UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :

Pierre LLOMPART

Soutenue le : **5 juin 2025**

Pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie informatique et théorique

Conception moléculaire par IA multitâche et exploration chémographique

THÈSE dirigée par :

M. MARCOU Gilles	Assistant-professeur**,** Université de Strasbourg
M. VARNEK Alexandre	Professeur, Université de Strasbourg
Mme. MINOLETTI Claire	Docteur, Sanofi R&D

RAPPORTEURS :

M. KIRCHMAIR Johannes	Professeur, Université de Vienne
Mme. LAGARDE Nathalie	Assistant professeur, Conservatoire National des Arts et Métiers (CNAM)

AUTRES MEMBRES DU JURY :

Mme. CAMPROUX Anne-Claude	Professeur, Université Paris Cité
Mme. KELLENBERGER Esther	Professeur, Université de Strasbourg
M. TABOUREAU Olivier	Professeur, Université Paris Cité

Abstract

This thesis is dedicated to advancing the role of in silico modeling in pharmaceutical research, addressing the persistent challenges of late-stage failures and inefficiencies in drug development. ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) testing often occurs too late in the pipeline, driving up costs and delaying progress. To mitigate these issues, in silico modeling, particularly early ADMET (eADMET) prediction, has become essential for streamlining decision-making in early drug discovery. However, the complexity of human biology, evolving assays, and data inconsistencies necessitate predictive models that are not only accurate but also adaptable and interpretable. This thesis presents a systematic approach to refining eADMET modeling through data curation, multi-task learning, large-scale applicability, and human–machine collaboration.

The first part of this work focuses on solubility modeling, emphasizing the challenges posed by assay variability and dataset inconsistencies regarding thermodynamic solubility. We also demonstrate that kinetic solubility data, contrary to common assumptions, can be reliably modeled when properly curated. A framework for solubility prediction is introduced, improving model accuracy and reproducibility. The second part investigates drug absorption modeling using multi-task learning (MTL). By leveraging shared patterns among related endpoints, MTL enhances predictive performance over single-task models. This approach is then expanded to ultra-large datasets encompassing ADMET and bioactivity measures. To further optimize lead selection, we explore collective intelligence strategies, comparing expert feedback to modeling at the late-stage optimization phase. Finally, the thesis examines the broader landscape of AI-driven drug discovery, critically assessing industry trends, overhyped claims, and the reality of AI's impact on pharmaceutical R&D.

The findings highlight the importance of high-quality data, rigorous validation, and interdisciplinary collaboration for sustainable AI adoption. This work underscores the necessity of flexible, interpretable, and data-driven in silico tools to enhance efficiency in modern drug discovery, ultimately aiding the search for safer and more effective medicines.

Acknowledgements

To those who have actively contributed, those who have subtly influenced, and the unfortunate few who wandered in here by accident, thank you. I imagined a PhD thesis to be a solitary endeavor. Yet, I realize it has been anything but a journey shaped by many hands, guided by many minds, and supported by many hearts.

First and foremost, my heartfelt gratitude goes to my academic advisors, those luminous minds who guided me with patience. Professor Alexandre Varnek, thank you for your supervision, your philosophy of research, and for seeing the true potential of scientific inquiry. Dr. Gilles Marcou, your dedication and time invested, along with the countless hours spent meticulously reviewing papers. Dr. Claire Minoletti, my mentor, your openness to creativity and innovation made this journey all the more fulfilling. Beyond my direct supervisors, I thank my thesis committee, Prof. Johannes Kirchmair, Dr. Nathalie Lagarde, Prof. Esther Kellenberger, for reviewing my thesis. Special thanks to Prof. Anne-Claude Camproux and Prof. Olivier Taboureau for introducing me to cheminformatics 6 years ago and witnessing the culmination of this journey.

This work would not have been the same without the support of my academic colleagues, whose presence turned long hours into something far more meaningful. Dr. Olga Klimchul, for your boundless energy, and infectious optimism. Dr. Dragos Horvath, your unique stories and acidic humor have been invaluable. Dr. Fanny Bonachera, for your openness and support.

Meanwhile, to my teammates in Vitry for shaping my industrial mind. Dr. Bruno Filoche-Rommé, beyond your contributions to our brainstorming sessions, you brought a whirlwind of ideas, truly a medchem Yoda. Dr. Paraskevi Gkeka, for your guidance, collaboration, and effortlessly enjoyable conversations. Dr. Jean-Philippe Rameau, our indefatigable BSO, you have been a driving force behind countless ideas and presentations. Dr. Nicolas Muzet, if not for the *manelle*, then certainly for your ever-sharp wit and sarcasm. And to Dr. Marc Bianciotto, Dr. David Papin, Dr. Corinne Terrier, Dr. Anna Gogolou, Dr. Kwame Amaning, Dr. Bruno Cornet, Dr. Hervé Minoux, Dr. Laurent Schio, Dr. Yann Foricher, you have each, in your own way, shaped this experience, and for that, I am deeply grateful.

To my fellow colleagues. Dr. Shamkhal Baybekov, for your *optimized* class, and unique way of seeing the world. Dr. Maxime Langevin, not just a scientist but a true icon, every discussion with you was enriching. Dr. Yuliana Zabolotna, for your warm welcome in the lab, the joyful moments, and your unwavering dedication. Dr. Alexey Orlov, Dr. Louis Plyer, Charlotte Chapellier, Dr. Francesco-Paolo Panei, Erik Yeghyan, Dr. Giuseppe Lamanna, Dr. Maxim Shevelev, Dr. Farah Asgarkhanova, Dr. William Bort, Aykhan Israfilli, Dr. Polina Oleneva, Dr. Vasilii Fastovsky, Dr. Mikhail Volkov, Dr. Kyrylo Klimenko, thank you for your companionship.

A special thank you to my interns, Jeanne Mithieux, for generating such good humor, and for your patience in handling my sometimes-over-the-top requests. Jianjun Tao, for our lively discussions. Margaux Dumon, Nicolas Simon, and those I had the pleasure of working with at Sanofi, Yanna Gautier, Mathilde Kretz, and William Vo-Thanh.

Beyond the walls of the lab, I am just as grateful to the friends who never wavered in their support. Sujith, for our balades, drinks, and late nights that inevitably left us in desperate need of a long hike. Pablo Mas, for these discussions as comforting as hot cocoa. To the Pikalyova mafia, Karina, for the insightful debates and great meals; Regina, for your boldness, dedication, and vibrant presence. And to Dr. Tagir Akhmetshin, your brilliance could bring back sunshine to Paris. Hope it will !

To my family, who has been the constant force underpinning my efforts and instilled in me the love of learning, I thank you.

And to Congcong, my love and partner in all things, thank you for walking every step of this winding road by my side. Your unwavering presence and support have made all the difference.

And finally, to our adversaries, time, doubt, and the occasional second reviewer, I acknowledge your efforts to derail us. Alas, we have prevailed.

With that, I want to acknowledge that this thesis is truly a collective achievement. My name may be on the cover, but so many people have left their mark on its pages. To everyone, those named here and those unnamed but not unappreciated, thank you for being part of my story. Wherever this road leads next, I carry with me the memory of your kindness, and for that, I am profoundly grateful.

Contents

Chapter 1. Résumé en Français

1.1. Introduction

La découverte de nouveaux médicaments est un processus complexe et exigeant en ressources, impliquant plusieurs étapes allant de l'identification initiale de la cible aux essais cliniques. Une phase critique est l'optimisation des composés leads, où les structures chimiques sont révisées pour améliorer l'efficacité, la sélectivité et les propriétés pharmacocinétiques tout en minimisant la toxicité potentielle.[1] Malgré des efforts considérables, les échecs liés aux propriétés ADMET (Absorption, Distribution, Métabolisme, Elimination et Toxicité) restent un obstacle majeur, avec des études récentes indiquant qu'environ 90 % des candidats-médicaments échouent lors du développement clinique en raison de problèmes d'efficacité et de sécurité.[2,3] Ces échecs contribuent à l'escalade des coûts du développement de médicaments, estimés entre 1,3 et 2,8 milliards de dollars par nouveau médicament.[4]

Cette thèse contribue à répondre à ces défis grâce à des techniques d'apprentissage automatique (Machine Learning, ML) et d'intelligence artificielle (IA) permettant l'analyse de données complexes, facilitant la prédiction des propriétés moléculaires, des interactions ligand-cible et des effets secondaires potentiels.[5,6] La prolifération de larges ensembles de données a certes révolutionné la découverte de médicaments, mais les prises de décision en découverte de médicaments sont d'autant plus impactées par les problématiques liées aux données, par exemple : les données incertaines ou incohérentes ou des modèles statistiques inadéquats.[7]

Cette thèse se concentre sur l'amélioration du processus de sélection des données et l'amélioration des méthodes de décision en développant des modèles computationnels pour la prédiction des propriétés ADMET, en explorant l'espace chimique à travers la cartographie topographique générative (Generative Topographique Mapping, GTM) et en exploitant des approches d'intelligence collective. Par l'intégration de données de haute qualité à des modèles robustes, ces travaux contribuent à améliorer les processus décisionnels, réduisant les taux d'attrition et optimisant les ressources investies dans le développement de nouveaux médicaments.

1.2. Méthodologies pour l'Apprentissage Automatique

Les modèles d'apprentissage automatique développés au cours de la découverte de médicaments reposent souvent sur les relations quantitatives structure-propriété (QSPR), établissant une fonction Y=f(X), où X est la représentation d'une entité chimique par des descripteurs moléculaires et Y, une propriété d'intérêt telle que la solubilité. Ces travaux ajoutent deux approches complémentaires à ces modèles QSPR : l'exploration rationnelle de l'espace chimique à l'aide de cartes GTM[8], et la définition de domaines d'applicabilité (Applicability Domain, AD) visant à estimer la pertinence des prédictions – Isolation Forest[9], par exemple.

Cette thèse présente par ailleurs, des approches récentes de modélisation QSPR reposant sur des réseaux de neurones artificiels basés sur des graphes (Graph Neural Network, GNN). Ces méthodes formulent des modèles dont les données chimiques utilisées en entrées se présentent sous forme de graphes moléculaires.[10] Nos travaux illustrent l'intérêt de ces méthodes, pour leurs performances, leur capacité à traiter des grands ensembles de données et pour la facilité avec laquelle des concepts élaborés, comme l'apprentissage multi-tâches, sont formulés.[11]

L'apprentissage multitâche (Multi Task Learning, MTL), est une extension de l'apprentissage monotâche (Single Task Learning, STL). La stratégie consiste à entrainer un unique modèle MTL sur plusieurs tâches connexes simultanément au lieu d'entraîner des modèles STL indépendants pour chacune de ces tâches. Un modèle multi-tâche est moins compliqué à entraîner car il ne faut fixer les valeurs des paramètres libres que de ce seul modèle MTL au lieu de devoir les fixer pour chaque modèle STL séparément.[12] Au cours de nos travaux, nous avons constaté que des synergies entre les tâches sont assez fréquentes pour accroître, globalement, la généralisation des modèles, des antagonismes pouvant également être observés, se traduisant par la détérioration des performances pour certaines tâches. Dans cette thèse, nous appliquons des méthodes MTL basées sur GNN pour développer des modèles prédictifs (**Figure 1**). Les modèles sont entraînés sur des ensembles de données soigneusement nettoyés pour assurer la qualité des prédictions, et diverses métriques, sont utilisées pour évaluer leurs performances.

1.3. Résultats & Discussions

La thèse s'articule autour de quatre volets : (1) la modélisation de la solubilité des composés chimiques, (2) le développement d'une approche multitâche pour prédire l'absorption des petites molécules, (3) l'extension de cette approche à grande échelle pour le profilage des propriétés ADMET et de nombreuses activités biologiques, et (4) l'application des modèles prédictifs dans l'optimisation des leads en chimie médicinale, comparée aux méthodes d'intelligence collective. Chaque section explore des méthodes innovantes pour affiner la prédiction des propriétés ADMET, avec une attention particulière portée au nettoyage des données et aux techniques de modélisation.

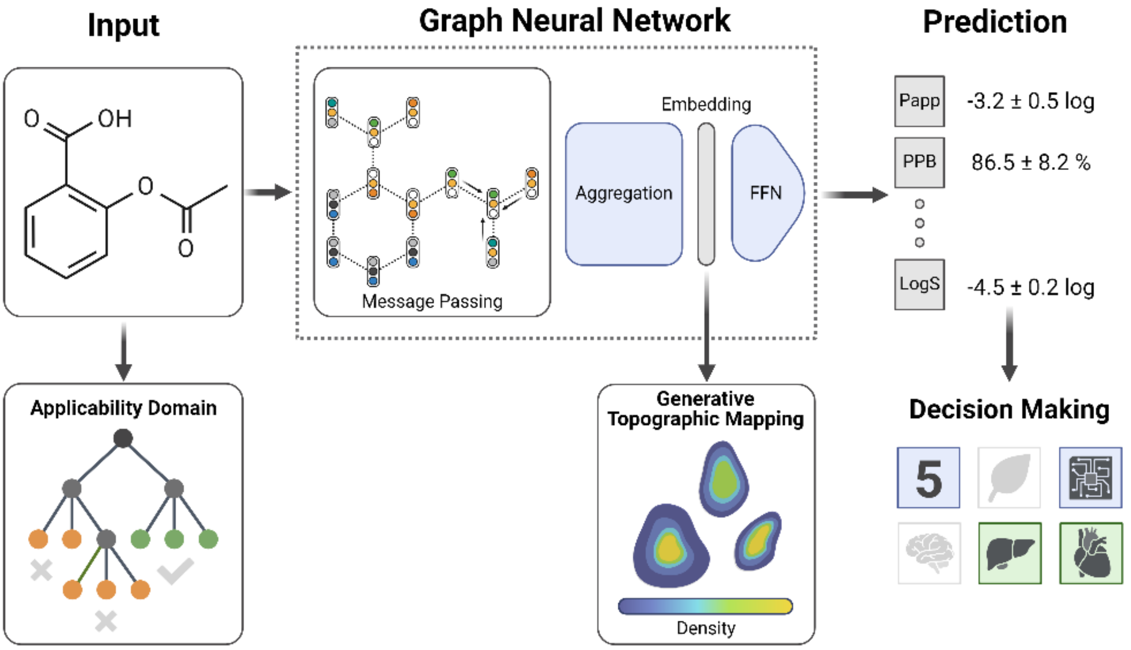


Figure 1 : Schéma du workflow de prédiction combinant un MTL GNN avec la GTM et l'évaluation du domaine d'applicabilité. Les structures d'entrée sont traitées via des couches de passage de messages (Message Passing), d'agrégation (Aggregation) et de propagation direct (Feed-Forward Network, FFN) pour prédire des propriétés telles que P_{app} (perméabilité apparente), PPB (liaison aux protéines plasmatiques) et LogS (log10 de la solubilité en molaire). L'AD garantit des prédictions dans un espace chimique valide, tandis que la GTM permet la visualisation et l'évaluation des caractéristiques des espaces chimique locaux.

Modélisation de la Solubilité des Composés Chimiques

Solubilité Thermodynamique

La prédiction précise de la solubilité aqueuse demeure un défi. Les modèles existants offrent souvent de bonnes performances sur les données d'entraînement mais échouent à se généraliser à de nouveaux composés. Nous avons compilé une liste exhaustive de jeux de données de solubilité, identifiant des sources négligées et des recouvrements. En nettoyant et en standardisant le jeu de données AqSolDB[13], nous avons créé un jeu de données de haute qualité, AqSolDBc, pour l'entraînement des modèles.

En utilisant à la fois des méthodes de forêts aléatoires (Random Forest, RF) et des GNN, nous avons développé des modèles prédictifs pour la solubilité aqueuse (**Figure 2a**). L'utilisation de ces modèles pour prédire de nouvelles données a révélé l’importance de définir leurs domaines d'applicabilité ; expliquant les performances décevantes des modèles mis en production en négligeant cet aspect (**Figure 2b**). Ces conclusions soulignent l'importance de la qualité des données et les défis liés à l'extrapolation au-delà du domaine d’entraînement.[14]

Solubilité Cinétique

Dans une perspective de criblage de molécules, la solubilité cinétique est plus pertinente que la solubilité thermodynamique car elle fixe les concentrations maximales auxquelles des échantillons peuvent être testées. L'analyse des données de solubilité cinétique et thermodynamique a confirmé les relations connues entre ces deux types de solubilité (**Figure 2c**). Les données de solubilité cinétique obtenues par différents protocoles se sont révélées cohérents, permettant ainsi de fusionner ces données en un jeu unique et exclusif pour entraîner des modèles prédictifs. Ces modèles renforcent la conclusion que la solubilité cinétique dépend moins de la méthode de mesure expérimentale que ce qui était initialement supposé (**Figure 2a**).[15]

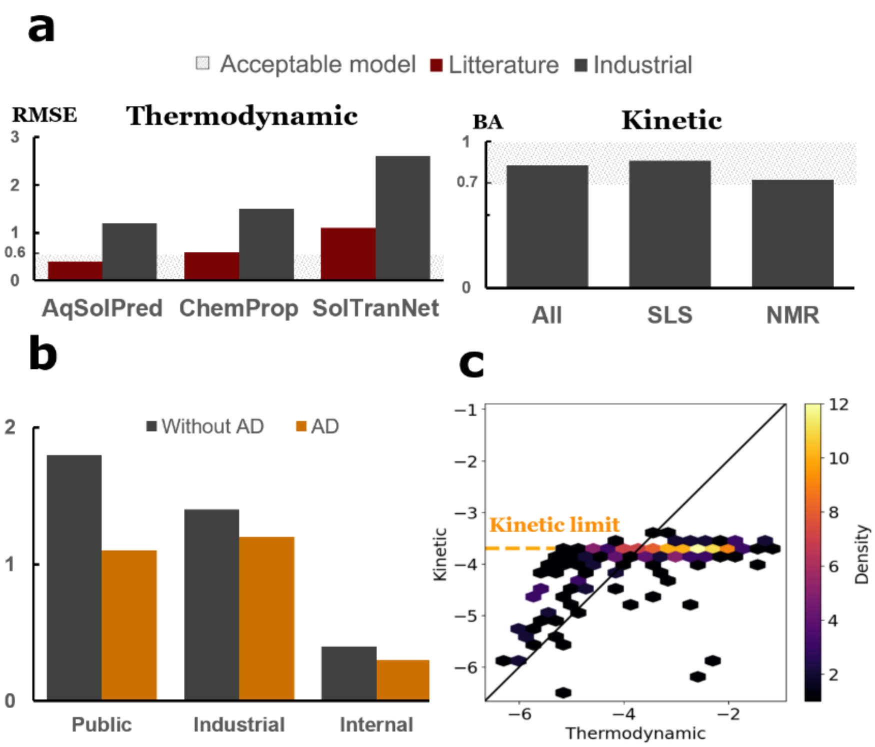


Figure 2 : Evaluation et analyses des modèles et valeurs de solubilité cinétique et thermodynamique. (a) Benchmark des performances des modèles public sur des données externes public et industrielles**.** Une zone grise définit les métriques nécessaires pour qu'un modèle soit considéré performant. **(b)** Performances en RMSE d'un modèle entrainé sur des données publiques et validé sur des données externes avec ou sans l'utilisation d'un domaine d'applicabilité. **(c)** Analyse comparative des valeurs de solubilité cinétiques et thermodynamiques. La couleur représente la densité des 186 composés du jeu de données de la Chimiothèque Nationale Essentielle, allant de faible (noir) à élever (jaune). La ligne pointillée orange indique la limite supérieure pour les mesures de solubilité cinétique (0,2 mM).

Approche Multitâche pour Prédire l'Absorption des Petites Molécules

L'absorption d'une petite molécule, influencée par la perméabilité et la solubilité, est un défi majeur en optimisation de leads. Nous avons modélisé la perméabilité via des approches multitâches pour améliorer la précision des prédictions d'absorption. Nos résultats confirment les facteurs clés influençant l'absorption. En comparant des modèles MTL et STL basés sur des GNN, les modèles MTL se sont révélés supérieurs pour les petits jeux de données. Pour tester l'apprentissage multitâche, nous avons introduit une tâche « leurre », l'énergie libre d'hydratation (HFE), qui n'a pas profité des synergies MTL, perdant même en performance par rapport au STL (**Figure 3**). Par ailleurs, les modèles basés sur des données publiques se généralisent mal aux données industrielles, à cause de disparités dans l'espace chimique, les conditions expérimentales et la qualité des données. Les modèles GNN STL se montrent plus robustes pour les grands jeux de données. Ces travaux soulignent l'importance de regrouper des tâches connexes pour optimiser les modèles multitâches.

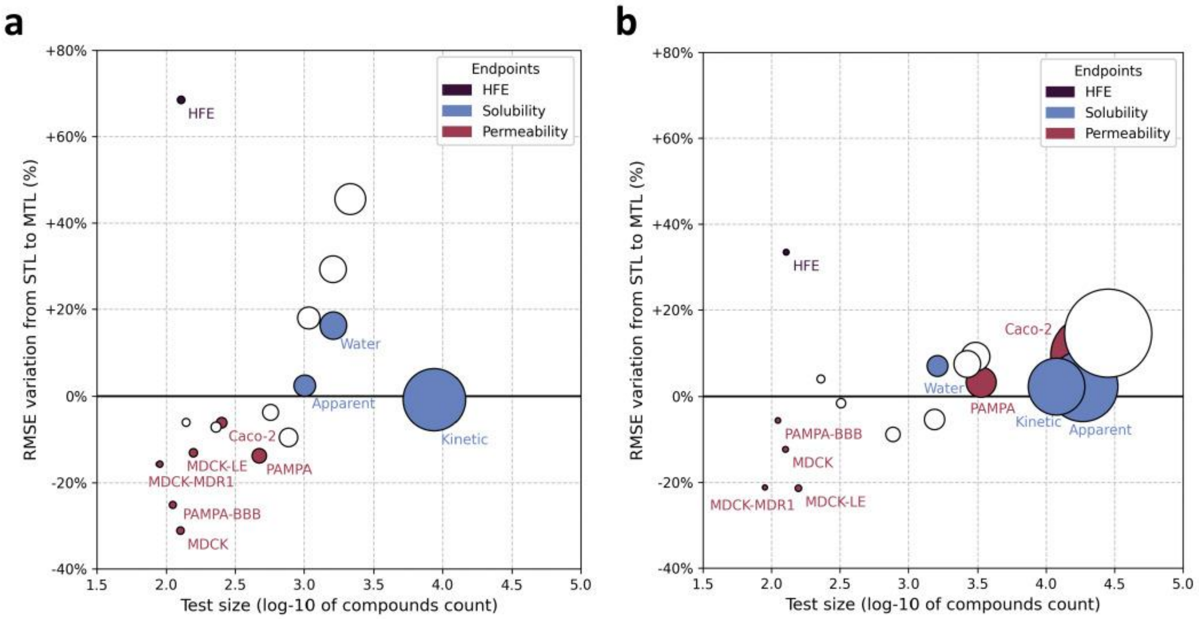


Figure 3 : Analyse des effets de synergie et d'antagonisme dans les modèles GNN MTL versus STL (a) Données publiques (b) Données industrielles. La taille des points est proportionnelle à la taille du jeu de données. Les points rouges représentent les endpoints de perméabilité, les points bleus ceux de solubilité, et les points noirs représentent le HFE, tâche neutre.

Profilage des Propriétés ADMET et des Activités Biologiques

Aujourd'hui, de nombreux jeux de données et serveurs web autorisent la modélisation des propriétés ADMET et de l'activité biologique, avec des benchmarks comme le Therapeutic Data Commons, démocratisant les thèmes de la découverte du médicament auprès des experts de l'apprentissage automatique. Les serveurs sont souvent redondants et mettent en œuvre des approches analogues. Pour contribuer à la production de modèles et de serveurs de prédictions plus fiables, nous avons travaillé un jeu de données de haute qualité spécifiquement orienté pour la découverte de médicaments, en compilant des données de BindingDB, OChem et ChEMBL. Nous avons méticuleusement nettoyé celles-ci en se basant sur les métadonnées expérimentales visant à améliorer leur cohérence et leur validité. Le modèle MTL GNN développé prédit simultanément plus de 2,000 activités biologiques et propriétés ADMET. Divers descripteurs moléculaires et algorithmes, dont la GNN, ont été évalués sur des données publiques et privées. Cette étude représente le modèle MTL de régression le plus large dans le domaine ADMET. Les performances sur chaque tâche sont comparables ou meilleures que celles des modèles mono-tâche. La Figure 4 montre des performances représentatives, avec des améliorations notables, comme pour pIC50 IL8. Les tâches avec peu de données bénéficient particulièrement du MTL, car le réseau de tâches est assez dense pour éviter des tâches isolées.

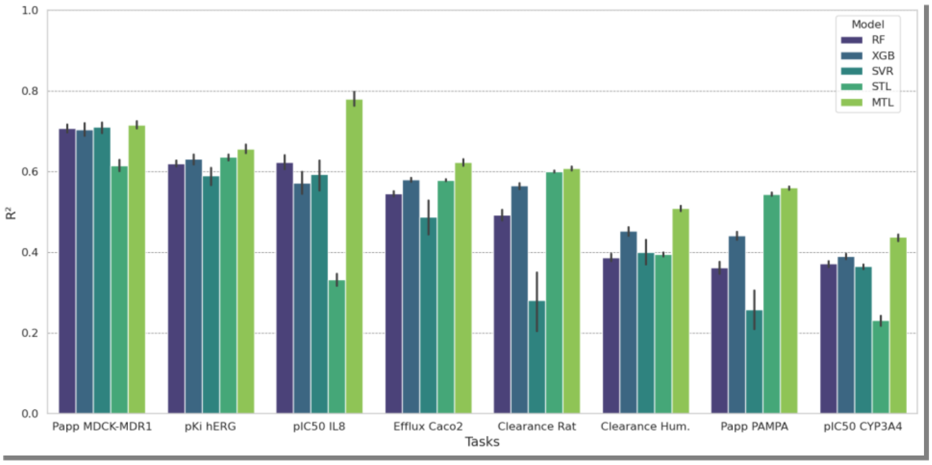


Figure 4 : Performance des différentes méthodes prédictives appliquées à un échantillon des propriétés modélisées. Les performances des modèles sur les données de tests sont représentées par le R^2. Chaque barre indique le R^2 moyen avec un intervalle d'erreur représentant la déviation standard.

L'Intelligence Collective et les Méthodes Automatisées dans l'Optimisation ADMET

Les campagnes d'optimisation en chimie médicinale sont souvent basées sur des méthodes prédictives et l'intuition des chimistes, influencées par leurs expériences et biais personnels. Afin d'introduire l'intelligence collective dans ce domaine, nous avons recueilli les réponses de 92 chercheurs de Sanofi à des questions d'optimisation de leads. Cela nous a permis d'analyser comment l'expertise et la confiance affectent les choix en conception moléculaire. Nos résultats montrent que l'intelligence collective améliore les taux de réussite pour des tâches courantes telles que la prédiction de l'hydrophobicité, la perméabilité et la solubilité (Figure 5). Toutefois, pour des phénomènes moins familiers, comme l'inhibition du canal cardiaque hERG, l'efficacité diminue. Nos conclusions suggèrent que l'intelligence collective pourrait constituer une voie prometteuse pour renforcer le processus de décision dans la découverte de médicaments.[16]

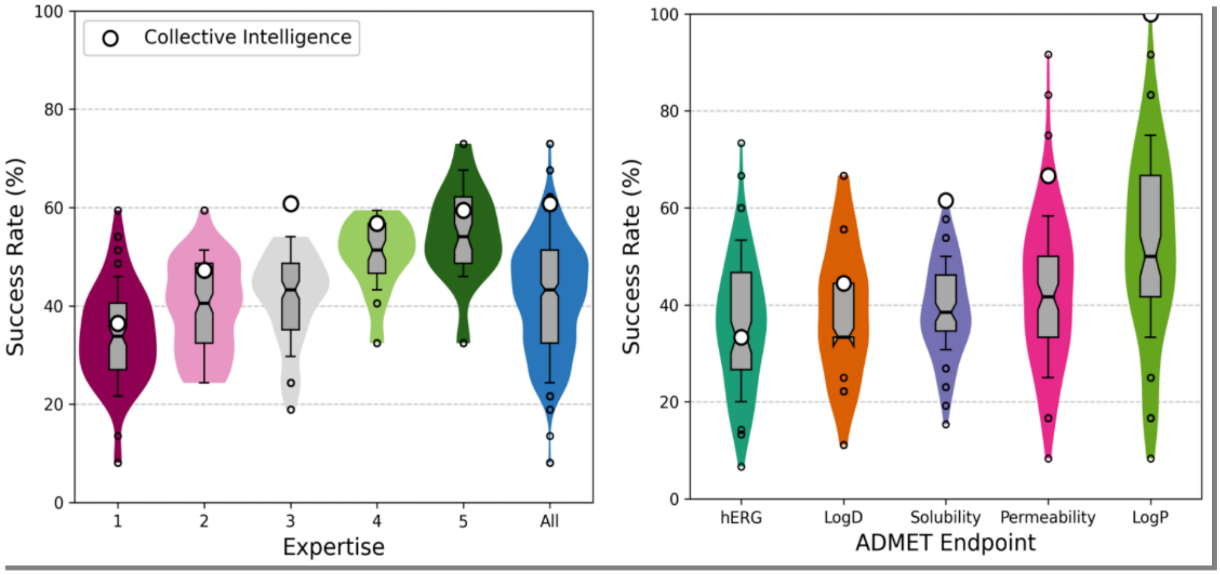


Figure 5 : Performances de l'intelligence collective par niveau d'expertise et propriétés ADMET. (Gauche) Diagrammes en violon du taux de succès (Success Rate, SR) par niveau d'expertise pour chaque groupe et tous les participants (en bleu). La médiane est représentée par ligne d'étranglement de la boite grise. Les SR collectifs sont montrés par des cercles pleins blancs. (Droite) Diagrammes en violon du SR pour chaque propriété.

1.4. Conclusion Générale

Cette thèse établit des bases solides pour l'intégration de l'apprentissage automatique dans la prédiction des propriétés ADMET, en proposant des contributions majeures à la modélisation moléculaire et à la réduction des échecs en recherche pharmaceutique. Les travaux mettent en avant l'importance du nettoyage des données et de la définition des domaines d'applicabilité pour garantir des prédictions fiables, en particulier pour des composés hors de l'espace chimique initial des modèles.

Les approches multitâches se sont révélées particulièrement efficaces pour exploiter les synergies entre propriétés connexes, notamment sur des données limitées, améliorant ainsi la généralisation des modèles et leurs performances globales. L'intégration des réseaux de neurones sur graphes et des outils de visualisation, comme la cartographie topographique générative, a renforcé la précision des prédictions tout en offrant des moyens d'explorer l'espace chimique.

Enfin, en associant modèles prédictifs et intelligence collective, cette thèse a démontré que l'expertise humaine peut complémenter les outils automatisés. Ces travaux posent les bases d'une modélisation ADMET hybride, combinant apprentissage profond et intelligence collective.

Chaque partie du travail a contribué à une meilleure compréhension et optimisation des processus de décision et de découverte de médicaments, en fournissant des outils et des ressources accessibles pour la recherche.

1.5. Références

(1) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

(2) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2019**, *20* (2), 273–286. https://doi.org/10.1093/biostatistics/kxx069.

(3) *Clinical Development Success Rates and Contributing Factors 2011-2020 | BIO*. https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020 (accessed 2024-11-18).

(4) Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323* (9), 844–853. https://doi.org/10.1001/jama.2020.1166.

(5) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat Rev Drug Discov* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(6) Martin, O. Artificial Intelligence in Drug Discovery and Development. *Advanced Sciences* **2021**, *3* (2), 1–10. https://doi.org/10.69610/j.as.20210822.

(7) Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discovery Today* **2021**, *26* (4), 1040–1052. https://doi.org/10.1016/j.drudis.2020.11.037.

(8) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998**, *21* (1), 203–224. https://doi.org/10.1016/S0925-2312(98)00043-5.

(9) Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*; 2008; pp 413–422. https://doi.org/10.1109/ICDM.2008.17.

(10) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* **2009**, *20* (1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.

(11) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

(12) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076. https://doi.org/10.1021/acs.jcim.7b00146.

(13) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds. *Sci Data* **2019**, *6* (1), 143. https://doi.org/10.1038/s41597-019-0151-1.

(14) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will We Ever Be Able to Accurately Predict Solubility? *Sci Data* **2024**, *11* (1), 303. https://doi.org/10.1038/s41597-024-03105-6.

(15) Baybekov, S.; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.-L.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Kinetic Solubility: Experimental and Machine-Learning Modeling Perspectives. *Molecular Informatics* **2024**, *43* (2), e202300216. https://doi.org/10.1002/minf.202300216.

(16) Llompart, P.; Amaning, K.; Bianciotto, M.; Filoche-Rommé, B.; Foricher, Y.; Mas, P.; Papin, D.; Rameau, J.-P.; Schio, L.; Marcou, G.; Varnek, A.; Moussaid, M.; Minoletti, C.; Gkeka, P. Harnessing Medicinal Chemical Intuition from Collective Intelligence. *Journal of Medicinal Chemistry* **2025**, https://doi.org/10.1021/acs.jmedchem.4c03066.

