse car cela n'a pas de sens physiquement. Ces règles dites d'analyse dimensionnelle interdisent certaines combinaisons de symboles mathématiques lors de l'écriture d'une équation physique et permettent de grandement réduire l'espace combinatoire.

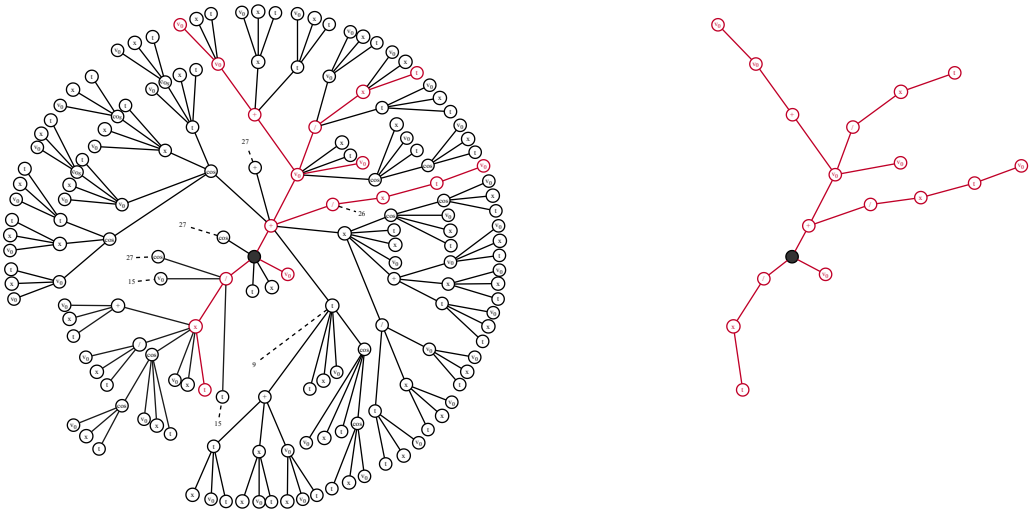


Figure A.3: Illustration de la réduction de l'espace combinatoire offerte par l'analyse dimensionnelle.

La méthode d'intelligence artificielle baptisée "PhySO" acronyme d'Optimisation Symbolique Physique élabore des milliers d'équations par seconde et apprend de façon autonome à formuler des équations de qualité croissante par essai erreur tout en capitalisant sur ces règles d'analyse dimensionnelle. Il convient de souligner l'absence de préjugés de ce type de méthode non supervisée quant à la configuration précise des équations recherchées. Ce type d'impartialité intrinsèque pourrait-il un jour conduire à

une recherche scientifique plus agnostique ?

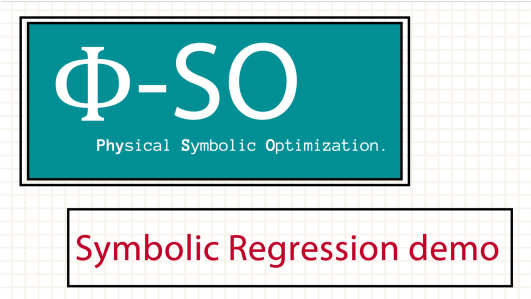


Figure A.4: [SR demo] (https://youtu.be/wubzZMkoTUY) PhySO, the artificial intelligence developed by the Franco-Australian collaboration, rediscovering the equation of the textbook harmonic oscillator from a dataset.

Laboratoires impliqués:

- Laboratoire CNRS : Observatoire astronomique de Strasbourg (ObAS) Tutelles : CNRS / Univ. Strasbourg
- Autre laboratoire : Agence scientifique nationale australienne, CSIRO

Pour en savoir plus:

Wassim Tenachi, Rodrigo Ibata, Foivos Diakogiannis, Deep symbolic regression for physics guided by units constraints: toward the automated discovery of physical laws, The Astrophysical Journal, 2023.

Contact:

- Rodrigo Ibata Directeur de recherche CNRS à l'Observatoire astronomique de Strasbourg rodrigo.ibata@astro.unistra.fr
- Wassim Tenachi Doctorant à l'Observatoire astronomique de Strasbourg wassim.tenachi@astro.unistra.fr

Université de Strasbourg

Wassim Tenachi

École doctorale Physique, chimie-physique ED 182 Université de Strasbourg

Symbolic Machine Learning for Physics & Astrophysics

Résumé

Nous explorons le potentiel novateur de l'apprentissage automatique symbolique dans les domaines de la physique et de l'astrophysique, afin de surmonter les limites d'interprétabilité des méthodes traditionnelles dans cette ère caractérisée par une profusion de données. Nous présentons Φ-SO, un paradigme d'Optimisation Symbolique Physique qui exploite l'apprentissage profond par renforcement pour générer des expressions symboliques analytiques directement à partir de données. Cette approche de régression symbolique (SR) atteint des performances de premier plan en intégrant l'analyse dimensionnelle et en facilitant l'exploitation de diverses réalisations d'une unique classe de phénomènes : une approche que nous nommons Class SR.

Nous nous penchons sur les enjeux liés à la matière noire à l'échelle galactique et identifions plusieurs nouveaux courants stellaires grâce aux données du satellite Gaia, complétées par des observations de suivi effectuées avec les télescopes INT et VLT. Nous mettons en lumière l'existence d'un courant polaire émanant du halo externe traversant le voisinage solaire, que nous baptisons Typhon. Enfin, nous proposons une approche pionnière d'apprentissage non supervisé pour déterminer de manière agnostique la distribution de la matière noire dans la Voie Lactée, à partir d'un cliché des coordonnées stellaires en employant des transformations canoniques.

Mots-clés: apprentissage automatique symbolique, apprentissage profond par renforcement, régression symbolique, matière noire, courants stellaires, Voie Lactée.

Résumé en anglais

We explore the transformative potential of symbolic machine learning in physics and astrophysics, seeking to overcome the interpretability challenges of traditional methods in the era of data abundance. We introduce Φ-SO, a Physical Symbolic Optimization framework that relies on deep reinforcement learning to extract analytical symbolic expressions directly from data. This symbolic regression (SR) framework achieves state-of-the-art performance by integrating physical dimensional analysis and enabling the exploitation of diverse realizations of a singular class of phenomena — an approach we dub Class SR.

Focusing on the dark matter challenges at the galactic scale, we uncover several new stellar streams from Gaia satellite data and perform follow-up observations using the INT and VLT telescopes. Notably, we discover a polar stream from the outer halo passing through the Solar neighborhood, which we dub Typhon. Finally, we propose a first observation-driven, unsupervised learning approach to agnostically constrain the dark matter distribution of the Milky Way from a snapshot of stellar coordinates using canonical transformations.

Keywords: symbolic machine learning, deep reinforcement learning, symbolic regression, dark matter, stellar streams, Milky Way