1.6. Liste des Présentations

Conférences Nationales

Titre	Événement	Auteurs	Lieu	Date	Type
Transforming molecular optimization using collective intelligence	Journées R&D de Sanofi	Llompart P., Amaning K., Bianciotto M., Filoche-Rommé B., Foricher Y., Mas P., Papin D., Rameau J.-P., Schio L., Marcou G., Varnek A., Moussaid M., Minoletti C., Gkeka P.	Vitry, France	31 Octobre 2024	Affiche
Harnessing Medicinal Chemical Intuition from Collective Intelligence	Museum National d'Histoire Naturelle de Paris	Llompart P., Amaning K., Bianciotto M., Filoche-Rommé B., Foricher Y., Mas P., Papin D., Rameau J.-P., Schio L., Marcou G., Varnek A., Moussaid M., Minoletti C., Gkeka P.	Paris, France	14 Juin 2024	Oral
Predicting solubility, but which one ?	Journées R&D de Sanofi	Llompart P., Baybekov S., Marcou G., Horvath D., Gizzi P., Galzi J., Ramos P., Saurel O., Bourban C., Minoletti C., Varnek A.	Vitry, France	19 Octobre 2023	Affiche
Predicting solubility, but which one ?	11ème Conférence de la Société Française de Chemoinformatique	Llompart P., Baybekov S., Marcou G., Horvath D., Gizzi P., Galzi J., Ramos P., Saurel O., Bourban C., Minoletti C., Varnek A.	Caen, France	5-6 Octobre 2023	Affiche
Will we ever be able to accurately predict solubility ?	Jour de l'Unité Mixte de Recherche 7140	Llompart P., Minoletti C., Baybekov S., Horvath D., Marcou D.,Varnek A.	Strasbourg, France	28 Avril 2023	Oral

Conférences Internationales

Titre	Événement	Auteurs	Lieu	Date	Type
Transforming molecular optimization using collective intelligence	American Chemical Society in Elevating Chemistry	Llompart P., Amaning K., Bianciotto M., Filoche-Rommé B., Foricher Y., Mas P., Papin D., Rameau J.-P., Schio L., Marcou G., Varnek A., Moussaid M., Minoletti C., Gkeka P.	Denver, USA	18-22 Août 2024	Affiche
Drug absorption, a multi-task solution to a multi-parametric problem	Ecole d'été de Chémoinformatique de Strasbourg	Llompart P., Marcou G., Minoletti C., Varnek A.	Strasbourg, France	24-28 Juin 2024	Affiche
Predicting solubility, but which one ?	1ère Ecole d'été en Chemoinformatique	Llompart P., Baybekov S., Marcou G., Horvath D., Gizzi P., Galzi J., Ramos P., Saurel O., Bourban C., Minoletti C., Varnek A.	Paris, France	26-28 Juin 2023	Affiche
Will we ever be able to accurately predict solubility ?	9ème Workshop Franco-Japonais sur les méthodes computationnelles en chimie	Llompart P., Minoletti C., Baybekov S., Horvath D., Marcou D.,Varnek A.	Strasbourg, France	24-25 Avril 2023	Affiche

1.7. Liste des Publications

1) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will We Ever Be Able to Accurately Predict Solubility? Scientific Data, 2024, 11. https://doi.org/10.1038/s41597-024-03105-6.

2) Baybekov, S.; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Kinetic Solubility: Experimental and Machine-learning Modeling Perspectives. Molecular Informatics, 2024, 43. https://doi.org/10.1002/minf.202300216.

3) Llompart, P.; Amaning, K.; Bianciotto, M.; Filoche-Rommé, B.; Foricher, Y.; Mas, P.; Papin, D.; Rameau, JP.; Schio, L.; Marcou, G.; Varnek, A.; Moussaid, M.; Minoletti, C.; Gkeka, E. Harnessing Medicinal Chemical Intuition from Collective Intelligence. Journal of Medicinal Chemistry, 2025. https://doi.org/10.1021/acs.jmedchem.4c03066.

Chapter 2. General Introduction

The pharmaceutical industry faces steep R&D costs and high late-stage failure rates. About 40% of attrition in the early 1990s was linked to poor pharmacokinetics. For a chemical to function as a drug, it must be absorbed, distributed to target areas, metabolized without losing activity, and eliminated effectively while limiting its toxicity (ADMET).[17] ADMET has long been crucial in drug development and has gained increased attention over the past 10 years. Early ADMET assessments reduced project failures to under 10% by 2000, yet the number of novel therapeutics approved by the FDA has been declining, as roughly half of drugs in development still fail due to pharmacokinetic deficiencies, and even approved drugs often present toxicology issues (**Figure 1**).[18,19] Hence, challenges persist, especially in areas like toxicology and clinical safety, necessitating improved and standardized toxicity testing methods.

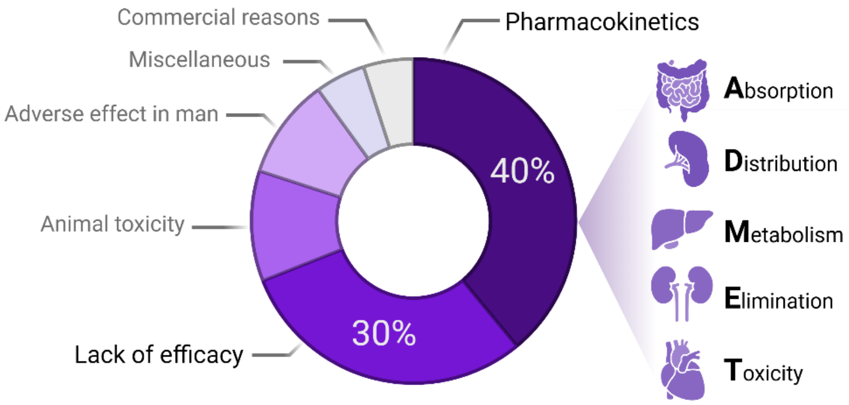


Figure 1: ***Main causes of failure in the drug development phase.***

Recently, several companies have integrated early ADMET considerations with systematic project management to address these challenges. By integrating early attrition strategies and lean modeling, the frameworks seek to reduce ADMET screening costs by preventing unnecessary investment in molecules with low probability of success.[20] For instance, one framework currently used by AstraZeneca has led to a jump in overall success rates (from candidate nomination to Phase III) from 4% to 19% and a shortening of optimization cycle times from 26 to 19 months.[21]

2.1. Drug Discovery & Development

Understanding how various properties affect drug candidates is crucial to improving their success rates. This chapter provides a comprehensive overview of the modern drug discovery and development process.

Overview

Drug discovery is a multi-phase process that spans over a decade and requires significant investments of time and money, often exceeding $2 billion.[4] It begins with identifying potential drug candidates through rational design and properties optimization before advancing to preclinical and clinical trials. The actual model of drug discovery has led to a sharp increase in R&D costs but no significant rise in the number of FDA-approved new molecular entities (NMEs) since the 1990s.[22] From 2009 onward, the cost of bringing a new drug to market has risen by 10% annually. Meanwhile, the investment required has fluctuated between $314 million and $2.8 billion, while the median market exclusivity period for first-in-class drugs has shortened from 10.2 years in the 1970s to 1.2 years in the late 1990s, highlighting the intensification of competition.[4,23]

Early Discovery of Drugs

The earliest forms of drug discovery relied on natural sources, with ancient civilizations utilizing plant extracts, minerals, and animal-derived substances for medicinal purposes. Natural products have been historically successful as antibiotics, chemotherapeutics, immunosuppressants, and crop protection agents.[24] Traditional medicine systems such as Ayurveda, Traditional Chinese Medicine, and Greek pharmacopeia documented the effects of bioactive substancess.[25] During the Renaissance and Enlightenment periods, the extraction and isolation of active ingredients became more refined. Advances in organic chemistry enabled the identification of alkaloids such as morphine, quinine, and strychnine. By the 19th century, Friedrich Wöhler's synthesis of urea in 1828 marked the advent of synthetic medicinal chemistry, proving that organic compounds could be artificially synthesized rather than solely derived from natural sources.[26]

Rise of Rational Design & High Throughput Screening

Between 1990 and 2010, pharmaceutical companies shifted away from natural product discovery in favor of rational design.[24] The 20th century ushered in a revolution in medicinal chemistry, transitioning from serendipity to systematic drug development. The introduction of X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) provided unprecedented insights into drug-target interactions. Simultaneously, the birth of the pharmaceutical industry accelerated the synthesis and screening of chemical libraries. Hence, numerous "targets" linked to diseases where aimed, which some more or less "druggable" (e.g., capacity to be selectively influenced by therapeutics) than others. The emergence of combinatorial chemistry and high-throughput screening (HTS) in the late 20th century dramatically increased the efficiency of drug discovery. Automation and robotics enabled researchers to screen thousands of compounds against biological targets, leading to significant advancements in fields such as oncology and infectious diseases.[27]

The Bid Data & Computational Drug Design Era

The 21st century brought molecular biology, genomics, and bioinformatics. Genomic sequencing has unveiled numerous potential drug targets, while advancements in artificial intelligence (AI) have facilitated structure-based drug design. As drug discovery becomes standardized, more actors are focusing on the same targets, increasing competition and pushing research toward novel territories such as RNA targeting, with the recently FDA-approved Branaplam and Risdiplam.[28] Moreover, the concept of "druggability" is becoming obsolete, as evidenced by RAS targets, considered undruggable during 30 years but have now seen approvals like Sotorasib.[29]

As of today, the drug discovery and development process unfold in three key phases: (**i**) exploratory research to identify and optimize active compounds targeting a specific biological mechanism; (**ii**) preclinical and early clinical testing, first on animals and then on healthy humans; and (**iii**) clinical trials on patients to assess efficacy and safety. If deemed effective and safe, the compound undergoes regulatory approval by authorities such as the FDA, obtaining market authorization. The technical aspects of this process are outlined (**Figure 2**).

From Diseases to Target Validation

The high attrition rate in drug development means that the cost of a single successful drug must cover the expenses of numerous failed candidates, making drug discovery an expensive and resource-intensive process. To sustain innovation and profitability, the pharmaceutical industry must continuously replenish its pipeline with promising targets.[30] This process begins with a molecular and cellular analysis of diseases, where researchers map dysregulated pathways to identify proteins or genes related to pathology. For example, targeting aberrant kinase signaling in cancer or excessive cytokine activity in autoimmune diseases has led to therapies like Tofacitinib, Ruxolitinib, and Osimertinib.[31]

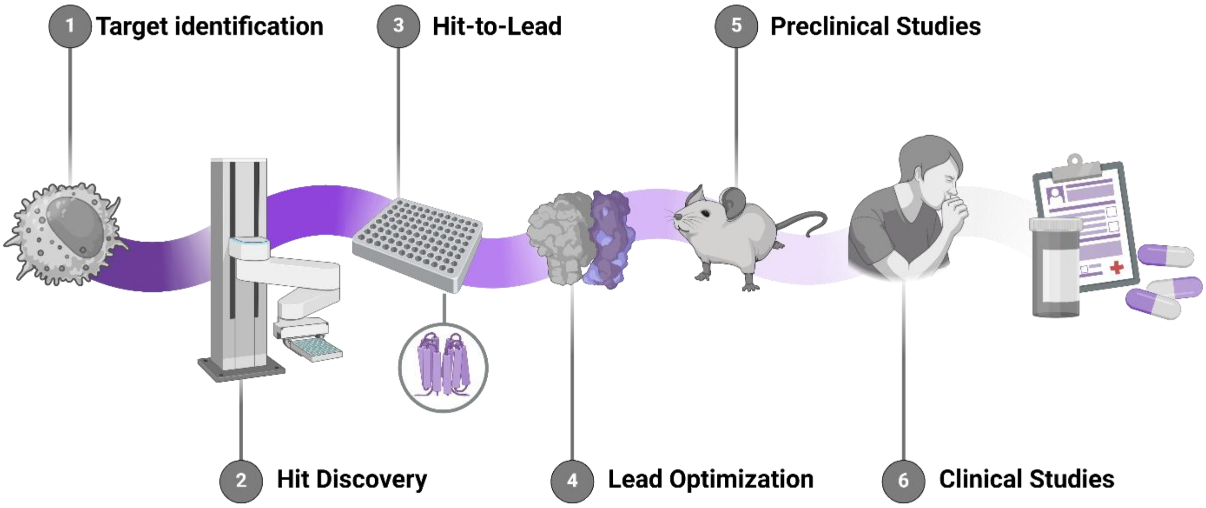


Figure 2: ***Drug Discovery and Development process.*** Schematic representation of the key stages in drug discovery and development. The process begins with Target Identification, selecting a biological target relevant to a disease. Hit Discovery identifies active compounds via high-throughput screening or computational methods. In Hit-to-Lead, candidates are refined for potency, selectivity, and drug-like properties. Lead Optimization enhances efficacy, safety, and pharmacokinetics. Preclinical Studies assess toxicity and pharmacodynamics in vitro and in vivo. Clinical Studies progress through Phase I (safety/dosage), Phase II (efficacy/side effects), and Phase III (confirmation/large-scale evaluation) before regulatory approval and market introduction.

Elucidating the Disease Mechanism

Extensive molecular profiling of diseased tissues and experimental models has revolutionized our understanding of pathological mechanisms. This allowed to identify aberrant tau phosphorylation as cause of neuronal function disruption linked to Alzheimer's disease.[32] Biomarkers (e.g., molecular indicators of disease presence or severity) has been employed to further validate the therapeutic relevance of pathways or protein families, enabling researchers to distinguish true disease drivers from incidental.[31]

Many therapeutic targets are proteins which reside within protein families sharing structural and functional similarities.[33] For instance, kinases are implicated in cancer, immunodeficiencies, viral infections, neurodegenerative diseases, diabetes, and inflammatory diseases with the RAS gene family.[34,35] Other notable examples include G-Protein-Coupled Receptors (GPCRs), which account for roughly 34% of approved drugs and ion channels that have been the target of drug development for the past 50 years (e.g., phenytoin, carbamazepine).[36,37]

Focusing on these target families capitalizes on a wealth of existing data. Strategies vary and may include inhibitors that block activity at orthosteric or allosteric sites, activators that stabilize specific conformations, or compounds that induce structural shifts. Depending on the target’s nature and accessibility, the therapeutics may either be a small molecule, peptide, antibody, or nucleic acid (e.g., siRNA, antisense oligonucleotides).

Target Identification & Validation

Once the target is identified, the validation consolidates its potential by assessing its “druggability”. Validation typically involves in vitro assays or animal models to determine the therapeutic strategy at the molecular level.[31] Without loss of generality and to fix ideas, the target can be a protein. The studies start from observations indicating the biological function, the role of the protein in a biological pathway or a physiological process. Endogenous or exogenous modulators can be discovered during this process, or mutations can indicate strategic molecular mechanisms that can be exploited.

Discovery of Potent Leads

The target identification is followed by the search for active hits. This process typically involves three core steps: **i**) Compound screening, often in a HTS format, to evaluate thousands or even millions of molecules for potency; **ii**) Hit identification, which selects the most promising active compounds from the initial screen; and **iii**) Hit validation, during which potency and specificity are confirmed through secondary assays (**Figure 3**).[38]

Hit Discovery

Hit discovery typically begins with HTS to test large libraries of compounds against a biological target. Over the past two decades, small-molecule drug discovery has been driven by HTS, with estimated hit rates as low as 0.01–1.00%, depending on factors such as the definition of a "hit", target nature, assay type, and the diversity of the compound pool.[7,39,40] In this screening phase, a small fraction of compounds may display measurable activity, although false positives often arise from experimental artifacts or nonspecific binding. Subsequent counter-screens and orthogonal assays help confirm genuine activity, filtering out spurious results.

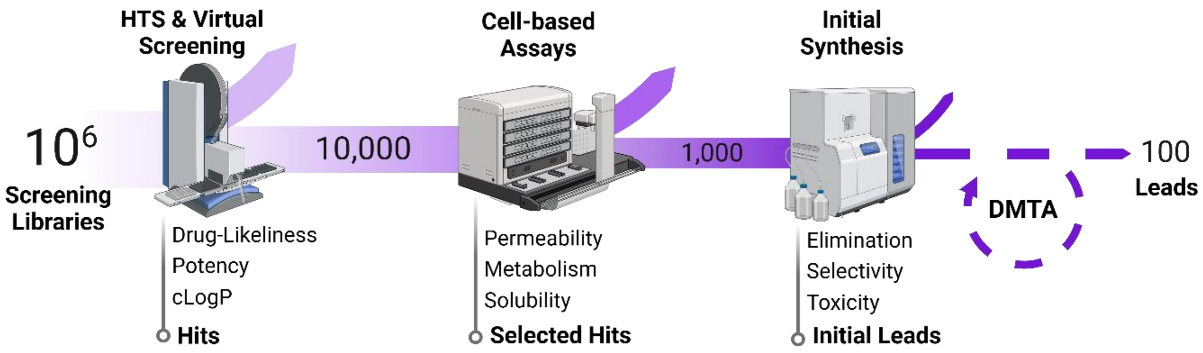


Figure 3: ***Screening process from hit to lead.*** Schematic representation of the screening workflow and the progressive reduction in library size at each stage. The process begins with high-throughput and/or virtual screening, leveraging miniaturized and cost-effective assays. Hits are refined through more complex and informative models such as cell-based assays. Selected compounds proceed to Initial Synthesis before entering the Design-Make-Test-Analyze (DMTA) cycle.

Typical HTS libraries contain up to 10^6-10^7 compounds, and up to 10^{10} compounds with DNA-encoded libraries (DEL), which still represent a minute portion of the chemical space.[41,42] As of 2015, 125 million compounds were commercial available; by 2025, this number exceeded 64.9 billion compounds, driven by the growth of the Enamine REAL collection.[43] These campaigns rely on “screening libraries”. Commercial libraries typically contain millions of purchasable compounds covering diverse chemical scaffolds, while in-house libraries are curated from proprietary research efforts. Nevertheless, the development and maintenance of large in-house libraries remain costly and are typically restricted to the chemical space of past projects, with little to no emphasis on exploring uncharted regions. In response, newer strategies have emerged such as fragment-based screening, DEL, and combinatorial chemistry to create smaller, more diverse libraries that require fewer resources yet maintain broad chemical coverage.[27,44] Regardless of the approach, screening efforts should ensure that hits show consistent activity, are not driven by assay artifacts or nonspecific effects, and do not bind closely related targets. Confirmation typically involves dose–response profiling and orthogonal assays. Although a hit may possess moderate potency (e.g., micromolar IC_{50} values when nanomolar is expected), it can still form the basis for a successful lead if it exhibits potential for on-target optimization and favorable ADMET characteristics (**Figure 4**).[19] Yet, over time, a general trend has also emerged toward the synthesis of larger and more lipophilic compounds, a phenomenon referred to as “molecular obesity”.[40,45,46] These compounds often exhibit high efficacy but poor pharmacokinetics and safety profiles.[47]

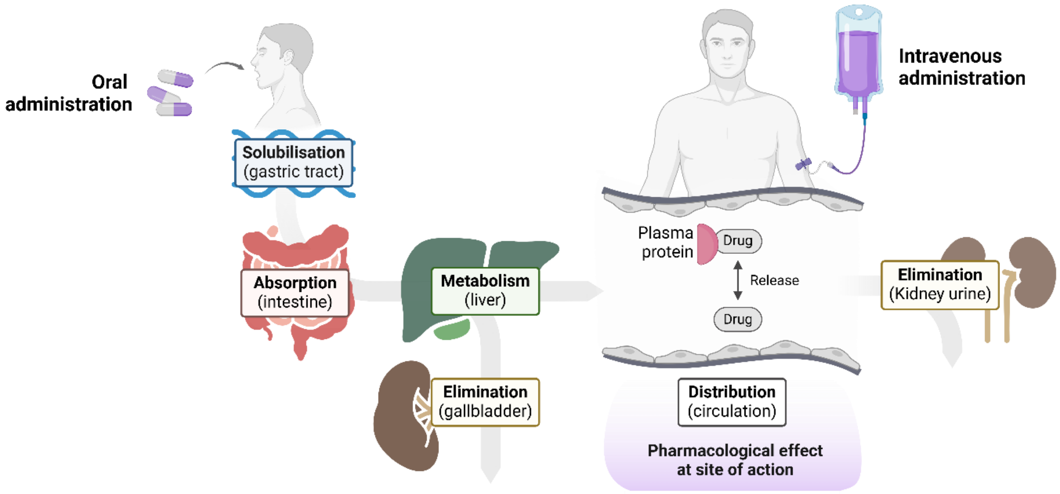

Oral administration
Intravenous administration
Solubilisation (gastric tract)
Absorption (intestine)
Metabolism (liver)
Elimination (gallbladder)
Plasma protein
Drug
Release
Drug
Distribution (circulation)
Pharmacological effect at site of action
Elimination (Kidney urine)

Figure 4: *Journey of xenobiotics from oral administration to elimination.*

Hit-to-Lead

Once a set of validated hits has been identified, the focus shifts to refining and optimizing the best candidates in an iterative optimization process, transforming hits into ”lead” compounds. Medicinal chemists structurally relate hits into congeneric series and analyze structure–activity relationships (SARs), the ways in which small structural modifications affect potency and selectivity (**Figure 5**). This stage initiates multiparametric optimization, a data-driven process aimed at simultaneously satisfying a set of predefined thresholds across several parameters (a.k.a. blueprint); including potency, selectivity, and ADMET properties.[48] In parallel, structure–property relationships (SPRs) examine eADMET endpoints to help identify liabilities before they become insurmountable in later development; for instance, screening for interactions with the hERG channel can mitigate the risk of QT prolongation, a common cause of drug monitoring and withdrawals (Terfénadine), similarly profiling unintended kinase inhibition can flag molecules with off-target toxicity, ensuring safer drug development pipelines. If such safety or pharmacokinetic issues cannot be resolved through structural modifications, the candidate will be discontinued. [49–51]

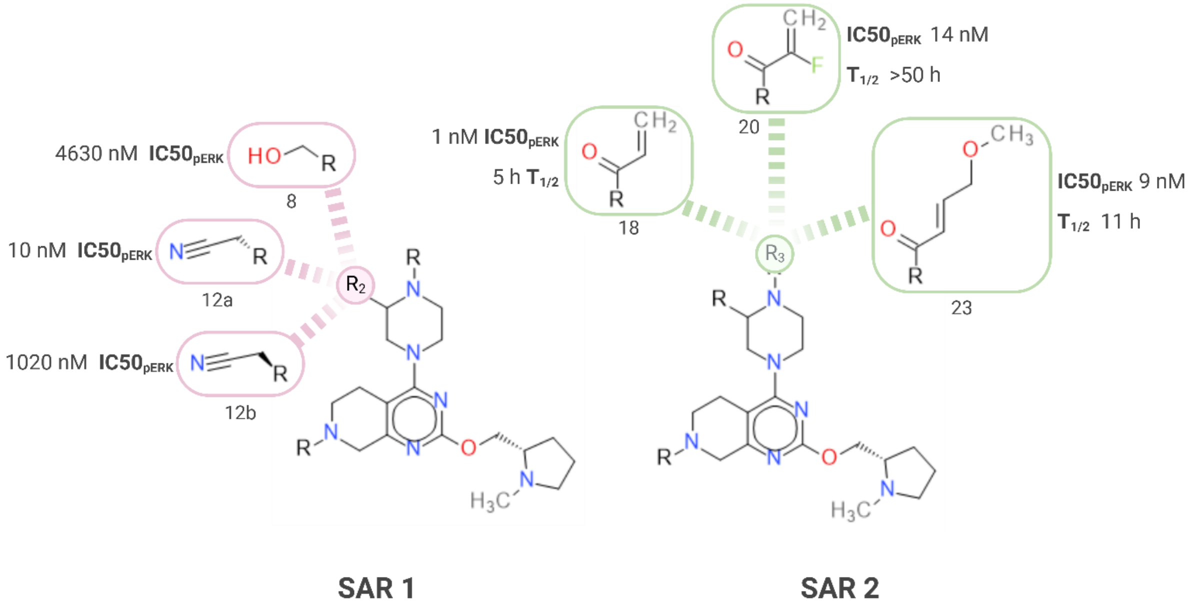


Figure 5: ***Structure-Activity Relationship (SAR) of the MRTX849 congeneric series.*** Illustrations of key molecular modifications and their impact on biological activity, highlighting structural features that contribute to potency and half-life.

Practically, hit-to-lead efforts follow a Design–Make–Test–Analyze (DMTA) cycle (**Figure 6**). In this cycle, chemists supported by modelers design structural modifications based on current SAR/SPR data, synthesize or acquire the new analogs, test them for biological activity and ADMET properties, and finally analyze the outcomes to inform the next design iteration. This method is an embodying application of active learning (AL).[52] Consequently, medicinal chemists face challenging multi-objective optimization (MOO) problems. This steps generally require support from medicinal chemists, chemical intuition, experimental data, and generative or predictive models. Despite its effectiveness, the DMTA cycle can be time-consuming, often taking weeks per iteration. Efforts to enhance efficiency include automated synthesis, simulations, active learning, computational profiling, organ-on-chip and ultimately automated laboratories, aiming to shorten cycles time that still exceed 4-8 weeks with more cost-efficient methods.[27,53] Compounds that balance potency, selectivity, and ADMET profiles emerge from this iterative optimization as ”leads”.

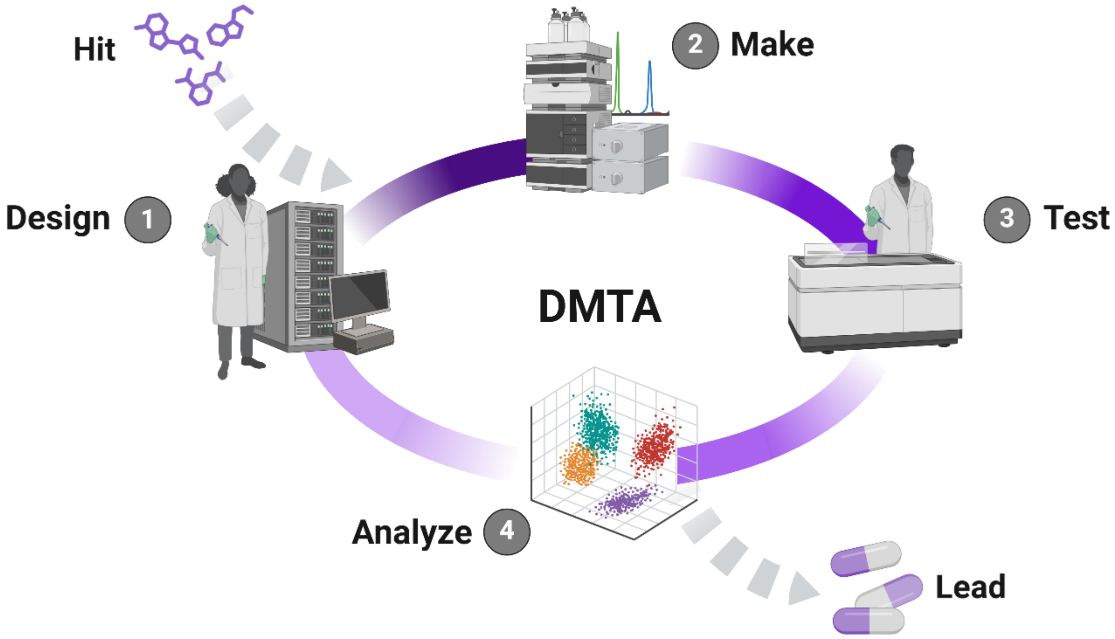


Figure 6: ***Overview of the DMTA (Design-Make-Test-Analyze) cycle.*** The process involves iterative refinement of molecular structures, synthesis of candidate compounds, experimental evaluation, and computational analysis to guide subsequent design steps. Each phase contributes to optimizing properties such as ADMET and bioactivity, ensuring data-driven decision-making in drug discovery.

Identifying a New Drug

Unlike earlier discovery phases, where the emphasis centered on finding any active compound, this stage places greater weight on the lead's properties essential for clinical success.[54]

Lead Optimization

Optimization efforts begin as soon as active compounds are identified. As leads with promising on-target activity emerge, optimization intensifies and profiles become more complex, incorporating additional safety endpoints. These steps refine properties essential for success in preclinical models and, ultimately, in humans. Medicinal chemists leverage MOO techniques, recognizing that bolstering one characteristic, such as improving permeability, may inadvertently result in an undesired hERG binding.[55] This optimization continues until a small number of compounds a.k.a. clinical candidates show maximal potential for efficacious, safe administration, at which point they transition to the more resource-intensive drug development stage. The late-stage transition from lead to preclinical and clinical candidates is subject to high attrition rates. They are usually due to toxicity, as preclinical models often lack translational relevance to humans. Hence, this requires mechanistic toxicology to understand and improve the transferability, with more than 50% of experimental work conducted in-house to ensure speed and flexibility with large companies adopting tiered toxicity screening: combining in silico, in vitro, and in vivo approaches to reduce late-stage failures.[56]

Drug Development

Drug development follows a series of evaluations to confirm a compound's therapeutic potential and safety before it can reach patients. Initially, preclinical studies use both in vitro assays and animal models to assess pharmacodynamics and pharmacokinetics, as well as short- and long-term toxicology. These tests characterize the compound's safety margins such as dose–exposure relationships. It is estimated that less than 10% of the initial chemical entities reach the market.[30]

Clinical trials begin with Phase I, which typically involves a small group of healthy volunteers to establish basic safety, tolerability, and pharmacokinetic profiles. Phase II enrolls patients with the target disease to refine dosing, gather preliminary efficacy data, and further detail the safety profile in a more relevant clinical context. By Phase III, the study population expands significantly to confirm clinical efficacy, identify side effects, and compare the new compound against existing therapeutic standards. Data accumulated from these trials are compiled into a regulatory dossier, such as a New Drug Application (NDA) for the Food and Drug Administration (FDA) or a Marketing Authorization Application (MAA) for the European Medicines Agency (EMA). Regulatory authorities allow the market launch of the drug and mandate post-marketing surveillance (pharmacovigilance) to monitor any long-term or low-incidence adverse effects.[57]

2.2. Computational Approaches in Drug Discovery

Data is the primary driver in drug discovery and development. In recent years, the acquisition of data has surged, not only in quantity, thanks to HTS and combinatorial chemistry, but also in quality, due to the standardization of assay protocols. This wealth has extended the application of AI. Breakthroughs such as AlphaFold, underscore its potential in supporting de novo design.[58] However, the effectiveness of AI is dependent on the availability of vast amounts of precise, high-quality data. In the chemical context, if the data is noisy or of poor quality, AI-driven decisions may yield false positives, leading to suboptimal decisions. Hence, AI outputs must be evaluated. Integrating these computational insights with expert human judgment is essential to ensure that decisions in drug discovery are both accurate and reliable.

Overview

AI-driven chemoinformatics enables tasks such as predicting assay outcomes, mapping chemical space, generating novel compounds, and optimizing molecular structures. Computer-assisted drug design (CADD) refines these in silico approaches, offering more targeted searches than traditional SAR methods. By identifying key molecular interactions and suggesting modifications to enhance activity and ADMET properties, CADD offer higher success rate to hit discovery.

A key application of these computational methods is Virtual Screening (VS), which facilitates hit identification by exploring large chemical libraries. Depending on target structure availability, VS is categorized into structure-based and ligand-based approaches, with hybrid methods integrating both when applicable. The success of hit identification depends not only on the strategy but also on the quality of the library and the complexity of the target itself. By applying AI-driven methods with optimized approaches, drug discovery can be made more efficient and precise (Figure 7).

Structure-based Drug Design

Structure-based drug design (SBDD) enables the identification of bioactive compounds by leveraging the three-dimensional structure of biological targets. Typically, SBDD begins with molecular docking, where chemical compounds (ligands) are positioned within the binding site of a target protein.[59] The binding affinity of each ligand is then approximated using empirical scoring functions. While this method can efficiently prioritize potential drug candidates, its success depends on the availability of high-resolution target structures and the accuracy of docking and scoring algorithms.[60,61]

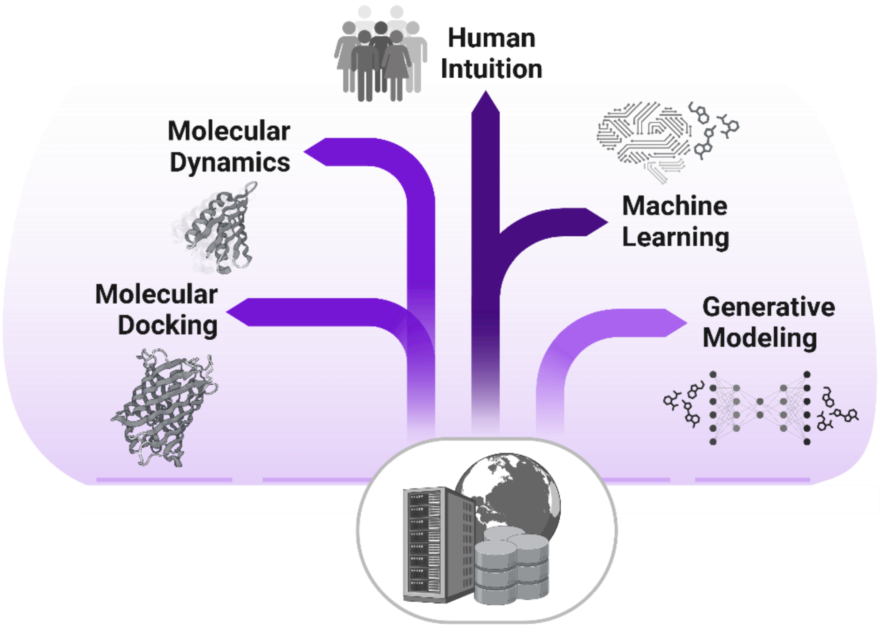


Figure 7: *Key computational strategies in drug discovery.*

Advances in Structural Data for SBDD

The effectiveness of structure-based approaches relies on access to accurate 3D representations of target proteins (**Table 1**). High-resolution structural data are ideally obtained through X-ray crystallography, NMR spectroscopy, or cryo-EM. However, many pharmaceutically relevant targets lack experimentally resolved structures. To address this, computational techniques such as homology modeling can be used to construct 3D models based on known protein structures.

Recent breakthroughs in protein structure prediction, particularly with AlphaFold and the newly introduced AlphaFold3 by Google DeepMind and Isomorphic Labs, have significantly expanded structural coverage. AF3 claims to outperform state-of-the-art docking tools in predicting protein-ligand interactions and offers superior accuracy in modeling protein-nucleic acid interactions compared to specialized predictors like RoseTTAFold. Additionally, recent advancements in structure resolution methods such as cryo-EM have provided unprecedented insights into previously elusive targets, further enhancing the potential of SBDD.[62–64]

Computational Strategies for Ultra-Large Libraries

The rapid growth of ultra-large chemical libraries, often exceeding billions of compounds, presents a computational challenge for traditional docking simulations. To overcome this, novel AL and deep docking workflows have been developed, integrating machine learning techniques with molecular docking to optimize compound selection and sampling efficiency. This allows for more targeted exploration of vast chemical spaces while maintaining computational feasibility.[65,66]

To further refine hit selection, additional more computationally intensive techniques such as binding site water analysis, diversity selection, and absolute free energy perturbation (AB-FEP) scoring can be employed. AB-FEP, for instance, provides highly accurate free energy of binding ($\Delta\Delta G$) estimations, improving the ranking of candidate molecules before experimental validation.

Table 1: ***Structure-based drug design (SBDD) methods.***

Method	Description	Pros	Cons
Molecular Docking $E_{binding} = E_{ligand-protein} - (E_{ligand} + E_{protein})$	Predicts binding using scoring functions. (Hit discovery)	Fast	Poor accuracy
Molecular Dynamics $F = ma$	Simulates ligand-target dynamics	Flexible Dynamic	Slow Force fields
Free Energy Perturbation (FEP) $\Delta G = -RT\ \ln K$	Calculates binding free energy (Lead optimization)	Accurate	Expensive Force fields
Fragment-Based Drug Design (FBDD) $LE = \frac{\Delta G}{\#\ heavey\ atoms}$	Identifies/optimizes small fragments	Novel scaffolds	Weak binders Iterative
Pharmacophore Modeling $d(A, B) \leq d^*$	Defines key activity features	Virtual screening	Needs actives

Where the variable,

$E_{binding}$ is the binding energy estimation (kcal/mol or kJ/mol),

ΔG is the free energy change (binding affinity),

R is the gas constant (1.987 cal/mol·K),

T is the temperature (K, Kelvin),

K is the binding constant (M^{-1}),

LE is the ligand efficiency (kcal/mol per heavy atom).

Ligand-based Drug Design

Ligand-based drug design (LBDD) is a fundamental approach for VS that relies on known bioactive compounds rather than direct structural information of a target protein. This method is particularly valuable when high-resolution target structures are unavailable, allowing researchers to predict activity, toxicity, and other pharmacokinetic properties based on molecular similarity principles (**Table 2**).

Two main ligand-based screening approaches exist, either based on (**i**) 2D similarity search and modeling or (**ii**) 3D shape search and pharmacophore modeling. 2D approaches are based on fingerprints, allowing rapid and scalable evaluation but lacking 3D information such as the binding poses. 3D methods are mainly employed to compare molecular conformations based on shape and volume. They are relevant to capture 3D signals such as chirality but suffer from inaccuracies of the physical models they are based on, and of their irrelevance when taking into account the complicated and unknown processes driving the interactions of the ligand to a target.

LBDD typically starts with one or more reference compounds with known biological activity. These molecules serve as templates for computational searches across large chemical libraries using 2D or 3D similarity screening. The underlying assumption is that structurally similar compounds often exhibit similar biological activity, a principle known as ″chemical promiscuity″. Once SAR data are available, in silico models can be developed, validated, and used for compound selection.[67]

These models rely on chemical representations to predict experimental endpoints through Quantitative Structure Activity/Property Relationship (QSAR/QSPR) modeling which has been essential in drug discovery for more than 60 years. QSAR relying on deep learning models a.k.a. Deep QSAR have emerged over the past 20 years, enabled by advances in neural networks, computational power, and large molecular databases.

Table 2: ***Ligand-based drug design (LBDD) methods.***

Method	Description	Pros	Cons
QSAR/QSPR $Y = f(X)$	Model experimental data from molecular descriptors	Predictive	Data quality sensitive
Similarity Searching $S = \frac{A \cap B}{A \cup B}$	Finds compounds similar to actives	Fast	Lacks novelty
Ligand-Based Virtual Screening (LBVS) Ranking & similarity scores	Screens libraries by ligand similarity	Quick filtering	Needs known ligands
Generative Models AI/ML-based molecule generation	Generates molecules using deep learning or combinatorial approaches	Novel scaffolds	Outputs can be unusuable

Where the variable,

X is the molecular descriptor,

Y is the predicted biological activity.

Machine Learning & QSAR

Drug design is a sampling problem where medicinal chemists select promising candidates from an unimaginably large pool of compounds. Random selection is impractical due to unfavorable odds.[52] Once SAR data are available, predictive models (e.g., Random Forest) and pharmacophore-based methods can be employed to virtually screen those libraries. Such strategies aim to avoid undesirable compounds and focus on ″activity islands″ containing attractive entities for specific drug discovery projects. Machine learning (ML)-guided virtual screening offers a promising support by rapidly evaluating compounds in silico, enabling the exploration of chemical libraries orders of magnitude larger than those accessible by traditional HTS.[68–71]

De-Novo Generation

Beyond screening existing libraries, generative models, viewed as a form of pattern matching in chemical design, allow the automated design of new compounds. Since the 1990s, computer-based de novo design methods have served as idea generators to support drug discovery.[72,73] Unlike earlier molecular design engines that relied on explicit chemical transformations, such as virtual reaction schemes based on reaction and assembly rules like fragment growing and linking; generative models represent chemical knowledge implicitly through statistical probabilities derived from data distributions. This means the "language" of these models is not traditional textbook chemistry but learned from training data. Chemical language models are able to design novel molecules and optimize bioactivity when guided with Reinforcement Learning (RL) methods. For instance, deep QSAR-guided generative model where shown to produce a RORγ inverse agonist (IC_{50} = 370 nM) and a PI3Kγ inhibitor (K_i = 63 nM).[65,74]

Requirements for Improved Methods

Despite significant scientific and technological progress, R&D productivity in drug discovery declined between 1950 and 2010, largely due to an over-reliance on high-throughput screening and limited predictive models.[75,76] These methods were hindered by restricted computational power, small and noisy datasets, and a narrow pool of available algorithms that failed to capture the complexity of biological systems. To address this, drug discovery is shifting from brute-force screening to intelligent, AI-enhanced strategies. Deep learning-driven QSAR, generative AI, and advanced computational models, when coupled with high-quality data, offer a more efficient and predictive approach. By integrating these tools, researchers can refine candidate selection, reduce failure rates, and maximize success, making the process both faster and more cost-effective.[40]

Chapter 3. Modeling for Drug Discovery

In this chapter, we provide an overview of the core concepts and techniques that underpin this work. We begin by discussing how molecular data are transformed into learnable representations for machine learning models. Next, we review the principal source of data, their standardization, and the methods to apply them in a drug discovery context, with a special focus on graph neural networks, multitask methods, and applicability domains.

3.1. Compound Representation & Databases

Training a predictive model requires a structured, tabular dataset that accurately describes each sample. Since molecules are composed of various atoms and bonds, specialized processing is necessary to render them compatible with machine learning algorithms. In this section, we review the common input formats in cheminformatics and the primary methods used for molecular featurization.

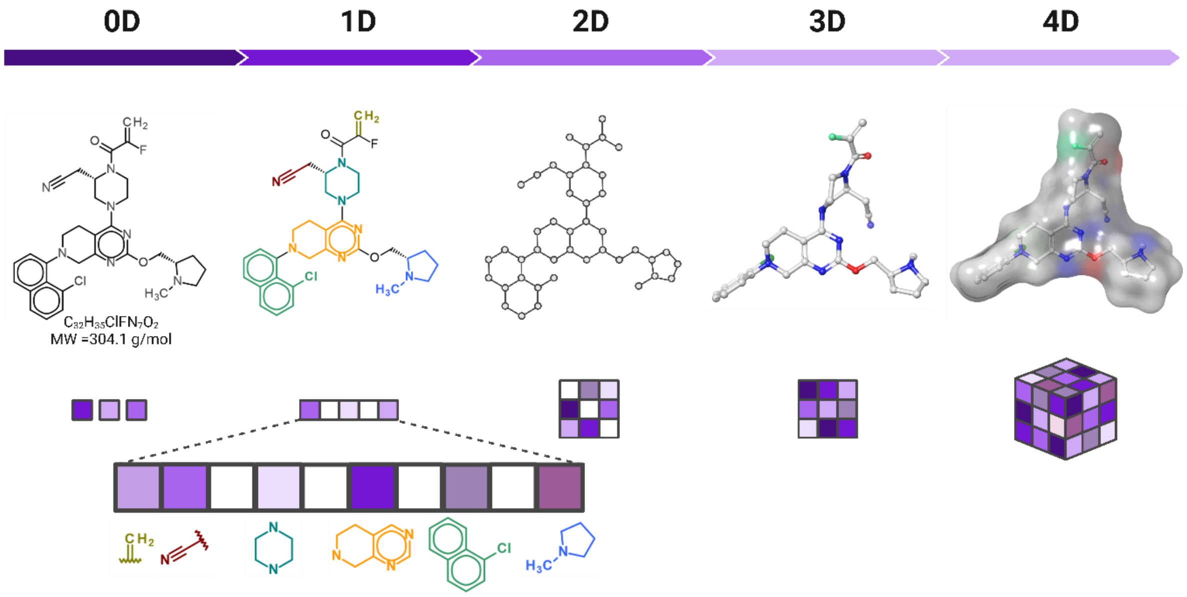


Figure 8: ***Representations of MRTX849 across different feature dimensions.*** (**0D**) Scalar properties such as molecular weight; (**1D**) feature-based description; (**2D**) graph-based representation; (**3D**) spatial conformation; and (**4D**) dynamic representation incorporating molecular flexibility and conformational ensembles.

Levels of Information

Molecular descriptions can be represented at varying levels of detail, ranging from coarse metrics (such as simple atom counts) to more precise multidimensional representations (e.g., 3D, 4D or 5D models). In the context of ligand-based modeling, we mainly focus on two-dimensional (2D) representations. The chosen level of detail involves a trade-off between the quantity of information provided, its distillation, and its overall accuracy (**Figure 8**).

SMILES Representation

The SMILES (Simplified Molecular Input Line Entry System) format is a one-dimensional, token-based representation widely used in the community. SMILES encodes a molecule's connectivity and bond types in a human-readable, lightweight string. It can also include annotations for stereochemistry, isotopic labels, or points of substitution. One limitation of SMILES is that a single compound can be represented by multiple valid strings, and small modifications to a SMILES string may result in an incorrect molecular structure. Canonical SMILES address this by applying deterministic rules to ensure a unique representation, using graph-based atom ranking, lexicographic sorting, and a standardized traversal path.

Graph-based Representation

Alternatively, molecules can be naturally represented as graphs, where atoms are depicted as nodes and bonds as edges. In this formalism, a molecule is modeled as a graph G(V, E), where V denotes the set of vertices, here atoms and E the set of edges, here bonds. Connectivity information is typically captured using an adjacency matrix, and atoms are assigned identifiers based on a chosen numbering scheme. Although graph-based representations provide an intuitive visualization and are well-suited for many machine learning applications, different numbering approaches may lead to variations in node labeling. Hence, variations exist across software implementations due to differences in ranking algorithms that can be answered using identifiers.

Chemical Identifiers

The IUPAC International Chemical Identifier (InChI) encodes molecular structures in a layered, deterministic format, capturing atomic composition, connectivity, hydrogenation, stereochemistry, and isotopic information. Unlike SMILES, InChI ensures a unique, software-independent representation for identical structures, enhancing data consistency. The InChI Key is a fixed-length, hashed version of InChI, optimized for fast indexing and searching. As a lossy transformation, it is non-reversible, preventing structure reconstruction while ensuring efficient molecular identification. When used alongside SMILES, it provides a robust dual-check system for confirming the uniqueness of chemical entries across molecular registries.

Synthesizable Compounds Libraries

Hit identification relies on screening large, diverse, and synthetically accessible libraries which have over the years increased exponentially in size (**Figure 9**). Ultra-large, make-on-demand collections now exceed 100 billion compounds.[77] High-throughput methods benefit from vast datasets, while computationally intensive approaches require smaller, curated selections. Larger screens explore broader chemical space but come with trade-offs in accuracy, cost, and efficiency.[78,79]

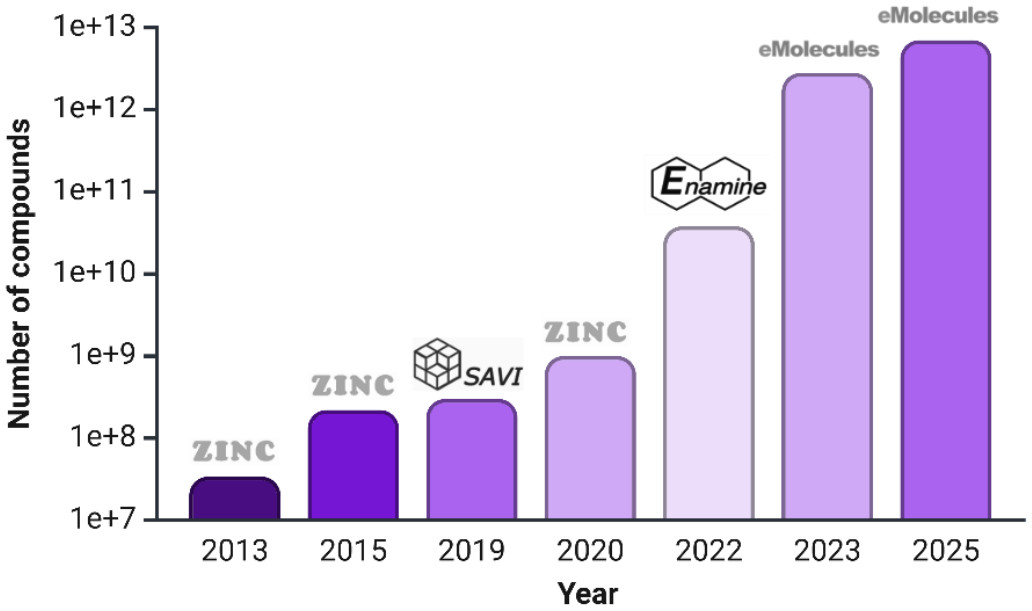


Figure 9: ***Growth of synthesizable small molecule libraries (2013–2025).*** Evolution of the largest available small molecule libraries over time.

Commercial Libraries

Commercial libraries like Enamine, ChemBridge, Life Chemicals, Asinex, Specs, Maybridge HitFinder, and Prestwick Chemical Library provide diverse, non-annotated compounds essential for hit discovery. The Enamine collection contains several chemical libraries based on combinatorial reaction of building blockS. They represent one of the largest libraries, with compounds deliverable under 2 weeks with a probability of synthesis of 80% or more. For instance, Enamine REAL database now offers 5.5 billion unique compounds, expandable to ~38 billion. As of today, Enamine represent the main distributor of screening libraries for HTS.[65]

ZINC

The ZINC database is a publicly available chemical library containing nearly 2 billion compounds, sourced from repositories such as PubChem, ChEMBL, and commercial vendors. The database grew from less than 1 million molecules in 2006 to more than 37 billion in 2024, a 50,000-fold increase. Each molecule is annotated with purchasability details, vendor sources, and key physicochemical properties such as molecular weight, LogP, hydrogen bond donors/acceptors, and rotatable bonds. ZINC also provides pre-generated 3D conformations for docking and structure-based virtual screening.[80]

Experimental Databases

Measuring compound properties requires costly, time-consuming assays prone to variability. Additionally, in vitro testing is restricted to stable, water-soluble compounds, limiting chemical space exploration. To address these challenges, public databases aggregate and standardize experimental data from publications, patents, and large screening campaigns. With over 20,000 to 30,000 new compounds published annually, various digital repositories provide structured access to medicinal chemistry data. PubChem, ChEMBL, and BindingDB offer quantitative bioactivity data, while DrugBank and KEGG provide binary interaction data. ChEMBL features expert-curated records, while BindingDB integrates data from literature and patents.

PubChem BioAssay

PubChem is the largest publicly accessible bioactivity database, containing approximately 230 million bioactivity records covering over 320 million compounds.[81] The database is structured into three main components. PubChem Substance stores chemical substance data submitted by contributors. PubChem Compound consolidates multiple substance records into unique compound entries using automated processes to reduce redundancy. PubChem BioAssay compiles bioactivity results, primarily from HTS and confirmatory assays. Despite its vast coverage, PubChem data is not curated, meaning that inconsistencies and errors may exist.

ChEMBL

ChEMBL is a publicly available bioactivity database maintained by the European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL).[82] Originally developed by Galapagos NV under the name StARlite, it was acquired by EBI in 2008 and has since evolved into one of the most widely used resources in drug discovery. As the second largest bioactivity database, ChEMBL contains over 15 million bioactivity records. Data is extracted from scientific literature, patents, and external databases. However, variability in experimental conditions necessitates further filtering before computational analysis.[24] The database undergoes regular annual updates, ensuring the continuous integration of new data. ChEMBL has been extensively used during the past 15 years. Numerous studies have curated its measurements to deliver publicly available datasets ready for QSAR.

BindingDB

BindingDB is a publicly available database dedicated to experimentally measured binding affinities of small molecules interacting with protein targets.[83] The database contains over 1 million binding affinity measurements, derived from both cell-based assays and isolated protein target assays. BindingDB provides data in 2D and 3D formats.

OChem

OChem is a publicly available database primarily focused on ADMET properties.[84] OChem provides a richly annotated collection of physicochemical and pharmacokinetic data. Users can filter and customize datasets based on specific metadata. OChem's dataset is partially orthogonal to ChEMBL, meaning it complements but does not entirely overlap with ChEMBL's bioactivity-focused records.

Benchmark Datasets

In recent years, several benchmark datasets have been developed for the ML community to evaluate predictive models, including Tox21, Therapeutic Data Commons, and MoleculeNet.[85–87] These datasets compile ADMET and bioactivity data obtained either from ChEMBL or from sources published over the past two decades. While they have been widely adopted for benchmarking by the machine learning and deep learning community, their quality and standardization remain suboptimal, limiting their direct applicability to real-world pharmaceutical challenges.[88]

3.2. Data Preprocessing & Featurization

Building a predictive model starts with data acquisition, assembling a curated dataset of chemical compounds with experimentally measured activity values.

Curation & Filtering

Datasets tailored to specific properties often undergo modeling, ensuring prior quality checks. Reviewed and published datasets reduce reliability concerns, but many ADMET datasets compile data from varied sources with inconsistent experimental details. Limited assay information and missing references hinder data verification, affecting quality. While smaller datasets may offer higher accuracy, they restrict chemical space coverage, potentially limiting model applicability.

Data Caveats

Obtaining homogeneous data remains a challenge due to the cost and time required for experimental measurements. When direct experimental determination is not feasible, researchers rely on public datasets, which vary in format and reliability. Some datasets are structured and tailored to specific biological activities, while others aggregate diverse experimental measurements from the literature. Regardless of the source, an ideal dataset must meet three essential criteria: reliability, homogeneity, and sufficient size. Reliable data minimizes errors introduced by transcription, structural misclassification, or inconsistencies between different sources. Errors can arise from automated structure conversion, manual transcription mistakes, or misclassification of molecular properties.[88] Additionally, datasets often contain duplicates, mixtures, or undefined stereochemistry, which introduce biases into machine learning models. To address these challenges, researchers employ cross-validation of activity measurements for the same compound across different sources, detect outliers based on significant deviations in predicted activity, and standardize assay protocols to improve comparability.

Good Practices

ADMET data is limited in experimental quantity. Yet, the dataset size must be sufficient to meaningfully represent the desired chemical space. Global models aim to cover a broad chemical landscape, while local models focus on congeneric chemical series. While no universal rule exists, studies indicate that dataset size significantly impacts model performance. Tropsha highlighted that excessively large datasets complicate model construction, whereas small datasets risk random correlations and overfitting.[89] To mitigate these risks, careful feature selection and validation techniques are necessary to prevent overfitting and ensure model reliability.

Structure standardization begins with the removal of unwanted molecular components, such as fragments, solvent molecules, and counterions, ensuring that only the primary molecular structure is retained. Neutralization and tautomer standardization are performed to create uniform molecular representations. The dominant tautomer is selected based on predefined rules to ensure structural consistency across the dataset.

Table 3: ***Molecular descriptors across different dimension levels.***

Dim.	Description	Examples	Pros	Cons
0D	Basic properties and counts	Molecular weight Atom/bond count Rings count	Simple Fast Interpretable	Can't distinguish isomers
1D	Drug-likeliness Physchem features	LogP, LogD, pKa Solubility	Key for ADMET endpoints	No structural context
2D	Graph-based Connectivity informations	ECFP, MACCS keys, Wiener/Balaban index	Encodes structures Key for QSAR	No 3D info Sensible to hashing
3D	Spatial and electronic descriptors	Volume, dipole HOMO-LUMO	Show conformation Reactivity info	Needs conformers Slow
4D	Dynamic and pharmacophore informations	Molecular Dynamics Pharmacophoric keys	Includes flexibility Realistic	Compute intensive

For datasets requiring pH-dependent ionization, tools are used to adjust structures at a physiological pH of 7.4 or specific pH depending on the assay conditions. Duplicate removal is then performed by generating canonical SMILES and InChI Key, allowing identification and elimination of redundant structures. To ensure high-quality predictive modeling, each compound in the dataset must have a unique value. A decision is made based on data reliability, either retaining the most experimentally validated value or using statistical aggregation (e.g., median value) if multiple valid measurements exist.

Featurization

The section below presents common featurization of chemical structures. The methods take as input a molecular graph and transform it in a vector of numerical information describing the physicochemical or structural properties of the molecules. In this study, we focus on 0D, 1D, and 2D descriptors, as the bioactive conformation of molecules is usually unknown. They are commonly used for similarity searches, virtual screening, and modeling (**Table 3**).

Structural Keys & Physicochemical Descriptors

Structural keys are coded into fingerprints based on a predefined dictionary of fragments. Fingerprints are numerical vectors that encode molecular structures by representing the presence or absence of specific substructures by bits. One of the most widely used structural key-based fingerprints is MACCS (Molecular ACCess System) keys, introduced by Molecular Design Limited (MDL).[90] This system defines 166 structural keys, optimized for molecular similarity searches (**Table 4**). An extended 320-bit version is also used for broader substructure coverage. These fingerprints assign each bit to a specific chemical motif. Physicochemical descriptors capture medicinal chemistry properties. They include simple parameters such as hydrogen bond donors, as well as more complex engineered descriptors like topological polar surface area. While interpretable, they lack detail.

Molecular Fingerprints

Fingerprints can be categorized into linear connectivity-based and circular topology-based representations. Linear Connectivity-Based Fingerprints enumerate all substructures by considering the shortest paths between connected atoms. Each path is hashed as a bit in the fingerprint. Circular fingerprints, also known as atom-centered fragments (ACF), define a sphere around a central atom and capture all neighboring atomic connections within a predefined radius. These are repeated for every atom. However, since a single bit can correspond to multiple fragments due to hash collisions, different substructures may be mapped to the same position. This redundancy introduce ambiguity, potentially misleading models. Some fingerprints go beyond binary encoding and track how often each substructure appears. These frequency-aware versions retain more structural information and improve performance in machine learning tasks by reducing ambiguity and enriching feature representation.

Table 4: ***Overview of molecular descriptor calculation tools.*** Summary of key computational tools for molecular descriptor calculation, categorized by license type, descriptor range, and features.

Tool	License	Type	Description
RDKit[91]	Open-source (BSD)	~200 descriptors (0D-3D)	Constitutional (MW, atom/bond counts), physicochemical (LogP, PSA, H-bond donors/acceptors), topological (connectivity indices), geometric (3D volume, shape), electronic (formal charge), fingerprints
CDK[92]	Open-source (LGPL)	~100 descriptors (0D-2D)	Constitutional, physicochemical, topological, substructure-based
MORDRED[93]	Open-source (MIT)	~1800 descriptors (0D-3D)	Constitutional, physicochemical, topological, electronic (HOMO-LUMO), geometric features
ISIDA[94]	Proprietary	Variable size (fragment-based, ~100-10,000)	Substructure descriptors including atom-centered fragments, bond-centered fragments, connectivity matrices
MOE[95].	Proprietary	~300 descriptors (0D-3D)	Constitutional, physicochemical, topological, electronic, geometric, and pharmacophore-based descriptors
Avalon[96]	Open-source	Variable size (~1000-10,000)	Topological, substructure-based, hashed fingerprints
PubChem[81]	Open-source	Large set (~700-1000)	Constitutional, physicochemical, topological, fingerprints
MACCS[97]	Open-source	166-bit	SMARTS-based substructure keys encoding molecular fragments and functional groups
AtomPairs[98]	Open-source	Variable size (~1000-10,000)	Connectivity , atom pairs, distance-based fragment pairs

The Morgan algorithm (1965) was initially developed to solve molecular isomorphism problems by iteratively assigning numerical identifiers to atoms. This laid the foundation for Extended Connectivity Fingerprints (ECFPs).[99] ECFPs encode substructures in binary vectors. ISIDA fragment descriptors have been developed as refined topological fingerprints that count fragment occurrences rather than binary presence.[94] ISIDA provides linear (sequence-based) and circular (ACF-based) fragmentation, offering. Various fingerprinting methods have emerged, including Avalon, PubChem and CDK fingerprints.[92,100]

Model Embedding

Molecular featurization can also be achieved through model embedding, where compounds are represented as continuous latent vectors generated from pretrained neural networks. Embeddings are learned representations optimized for predicting target endpoints (**Figure 10**). However, they typically have high dimensionality and non-linear relationships, making them harder to interpret. Additionally, embedding may struggle with out-of-distribution molecules. Several popular architectures have been explored for molecular embedding, including Graph Neural Networks (GNNs), Variational Autoencoders (VAEs), and Transformer-based models. Examples include Mol2Vec[101], which adapts word embeddings, and ChemBERTa[102], a bidirectional Transformer-based molecular representation model

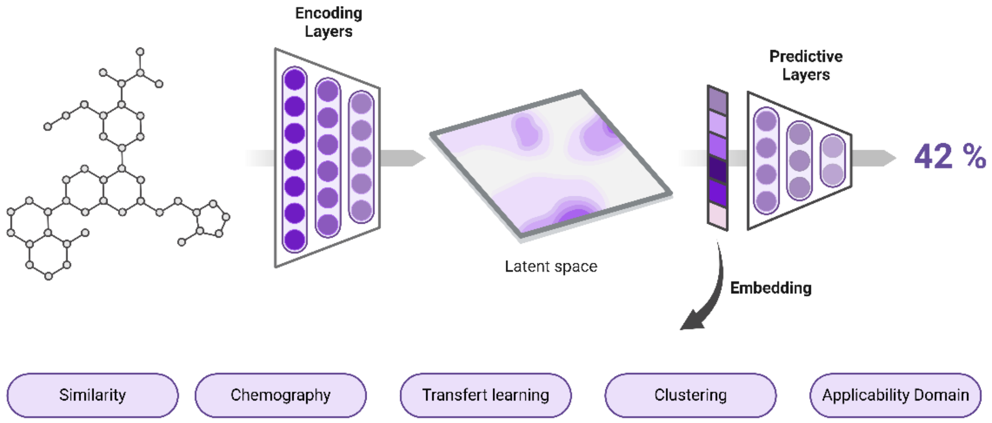


Figure 10: ***From molecular graph to predictive applications methods.*** Representation of the transformation from a molecular graph to predictive applications.

3.3. Chemical Space Cartography

In this section, we discuss the various methods of representation of a chemical space on a 2D maps. Chemical space, an abstract landscape where molecules are positioned on a map based on similarity.

Linear

Unsupervised methods like Principal Component Analysis (PCA) reduce dimensionality by transforming correlated descriptors into uncorrelated principal components. PCA optimally reconstructs the dataset using linear combinations of features and captures maximum variance along the first components. However, it does not exploit potentially simpler non-linear structures. The first principal components capture the maximum variance in the dataset, but its reliance on orthogonal transformations limits its ability to capture complex non-linear relationships.[103]

Non-Linear

t-SNE

Non-linear dimensionality reduction methods are widely used to map high-dimensional chemical spaces into interpretable 2D projections. t-SNE is a probabilistic, non-linear technique that projects high-dimensional data into lower dimensions while preserving local relationships. It converts Euclidean distances between molecules into conditional probabilities, representing the likelihood of two points being neighbors. The algorithm then minimizes the difference between probability distributions in the high- and low-dimensional spaces, ensuring that molecular neighbors remain close in the 2D map. While effective at capturing local structures, t-SNE can struggle with global relationships and is computationally expensive for large datasets.[104]

UMAP

UMAP constructs a graph-based representation of data, optimizing embeddings via Riemannian geometry and fuzzy simplicial sets. It preserves both local and global structures more effectively than t-SNE, offers faster computations, and supports out-of-sample projections through a learned transformation function.[105]

Generative Topographic Mapping

GTM is a probabilistic dimensionality reduction method introduced by Bishop et al. (1998) as an extension of Self-Organizing Maps (SOM).[106,107] Unlike t-SNE and UMAP, which rely on stochastic neighbor embeddings, GTM models high-dimensional data as a continuous manifold embedded in a lower-dimensional space (**Figure 11**). GTM represents the high-dimensional descriptor space using a grid of nodes, where each node is assigned a set of Gaussian Radial Basis Functions (RBFs). These functions define a smooth, flexible surface that is trained to match the distribution of molecular data. The manifold is iteratively adjusted to fit the densest regions of the dataset, ensuring that molecules with similar properties remain close on the 2D projection. Once trained, molecules are mapped onto the manifold and subsequently unfolded into a readable 2D representation. Instead of direct point placement like in t-SNE or UMAP, GTM assigns responsibility scores to molecules, indicating their association with different nodes on the grid. However, it requires careful tuning of grid resolution and RBF parameters to achieve optimal performance.[108]

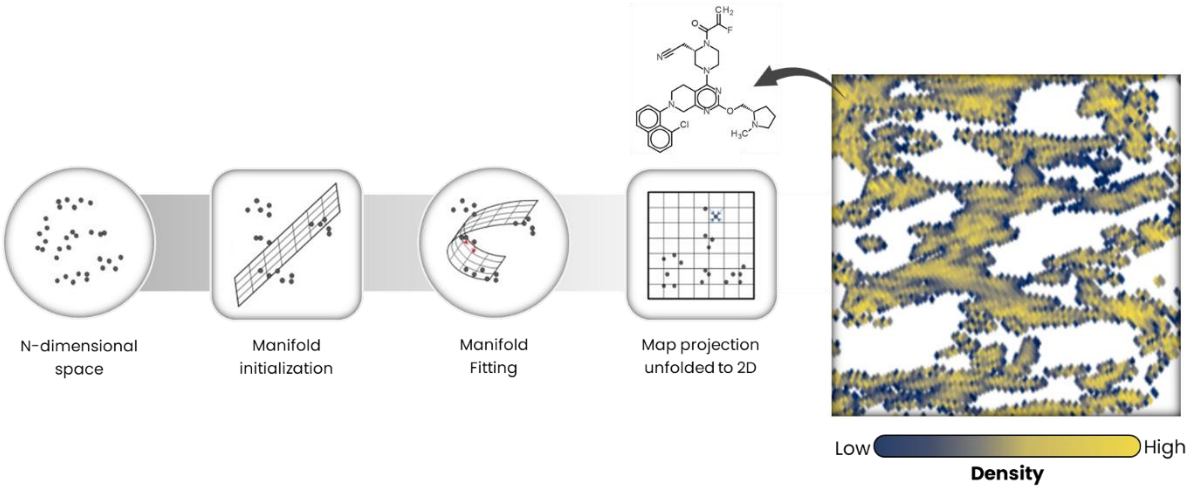


Figure 11: ***Generative Topographic Mapping (GTM) training workflow.*** The GTM fits a manifold (a square bounded bidimensional geometric object) on the dataset embedded in high dimensional space. Each data point, corresponding to molecules, is explained as a sample of a normal distribution centered on the manifold. Conversely, each position on the manifold is associated to a density of molecules in the chemical space. Sampling on the GTM means sampling the chemical structures from a region of interest of the chemical space

3.4. QSPR Modeling

The SAR and SPR concepts originated in 1868 when Crum-Brown and Fraser[109] proposed that a compound's biological activity is directly linked to its chemical structure. This idea was further developed by Richet[110], Meyer[111], and Overton[112], who demonstrated strong correlations between molecular properties and biological effects. The modern QSAR era began in 1964 with Hansch et al., who formulated the first mathematical model predicting partition coefficients using electronic descriptors, setting the foundation for systematic structure-activity modeling.[113]

QSPR/QSAR models aim to predict a chemical property (Y) from molecular descriptors (X) using a mathematical function f(X). These models enable researchers to estimate biological activity, toxicity, or physicochemical properties of untested compounds. A functional QSAR model requires three essential components: (**i**) a dataset containing experimental values for the target property, (**ii**) a set of molecular descriptors that encode relevant structural and physicochemical features, and (**iii**) a statistical or machine learning approach to infer the relationship between X and Y. To ensure scientific rigor and regulatory acceptance, the OECD (Organisation for Economic Co-operation and Development) has proposed a set of five widely recognized principles for QSAR model validation. These principles emphasize the importance of: (1) a defined endpoint relevant to human or environmental health, (2) an unambiguous algorithm that is transparent and reproducible, (3) a clearly described domain of applicability indicating where the model makes reliable predictions, (4) robust performance metrics evaluated through internal and external validation procedures, and (5) mechanistic interpretability where possible. These guidelines have become a cornerstone in the development of regulatory-compliant predictive models for chemical safety assessment.

While traditional QSPR models have proven valuable, their scope can be limited by the assumption of linear relationships between molecular descriptors and properties. This is where machine learning approaches such as decision tress come into play, offering a more flexible, non-linear approach to modeling that can better capture the complexities of structure-activity relationships.

Machine Learning

Decision Trees

Decision tree training follows a recursive partitioning approach.[114] At each node, a descriptor (X) is selected as a splitting criterion, based on its ability to minimize the variance of the target property (Y) within each subset (**Figure 12**). This process continues until a stopping condition is met, such as a minimum number of data points per leaf or a threshold on variance reduction. The splitting threshold for each descriptor is determined by optimizing the sum of squared errors, ensuring that each branch maximally explains the variance in Y. To prevent overfitting, decision trees undergo pruning, where unnecessary branches that do not significantly improve prediction accuracy are removed. Overly complex trees can lead to models that fit noise rather than true relationships, reducing their generalizability. Conversely, trees that are too shallow may suffer from underfitting, failing to capture meaningful patterns in the data. Decision trees are valued for their interpretability and ability to handle non-linear relationships without requiring extensive data preprocessing.

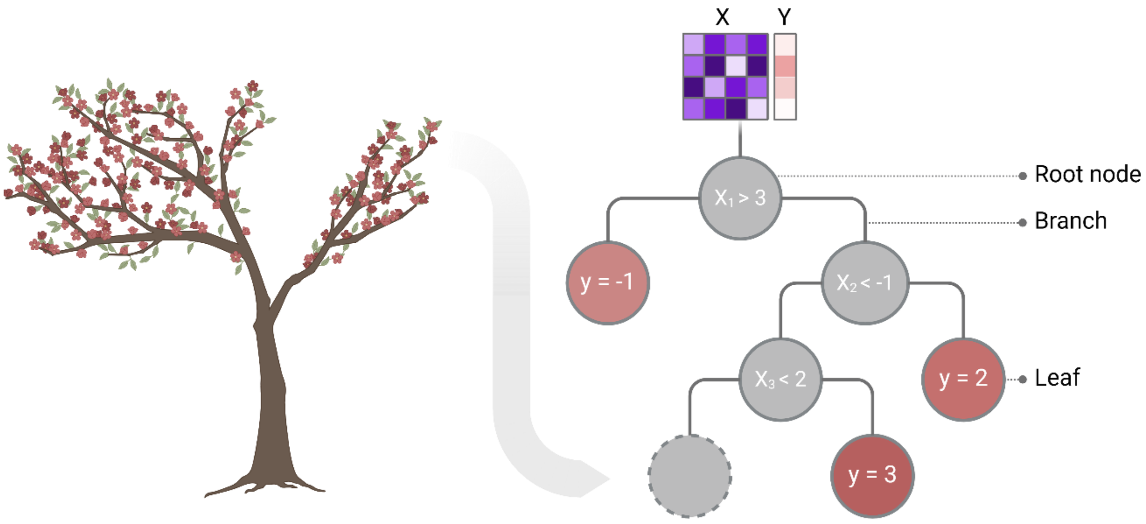


Figure 12: ***Decision tree structure for classification or regression.*** Input features X and Y are evaluated at the root node, where the first split occurs based on a decision criterion. Branches represent conditions leading to further nodes. At the end of each path, leaf nodes contain the final predictions or classifications.

Random Forest

To overcome the limitations of single decision trees, ensemble learning techniques aggregate multiple models to improve predictive accuracy and reduce variance.[115] Random Forests use bootstrap aggregation (bagging), where multiple decision trees are trained on random subsets of the training data and molecular descriptors (**Figure 13**). Each tree makes an independent prediction, and the final output is obtained by averaging the predictions (in regression) or taking a majority vote (in classification). The randomness introduced in feature selection and data sampling reduces overfitting, leading to better generalization.

Gradient Boosting

Unlike Random Forest, Gradient Boosting improves model performance through sequential training, where each tree corrects the errors of the previous one. Unlike bagging, which trains models independently, boosting assigns higher weights to mispredicted samples, ensuring that each subsequent tree focuses more on difficult cases. This makes boosting methods like AdaBoost[116] and XGBoost[117] highly effective for improving accuracy, especially in noisy datasets. Gradient Boosting models tend to outperform Random Forests in many regression tasks but are more computationally expensive and prone to overfitting if not properly regularized.

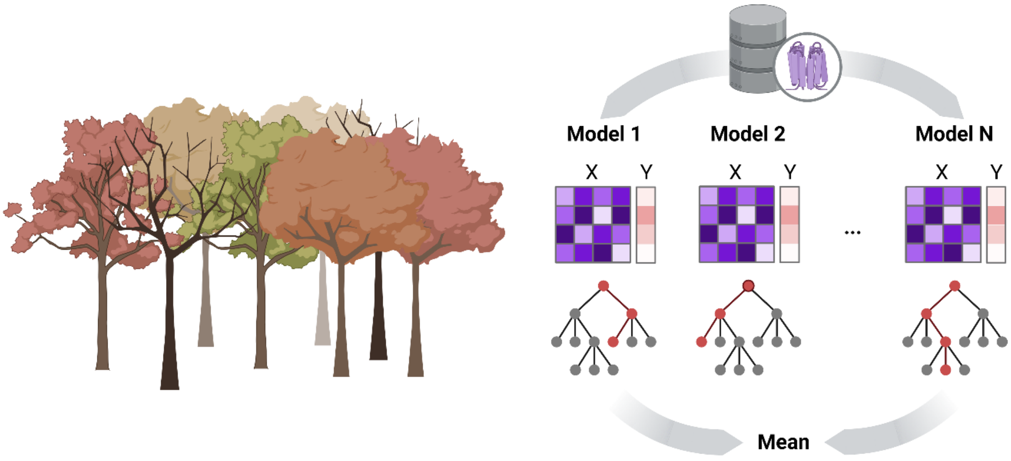


Figure 13: ***Random Forest ensemble prediction.*** Each tree generates an individual prediction, the final output is obtained by aggregating the predictions from all trees.

Support Vector Machine

Support Vector Machines (SVM) can effectively handle non-linearly separable problems using kernel functions, which project the input data into a higher-dimensional space where a linear separation is possible. This transformation is known as the kernel trick, first introduced by Boser, Guyon, and Vapnik[118] (1995) to extend the applicability of SVM to more complex datasets. By mapping the original descriptor space into a Hilbert space, SVMs can model intricate relationships between molecular descriptors and biological activity. A kernel function computes a similarity measure between data points in the transformed space without explicitly computing the high-dimensional representation, making SVM computationally efficient, yet non-parallelizable. Common kernel types include the linear kernel, the polynomial kernel, and the RBF kernel, which models complex non-linear relationships by emphasizing local descriptor similarities.

k-Nearest Neighbors

The k-Nearest Neighbors (k-NN) algorithm is a non-parametric method.[119] It assigns a predicted value to a molecule based on the properties of its k most similar compounds in a multidimensional descriptor space (**Figure 14**). Molecular similarity is typically determined using Euclidean distance, or Tanimoto similarity, with the latter being more common for chemical fingerprints. In k-NN regression, the predicted activity of a molecule is computed as the weighted or unweighted average of its k-nearest neighbors' activity values. In the weighted approach, closer neighbors have a greater influence, with weights typically inversely proportional to the distance.

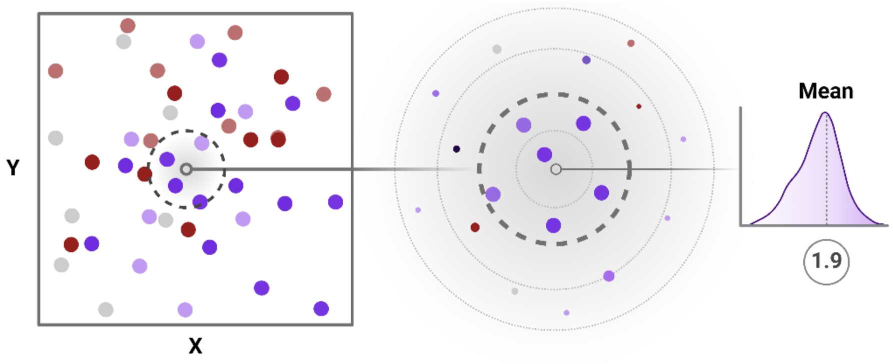


Figure 14: *k-Nearest Neighbors (k-NN) prediction approach.*

Deep Learning

Neural networks are computational models inspired by the structure and function of biological neural networks in animal brains. These models consist of artificial neurons that process and transmit signals through weighted connections, like synapses in the human brain. Each neuron receives numerical inputs, applies a non-linear activation function, and transmits the output to the next layer. The strength of each connection is determined by trainable weights, which are adjusted during the learning process. Neurons are typically organized into three types of layers. The input layer receives raw data, such as molecular descriptors. The hidden layers process this data through multiple transformations, extracting meaningful representations, and the output layer produces the final predictions, such as the bioactivity of a molecule.

Deep Neural Networks

A neural network is considered deep when it contains two or more hidden layers. These Deep Neural Networks (DNNs) have become essential tools in drug discovery due to their capacity to learn complex, non-linear relationships from high-dimensional data.

By stacking multiple layers of processing units, DNNs progressively extract increasingly abstract features from raw inputs, such as molecular descriptors, fingerprints, or physicochemical properties, transforming low-level information into high-level representations.

Each layer in a DNN applies a learned affine transformation followed by a non-linear activation function, such as ReLU or tanh. This sequential structure implements a composition of functions, allowing the network to approximate intricate mappings between input and output spaces.[120] The typical architecture includes an input layer, one or more hidden layers for intermediate representation learning, and an output layer that delivers the final prediction, for example, a molecular property or bioactivity score.

Training a DNN involves minimizing a loss function that quantifies the discrepancy between predicted and true values. This is achieved through a sequence of operations: forward propagation computes predictions across the network; the loss is then calculated using a function like Mean Squared Error for regression; and backpropagation computes gradients of the loss with respect to each parameter. These gradients guide weight updates through optimization algorithms such as stochastic gradient descent (SGD) descent[121] or Adam[122], repeated over multiple epochs until convergence or early stopping.

DNNs have achieved state-of-the-art performance in several key areas of drug discovery, including molecular property prediction, de novo molecular generation, virtual screening, and ADMET profiling.[123–125]

Graph Neural Networks

Molecular compounds being described as graphs, calculated molecular descriptors or fingerprints are an intermediate step that can be integrated in an end-to-end modeling approach. In this frame, Graph Neural Networks (GNN) have emerged as an attractive technology since 2016.[126–129] GNNs operate through several essential mechanisms. Node embeddings assign feature vectors to atoms representing properties such as atomic number and hybridization state. Message passing allows nodes to exchange information with their neighbors to refine representations. Aggregation functions such as mean pooling, sum pooling, or attention-based weighting combine node information. Graph convolutional layers extract structural features from the molecular graph.

GNNs have outperformed traditional molecular fingerprints such as Morgan fingerprints in predicting molecular properties, bioactivity, and drug-likeness. A notable example is AttentiveFP, which achieves state-of-the-art performance by incorporating self-attention mechanisms, allowing the model to focus on relevant molecular substructures during learning.[130] GNNs are now widely used in virtual screening for drug candidates, chemical space exploration for molecular diversity analysis, and ADMET property prediction.

Multi-Output

Ensemble Modeling

Predictive models trained on the same task often produce different outputs due to variations in training data, descriptor sets, hyperparameters, or learning algorithms. To improve accuracy, robustness, and generalization, ensemble modeling aggregates multiple predictions, reducing biases toward the dataset composition and model variance, integrating more values in the output. Two common approaches are majority voting for classification, where the most frequently predicted class is selected, and averaging predictions for regression, which smooths variations between models.

Beyond these basic strategies, more advanced ensemble techniques exist. Bagging (Bootstrap Aggregating) improves stability by training multiple models on different bootstrapped subsets of data, as seen in Random Forests. Boosting iteratively adjusts model weights to enhance weak predictions, with popular implementations including Gradient Boosting and XGBoost. Stacking combines multiple models by training a meta-model on their predictions. These ensemble approaches are widely applied in QSAR modeling, molecular property prediction, and ADMET screening, where consensus models generally outperform individual models in terms of accuracy and reliability.[131,132]

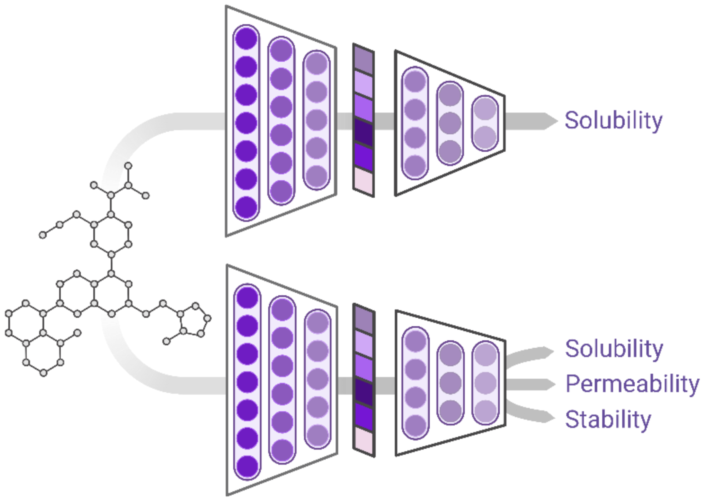


Figure 15: ***Comparative representation of a Single-Task to a Multi-Task predictive Graph Neural Network.***

Multi-Task Learning

Multi-Task Learning (MTL) is a machine learning paradigm where a model is trained to perform multiple related tasks simultaneously, leveraging shared representations to improve generalization across tasks. Instead of training separate models for each task, MTL introduces an inductive bias that allows knowledge transfer, particularly beneficial when tasks share underlying patterns (**Figure 15**). This approach is commonly used in various domains, where learning from related objectives enhances predictive performance and reduces overfitting. It has demonstrated improved accuracy against standard approach for QSAR modeling.[65]

One of the key advantages of MTL is its ability to maximize the use of limited datasets. In domains such as ADMET prediction, QSAR modeling, and bioactivity profiling, data availability is often sparse. By training a model on multiple related objectives, MTL allows it to extract meaningful features even from small datasets, improving performance compared to single-task models. Studies have shown that MTL consistently outperforms single-task QSAR models when predicting related endpoints, such as different toxicity measures or multiple kinase inhibitions, by capturing common molecular features across tasks.[133]

A critical aspect of MTL is determining the relative importance of tasks in the loss function: different tasks may have varying scales, convergence rates, or signal to noise ratio, for instance. We use task weighting to ensure that no single task dominates the optimization process. Here, the overall loss function in MTL is typically a weighted sum of individual task losses:

$$L_{MTL} = \sum_{i=1}^{n} w_i L_i$$

where L_i represents the loss for task i and w_i is the weight assigned to that task. Choosing appropriate task weights is crucial because improper weighting can lead to model bias, where one task is prioritized over others, leading to suboptimal generalization. Proper task weighting significantly impacts model stability, convergence, and predictive performance, particularly when dealing with imbalanced datasets or tasks with different difficulty levels (**Table 5**).

Table 5: ***Loss weighting methods for multi-task modeling.***

Method	Description	Advantages	Limitations
Fixed Weights[134]	Manually assigned constant weights	Simple Account for data scale	Lacks adaptability
Uncertainty-Based Weighting[134]	Adjusts task weights based on uncertainty	Prevents imbalance	Sensitive to data scale Requires tuning
GradNorm[135]	Adjusts weights to equalize gradient magnitudes	Dynamic learning	Requires tuning
Pareto Optimization[136]	Finds an optimal balance between task losses	Adaptative Learns trade-offs.	Computation intensive
Meta-Learning-Based Weighting[137]	Uses models to adjust task weights dynamically	Adaptative	Requires tuning Computation intensive

Despite its advantages, MTL is not always beneficial. When tasks are too unrelated or exhibit conflicting learning objectives, negative transfer can occur, where the model's performance degrades due to interference from irrelevant tasks. Proper task selection and dataset curation are crucial to prevent such issues. Additionally, designing an effective MTL model as an artificial neural network requires careful tuning of architectural components, such as shared vs. task-specific layers, weighting of different loss functions, and handling of imbalanced task distributions.[12,138]

Model Validation

A robust validation strategy is essential to ensure QSAR models are predictive and generalizable. Poor validation can lead to overfitting, underfitting, or spurious correlations, compromising model reliability. Ensuring the reliability of a QSAR model requires a rigorous validation strategy. A poorly validated model may appear accurate on training data but fail on new compounds due to overfitting, underfitting, or spurious correlations. To avoid these pitfalls, validation must assess both predictive accuracy and generalizability.

External Validation & Cross-Validation

The gold standard for evaluating model performance is testing on an independent external set, ensuring predictions are not biased by the training data. To maximize reliability, the test set should include structurally diverse compounds within the model's applicability domain. When data is limited, cross-validation (CV) provides an estimate of model stability. The most common approach, k-fold CV, partitions data into k subsets, iteratively training on k-1 folds and testing on the remaining one. Leave-one-out CV further maximizes data use but tends to over-estimate the generalization performances while being computationally intensive. Stratified k-fold CV is recommended to control the instability of the performance measures in presence of imbalanced datasets.

Bias, Variance, and Overfitting

A model's error stems from bias (systematic underestimation of complexity) and variance (excessive sensitivity to training data). Underfitting occurs when a model is too simplistic, failing to capture structure-activity relationships, mostly because of a lack of expressivity of the concept used to fit the data (too few molecular descriptors, for instance). Overfitting, in contrast, arises when a model memorizes training data rather than learning general patterns. This is common when the number of molecular descriptors exceeds the number of compounds, and the machine learning algorithm is insufficiently regularized. A practical guideline for multi-linear regression is maintaining a sample-to-descriptor ratio of at least 3:1 to reduce spurious correlations.[139,140]

Metrics

Model evaluation relies on quantitative metrics that compare predicted values to experimental data, ensuring the accuracy of a model's performance (**Table 6**). For regression models, several key indicators are used. The coefficient of determination (R^2) measures how well the model explains the variance in experimental data, with values close to 1 indicating a strong fit. The Root Mean Squared Error (RMSE) assesses the dispersion of prediction errors, where lower RMSE values indicate better performance and a closer alignment between predicted and actual values. As RMSE is very sensitive to outliers, we supplement it with the Mean Absolute Error (MAE). Using both provides insight into the distribution of errors, offering an intuitive measure of the average deviation between predictions and experimental results.

Table 6: ***Performance metrics for regression models.***

Metric	Equation
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

Where the variable,

n is the total number of samples in the dataset.

y_i is the actual value (ground truth) for the i^{th} sample.

$\hat{y}_i$ is the predicted value for the i^{th} sample.

$\bar{y}$ is the mean of the actual value.

Applicability Domain

The applicability domain (AD) is a key concept in QSAR modeling, determining the chemical space where a model's predictions remain reliable. Since models are trained on a limited set of molecules, their predictive accuracy depends on whether new compounds fall within the structural and property range covered. Predictions for molecules outside this domain tend to be unreliable due to a lack of learned patterns to generalize from. Various approaches exist to define the AD.

Table 7 ***Main range-based applicability domain methods.***

Method	Description	Pros	Cons
Bounding Box[141] $X_{min} \leq X \leq X_{max}$	Defines AD by setting min and max limits for each descriptor.	Simple Fast Intuitive	Too rigid No variability
Z-Score Method $X \in [\mu - k\sigma, \mu + k\sigma]$	Defines AD based on standard deviations from the mean.	Handles variability	Assumes normality Skewness sensitive
Percentile-Based Range $P_{low} \leq X \leq X_{high}$	Uses percentile cutoffs (e.g., 5th–95th) to define a feature-wise range.	Outlier-resistant	May exclude useful data

Where the variable,

X_i is the feature value of the sample.

σ is the standard deviation of the feature.

μ is the mean of the feature in training data.

k is a defined threshold.

Table 8: ***Main distance-based applicability domain methods.***

Method	Description	Pros	Cons
Leverage Method	Measures how influential a sample is in feature space	Detects extrapolation	Requires matrix inversion Assumes linearity Sensitive to collinearity
Mahalanobis Distance	Measures distance from mean	Captures correlation	Assumes Gaussian distribution Computation intense
Euclidean Distance	Straight-line distance to mean	Simple Interpretable	Ignores correlation Scaling-sensitive Equal feature weighting

Range-based

Range-based methods establish the AD by setting minimum and maximum thresholds for molecular descriptors based on values observed in the training set (**Table 7**). A molecule is considered within the AD if all its descriptor values fall within these limits. The most common techniques in this category include bounding box and PCA bounding box approaches.[142] These methods are computationally efficient and easy to implement. Additionally, they ignore sparsely populated regions in the descriptor space, increasing the risk of unreliable predictions. As a result, range-based methods tend to be highly restrictive.

Distance-based

Distance-based approaches define the AD by measuring the proximity of new molecules to those in the training set. These methods are widely used.[143] One common approach is centroid-based distances, where molecules are compared to a central reference point of the training set. Several distance metrics are used in this context (**Table 8**). These methods establish a threshold beyond which molecules are considered outside the AD.

Model-based

Model-based approaches define the AD by leveraging ML techniques that analyze patterns, anomalies, and outliers within the training dataset. Unlike geometric or distance-based methods, these approaches rely on statistical learning to determine whether a new molecule belongs to the known chemical space. Various methods exist to define the model.

Isolation Forest

The Isolation Forest (IF) method is designed for anomaly detection by constructing an ensemble of randomly partitioned decision trees (**Figure 16**).[9] The underlying principle is that outliers require fewer splits to be isolated, whereas inliers need more. The process involves randomly segmenting the dataset into decision trees and recording the depth at which each molecule is isolated. Molecules with shorter path lengths are considered outliers, and a predefined threshold determines whether a molecule belongs to the applicability domain. This method is efficient for large datasets, robust to irrelevant features, and non-parametric, making it flexible for different datasets. The AD based on IF are softer (accepting more compounds in the AD) or harder (refusing more compounds in the AD) depending on the contamination parameter. It defines the proportion of isolated instances in the training set.

One -Class Support Vector Machine

The One-Class Support Vector Machine (OcSVM) is an extension of SVM tailored for unsupervised anomaly detection.[144] It learns a decision boundary around the training data, classifying new samples as inside or outside the applicability domain. The model constructs a hypersphere or hyperplane enclosing most training data points, using kernel functions such as the RBF to capture complex relationships between molecular descriptors. This method is particularly effective for high-dimensional data, for non-vectorial data and non-linearly separable datasets. It is sensitive to kernel selection and is controlled by the kernel parameters, if any, as well as the cost, regularizing the boundary. The algorithm may be computationally expensive for large datasets because of the kernel estimation.

Local Outlier Factor

The Local Outlier Factor (LOF) algorithm assesses whether a molecule lies within the applicability domain by comparing its local data density to that of its neighbors.[145] Unlike global distance-based methods, LOF focuses on the immediate chemical environment of each molecule. It begins by identifying the k nearest neighbors of a given molecule and computing the local density based on their proximity. This local density is then compared to the densities of the neighbors themselves. If the molecule resides in a much sparser region than its neighbors, it is flagged as an outlier. In AD applications, molecules with similar densities to the training data are considered in-domain, while those with significantly lower densities are flagged as outside the domain. The choice of k is crucial: a smaller k allows for higher sensitivity to rare patterns but may increase noise, while a larger k stabilizes the assessment but may overlook subtle deviations.

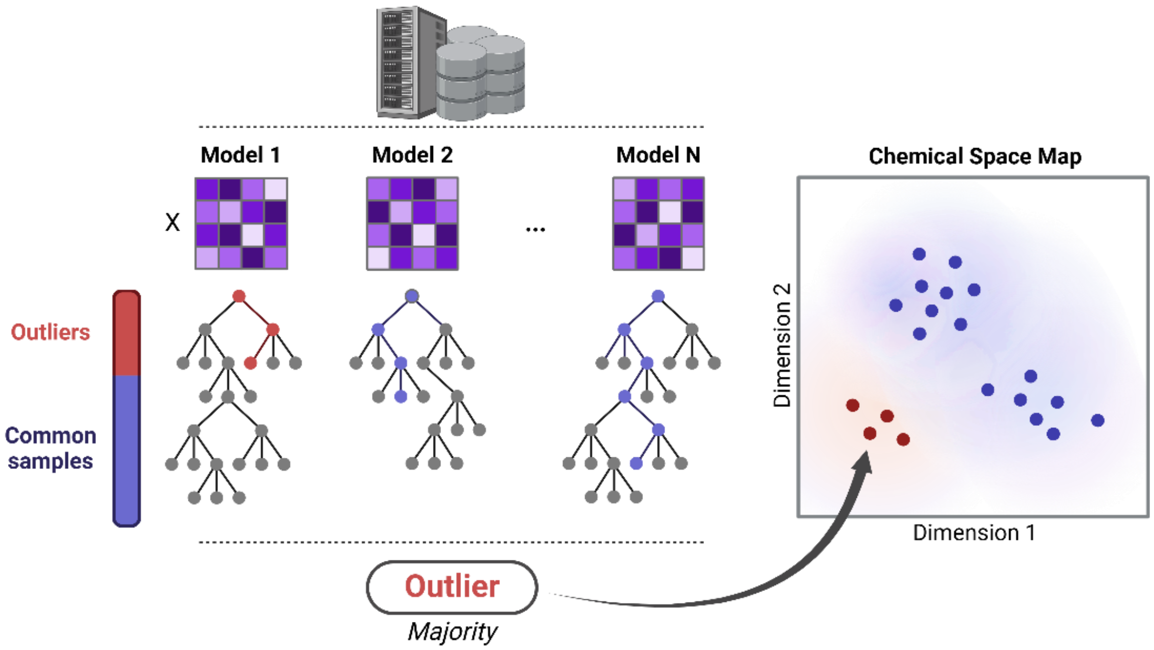


Figure 16: ***Mechanism of prediction by the Isolation Forest.*** The model consists of multiple randomly partitioned decision trees, where each tree recursively splits the data until individual points are isolated. Outliers are identified as compounds requiring fewer splits to be isolated, meaning they appear closer to the root in most trees. The final anomaly score is computed by aggregating the outputs across all trees, with compounds classified as outliers receiving consistently high anomaly scores.

Chapter 4. Modeling of Solubility

The ability of a drug candidate to dissolve in an aqueous environment is a critical factor influencing its development, formulation, and therapeutic success. Solubility dictates bioavailability, impacts pharmacokinetics, and directly affects dosing requirements. As illustrated by the Mayer-Overton Rule more the 100 years ago, the drug potency correlates with the oil solubility, supporting the importance of molecular solubility and its relation with drug solution and target interaction.[33]

Poor aqueous solubility is a major cause of late-stage failures in drug development, as it can lead to suboptimal absorption, necessitate high-dose, and thus increase the risk of toxicity where high lipophilicity is usually linked with increased risk of hERG blocking (cardiotoxicity) and phospholipidosis. Consequently, solubility screening is an essential step in early-stage drug discovery.

Despite its importance, solubility prediction remains a significant challenge. Experimental measurements vary across laboratories due to methodological differences, batch effects, and inconsistencies in reporting standards. Many solubility datasets result from iterative aggregation, often incorporating low-quality measurements, computational predictions, or data that do not adhere to OECD guidelines.[146] This lack of standardization complicates the development of reliable predictive models.

In this chapter, we investigate the challenges of modeling solubility. While thermodynamic solubility is crucial for late-stage development, kinetic solubility is widely used in early screening but suffers from high protocol sensitivity and interlaboratory variability. We analyze the relationship between these two solubility types, demonstrating their poor correlation and the need for separate predictive models. By benchmarking existing approaches, we reveal the limitations of current models is due to inconsistent data curation and measurement variability. Contrary to expectations, we find that kinetic solubility datasets show higher reproducibility than assumed, allowing for the development of accurate QSPR models. We propose a workflow to improve solubility predictions, providing curated datasets and models to enhance their reliability and utility in drug discovery.

4.1. Thermodynamic Solubility

Introduction

Machine learning has shown promise in solubility prediction, with recent models achieving seemingly strong performance. However, their reliability in prospective applications is often overestimated. Many models rely on overlapping training sets, leading to overfitting rather than genuine generalizability. Additionally, the applicability domain of these models is rarely well-defined, further limiting their practical use in real-world drug discovery. In this section, we investigate these challenges by analyzing over two decades of thermodynamic solubility datasets, computational models, and acquisition methods (**Figure 17**). We explore historical dataset curation practices, evaluate data quality, and assess the generalization capacity of state-of-the-art solubility prediction models.

Main Terminology

Thermodynamic solubility is the maximum concentration of a compound that dissolves in a solvent at equilibrium, typically used in late-stage drug development due to its reproducibility and relevance for formulation.

Applicability domain is the descriptor space within which a predictive model is expected to provide reliable results, ensuring that predictions are made only for compounds similar to those in the training data.

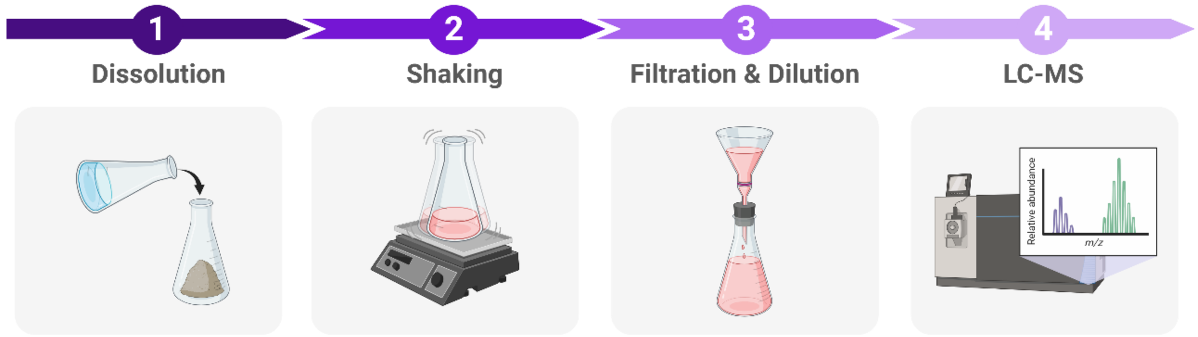
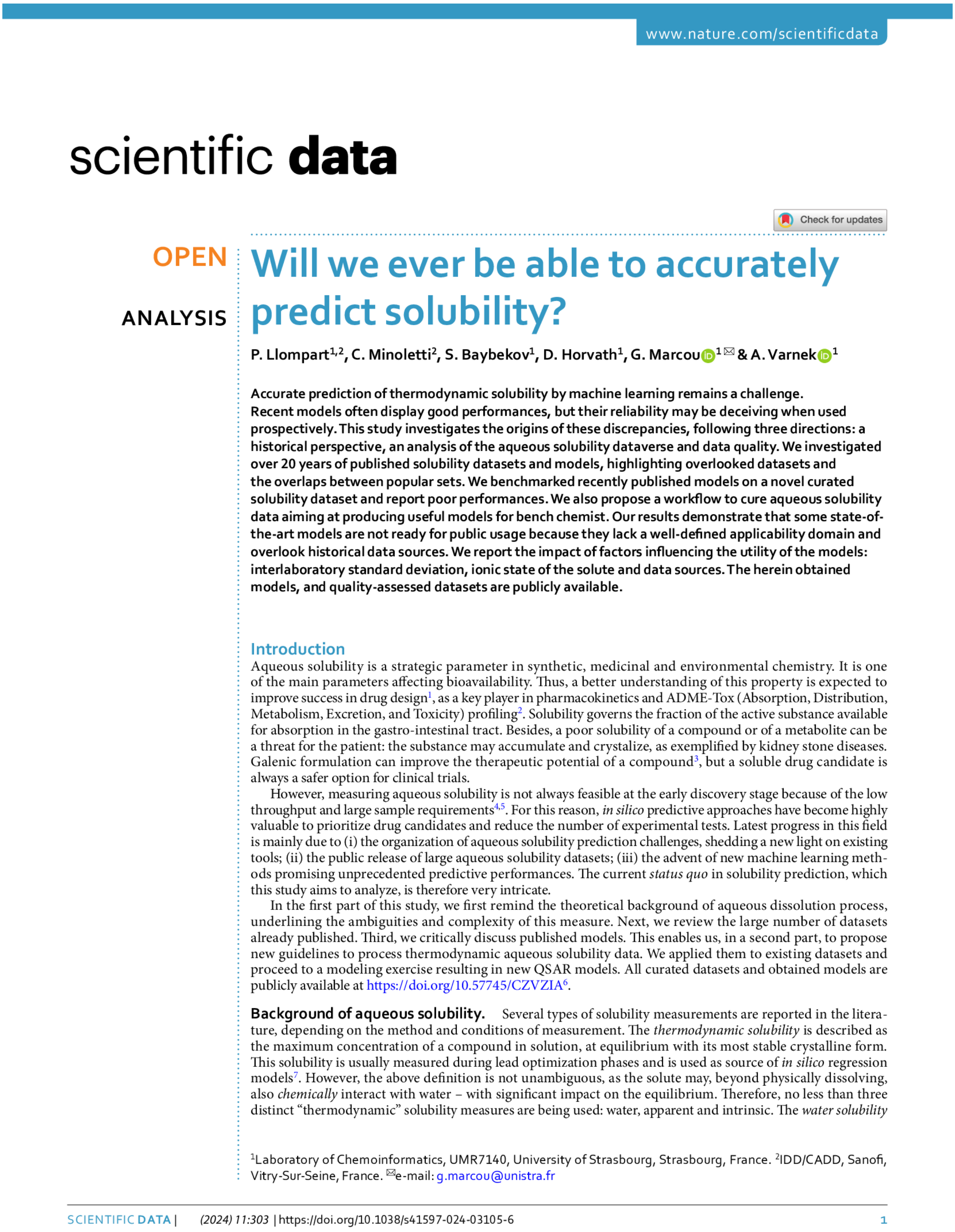


Figure 17: ***Thermodynamic solubility measurement using the shake flask method.*** The process begins with compound dissolution, where the molecule is introduced into a solvent (typically water or a buffer) and shaken to reach equilibrium. The solution is then filtered, followed by quantification of the dissolved compound using techniques such as UV-Vis spectroscopy, HPLC, or LC-MS.



www.nature.com/scientificdata

scientific data

OPEN

ANALYSIS

Will we ever be able to accurately predict solubility?

P. Llompart[1,2], C. Minoletti[2], S. Baybekov[1], D. Horvath[1], G. Marcou[1] ✉ & A. Varnek[1]

Accurate prediction of thermodynamic solubility by machine learning remains a challenge. Recent models often display good performances, but their reliability may be deceiving when used prospectively. This study investigates the origins of these discrepancies, following three directions: a historical perspective, an analysis of the aqueous solubility dataverse and data quality. We investigated over 20 years of published solubility datasets and models, highlighting overlooked datasets and the overlaps between popular sets. We benchmarked recently published models on a novel curated solubility dataset and report poor performances. We also propose a workflow to cure aqueous solubility data aiming at producing useful models for bench chemist. Our results demonstrate that some state-of-the-art models are not ready for public usage because they lack a well-defined applicability domain and overlook historical data sources. We report the impact of factors influencing the utility of the models: interlaboratory standard deviation, ionic state of the solute and data sources. The herein obtained models, and quality-assessed datasets are publicly available.

Introduction

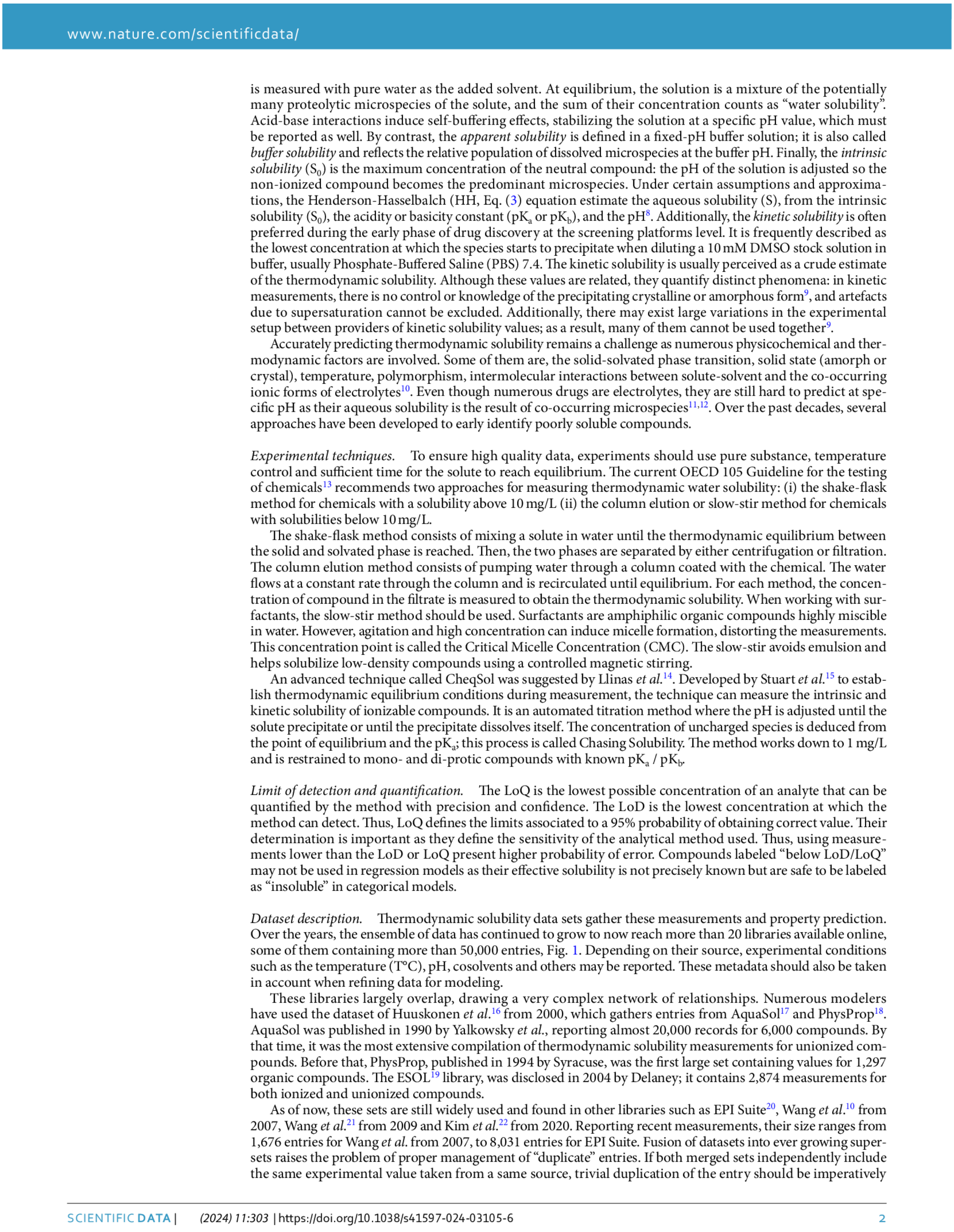
Aqueous solubility is a strategic parameter in synthetic, medicinal and environmental chemistry. It is one of the main parameters affecting bioavailability. Thus, a better understanding of this property is expected to improve success in drug design[1], as a key player in pharmacokinetics and ADME-Tox (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiling[2]. Solubility governs the fraction of the active substance available for absorption in the gastro-intestinal tract. Besides, a poor solubility of a compound or of a metabolite can be a threat for the patient: the substance may accumulate and crystalize, as exemplified by kidney stone diseases. Galenic formulation can improve the therapeutic potential of a compound[3], but a soluble drug candidate is always a safer option for clinical trials.

However, measuring aqueous solubility is not always feasible at the early discovery stage because of the low throughput and large sample requirements[4,5]. For this reason, *in silico* predictive approaches have become highly valuable to prioritize drug candidates and reduce the number of experimental tests. Latest progress in this field is mainly due to (i) the organization of aqueous solubility prediction challenges, shedding a new light on existing tools; (ii) the public release of large aqueous solubility datasets; (iii) the advent of new machine learning methods promising unprecedented predictive performances. The current *status quo* in solubility prediction, which this study aims to analyze, is therefore very intricate.

In the first part of this study, we first remind the theoretical background of aqueous dissolution process, underlining the ambiguities and complexity of this measure. Next, we review the large number of datasets already published. Third, we critically discuss published models. This enables us, in a second part, to propose new guidelines to process thermodynamic aqueous solubility data. We applied them to existing datasets and proceed to a modeling exercise resulting in new QSAR models. All curated datasets and obtained models are publicly available at https://doi.org/10.57745/CZVZIA[6].

Background of aqueous solubility. Several types of solubility measurements are reported in the literature, depending on the method and conditions of measurement. The *thermodynamic solubility* is described as the maximum concentration of a compound in solution, at equilibrium with its most stable crystalline form. This solubility is usually measured during lead optimization phases and is used as source of *in silico* regression models[7]. However, the above definition is not unambiguous, as the solute may, beyond physically dissolving, also *chemically* interact with water – with significant impact on the equilibrium. Therefore, no less than three distinct “thermodynamic” solubility measures are being used: water, apparent and intrinsic. The *water solubility*

[1]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg, France. [2]IDD/CADD, Sanofi, Vitry-Sur-Seine, France. ✉e-mail: g.marcou@unistra.fr



is measured with pure water as the added solvent. At equilibrium, the solution is a mixture of the potentially many proteolytic microspecies of the solute, and the sum of their concentration counts as "water solubility". Acid-base interactions induce self-buffering effects, stabilizing the solution at a specific pH value, which must be reported as well. By contrast, the *apparent solubility* is defined in a fixed-pH buffer solution; it is also called *buffer solubility* and reflects the relative population of dissolved microspecies at the buffer pH. Finally, the *intrinsic solubility* (S_0) is the maximum concentration of the neutral compound: the pH of the solution is adjusted so the non-ionized compound becomes the predominant microspecies. Under certain assumptions and approximations, the Henderson-Hasselbalch (HH, Eq. (3) equation estimate the aqueous solubility (S), from the intrinsic solubility (S_0), the acidity or basicity constant (pK_a or pK_b), and the pH[8]. Additionally, the *kinetic solubility* is often preferred during the early phase of drug discovery at the screening platforms level. It is frequently described as the lowest concentration at which the species starts to precipitate when diluting a 10 mM DMSO stock solution in buffer, usually Phosphate-Buffered Saline (PBS) 7.4. The kinetic solubility is usually perceived as a crude estimate of the thermodynamic solubility. Although these values are related, they quantify distinct phenomena: in kinetic measurements, there is no control or knowledge of the precipitating crystalline or amorphous form[9], and artefacts due to supersaturation cannot be excluded. Additionally, there may exist large variations in the experimental setup between providers of kinetic solubility values; as a result, many of them cannot be used together[9].

Accurately predicting thermodynamic solubility remains a challenge as numerous physicochemical and thermodynamic factors are involved. Some of them are, the solid-solvated phase transition, solid state (amorph or crystal), temperature, polymorphism, intermolecular interactions between solute-solvent and the co-occurring ionic forms of electrolytes[10]. Even though numerous drugs are electrolytes, they are still hard to predict at specific pH as their aqueous solubility is the result of co-occurring microspecies[11,12]. Over the past decades, several approaches have been developed to early identify poorly soluble compounds.

Experimental techniques. To ensure high quality data, experiments should use pure substance, temperature control and sufficient time for the solute to reach equilibrium. The current OECD 105 Guideline for the testing of chemicals[13] recommends two approaches for measuring thermodynamic water solubility: (i) the shake-flask method for chemicals with a solubility above 10 mg/L (ii) the column elution or slow-stir method for chemicals with solubilities below 10 mg/L.

The shake-flask method consists of mixing a solute in water until the thermodynamic equilibrium between the solid and solvated phase is reached. Then, the two phases are separated by either centrifugation or filtration. The column elution method consists of pumping water through a column coated with the chemical. The water flows at a constant rate through the column and is recirculated until equilibrium. For each method, the concentration of compound in the filtrate is measured to obtain the thermodynamic solubility. When working with surfactants, the slow-stir method should be used. Surfactants are amphiphilic organic compounds highly miscible in water. However, agitation and high concentration can induce micelle formation, distorting the measurements. This concentration point is called the Critical Micelle Concentration (CMC). The slow-stir avoids emulsion and helps solubilize low-density compounds using a controlled magnetic stirring.

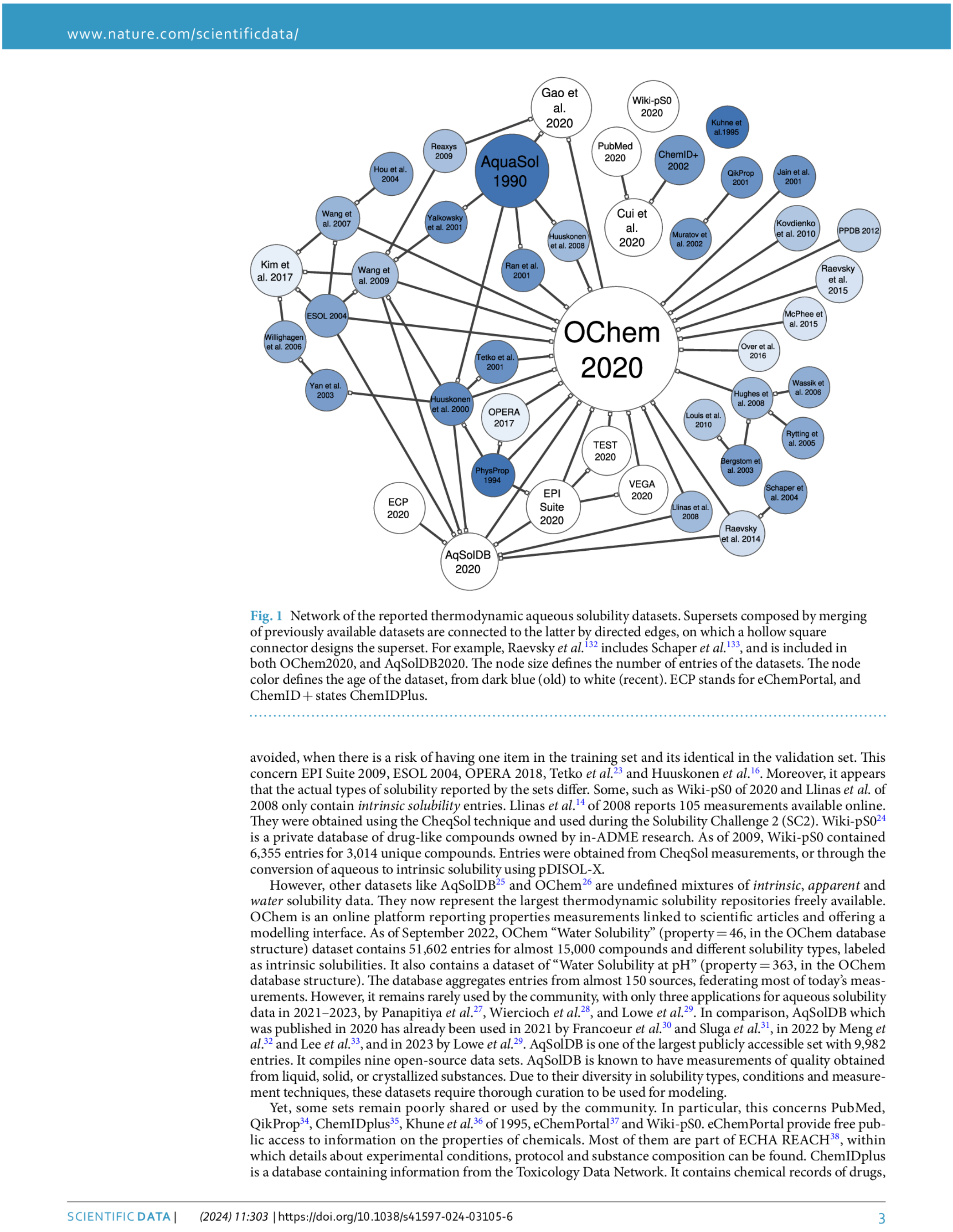
An advanced technique called CheqSol was suggested by Llinas *et al.*[14]. Developed by Stuart *et al.*[15] to establish thermodynamic equilibrium conditions during measurement, the technique can measure the intrinsic and kinetic solubility of ionizable compounds. It is an automated titration method where the pH is adjusted until the solute precipitate or until the precipitate dissolves itself. The concentration of uncharged species is deduced from the point of equilibrium and the pK_a; this process is called Chasing Solubility. The method works down to 1 mg/L and is restrained to mono- and di-protic compounds with known pK_a / pK_b.

Limit of detection and quantification. The LoQ is the lowest possible concentration of an analyte that can be quantified by the method with precision and confidence. The LoD is the lowest concentration at which the method can detect. Thus, LoQ defines the limits associated to a 95% probability of obtaining correct value. Their determination is important as they define the sensitivity of the analytical method used. Thus, using measurements lower than the LoD or LoQ present higher probability of error. Compounds labeled "below LoD/LoQ" may not be used in regression models as their effective solubility is not precisely known but are safe to be labeled as "insoluble" in categorical models.

Dataset description. Thermodynamic solubility data sets gather these measurements and property prediction. Over the years, the ensemble of data has continued to grow to now reach more than 20 libraries available online, some of them containing more than 50,000 entries, Fig. 1. Depending on their source, experimental conditions such as the temperature (T°C), pH, cosolvents and others may be reported. These metadata should also be taken in account when refining data for modeling.

These libraries largely overlap, drawing a very complex network of relationships. Numerous modelers have used the dataset of Huuskonen *et al.*[16] from 2000, which gathers entries from AquaSol[17] and PhysProp[18]. AquaSol was published in 1990 by Yalkowsky *et al.*, reporting almost 20,000 records for 6,000 compounds. By that time, it was the most extensive compilation of thermodynamic solubility measurements for unionized compounds. Before that, PhysProp, published in 1994 by Syracuse, was the first large set containing values for 1,297 organic compounds. The ESOL[19] library, was disclosed in 2004 by Delaney; it contains 2,874 measurements for both ionized and unionized compounds.

As of now, these sets are still widely used and found in other libraries such as EPI Suite[20], Wang *et al.*[10] from 2007, Wang *et al.*[21] from 2009 and Kim *et al.*[22] from 2020. Reporting recent measurements, their size ranges from 1,676 entries for Wang *et al.* from 2007, to 8,031 entries for EPI Suite. Fusion of datasets into ever growing supersets raises the problem of proper management of "duplicate" entries. If both merged sets independently include the same experimental value taken from a same source, trivial duplication of the entry should be imperatively



Gao et al. 2020
Wiki-pS0 2020
Kuhne et al.1995
PubMed 2020
ChemID+ 2002
QikProp 2001
Jain et al. 2001
Reaxys 2009
AquaSol 1990
Hou et al. 2004
Wang et al. 2007
Yalkowsky et al. 2001
Huuskonen et al. 2008
Cui et al. 2020
Muratov et al. 2002
Kovdienko et al. 2010
PPDB 2012
Kim et al. 2017
Wang et al. 2009
Ran et al. 2001
Raevsky et al. 2015
ESOL 2004
OChem 2020
McPhee et al. 2015
Willighagen et al. 2006
Over et al. 2016
Tetko et al. 2001
Wassik et al. 2006
Yan et al. 2003
Huuskonen et al. 2000
OPERA 2017
Louis et al. 2010
Hughes et al. 2008
Rytting et al. 2005
TEST 2020
PhysProp 1994
Bergstom et al. 2003
VEGA 2020
ECP 2020
EPI Suite 2020
Schaper et al. 2004
Llinas et al. 2008
Raevsky et al. 2014
AqSolDB 2020

Fig. 1 Network of the reported thermodynamic aqueous solubility datasets. Supersets composed by merging of previously available datasets are connected to the latter by directed edges, on which a hollow square connector designs the superset. For example, Raevsky *et al.*[132] includes Schaper *et al.*[133], and is included in both OChem2020, and AqSolDB2020. The node size defines the number of entries of the datasets. The node color defines the age of the dataset, from dark blue (old) to white (recent). ECP stands for eChemPortal, and ChemID + states ChemIDPlus.

avoided, when there is a risk of having one item in the training set and its identical in the validation set. This concern EPI Suite 2009, ESOL 2004, OPERA 2018, Tetko *et al.*[23] and Huuskonen *et al.*[16]. Moreover, it appears that the actual types of solubility reported by the sets differ. Some, such as Wiki-pS0 of 2020 and Llinas *et al.* of 2008 only contain *intrinsic solubility* entries. Llinas *et al.*[14] of 2008 reports 105 measurements available online. They were obtained using the CheqSol technique and used during the Solubility Challenge 2 (SC2). Wiki-pS0[24] is a private database of drug-like compounds owned by in-ADME research. As of 2009, Wiki-pS0 contained 6,355 entries for 3,014 unique compounds. Entries were obtained from CheqSol measurements, or through the conversion of aqueous to intrinsic solubility using pDISOL-X.

However, other datasets like AqSolDB[25] and OChem[26] are undefined mixtures of *intrinsic*, *apparent* and *water* solubility data. They now represent the largest thermodynamic solubility repositories freely available. OChem is an online platform reporting properties measurements linked to scientific articles and offering a modelling interface. As of September 2022, OChem "Water Solubility" (property = 46, in the OChem database structure) dataset contains 51,602 entries for almost 15,000 compounds and different solubility types, labeled as intrinsic solubilities. It also contains a dataset of "Water Solubility at pH" (property = 363, in the OChem database structure). The database aggregates entries from almost 150 sources, federating most of today's measurements. However, it remains rarely used by the community, with only three applications for aqueous solubility data in 2021–2023, by Panapitiya *et al.*[27], Wiercioch *et al.*[28], and Lowe *et al.*[29]. In comparison, AqSolDB which was published in 2020 has already been used in 2021 by Francoeur *et al.*[30] and Sluga *et al.*[31], in 2022 by Meng *et al.*[32] and Lee *et al.*[33], and in 2023 by Lowe *et al.*[29]. AqSolDB is one of the largest publicly accessible set with 9,982 entries. It compiles nine open-source data sets. AqSolDB is known to have measurements of quality obtained from liquid, solid, or crystallized substances. Due to their diversity in solubility types, conditions and measurement techniques, these datasets require thorough curation to be used for modeling.

Yet, some sets remain poorly shared or used by the community. In particular, this concerns PubMed, QikProp[34], ChemIDplus[35], Khune *et al.*[36] of 1995, eChemPortal[37] and Wiki-pS0. eChemPortal provide free public access to information on the properties of chemicals. Most of them are part of ECHA REACH[38], within which details about experimental conditions, protocol and substance composition can be found. ChemIDplus is a database containing information from the Toxicology Data Network. It contains chemical records of drugs, pesticides, pollutants, and toxins. Although relatively vintage, these datasets are overlooked resources that contain a wealth of experimental data.

Solubility prediction. Predictive approaches are either based on theorical equations or Machine Learning (ML) methods, including Neural Networks (NN). The few approaches based on first principles are mainly applied to estimate the solvation energy changes associated with a solute transitioning from its solid state to its solvated state.

From a thermodynamic point of view, solubilization can be managed in one or two steps starting from a solid material. It can either be by sublimation from solid to gas or by fusion from solid to liquid, followed with an energy transfer to water. Hence, in 1965 Irmann[39] coupled the entropy of fusion (ΔS_m) to the melting point (MP) through a group contribution approach to predict water solubility. Then, in 1968, Hansch *et al.*[40] found that the water solubility of organic liquid compounds was linearly dependent to the octanol/water partition coefficient (Log $P_{o/w}$). Yalkowsky *et al.*[41] combined these results in 1980 to develop the General Solubility Equation (GSE) and estimate the base-10 logarithm of water solubility $Log_{10}S_w$ using the MP and Log $P_{o/w}$ - see Eq. (1).

$$Log_{10}(S_w) = 0.5 - 0.01 \cdot (MP - 25) - LogP_{o/w} \tag{1}$$

The equation is restrained to solid nonelectrolytes, but it usually performs well (RMSE: 0.7–0.8 log) when employed with experimental values[42]. Here, an electrolyte is a chemical substance that produces mobile charges. As most drugs are electrolytes, only few are covered by the GSE. Also, High Throughput Screening (HTS) does not usually include the measurement of MP and Log $P_{o/w}$, which are thus replaced by predicted values. Their use can introduce major discrepancies in the estimation of thermodynamic solubility, not to mention that the prediction of MP represents itself a challenge. Thus, the GSE is not practically useful for large-scale predictions.

20 years of solubility modelling. Most of today's models are Quantitative Structure Property Relationship (QSPR). These methods seek to find a mathematical function expressed as Y = f(X) where X defines a set of N molecular descriptors [D_1, D_2, …, D_N] to correlate to the response value Y. Of course, the inner representation of a chemical graph by a GNN (Graph Neural Network) is no different. In our case, this Y value is the base-10 logarithm of the molar measurement of thermodynamic solubility, expressed as $Log_{10}(S)$.

Machine learning methods are mainly used to develop regression models leveraged on the compound's topological, electronic, structural 2D/3D features, and molecular fragment counts. Models are then optimized using many ML methods to best fit the descriptors set. Recently, feature-based NN, graph-based NN (GNN) and structural attention methods have been used to develop powerful solubility predictive models. Tables 1 to 3 report a representative but not exhaustive list of aqueous solubility models developed over the last 20 years. It aims to highlight significant trends and achievements in this area. While the table includes models using diverse methods, caution is advised regarding overly optimistic performances. Depending on the data and approach employed, three periods can be distinguished. Prior to 2008, models were trained on vast datasets such as AquaSol, PhysProp and their aggregation, Huuskonen *et al.*[16] (Table 1). Few methods (ANN, SVM, MLR and theorical equations) were applied as the most decisive parameter of one's ML model performance was the size and diversity of its training set. From them, two lessons can be shared:

- The relationship between solubility and the classical descriptors used here tends to be largely non-linear. Therefore, in this context, ANNs clearly outperformed linear regression.
- The prediction performances are limited by the quality of the experimental data. It is usually measured using the Inter-laboratory Standard Deviation (*SDi*) - Eq. (2). It is considered as a lower limit for theoretical prediction accuracy, and it was pointed out that the *SDi* can reach up to 1.0 log unit.

$$SDi = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}} \tag{2}$$

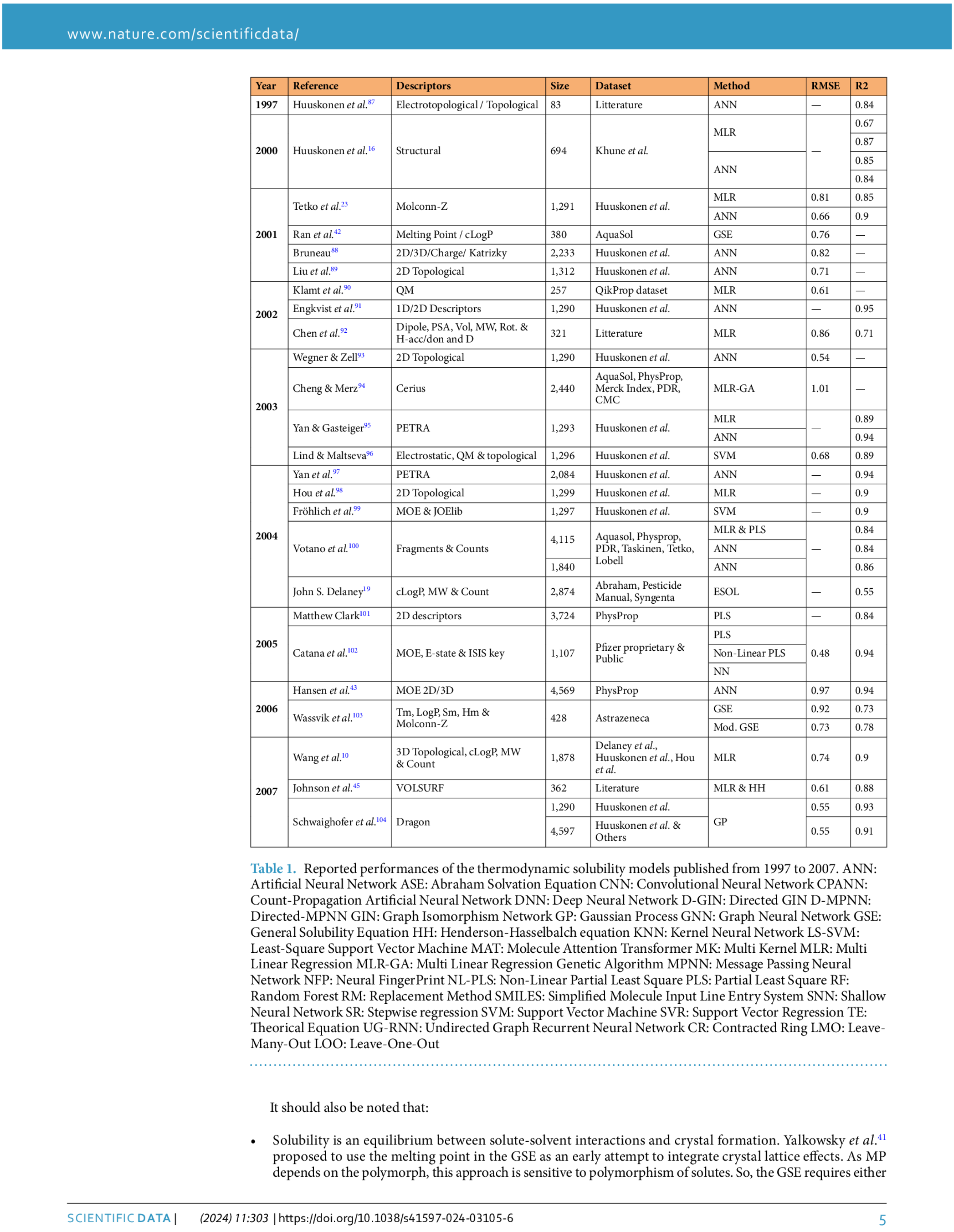
The *SDi* depends of the average value $\bar{x}$ of the *n* replicated measures, x_i.

Few attempts were also made to predict[43] the *intrinsic* solubility using the HH equation. An ANN was trained on PhysProp to obtain the predicted aqueous solubility. Acidity and basicity constants (pK_a and pK_b) required by HH were estimated by pKaPlugIn from ChemAxon[44]. The HH equation depends on the ionization state of the compounds and can thus be used by Hansen's combined model to compute the *intrinsic* solubility ($Log\,(S_0)$) as a function of pH – see Eq. (3).

$$Log\,(S_w) = Log\,(S_0) + (1 + 10^{(pH - pK_a)} + 10^{(pK_b - pH)}) \tag{3}$$

In 2007, Johnson *et al.*[45] renewed this approach by postulating an *ansatz* describing the *intrinsic* solubility as a function of the pK_a, pK_b, pH and, crystal packing χ_{pack} and degree of ionization F_I – see Eq. (4). The influence of the crystal lattice on the solubility were simulated by a molecular dynamics simulation[45].

$$Log\,(S_{pH}) = Log\,(S_0) + \min\left[Log\left(10^{\sum_i^{N_{acids}}(pH - pKa_i) + \sum_j^{N_{bases}}(pKb_j - pH) + 1}\right), 4.25\right] - \chi_{pack} \cdot e^{-F_I} \tag{4}$$



Year	Reference	Descriptors	Size	Dataset	Method	RMSE	R2
1997	Huuskonen *et al.*[87]	Electrotopological / Topological	83	Litterature	ANN	—	0.84
2000	Huuskonen *et al.*[16]	Structural	694	Khune *et al.*	MLR	—	0.67
							0.87
					ANN		0.85
							0.84
2001	Tetko *et al.*[23]	Molconn-Z	1,291	Huuskonen *et al.*	MLR	0.81	0.85
					ANN	0.66	0.9
	Ran *et al.*[42]	Melting Point / cLogP	380	AquaSol	GSE	0.76	—
	Bruneau[88]	2D/3D/Charge/ Katrizky	2,233	Huuskonen *et al.*	ANN	0.82	—
	Liu *et al.*[89]	2D Topological	1,312	Huuskonen *et al.*	ANN	0.71	—
2002	Klamt *et al.*[90]	QM	257	QikProp dataset	MLR	0.61	—
	Engkvist *et al.*[91]	1D/2D Descriptors	1,290	Huuskonen *et al.*	ANN	—	0.95
	Chen *et al.*[92]	Dipole, PSA, Vol, MW, Rot. & H-acc/don and D	321	Litterature	MLR	0.86	0.71
2003	Wegner & Zell[93]	2D Topological	1,290	Huuskonen *et al.*	ANN	0.54	—
	Cheng & Merz[94]	Cerius	2,440	AquaSol, PhysProp, Merck Index, PDR, CMC	MLR-GA	1.01	—
	Yan & Gasteiger[95]	PETRA	1,293	Huuskonen *et al.*	MLR	—	0.89
					ANN		0.94
	Lind & Maltseva[96]	Electrostatic, QM & topological	1,296	Huuskonen *et al.*	SVM	0.68	0.89
2004	Yan *et al.*[97]	PETRA	2,084	Huuskonen *et al.*	ANN	—	0.94
	Hou *et al.*[98]	2D Topological	1,299	Huuskonen *et al.*	MLR	—	0.9
	Fröhlich *et al.*[99]	MOE & JOElib	1,297	Huuskonen *et al.*	SVM	—	0.9
	Votano *et al.*[100]	Fragments & Counts	4,115	Aquasol, Physprop, PDR, Taskinen, Tetko, Lobell	MLR & PLS	—	0.84
					ANN		0.84
			1,840		ANN		0.86
	John S. Delaney[19]	cLogP, MW & Count	2,874	Abraham, Pesticide Manual, Syngenta	ESOL	—	0.55
2005	Matthew Clark[101]	2D descriptors	3,724	PhysProp	PLS	—	0.84
	Catana *et al.*[102]	MOE, E-state & ISIS key	1,107	Pfizer proprietary & Public	PLS	0.48	0.94
					Non-Linear PLS		
					NN		
2006	Hansen *et al.*[43]	MOE 2D/3D	4,569	PhysProp	ANN	0.97	0.94
	Wassvik *et al.*[103]	Tm, LogP, Sm, Hm & Molconn-Z	428	Astrazeneca	GSE	0.92	0.73
					Mod. GSE	0.73	0.78
2007	Wang *et al.*[10]	3D Topological, cLogP, MW & Count	1,878	Delaney *et al.*, Huuskonen *et al.*, Hou *et al.*	MLR	0.74	0.9
	Johnson *et al.*[45]	VOLSURF	362	Literature	MLR & HH	0.61	0.88
	Schwaighofer *et al.*[104]	Dragon	1,290	Huuskonen *et al.*	GP	0.55	0.93
			4,597	Huuskonen *et al.* & Others		0.55	0.91

Table 1. Reported performances of the thermodynamic solubility models published from 1997 to 2007. ANN: Artificial Neural Network ASE: Abraham Solvation Equation CNN: Convolutional Neural Network CPANN: Count-Propagation Artificial Neural Network DNN: Deep Neural Network D-GIN: Directed GIN D-MPNN: Directed-MPNN GIN: Graph Isomorphism Network GP: Gaussian Process GNN: Graph Neural Network GSE: General Solubility Equation HH: Henderson-Hasselbalch equation KNN: Kernel Neural Network LS-SVM: Least-Square Support Vector Machine MAT: Molecule Attention Transformer MK: Multi Kernel MLR: Multi Linear Regression MLR-GA: Multi Linear Regression Genetic Algorithm MPNN: Message Passing Neural Network NFP: Neural FingerPrint NL-PLS: Non-Linear Partial Least Square PLS: Partial Least Square RF: Random Forest RM: Replacement Method SMILES: Simplified Molecule Input Line Entry System SNN: Shallow Neural Network SR: Stepwise regression SVM: Support Vector Machine SVR: Support Vector Regression TE: Theorical Equation UG-RNN: Undirected Graph Recurrent Neural Network CR: Contracted Ring LMO: Leave-Many-Out LOO: Leave-One-Out

It should also be noted that:

- Solubility is an equilibrium between solute-solvent interactions and crystal formation. Yalkowsky *et al.*[41] proposed to use the melting point in the GSE as an early attempt to integrate crystal lattice effects. As MP depends on the polymorph, this approach is sensitive to polymorphism of solutes. So, the GSE requires either

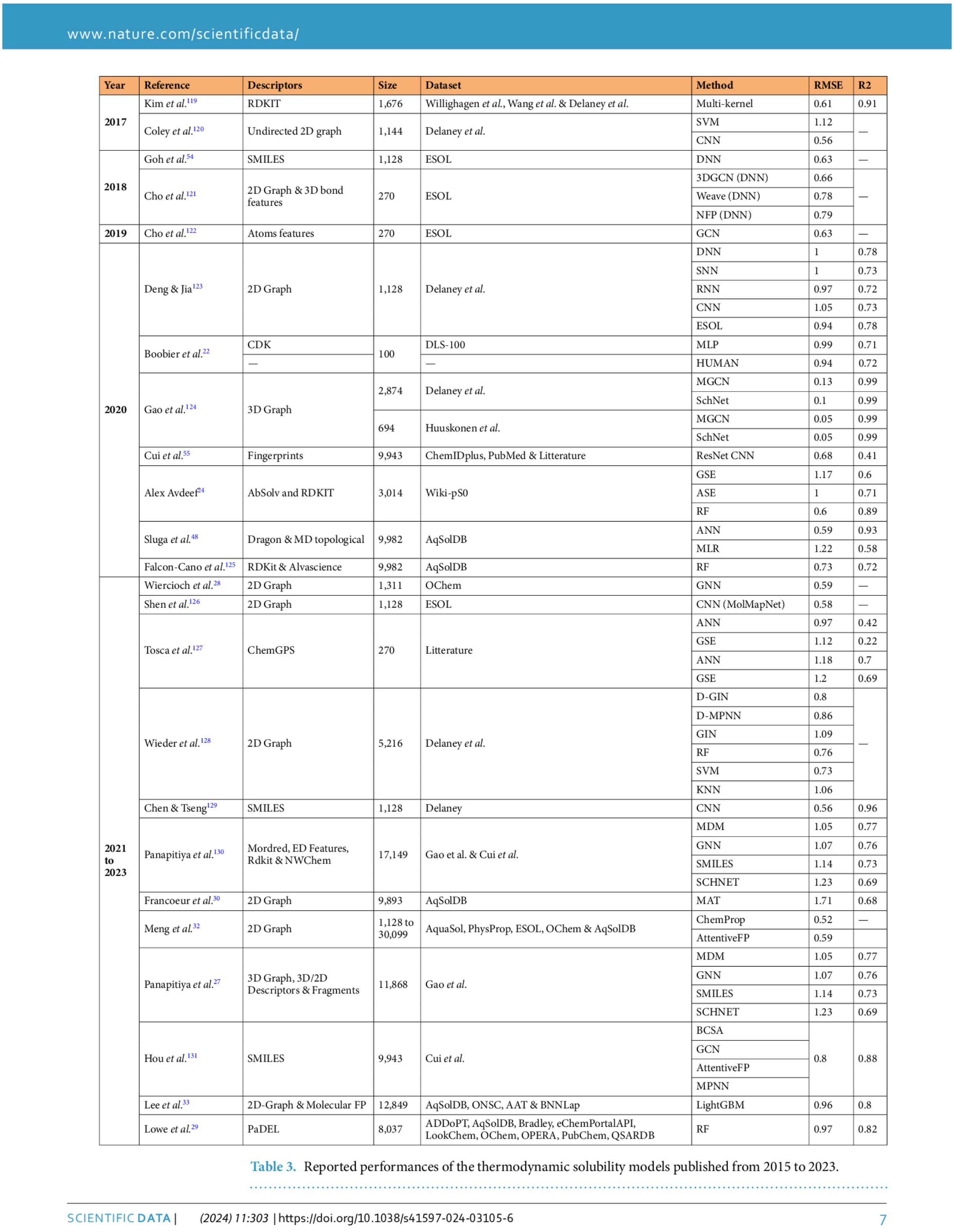
Year	Reference	Descriptors	Size	Dataset	Method	RMSE	R2
2008	Cheung *et al.*[105]	MOE	110	Litterature	MLR	—	0.9
					ANN		0.85
	Duchowicz *et al.*[106]	Dragon	166	Merck Index	RM	—	0.85
	Huuskonen *et al.*[48]	DayLight	191	AquaSol, Merck Index, ChemFinder & PhysProp	MLR	—	0.8
	Hughes *et al.*[107]	cLopP & Tm	237	Bergström *et al.*, Rytting *et al.* & Wassvik *et al.*	MLR	1.03	0.63
					SVM		
	Zhou *et al.*[49]	ECFP	1,299	Huuskonen *et al.*	PLS	0.71	0.85
	Husskonen *et al.*[48]	cLogP & Counts	365	AquaSol	MLR	—	0.87
	Du-Cuny *et al.*[108]	LogP, Fragments & Index	2,473	Roche proprietary	PLS	0.42	0.84
	Obrezanova *et al.*[109]	ATC, logP, Volume & MW	592	Syracuse	GP	0.71	0.88
2009	Wang *et al.*[21]	ATC,ClogP, MW	4,874	Delaney *et al.* & Huuskonen *et al.*	MLR	0.98	0.83
	Hewitt *et al.*[53]	LogP, Tb & Dragon	104	SC1	MLR	0.95	0.74
					ANN	1.51	0.79
	Duchowicz & Castro[110]	Dragon	145	Merck Index	MLR	0.9	0.76
2010	Ghafourian & Bozorgi[111]	ACD-Labs & TSAR 3D	141	Rytting *et al.*	SR	0.71	—
	Muratov *et al.*[112]	2D Simplex	290	Klampt *et al.*	PLS	—	0.81
	Cao *et al.*[65]	Dragon	225	Llinas *et al.* & Merck Index	SVR	—	0.74
	Jain & Yalkowsky[113]	Activity coefficients, Melting Entropy & MP	883	AquaSol & EPA	TE	—	0.73
	Eric *et al.*[114].	CODESSA	319	Rytting *et al.*	MLR	0.96	0.66
	Louis *at al*[115]	Marvin & Karselson	74	Bergstrom *et al.* & others	MLR	0.8	0.55
					ANN	0.74	0.59
					SVM	0.83	0.53
	Fatemi *et al.*[116]	LFER from ADME Boxes	145	Duchowicz *et al.*	MLR	0.92	0.71
					LS-SVM	0.73	0.85
					ANN	0.75	0.72
2012	Chevillard *et al.*[64]	MOE, ADMET predictor & ISIDA	4,897	PhysProp, Huuskonen *et al.* & SC1	RF	0.51	0.62
						0.72	0.56
						0.89	0.23
	Slavica *et al.*[50]	CODESSA	374	Eric Slavica *et al.*	CPANN	0.68	—
2013	Lusci *et al.*[47]	2D Graph	1,144	Delaney *et al.*	UG-RNN	0.58	0.92
					UG-RNN-CR	0.79	0.86
					UG-RNN + logP	0.61	0.91
					UG-RNN-CR + log P	0.63	0.91
					2D kernel	0.61	0.91
	Salahinejad *et al.*[117]	VOLSURF, CPSA, Energy lattice and Sublimation enthalpie	4,376	PhysProp	MLR	—	0.9
2014	McDonagh *et al.*[118]	CDK	100	CSD	PLS	1.08	—
					RF	0.93	
					SVR	1.17	

Table 2. Reported performances of the thermodynamic solubility models published from 2008 to 2014.

an experimental knowledge of the MP of the solutes or a precise knowledge of the polymorph. In both cases, it may be easier to measure the solubility directly.

- Additionally, the solubility of a compound is highly dependent on its acid-base properties, particularly when the solution pH is within 2 log units of the compound's pK_a. Any errors in estimating pKa can lead to large deviations in solubility values. Thus, it may be safer to rely on experimental determination for these properties rather than trying to estimate them in QSPR models.

The abundance of modeling approaches motivated Llinas *et al.*[14] to organize in 2008 the *Solubility Challenge* (SC1). Its goal was to correctly predict the intrinsic solubility from 32 compounds using a given training set of 100 compounds. The challenge data covered a wide and high range in measurements, from 0.5 to 3.0 log unit. To predict it, participants used the full range of existing methods. Models' performances highlighted difficulties in the prediction of highly and poorly soluble compounds. Overall, only about one-third of the compounds were correctly predicted by the best performing models, with the lower RMSE around 0.6 log[46]. SC1 sparked debates on how to enhance the predictive methods as well as the quality of the measurements. It also triggered



www.nature.com/scientificdata/

Year	Reference	Descriptors	Size	Dataset	Method	RMSE	R2
2017	Kim *et al.*[119]	RDKIT	1,676	Willighagen *et al.*, Wang *et al.* & Delaney *et al.*	Multi-kernel	0.61	0.91
	Coley *et al.*[120]	Undirected 2D graph	1,144	Delaney *et al.*	SVM	1.12	—
					CNN	0.56	
2018	Goh *et al.*[54]	SMILES	1,128	ESOL	DNN	0.63	—
	Cho *et al.*[121]	2D Graph & 3D bond features	270	ESOL	3DGCN (DNN)	0.66	—
					Weave (DNN)	0.78	
					NFP (DNN)	0.79	
2019	Cho *et al.*[122]	Atoms features	270	ESOL	GCN	0.63	—
2020	Deng & Jia[123]	2D Graph	1,128	Delaney *et al.*	DNN	1	0.78
					SNN	1	0.73
					RNN	0.97	0.72
					CNN	1.05	0.73
					ESOL	0.94	0.78
	Boobier *et al.*[22]	CDK	100	DLS-100	MLP	0.99	0.71
		—		—	HUMAN	0.94	0.72
	Gao *et al.*[124]	3D Graph	2,874	Delaney *et al.*	MGCN	0.13	0.99
					SchNet	0.1	0.99
			694	Huuskonen *et al.*	MGCN	0.05	0.99
					SchNet	0.05	0.99
	Cui *et al.*[55]	Fingerprints	9,943	ChemIDplus, PubMed & Litterature	ResNet CNN	0.68	0.41
	Alex Avdeef[24]	AbSolv and RDKIT	3,014	Wiki-pS0	GSE	1.17	0.6
					ASE	1	0.71
					RF	0.6	0.89
	Sluga *et al.*[48]	Dragon & MD topological	9,982	AqSolDB	ANN	0.59	0.93
					MLR	1.22	0.58
	Falcon-Cano *et al.*[125]	RDKit & Alvascience	9,982	AqSolDB	RF	0.73	0.72
2021 to 2023	Wiercioch *et al.*[28]	2D Graph	1,311	OChem	GNN	0.59	—
	Shen *et al.*[126]	2D Graph	1,128	ESOL	CNN (MolMapNet)	0.58	—
	Tosca *et al.*[127]	ChemGPS	270	Litterature	ANN	0.97	0.42
					GSE	1.12	0.22
					ANN	1.18	0.7
					GSE	1.2	0.69
	Wieder *et al.*[128]	2D Graph	5,216	Delaney *et al.*	D-GIN	0.8	—
					D-MPNN	0.86	
					GIN	1.09	
					RF	0.76	
					SVM	0.73	
					KNN	1.06	
	Chen & Tseng[129]	SMILES	1,128	Delaney	CNN	0.56	0.96
	Panapitiya *et al.*[130]	Mordred, ED Features, Rdkit & NWChem	17,149	Gao et al. & Cui *et al.*	MDM	1.05	0.77
					GNN	1.07	0.76
					SMILES	1.14	0.73
					SCHNET	1.23	0.69
	Francoeur *et al.*[30]	2D Graph	9,893	AqSolDB	MAT	1.71	0.68
	Meng *et al.*[32]	2D Graph	1,128 to 30,099	AquaSol, PhysProp, ESOL, OChem & AqSolDB	ChemProp	0.52	—
					AttentiveFP	0.59	
	Panapitiya *et al.*[27]	3D Graph, 3D/2D Descriptors & Fragments	11,868	Gao *et al.*	MDM	1.05	0.77
					GNN	1.07	0.76
					SMILES	1.14	0.73
					SCHNET	1.23	0.69
	Hou *et al.*[131]	SMILES	9,943	Cui *et al.*	BCSA	0.8	0.88
					GCN		
					AttentiveFP		
					MPNN		
	Lee *et al.*[33]	2D-Graph & Molecular FP	12,849	AqSolDB, ONSC, AAT & BNNLap	LightGBM	0.96	0.8
	Lowe *et al.*[29]	PaDEL	8,037	ADDoPT, AqSolDB, Bradley, eChemPortalAPI, LookChem, OChem, OPERA, PubChem, QSARDB	RF	0.97	0.82

Table 3. Reported performances of the thermodynamic solubility models published from 2015 to 2023.

SCIENTIFIC DATA | *(2024) 11:303* | https://doi.org/10.1038/s41597-024-03105-6 7 the development of numerous models by the community, for which estimating the quality of the data took precedence over enhancing accuracy.

These methods employed novel neural network architectures (Table 2). For instance, Lusci *et al.*[47] introduced in 2013 a method based on Undirected Graphs (UG). Their approach was applied with a 10-fold internal Cross-Validation (CV) to ESOL, Llinas *et al*. 2008, and Huuskonen *et al.*[16] and reached a low RMSE of 0.58 log. Number of other approaches were introduced during this period: MLR by Huuskonen *et al.*[48] in 2008, PLS by Zhou *et al.*[49] in 2008, MLR by Wang *et al.*[21] in 2009 and CPANN by Eric *et al.*[50] in 2012.

This raise of powerful machine-learning methods available motivated Llinas and Avdeef[51] to organize a second *Solubility Challenge* (SC2) in 2019. This time, they invited participants to apply their own models to 2 datasets. Set 1 consisted of 100 druglike compounds with an average *SDi* of 0.17 log. Set 2 contained 32 molecules with an average *SDi* of 0.62 log. Participants were asked to use their own training set. No significant improvements were found compared to the SC1[52]. Every method worked equally well and achieved a minimal RMSE of 0.70 log[14,51,53].

The current period is marked by a trend of deep learning architecture and molecular embedding inputs emerged (Table 3). In 2018, Goh *et al.*[54] introduced SMILE2vec, the first interpretable DNN to use SMILES for chemical property prediction. The developed NN was inspired by Word2Vec, a DL technique commonly used in NLP research. By comparing the performance of different Bayesian optimization techniques for hyperparameter tuning on the ESOL dataset, they were able to identify the most effective architecture, CNN-GRU. Applied to ESOL validation set, their model achieved a RMSE of 0.63 log and demonstrated interpretability by highlighting chemical functions, using a residual NN as a mask to identify important characters from the input. Their model accuracy outperformed feature-based methods.

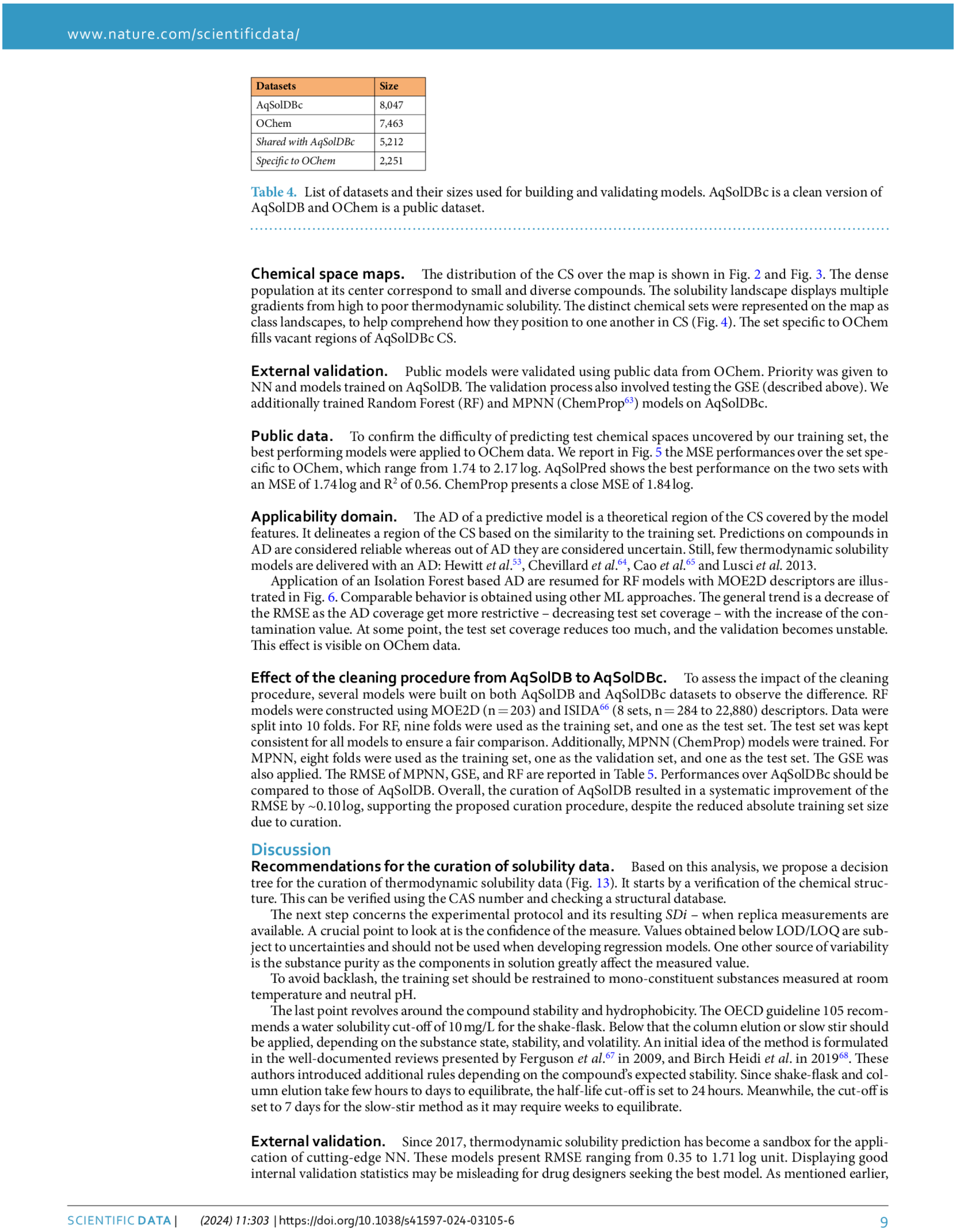
A similar approach was conducted by Cui *et al.*[55] in 2020 by adapting the well-known ResNet to accept PubChem fingerprints as input. They constructed N-layers (N = 14, 20, or 26) CNN models based on the architecture of ResNet. Models were evaluated with a 10-fold CV on 9,943 compounds from ChemIDplus and PubMed. They achieved a RMSE of 0.68 log, highlighting the advantage of going deeper. However, this is in contradiction with Francoeur *et al.*[30] results from 2021, concluding that smaller networks performed better.

In their study, Francoeur *et al*. optimized a Molecular Attention Transformer (MAT) to predict aqueous solubility from SMILES representation, called SolTranNet. Their method is based on the MAT architecture developed by Maziarka *et al.*[56] MAT functions by applying self-attention to a molecular graph where each node is defined as a feature vector. Vectors are then combined with the adjacency matrix before being fed to the NN layers. The MAT hyperparameters were optimized by minimizing the RMSE of an AqSolDB subset. To validate their model, SolTranNet was applied to three different test sets: the SC2 test set, Cui *et al*. 2020 dataset, and Boobier *et al.*[22] 2017 dataset, resulting in RMSE values of 1.295, 0.813 and 0.845 log, respectively. SolTranNet has comparable performance to current ML models. However, Francoeur *et al.*[30] points out that the small size of the community test sets limits the conclusions to be drawn from their reported performances. Even when trained over large sets, models may not be generalizable to other datasets, especially those from specific domains, such as compounds of pharmaceutical interest, as also mentioned in Lovrić *et al.*[57].

We hypothesized that the performances published might be optimistic, because of: (i) inaccurate delimitation or failure from the applicability domain, if defined, and (ii) lack of independent external validation sets. Yet, caution is warranted when comparing model efficacy across studies, given the significant variability in test sets and methodologies. As of now, numerous models are still published without validation on completely independent sets. Different validation strategies, such as internal and external, can be distinguished, varying in levels of rigor. Internal validation makes use of the same data from which the model was fitted. External validation requires an independent dataset to correctly assess the model's reproducibility and generalizability, and thus application to other chemical spaces (CS). However, it's a common misconception that splitting a dataset into a training and a validation set (random split or k-fold CV) is sufficient, especially with GNN where data leakage can happen. Data leakage occurs when information from the test set is used in the training process, which can lead to biased performance assessment of the model. In CV, the test sets are independent to some extent[58] but the training set largely overlap. In the case of GNN, this can happen if the GNN has seen test set chemical structures during the pre-training process. This problem has been discussed in various studies, offering alternative validation techniques as potential solutions[59]. Despite these criticisms, the efficacy of cross-validation remains undiscussed, as empirically demonstrated in works by Breiman & Spector[60] and further supported theoretically by Vapnik[61]. The importance of the test set size, coverage and quality is supported by Francoeur *et al.*[30]. Ideally, this set should be meaningful and be excluded from the model training to ensure realistic performances. For instance, Cui *et al.* in 2020 validated their DNN models on two small test sets of 62, and 5 compounds, obtaining RMSE of 0.681 and 0.689 *LogS* unit, respectively. These test sets are arguably small, but the former was aggregated from recent literature while the second was composed of new in-house data. In this publication, models' performances were also compared to human expert performances. This contrasted with previously reported results in Boobier *et al.* in 2017. In this study, models were trained and tested on 100 compounds from the DLS-100 dataset which regroup S_0 entries, mostly from Llinas *et al*. 2008 and Rytting *et al.*[62]. Data were used following a train/test split of 75/25 compounds. As a result, humans performed equally as ML models with a RMSE of 1.087 for the former against 1.140 log for the later.

Results

Data. For this study, we used two public thermodynamic solubility datasets: AqSolDBc (our clean version of AqSolDB) and OChem. Our intent was to externally validate models trained on AqSolDBc by testing them over public data. Datasets are resumed in the Table 4.



Datasets	Size
AqSolDBc	8,047
OChem	7,463
Shared with AqSolDBc	5,212
Specific to OChem	2,251

Table 4. List of datasets and their sizes used for building and validating models. AqSolDBc is a clean version of AqSolDB and OChem is a public dataset.

Chemical space maps. The distribution of the CS over the map is shown in Fig. 2 and Fig. 3. The dense population at its center correspond to small and diverse compounds. The solubility landscape displays multiple gradients from high to poor thermodynamic solubility. The distinct chemical sets were represented on the map as class landscapes, to help comprehend how they position to one another in CS (Fig. 4). The set specific to OChem fills vacant regions of AqSolDBc CS.

External validation. Public models were validated using public data from OChem. Priority was given to NN and models trained on AqSolDB. The validation process also involved testing the GSE (described above). We additionally trained Random Forest (RF) and MPNN (ChemProp[63]) models on AqSolDBc.

Public data. To confirm the difficulty of predicting test chemical spaces uncovered by our training set, the best performing models were applied to OChem data. We report in Fig. 5 the MSE performances over the set specific to OChem, which range from 1.74 to 2.17 log. AqSolPred shows the best performance on the two sets with an MSE of 1.74 log and R^2 of 0.56. ChemProp presents a close MSE of 1.84 log.

Applicability domain. The AD of a predictive model is a theoretical region of the CS covered by the model features. It delineates a region of the CS based on the similarity to the training set. Predictions on compounds in AD are considered reliable whereas out of AD they are considered uncertain. Still, few thermodynamic solubility models are delivered with an AD: Hewitt *et al.*[53], Chevillard *et al.*[64], Cao *et al.*[65] and Lusci *et al.* 2013.

Application of an Isolation Forest based AD are resumed for RF models with MOE2D descriptors are illustrated in Fig. 6. Comparable behavior is obtained using other ML approaches. The general trend is a decrease of the RMSE as the AD coverage get more restrictive – decreasing test set coverage – with the increase of the contamination value. At some point, the test set coverage reduces too much, and the validation becomes unstable. This effect is visible on OChem data.

Effect of the cleaning procedure from AqSolDB to AqSolDBc. To assess the impact of the cleaning procedure, several models were built on both AqSolDB and AqSolDBc datasets to observe the difference. RF models were constructed using MOE2D (n = 203) and ISIDA[66] (8 sets, n = 284 to 22,880) descriptors. Data were split into 10 folds. For RF, nine folds were used as the training set, and one as the test set. The test set was kept consistent for all models to ensure a fair comparison. Additionally, MPNN (ChemProp) models were trained. For MPNN, eight folds were used as the training set, one as the validation set, and one as the test set. The GSE was also applied. The RMSE of MPNN, GSE, and RF are reported in Table 5. Performances over AqSolDBc should be compared to those of AqSolDB. Overall, the curation of AqSolDB resulted in a systematic improvement of the RMSE by ~0.10 log, supporting the proposed curation procedure, despite the reduced absolute training set size due to curation.

Discussion

Recommendations for the curation of solubility data. Based on this analysis, we propose a decision tree for the curation of thermodynamic solubility data (Fig. 13). It starts by a verification of the chemical structure. This can be verified using the CAS number and checking a structural database.

The next step concerns the experimental protocol and its resulting *SDi* – when replica measurements are available. A crucial point to look at is the confidence of the measure. Values obtained below LOD/LOQ are subject to uncertainties and should not be used when developing regression models. One other source of variability is the substance purity as the components in solution greatly affect the measured value.

To avoid backlash, the training set should be restrained to mono-constituent substances measured at room temperature and neutral pH.

The last point revolves around the compound stability and hydrophobicity. The OECD guideline 105 recommends a water solubility cut-off of 10 mg/L for the shake-flask. Below that the column elution or slow stir should be applied, depending on the substance state, stability, and volatility. An initial idea of the method is formulated in the well-documented reviews presented by Ferguson *et al.*[67] in 2009, and Birch Heidi *et al.* in 2019[68]. These authors introduced additional rules depending on the compound's expected stability. Since shake-flask and column elution take few hours to days to equilibrate, the half-life cut-off is set to 24 hours. Meanwhile, the cut-off is set to 7 days for the slow-stir method as it may require weeks to equilibrate.

External validation. Since 2017, thermodynamic solubility prediction has become a sandbox for the application of cutting-edge NN. These models present RMSE ranging from 0.35 to 1.71 log unit. Displaying good internal validation statistics may be misleading for drug designers seeking the best model. As mentioned earlier,

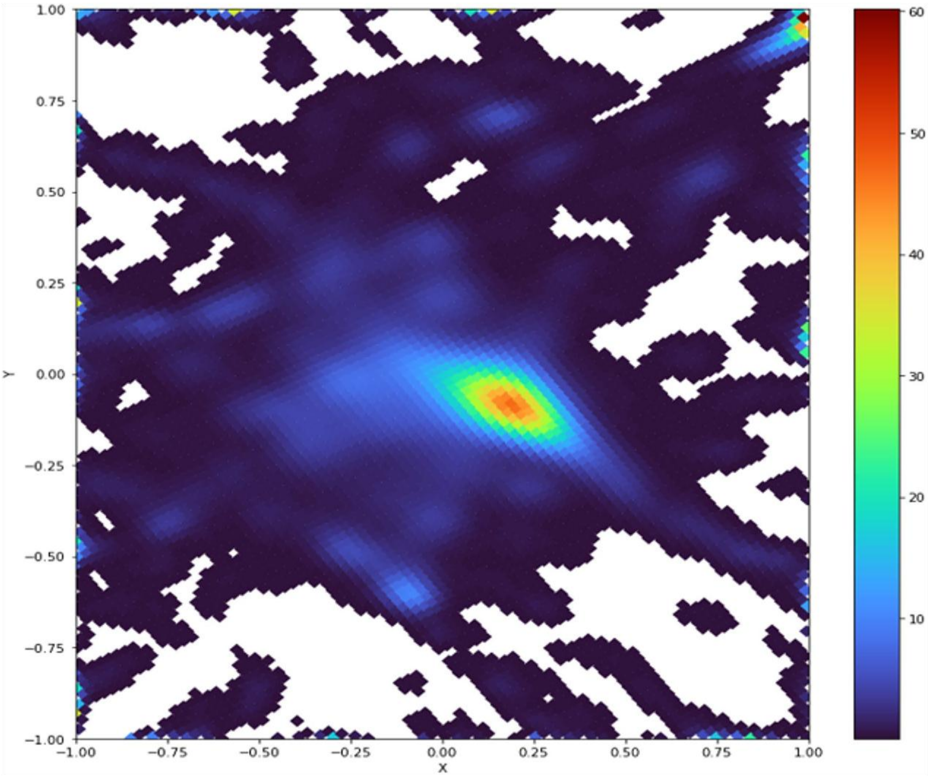


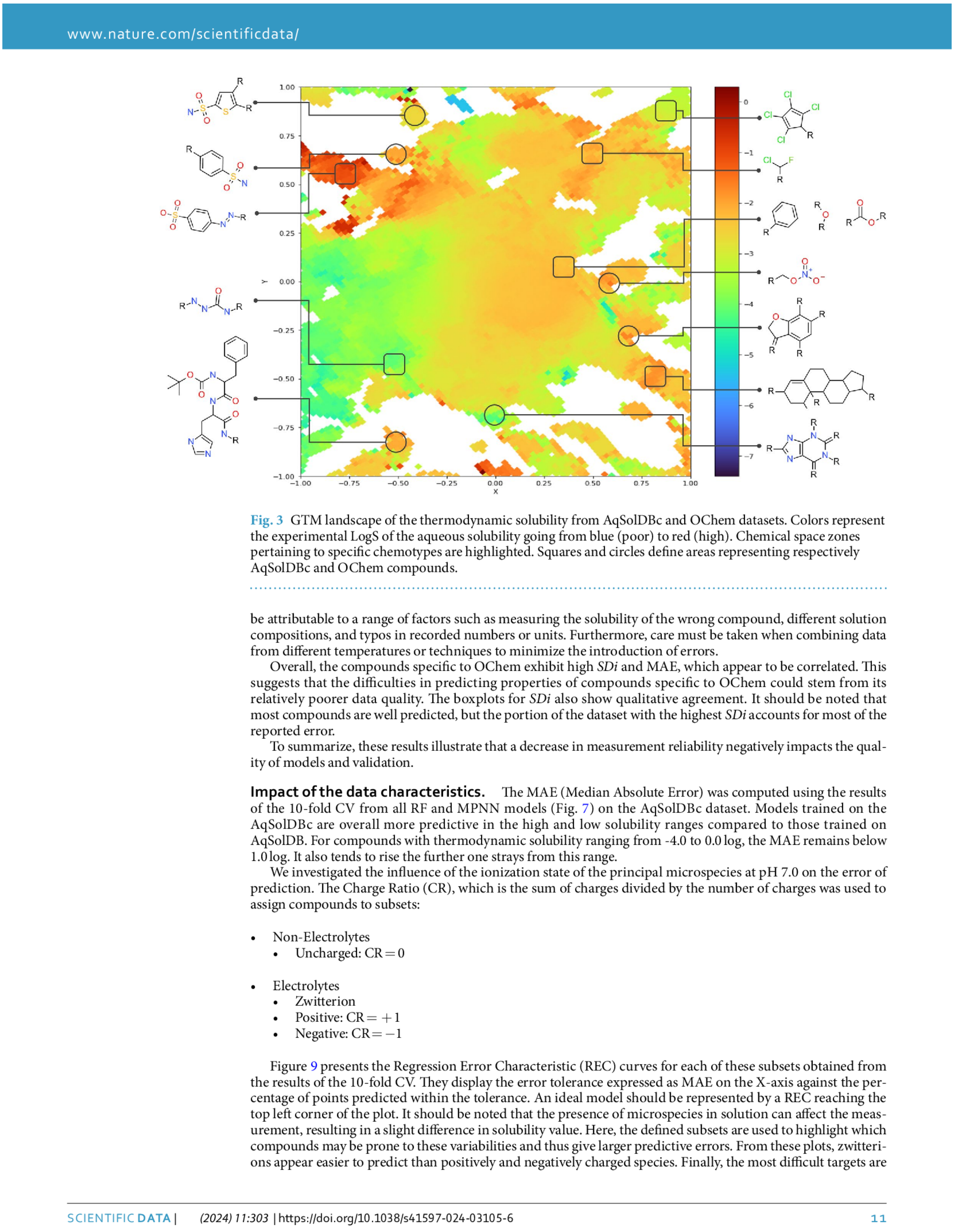
Fig. 2 GTM density landscape of the chemical space jointly covered by AqSolDBc and OChem. White spaces are unpopulated areas. Colors represent the number of molecules per nodes, from blue (low) to red (high).

these models often lack extensive external validation, and thus their performance should be considered with skepticism, particularly when applied to New Chemical Entities.

Public data. To confirm the difficulty of predicting test chemical spaces uncovered by our training set, the best performing models were applied to OChem data. The relevance of previously performed external validation may be questioned. For instance, evaluating performances using sets too small, internal, or distant from a target application (i.e. pharmaceutical data) may be an issue. Validation sets, which are meant to evaluate models in the context of their specific characteristics, should be carefully chosen based on their composition, diversity, size, and quality. It is important to note that each external test set presents its own challenges due to its peculiarities (size, diversity, predominance of various chemotypes, etc.), and past success on external validation does not guarantee future performance on different test sets. Moreover, Neural Network architectures do not display any breakthrough performances. As hypothesized previously, certain prediction errors may be avoided by using an Applicability Domain (AD) with published models.

Inter-laboratory standard deviation. The other possible source of prediction error could be the presence of poorly reproducible or variable training data. If the thermodynamic solubility is not known with sufficient accuracy or exhibits significant variability, it can introduce uncertainty into the models and distort their assessment. We analyzed the *SDi* of the OChem sets and the Median Average Error (MAE) of the set specific to OChem. The MAE is the median of the absolute difference between predictions and measurements for a given compound. Here we discuss MAE using results from a 10-fold cross-validation of ChemProp on OChem data, as a representative example model.

As OChem comprises datasets from various sources, the independent quality of each source can be investigated. To do so, the distributions of the *SDi* are confronted to the source of their entries (Fig. 8). The X-axis defines the source datasets found in OChem. To better highlight the quality of AqSolDBc, the set specific to OChem and shared with AqSolDBc are displayed as separated boxes. It is important to note that errors could



1.00
0.75
0.50
0.25
Y 0.00
−0.25
−0.50
−0.75
−1.00
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
X
0
−1
−2
−3
−4
−5
−6
−7

Fig. 3 GTM landscape of the thermodynamic solubility from AqSolDBc and OChem datasets. Colors represent the experimental LogS of the aqueous solubility going from blue (poor) to red (high). Chemical space zones pertaining to specific chemotypes are highlighted. Squares and circles define areas representing respectively AqSolDBc and OChem compounds.

be attributable to a range of factors such as measuring the solubility of the wrong compound, different solution compositions, and typos in recorded numbers or units. Furthermore, care must be taken when combining data from different temperatures or techniques to minimize the introduction of errors.

Overall, the compounds specific to OChem exhibit high *SDi* and MAE, which appear to be correlated. This suggests that the difficulties in predicting properties of compounds specific to OChem could stem from its relatively poorer data quality. The boxplots for *SDi* also show qualitative agreement. It should be noted that most compounds are well predicted, but the portion of the dataset with the highest *SDi* accounts for most of the reported error.

To summarize, these results illustrate that a decrease in measurement reliability negatively impacts the quality of models and validation.

Impact of the data characteristics. The MAE (Median Absolute Error) was computed using the results of the 10-fold CV from all RF and MPNN models (Fig. 7) on the AqSolDBc dataset. Models trained on the AqSolDBc are overall more predictive in the high and low solubility ranges compared to those trained on AqSolDB. For compounds with thermodynamic solubility ranging from -4.0 to 0.0 log, the MAE remains below 1.0 log. It also tends to rise the further one strays from this range.

We investigated the influence of the ionization state of the principal microspecies at pH 7.0 on the error of prediction. The Charge Ratio (CR), which is the sum of charges divided by the number of charges was used to assign compounds to subsets:

- Non-Electrolytes
  - Uncharged: CR = 0
- Electrolytes
  - Zwitterion
  - Positive: CR = +1
  - Negative: CR = −1

Figure 9 presents the Regression Error Characteristic (REC) curves for each of these subsets obtained from the results of the 10-fold CV. They display the error tolerance expressed as MAE on the X-axis against the percentage of points predicted within the tolerance. An ideal model should be represented by a REC reaching the top left corner of the plot. It should be noted that the presence of microspecies in solution can affect the measurement, resulting in a slight difference in solubility value. Here, the defined subsets are used to highlight which compounds may be prone to these variabilities and thus give larger predictive errors. From these plots, zwitterions appear easier to predict than positively and negatively charged species. Finally, the most difficult targets are

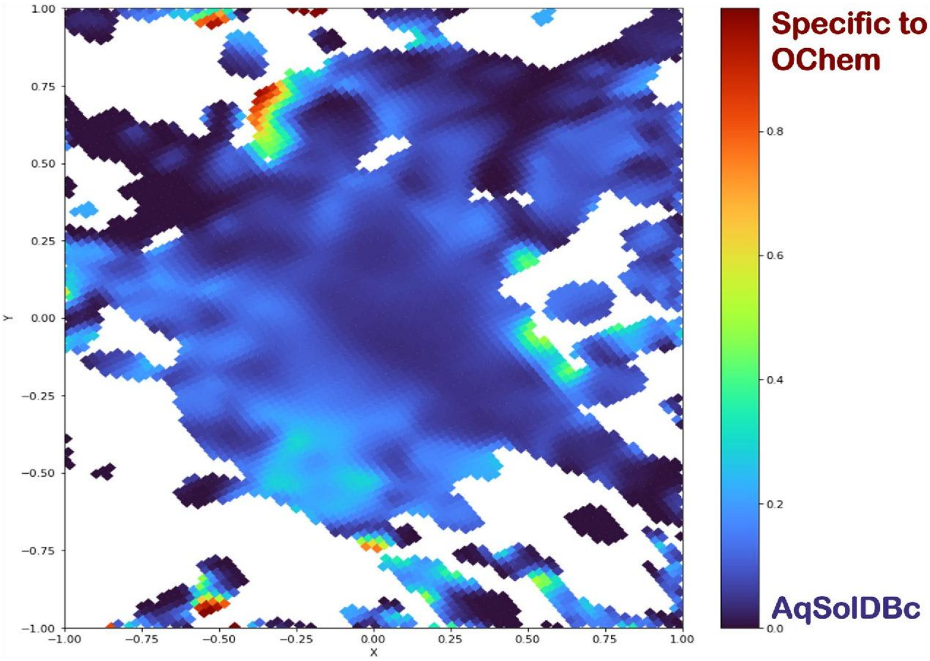


Fig. 4 Class landscape of the test sets versus the training set, AqSolDBc. The color represents the proportion of compounds from each dataset. Blue regions are populated with structures from AqSolDBc. White spaces are unpopulated areas and red spaces are from compounds specific to OChem datasets.

uncharged species. This is probably due to the fact that most poorly soluble species are actually uncharged, and some neutral species may be incorrectly identified as uncharged by the machine learner for rare groups.

Since AqSolDB and AqSolDBc are aggregations of public datasets, it was also possible to study the influence of data sources on the measured performances of the models (Fig. 10). The Huuskonen dataset is certainly the easiest data collection to predict. The largest errors are observed on the Raevsky, EPI Suite 2020 and, mostly eChemPortal 2020 datasets. The eChemPortal provides a lot of input data to AqSolDB, but it appears that they might be a large source of erroneous entries. Therefore, the eChemPortal dataset requires a closer look which is out of the scope of this study.

Hard-to-predict compounds. Finally, the information concerning the 20 hardest-to-predict compounds (having the largest MAE) from AqSolDBc are reported in Table 6 and Fig. 11. Most of them are hydrophobic compounds from eChemPortal and measured using the shake-flask method. However, the OECD 105 advises to use the column elution with poorly soluble molecules. The usual lack of confidence over poorly soluble substance can be partially explained by the non-respect of the OECD.

Interpretation of the model. To evaluate the contribution of each atom into the modelled solubility, we employed ColorAtom[69,70]. This interface employed our RF model based on ISIDA fragment descriptors to produce chemical structures where each atom bears an atomic contribution of the value calculated by the model. The 20 hardest-to-predict compounds were passed on ColorAtom. Their colored structures are reported Fig. 12. As expected, the polar parts of the molecules are usually colored in blue (high solubilization) whereas aromatic and aliphatic moieties are in red (poor solubilization).

Key results. In our study, we conducted an extensive analysis of thermodynamic solubility using two datasets: AqSolDBc and OChem. Our findings underscored the complexities and challenges of solubility prediction, but also highlighted potential strategies for improvement.

The mapping of chemical space revealed a diverse range of the solubility subspaces, highlighting the value of using diverse and complementary datasets. Despite the diversity of data, external validation revealed that all models struggled. This finding underscored the importance of model refinement and the need to consider the applicability domain when applying models to novel data. Moreover, the curation of AqSolDB into AqSolDBc significantly improved the RMSE, showing that data cleaning procedures can substantially enhance prediction accuracy.

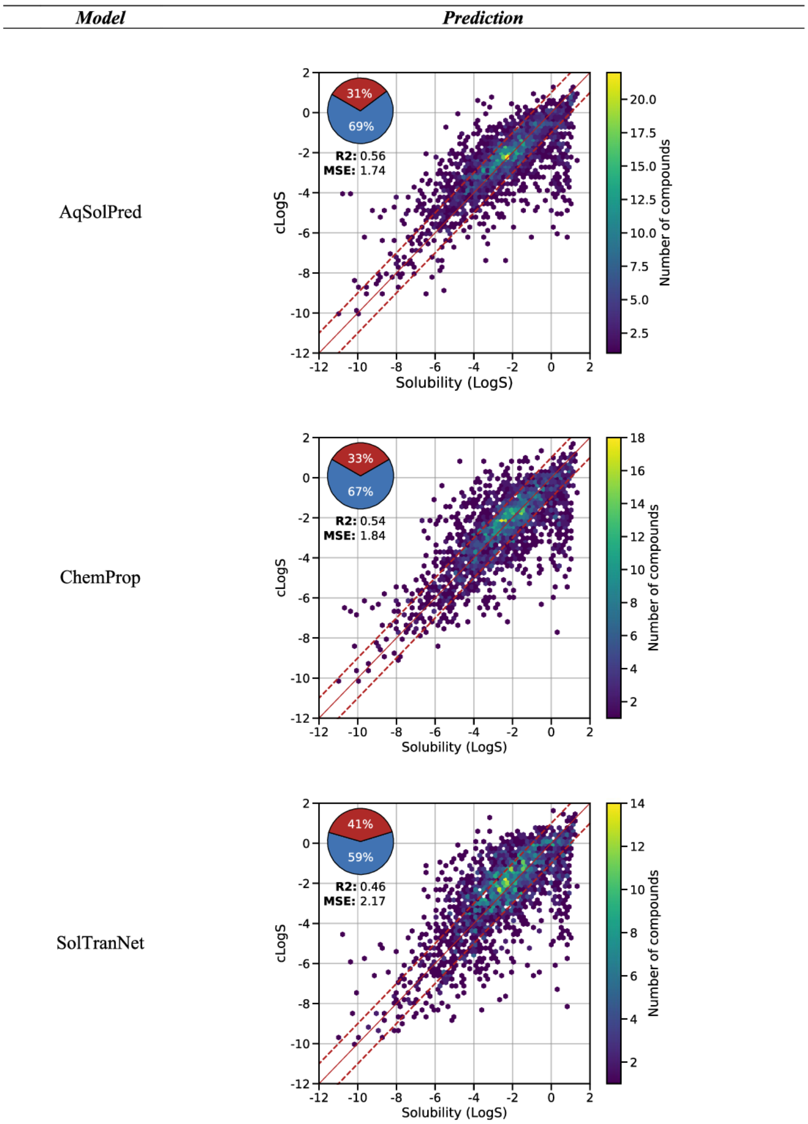


Fig. 5 Predicted thermodynamic solubility against experimental solubility for the set specific to OChem. The red line represents a ± 1.0 log interval. The hexbins represent the density of points in the plot.

Our study also revealed that inter-laboratory variability and the source of data can significantly influence model performance. This highlights the importance of measurement reliability and stringent data validation procedures, raising questions about the quality of datasets like eChemPortal.

Our study corroborates the findings of Lowe *et al.*[29], emphasizing the complexity and challenges in solubility prediction across diverse chemical spaces. We found that RF models provide a balanced and interpretable framework. The model's interpretation underscored the essential role of fragment-based modeling approaches in elucidating the underlying mechanisms of the predictions. These insights underline the importance of the application of OECD[68] principles for enhancing predictive accuracy and interpretability. Additionally, we investigated

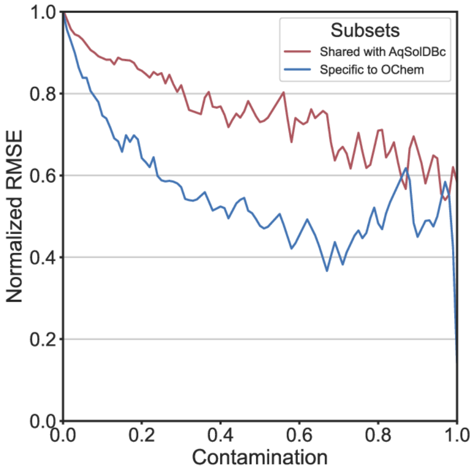


Fig. 6 Performance of the RF model (MOE2D) using the IsolationForest Applicability Domain. Performances were computed for each increment of the contamination parameter, from 0.0 to 0.99. Normalized RMSE is the external validation RMSE at contamination X divided by the RMSE at contamination zero.

Dataset	RF MOE2D	RF ISIDA	MPNN ChemProp	GSE (Eq. 1)
AqSolDBc	0.78	0.91	0.79	1.86
AqSolDB	0.86	0.99	0.89	2.05

Table 5. Root-Mean Squared Error (RMSE) of the RF, MPNN and GSE through 10-fold CV on AqSolDBc & AqSolDB. Colors are ranged from green (low RMSE) to red (high RMSE).

the 20 hardest-to-predict compounds, most of which were hydrophobic and measured using unsuitable methods. This underscored the need of carefully selecting entries based on their experimental procedure, to which we answered by delivering a decision tree for the curation of solubility data.

Overall, our findings indicate that while advancements have been made in the field of solubility prediction, challenges remain. These insights offer valuable guidance for future research and model refinement.

Summary. Published solubility models often display attractive performances. However, these same models very often fail in prospective predictions. This work aimed at clarifying the reasons for these repeated failures.

First, we compiled a comprehensive list of solubility datasets and highlighted their interconnections. It appears that some data sources are overlooked and others frequently aggregated.

Second, we observed that the use of sophisticated neural network architectures did not lead to any breakthrough, although major scientific discussions were triggered by both solubility challenges 1 and 2.

Third, when applied to an external public dataset, all models performed poorly. This is probably due to an applicability domain issue.

Fourth, we conducted a thorough reevaluation of the popular AqSolDB dataset to address potential inconsistencies and improve its quality. Our analysis led to the creation of a new version of the dataset, which exhibits improved internal consistency by ensuring that the data points are more reliable and better adhere to the principles of solubility prediction. This revised dataset allows for a more accurate assessment of factors that impact the performance of solubility prediction models, ultimately leading to better model development and evaluation. This allowed us to observe the influence of factors impacting the performances of the models: the laboratory standard deviation, the ionic state of the solute, and the source of the solubility data. It appears that the eChemPortal probably contains some corrupted data and requires careful data cleaning.

Lastly, we provide a thoroughly curated version of AqSolDB called AqSolDBc, obtained following a decision tree based on experimental conditions. With these rules, we hope to offer a correct way to curate aqueous solubility data. This set was used to train RF and MPNN models for solubility prediction and IsolationForest models for Applicability Domain. Models trained on public data, applied during this project are publicly available (https://chematlas.chimie.unistra.fr/WebTools/predictor_solubility.php).

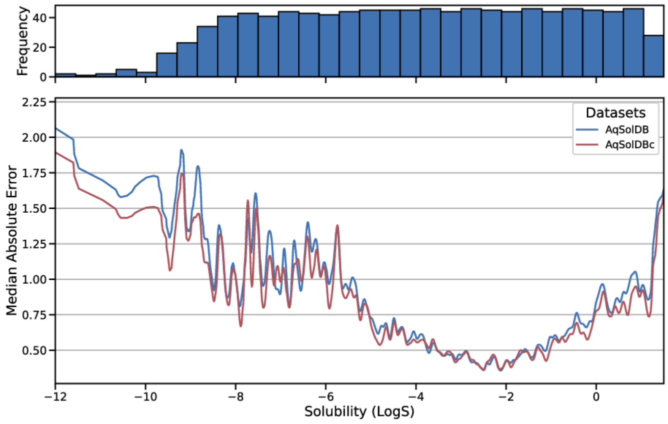


Fig. 7 Comparison of the MAE from AqSolDB and AqSolDBc. MAE from the 10-fold CV computed over all models for AqSolDB (blue) and AqSolDBc (red) against the solubility range.

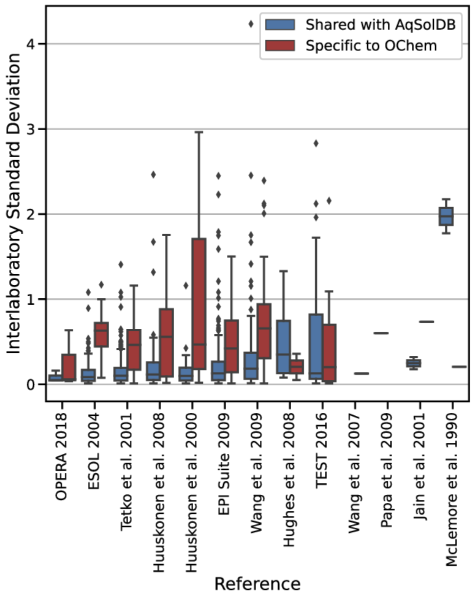


Fig. 8 Boxplots of the experimental standard deviation (*SDi*) of compounds in the OChem database. Data shared with AqSolDB (blue) are also present in AqSolDBc, and data specific to OChem (red) are absent from AqSolDBc. Boxplots are restrained to $SDi > 0.01$ log.

Methods

Data curation. For these approaches to produce accurate predictions over a vast CS, a high quality and diversified training set is a must. However, preserving accurate measurements necessitates accounting for experimental variability, often evaluated with the *SDi*. Experimental thermodynamic solubility data can have inaccuracies up to 1.5 log, according to John C. Dearden[71]. Additionally, Llinas *et al.* reported that measurements between laboratories may vary by 0.5 to 0.6 log. Poor reproducibility can be the consequence of unintentional mistakes brought on by combining entries with heterogenous conditions, or of poor quality[52].

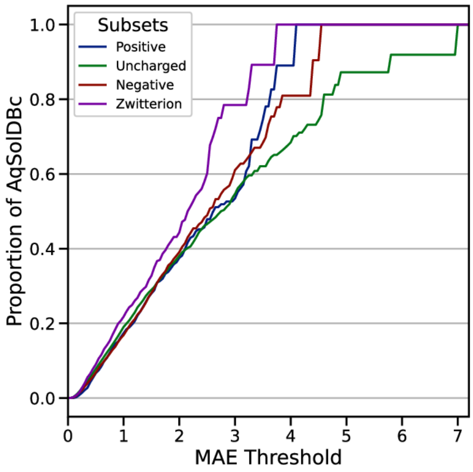


Fig. 9 REC curve for each AqSolDBc subset corresponding to the major microspecies at pH7.0: Uncharged, Zwitterion, Negative and Positive ions. The y-axis is the proportion of AqSolDBc predicted better than a threshold MAE value on the x-axis; MAE in log from the 10-fold CV computed over all models for AqSolDBc.

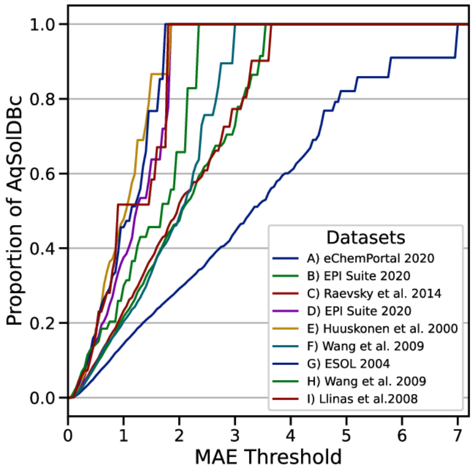


Fig. 10 REC curve of each of the 9 AqSolDB data source. The y-axis is the proportion of AqSolDBc predicted better than a threshold MAE value on the x-axis; MAE from the 10-fold CV computed over all models for AqSolDBc.

In the following, we propose a guideline for the improvement of thermodynamic solubility data set quality, which we applied to AqSolDB. This dataset, aggregated by Sorkun *et al.*[25] in 2020, was chosen for its size, diversity, and well referenced entries. To curate AqSolDB and obtain an experimentally homogenous library, we followed the flowchart illustrated in Fig. 13. Chemaxon's JChem[72] software was employed for structural database standardization. In case of ambiguities, chemical structures were verified in ChemSpider[73] to benefit from its crowd sourced annotations. When possible, these structures were also searched in the CSD where the values of bond lengths, angles and torsions help to disambiguate the nature of chemical functions. CAS numbers were verified using SciFinder[74] before using them to retrieve manually described experimental conditions from eChemPortal[75], EPI Suite[20], and PubChem[76] if available. Overall, 608 entries containing partial records on start and final pH, measurement limitation, composition, origin, stability, or cosolvents were reported (Fig. 14). The forementioned experimental conditions and their importance to modelers are discussed.

pH sensitive species. The thermodynamic solubility of ionizable compounds strongly depends on the pH and the presence of buffer or ions. These factors can influence the microspecies equilibrium by interacting with the solute. For instance, the counter-ion effect can increase, or decrease this solubility. Therefore, several control steps are recommended:

- Verifying the validity of the reported salt structure using its CAS number. This is manageable using the SciFinder[74] database and verifying when possible, in the Cambridge Structural Database[77] (CSD).
- Selecting measurements without buffer, added acids/bases, cosolvents and surfactants.
- Restraining the data to entries reporting a final $pH = 7 \pm 1$.

Ionized compounds obtained through standardization should correspond to the major microspecies in solution. The microspecies distributions have been obtained using ChemAxon pKa Plugin[44]. Compounds presenting too many microspecies (more than 4) and those with uncertain major microspecies at pH 7.0 have been excluded, because we could not decide which structure to use for modeling.

Overall, 399 entries from AqSolDB obtained in the presence of buffer, cosolvent, or undesirable pH were excluded. Five entries were also deemed uncertain for having ionized structures different from the major microspecies or poor microspecies distribution.

Substance composition. Water solubility is a property of pure compounds. However, it is sometimes reported for substances. Pure compounds solubilities cannot be considered together with complex substances solubilities. The European Chemical Agency[38] describes three types of substance:

- UVCB (Unknown or Variable composition, Complex reaction productions or biological materials), contain several chemicals without a complete understanding of their identity. Their composition is variable and often unknown. They usually originate from industrial processes or biological extracts.
- Multi-constituent, account for a mix of known chemicals and impurities. Reported ingredients should represent 10% to 80% of the substance.
- Mono-constituent refers to a solute that only contains one major component with up to 20% impurity. However, this level of purity is still high and can have a significant impact on solubility, bioactivity, and other important factors. It should be noted that such a high level of impurities can negatively affect the results and should be taken into consideration during their interpretation.

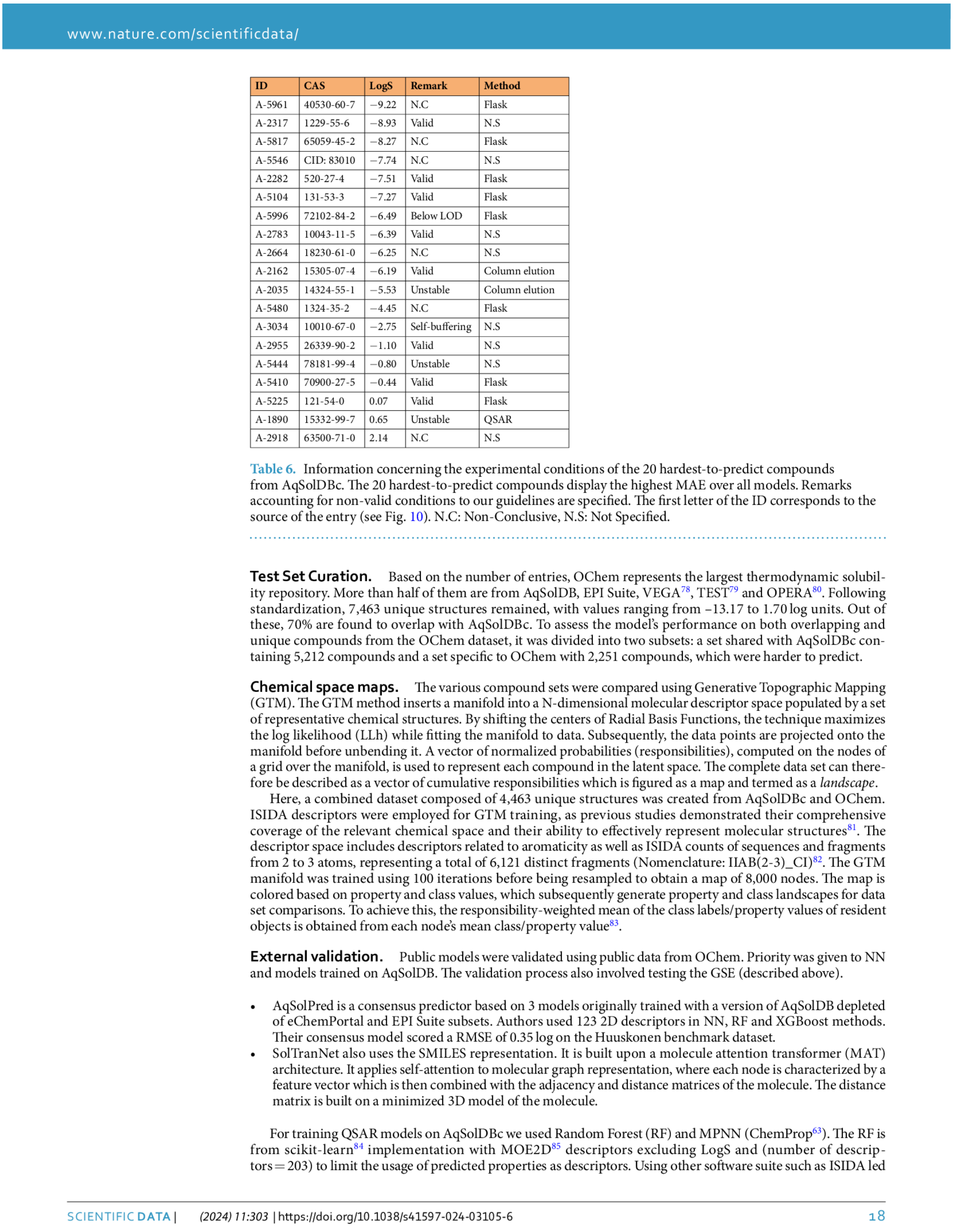
Ninety-nine entries from AqSolDB were found and eliminated for being UVCB, or multi-constituent substances (Fig. 14).

Unstable species. Chemical stability is related to the degradation processes. In solution, the compound can be subject to hydrolysis, hydration (R-(C=O)-R' → $R\text{-}C(OH)_2\text{-}R'$), photolysis, oxidation, biodegradation, and polymerization. These are generally dependent on the pH and temperature. The hydrolysis represents the most difficult ones to avoid during experimentation. Solubility test systems can limit photolysis by using amber glass bottles, aluminum or be done in the dark. Oxidation can be limited by working under anaerobic conditions, through nitrogen or argon flushing or by limiting the air headspace. Chemicals for which hydrolysis rapidly occur should be excluded to avoid measurements altered by reaction products. Care should be taken with compounds containing reactive functional groups such as mono- and poly- halogenated aliphatic (alkyl halides), epoxides, organophosphorus esters, carboxylic acid esters, carbamates, nitriles, organometallic, and peroxides. The Degradation Time (DT50) can be used to investigate the compounds stability. The DT50 is the period after which half of the original amount of chemical is degraded. Hydrophilic compounds with a DT50 lower than 24 hours and hydrophobic with a DT50 lower than 7 days should be discarded[68]. We identified 52 such entries in AqSolDB. Reversible reactions with water, such as hydration of activated aldehydes or internal hemiacetal formation in sugars are not *de facto* signaling compound instability but are sources of prediction error because the actual "solute" structures differ from the input standard form of the molecule.

Other errors. We identified 17 suspicious entries in AqSolDB resulted from either averaging measurement of similar chemicals or predictions with ML methods. In our opinion, such values should not be used for model building. Lastly, the experimental procedure may be biased. For example, two entries were discarded because the calibration of instruments was performed under different conditions than used to run the test samples.

Duplicate measurements. A common outcome of datasets aggregation is the occurrence of duplicated measurements. Managing them is a chance to investigate uncertainties. However, it is desirable to maintain one value per structure, preferably the median. This only make sense when reported values are relatively close. When there are only two very different values, or there are two or three clusters of different values associated to compounds with the same InChI Key, the median or average value becomes meaningless. Such cases are filtered out by a $SDi > 0.5$ log threshold.

The result of this process to the AqSolDB is labeled AqSolDBc in the following.



ID	CAS	LogS	Remark	Method
A-5961	40530-60-7	−9.22	N.C	Flask
A-2317	1229-55-6	−8.93	Valid	N.S
A-5817	65059-45-2	−8.27	N.C	Flask
A-5546	CID: 83010	−7.74	N.C	N.S
A-2282	520-27-4	−7.51	Valid	Flask
A-5104	131-53-3	−7.27	Valid	Flask
A-5996	72102-84-2	−6.49	Below LOD	Flask
A-2783	10043-11-5	−6.39	Valid	N.S
A-2664	18230-61-0	−6.25	N.C	N.S
A-2162	15305-07-4	−6.19	Valid	Column elution
A-2035	14324-55-1	−5.53	Unstable	Column elution
A-5480	1324-35-2	−4.45	N.C	Flask
A-3034	10010-67-0	−2.75	Self-buffering	N.S
A-2955	26339-90-2	−1.10	Valid	N.S
A-5444	78181-99-4	−0.80	Unstable	N.S
A-5410	70900-27-5	−0.44	Valid	Flask
A-5225	121-54-0	0.07	Valid	Flask
A-1890	15332-99-7	0.65	Unstable	QSAR
A-2918	63500-71-0	2.14	N.C	N.S

Table 6. Information concerning the experimental conditions of the 20 hardest-to-predict compounds from AqSolDBc. The 20 hardest-to-predict compounds display the highest MAE over all models. Remarks accounting for non-valid conditions to our guidelines are specified. The first letter of the ID corresponds to the source of the entry (see Fig. 10). N.C: Non-Conclusive, N.S: Not Specified.

Test Set Curation. Based on the number of entries, OChem represents the largest thermodynamic solubility repository. More than half of them are from AqSolDB, EPI Suite, VEGA[78], TEST[79] and OPERA[80]. Following standardization, 7,463 unique structures remained, with values ranging from −13.17 to 1.70 log units. Out of these, 70% are found to overlap with AqSolDBc. To assess the model's performance on both overlapping and unique compounds from the OChem dataset, it was divided into two subsets: a set shared with AqSolDBc containing 5,212 compounds and a set specific to OChem with 2,251 compounds, which were harder to predict.

Chemical space maps. The various compound sets were compared using Generative Topographic Mapping (GTM). The GTM method inserts a manifold into a N-dimensional molecular descriptor space populated by a set of representative chemical structures. By shifting the centers of Radial Basis Functions, the technique maximizes the log likelihood (LLh) while fitting the manifold to data. Subsequently, the data points are projected onto the manifold before unbending it. A vector of normalized probabilities (responsibilities), computed on the nodes of a grid over the manifold, is used to represent each compound in the latent space. The complete data set can therefore be described as a vector of cumulative responsibilities which is figured as a map and termed as a *landscape*.

Here, a combined dataset composed of 4,463 unique structures was created from AqSolDBc and OChem. ISIDA descriptors were employed for GTM training, as previous studies demonstrated their comprehensive coverage of the relevant chemical space and their ability to effectively represent molecular structures[81]. The descriptor space includes descriptors related to aromaticity as well as ISIDA counts of sequences and fragments from 2 to 3 atoms, representing a total of 6,121 distinct fragments (Nomenclature: IIAB(2-3)_CI)[82]. The GTM manifold was trained using 100 iterations before being resampled to obtain a map of 8,000 nodes. The map is colored based on property and class values, which subsequently generate property and class landscapes for data set comparisons. To achieve this, the responsibility-weighted mean of the class labels/property values of resident objects is obtained from each node's mean class/property value[83].

External validation. Public models were validated using public data from OChem. Priority was given to NN and models trained on AqSolDB. The validation process also involved testing the GSE (described above).

- AqSolPred is a consensus predictor based on 3 models originally trained with a version of AqSolDB depleted of eChemPortal and EPI Suite subsets. Authors used 123 2D descriptors in NN, RF and XGBoost methods. Their consensus model scored a RMSE of 0.35 log on the Huuskonen benchmark dataset.
- SolTranNet also uses the SMILES representation. It is built upon a molecule attention transformer (MAT) architecture. It applies self-attention to molecular graph representation, where each node is characterized by a feature vector which is then combined with the adjacency and distance matrices of the molecule. The distance matrix is built on a minimized 3D model of the molecule.

For training QSAR models on AqSolDBc we used Random Forest (RF) and MPNN (ChemProp[63]). The RF is from scikit-learn[84] implementation with MOE2D[85] descriptors excluding LogS and (number of descriptors = 203) to limit the usage of predicted properties as descriptors. Using other software suite such as ISIDA led

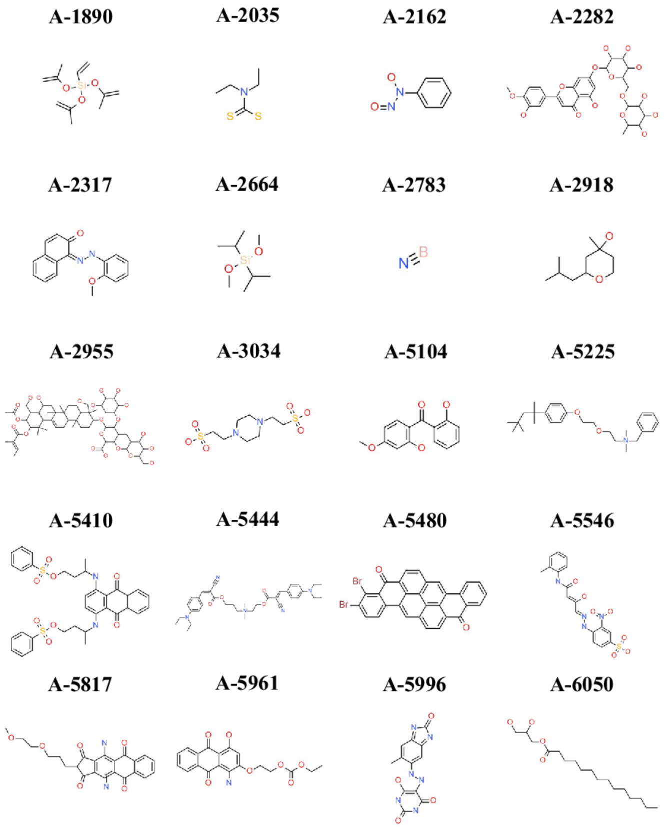


Fig. 11 Structures and compound ID from the 20 hardest-to-predict compounds from AqSolDBc. The first letter of the ID corresponds to the source of the entry (see Fig. 10).

to similar results. We also used OChem models (*LogPo/w*: *ALOGPS 2.1*, 2016; MP: *Best estate*, 2015) to predict *LogPo/w* and MP and used the computed values as input to the GSE. The ChemProp MPNN model is a Directed Message Passing Neural Network (D-MPNN) renowned for producing reliable predictive models of chemical properties. Finally, ChemProp was used alone and in consensus with AqSolPred.

The consensus prediction was conducted to improve the applicability of AqSolPred as it was trained with a version of AqSolDB lacking eChemPortal and EPI Suite. Following the guidelines shared by the authors, models were used as intended: the performances announced were retrieved. Models were applied to 7,463 compounds from OChem.

Applicability domain. We used Isolation Forest[86] models as AD to verify our hypothesis. The Isolation Forest method constructs an ensemble of trees for a given dataset. During the tree-building process, each tree is grown by recursively selecting a random feature and a random split value between the minimum and maximum values of the selected feature to partition the observations. Instances with short average path lengths within the trees are identified as outliers. The essence of the Isolation Forest algorithm lies in this random partitioning to identify outliers. The IsolationForest models were trained with AqSolDBc (MOE2D descriptors, n = 203) using scikit-learn[84] with an increasing contamination parameter, from 0.0 to 0.99.

The contamination parameter defines the expected proportion of outliers within the training set and is used by the Isolation Forest as a threshold to discriminate outliers from inliers. In other words, a contamination of 0 corresponds to a 100% coverage of the applicability domain (no molecule rejected) and a contamination of 1 corresponds to a 0% coverage of the applicability domain (all molecule rejected). OChem's set was applied to these models. The RMSE from the compounds within the AD was computed for each incrementation of the contamination Fig. 15.

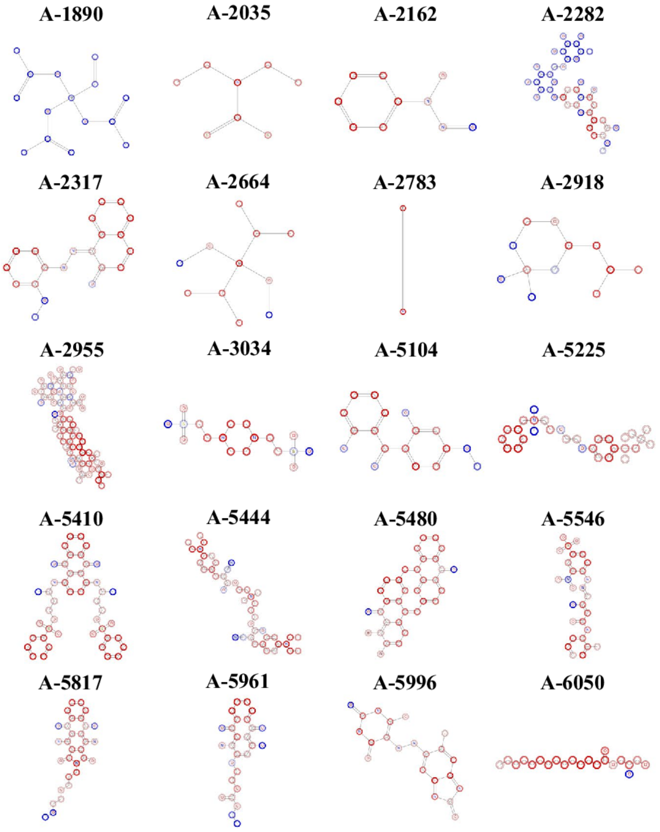


Fig. 12 Structures and compound ID from the 20 hardest-to-predict compounds colored using ColorAtom. Coloration of compounds according to the fragment-based RF model. Red and blue regions correspond, respectively, to negative and positive contributions to LogS. Dark colors correspond to large positive or negative atomic contributions.

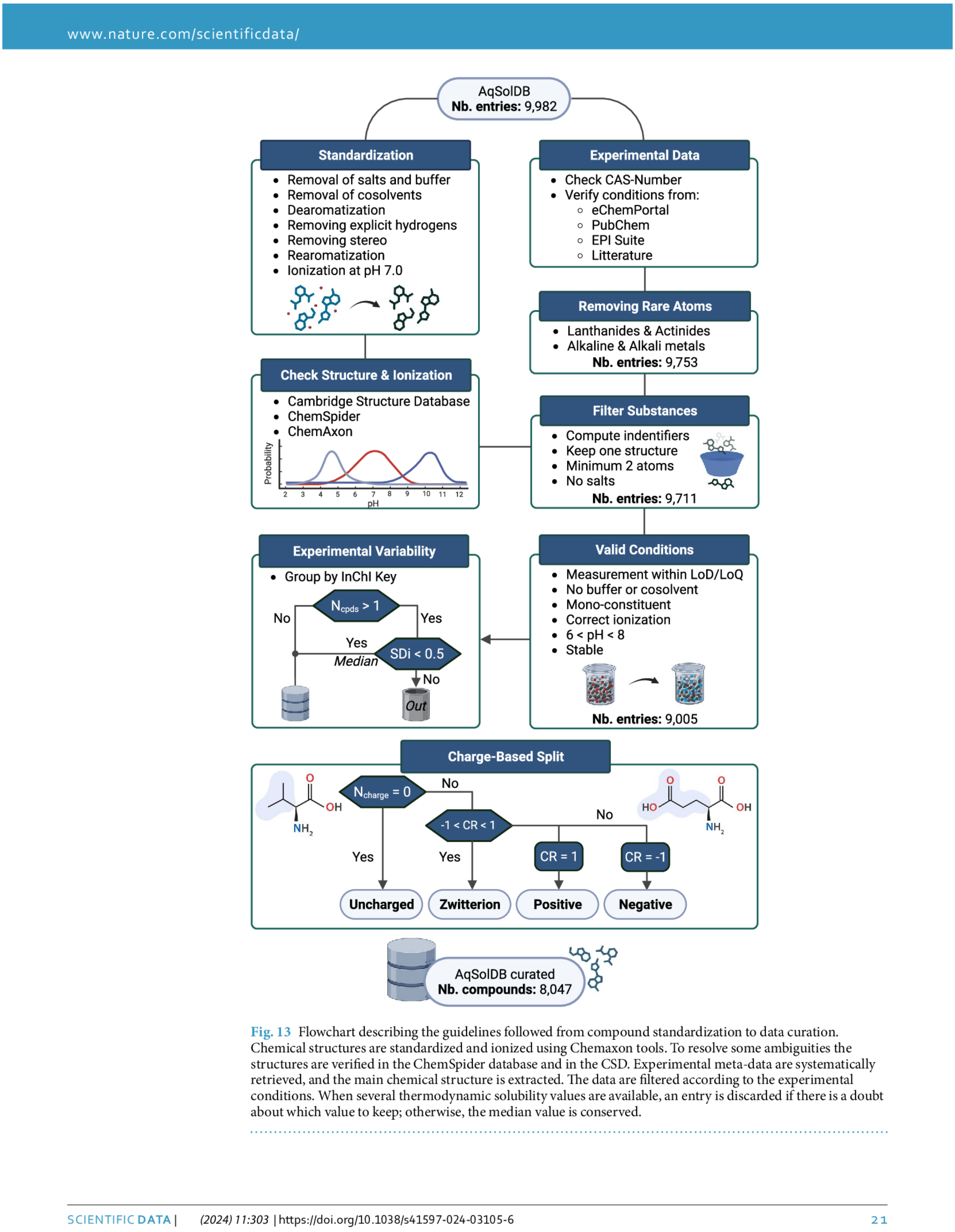
Data availability

The authors declare that the data supporting the findings of this study are available free of charge[6]. The repository features multiple datasets that have been curated for this research. The repository contains the following files:

File **AqSolDBc.csv**

Curated data from the AqSolDB. The available columns are:

- *ID* Compound ID (string)
- *InChI* InChI code of the chemical structure (string)
- *Solubility* Mole/L logarithm of the thermodynamic solubility in water at pH 7 (+/−1) at ~300 K (float)
- *SMILEScurated* Curated SMILES code of the chemical structure (string)
- *SD* Standard laboratory Deviation, default value: −1 (float)
- *Group* Data quality label imported from AqSolDB (string)
- *Dataset* Source of the data point (string)
- *Composition* Purity of the substance: mono-constituent, multi-constituent, UVCB (Categorical)
- *Error* Identifier error on the data point, default value: None (String)
- *Charge* Estimated formal charge of the compound at pH 7: Positive, Negative, Zwiterion, Uncharged (Categorical)



www.nature.com/scientificdata/

AqSolDB
Nb. entries: 9,982

Standardization

- Removal of salts and buffer
- Removal of cosolvents
- Dearomatization
- Removing explicit hydrogens
- Removing stereo
- Rearomatization
- Ionization at pH 7.0

Experimental Data

- Check CAS-Number
- Verify conditions from:
  - eChemPortal
  - PubChem
  - EPI Suite
  - Litterature

Removing Rare Atoms

- Lanthanides & Actinides
- Alkaline & Alkali metals

Nb. entries: 9,753

Check Structure & Ionization

- Cambridge Structure Database
- ChemSpider
- ChemAxon

Filter Substances

- Compute indentifiers
- Keep one structure
- Minimum 2 atoms
- No salts

Nb. entries: 9,711

Experimental Variability

- Group by InChI Key

Valid Conditions

- Measurement within LoD/LoQ
- No buffer or cosolvent
- Mono-constituent
- Correct ionization
- 6 < pH < 8
- Stable

Nb. entries: 9,005

Charge-Based Split

AqSolDB curated
Nb. compounds: 8,047

Fig. 13 Flowchart describing the guidelines followed from compound standardization to data curation. Chemical structures are standardized and ionized using Chemaxon tools. To resolve some ambiguities the structures are verified in the ChemSpider database and in the CSD. Experimental meta-data are systematically retrieved, and the main chemical structure is extracted. The data are filtered according to the experimental conditions. When several thermodynamic solubility values are available, an entry is discarded if there is a doubt about which value to keep; otherwise, the median value is conserved.

SCIENTIFIC DATA | (2024) 11:303 | https://doi.org/10.1038/s41597-024-03105-6 21

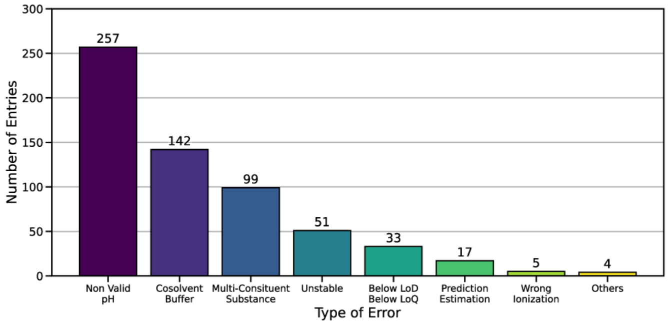


Fig. 14 Number of non-valid entries in AqSolDB identified with the help of the meta-data of measurement.

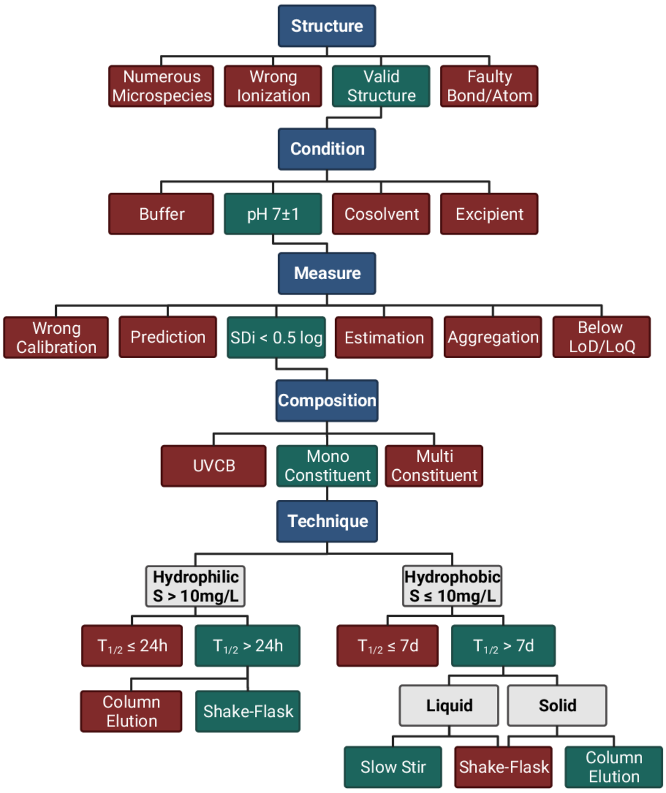


Fig. 15 Decision tree proposed for the curation of thermodynamic solubility data. Red nodes define non-valid conditions or chemical states, and green nodes account for correct entries.

File **OChemUnseen.csv**

Solubility data from OChem, curated and orthogonal to AqSolDB. The available columns are:

- *SMILES* Curated SMILES code of the chemical structure (string)

- *LogS* Mole/L logarithm of the thermodynamic solubility in water at pH 7 (+/−1) (float)

File **OChemOverlapping.csv**

Solubility data from OChem, curated; chemical structures are also present inside AqSolDB. The available columns are:

- *SMILES* Curated SMILES code of the chemical structure (string)
- *LogS* Mole/L logarithm of the thermodynamic solubility in water at pH 7 (+/−1) (float)

File **OChemCurated.csv**

Solubility data from OChem, curated. The available columns are:

- *ID* Compound ID (string)
- *Name* Compound name (string)
- *SMILES* Curated SMILES code of the chemical structure (string)
- *SDi* Standard laboratory Deviation, default value: −1 (float)
- *Reference* Unformated bibliographic reference which the data point is originating from (string)
- *LogS* Mole/L logarithm of the thermodynamic solubility in water at pH 7 (+/−1) (float)
- *EXTERNALID* Compound ID as appearing in its data source, default value: None (string)
- *CASRN* CAS number of the compound, default value: None (string)
- *ARTICLEID* Source ID linked to the column Reference (string)
- *Temperature* Temperature of the measure, in K (float)

Code availability

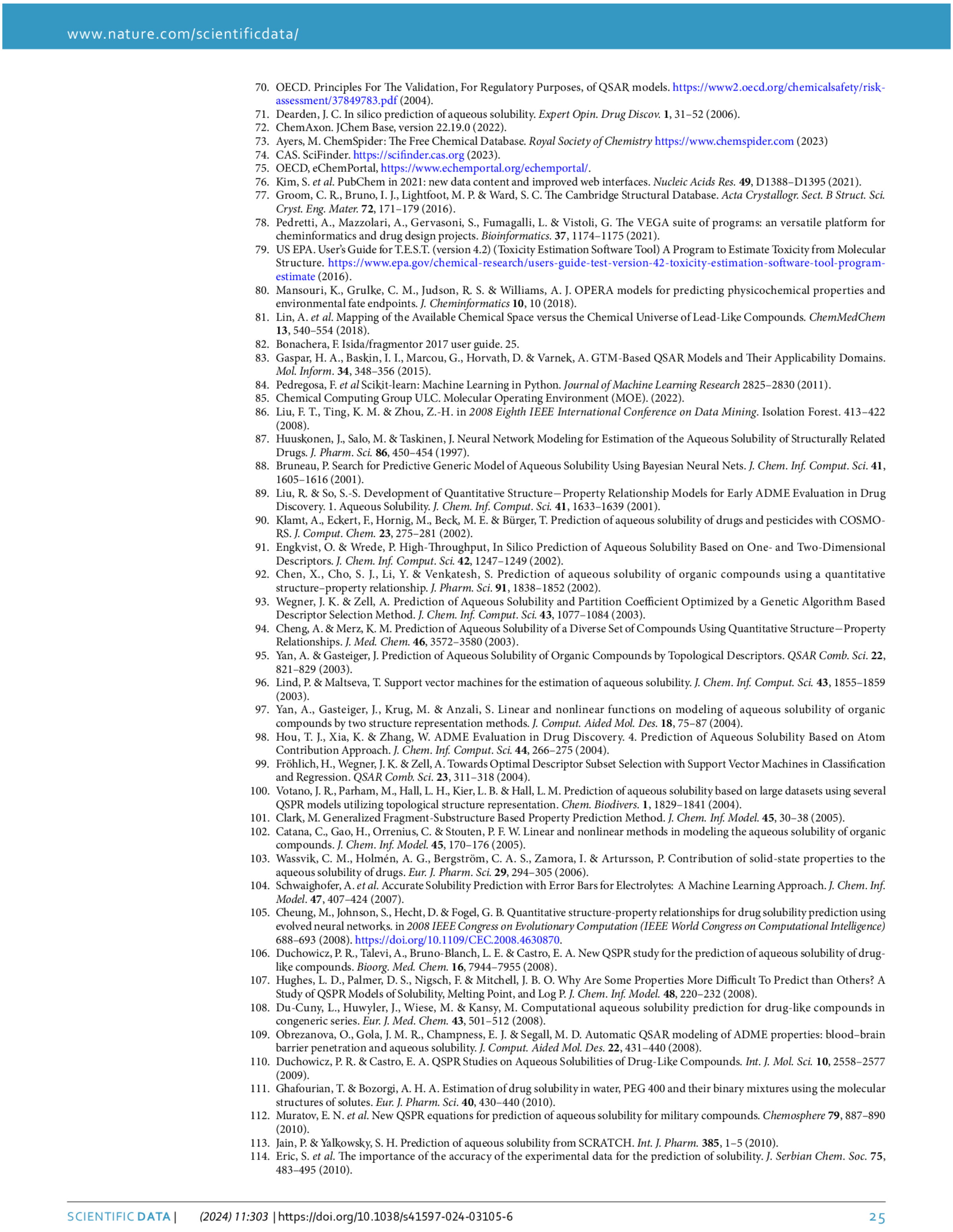
No custom code has been used.

Received: 1 September 2023; Accepted: 29 February 2024;
Published online: 18 March 2024

References

1. Kennedy, T. Managing the drug discovery/development interface. *Drug Discov. Today* **2**, 436–444 (1997).
2. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–716 (2004).
3. Millard, J., Alvarez-Núñez, F. & Yalkowsky, S. Solubilization by cosolvents. Establishing useful constants for the log-linear model. *Int. J. Pharm.* **245**, 153–166 (2002).
4. Jouyban, A. & Abolghassemi Fakhree, M. A. Solubility prediction methods for drug/drug like molecules. *Recent Pat. Chem. Eng.* **1**, 220–231 (2008).
5. van de Waterbeemd, H. Improving compound quality through in vitro and in silico physicochemical profiling. *Chem. Biodivers.* **6**, 1760–1766 (2009).
6. Llompart, P. *et al* Will we ever be able to accurately predict solubility? *Recherche Data Gouv* https://doi.org/10.57745/CZVZIA (2023)
7. Wang, J. & Hou, T. Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screen.* **14**, 328–338 (2011).
8. Elder, D. P., Holm, R. & Diego, H. L. Use of pharmaceutical salts and cocrystals to address the issue of poor solubility. *Int. J. Pharm.* **453**, 88–100 (2013). de.
9. Saal, C. & Petereit, A. C. Optimizing solubility: Kinetic versus thermodynamic solubility temptations and risks. *Eur. J. Pharm. Sci.* **47**, 589–595 (2012).
10. Wang, J. *et al*. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **47**, 1395–1404 (2007).
11. Johnson, S. R. & Zheng, W. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **8**, E27–E40 (2006).
12. Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discov. Today* **10**, 289–295 (2005).
13. OECD. Test No. 105: Water Solubility. *OECD Guidelines for the Testing of Chemicals, Section 1* https://read.oecd-ilibrary.org/environment/test-no-105-water-solubility_9789264069589-en (1995).
14. Llinàs, A., Glen, R. C. & Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **48**, 1289–1303 (2008).
15. Stuart, M. & Box, K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases. *Anal. Chem.* **77**, 983–990 (2005).
16. Huuskonen, J., Rantanen, J. & Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **35**, 1081–1088 (2000).
17. Yalkowsky, RM & Dannenfleser, SH. Aquasol database of aqueous solubility. Version 5. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/5348039 (2009).
18. Bloch, D. Computer Software Review. Review of PHYSPROP Database (Version 1.0). *ACS Publications* https://pubs.acs.org/doi/pdf/10.1021/ci00024a602 (2004) https://doi.org/10.1021/ci00024a602.
19. Dalanay, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).
20. US EPA. EPI Suite. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface
21. Wang, J., Hou, T. & Xu, X. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **49**, 571–581 (2009).
22. Boobier, S., Hose, D. R. J., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **11**, 5753 (2020).
23. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).
24. Avdeef, A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. *ADMET DMPK* **8**, 29 (2020).
25. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci. Data* **6**, 143 (2019).

26. Sushko, I. *et al*. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **25**, 533–554 (2011).
27. Panapitiya, G. *et al*. Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **7**, 15695–15710 (2022).
28. Wiercioch, M. & Kirchmair, J. Dealing with a data-limited regime: Combining transfer learning and transformer attention mechanism to increase aqueous solubility prediction performance. *Artif. Intell. Life Sci.* **1**, 100021 (2021).
29. Lowe, C. N. *et al.* Transparency in Modeling through Careful Application of OECD's QSAR/QSPR Principles via a Curated Water Solubility Data Set. *Chem. Res. Toxicol.* **36**, 465–478 (2023).
30. Francoeur, P. G. & Koes, D. R. SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction. *J. Chem. Inf. Model.* **61**, 2530–2536 (2021).
31. Sluga, J., Venko, K., Drgan, V. & Novič, M. QSPR Models for Prediction of Aqueous Solubility: Exploring the Potency of Randić-type Indices. *Croat. Chem. Acta* **93** (2020).
32. Meng, J. *et al.* Boosting the predictive performance with aqueous solubility dataset curation. *Sci. Data* **9**, 71 (2022).
33. Lee, S. *et al.* Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **7**, 12268–12277 (2022).
34. Schrödinger. QikProp. (2015).
35. United States National Library of Medicine. ChemIDplus advanced. https://pubchem.ncbi.nlm.nih.gov/source/ChemIDplus (2011).
36. Kühne, R., Ebert, R.-U., Kleint, F., Schmidt, G. & Schüürmann, G. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* **30**, 2061–2077 (1995).
37. OECD. eChemPortal: The Global Portal to Information on Chemical Substances, https://www.echemportal.org/echemportal/ (2023).
38. European Chemicals Agency. ECHA. https://echa.europa.eu/fr/ (2023).
39. Irmann, F. Eine einfache Korrelation zwischen Wasserlöslichkeit und Struktur von Kohlenwasserstoffen und Halogenkohlenwasserstoffen. *Chem. Ing. Tech.* **37**, 789–798 (1965).
40. Hansch, C., Quinlan, J. E. & Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **33**, 347–350 (1968).
41. Yalkowsky, S. H. & Valvani, S. C. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **69**, 912–922 (1980).
42. Ran, Y. & Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**, 354–357 (2001).
43. Hansen, N. T., Kouskoumvekaki, I., Jørgensen, F. S., Brunak, S. & Jónsdóttir, S. Ó. Prediction of pH-Dependent Aqueous Solubility of Druglike Molecules. *J. Chem. Inf. Model.* **46**, 2601–2609 (2006).
44. ChemAxon. Marvin. https://chemaxon.com/products/marvin (2023).
45. Johnson, S. R., Chen, X.-Q., Murphy, D. & Gudmundsson, O. A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects. *Mol. Pharm.* **4**, 513–523 (2007).
46. Hopfinger, A. J., Esposito, E. X., Llinàs, A., Glen, R. C. & Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *ACS Publications* https://pubs.acs.org/doi/pdf/10.1021/ci800436c (2008).
47. Lusci, A., Pollastri, G. & Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53**, 1563–1575 (2013).
48. Huuskonen, J., Livingstone, D. J. & Manallack, D. T. Prediction of drug solubility from molecular structure using a drug-like training set. *SAR QSAR Environ. Res.* **19**, 191–212 (2008).
49. Zhou, D., Alelyunas, Y. & Liu, R. Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility. *J. Chem. Inf. Model.* **48**, 981–987 (2008).
50. Erić, S., Kalinić, M., Popović, A., Zloh, M. & Kuzmanovski, I. Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks. *Int. J. Pharm.* **437**, 232–241 (2012).
51. Llinas, A. & Avdeef, A. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD $\sim$ 0.17 log) and Loose (SD $\sim$ 0.62 log) Test Sets. *J. Chem. Inf. Model.* **59**, 3036–3040 (2019).
52. Llinas, A., Oprisiu, I. & Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **60**, 4791–4803 (2020).
53. Hewitt, M. *et al.* In silico prediction of aqueous solubility: the solubility challenge. *J. Chem. Inf. Model.* **49**, 2572–2587 (2009).
54. Goh, G. B., Hodas, N., Siegel, C. & Vishnu, A. SMILES2vec: Predicting Chemical Properties from Text Representations. Preprint at arXiv:1712.02034 (2018).
55. Cui, Q. *et al.* Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **10** (2020).
56. Maziarka, Ł. et al. *Molecule Attention Transformer*. (2020).
57. Lovrić, M. *et al.* Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *J. Chemom.* **35**, e3349 (2021).
58. Kohavi, R. & Wolpert, D. H. in *International Conference on Machine Learning* Bias Plus Variance Decomposition for Zero-One Loss Function (1996).
59. Dwork, C. *et al.* The reusable holdout: Preserving validity in adaptive data analysis. *Science* **349**, 636–638 (2015).
60. Breiman, L. & Spector, P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int. Stat. Rev. Rev. Int. Stat.* **60**, 291–319 (1992).
61. Rao, R. B., Fung, G. & Rosales, R. in *Proceedings of the 2008 SIAM International Conference on Data Mining (SDM)* On the Dangers of Cross-Validation. An Experimental Evaluation. 588–596 (Society for Industrial and Applied Mathematics, 2008).
62. Rytting, E., Lentz, K. A., Chen, X. Q., Qian, F. & Vakatesh S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS J.* **7**, E78–105, https://doi.org/10.1208/aapsj070110 (2005).
63. Heid, E. *et al.* Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **64**, 9–17, https://doi.org/10.1021/acs.jcim.3c01250 (2024).
64. Chevillard, F. *et al.* In Silico Prediction of Aqueous Solubility: A Multimodel Protocol Based on Chemical Similarity. *Mol. Pharm.* **9**, 3127–3135 (2012).
65. Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Chen, X. & Li, H.-D. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J. Chemometrics.* **24**, 584–595 (2010).
66. Ruggiu, F., Marcou, G., Varnek, A. & Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **29**, 855–868 (2010).
67. Ferguson, A. L., Debenedetti, P. G. & Panagiotopoulos, A. Z. Solubility and Molecular Conformations of n-Alkane Chains in Water. *J. Phys. Chem. B* **113**, 6405–6414 (2009).
68. Birch, H., Redman, A. D., Letinski, D. J., Lyon, D. Y. & Mayer, P. Determining the water solubility of difficult-to-test substances: A tutorial review. *Anal. Chim. Acta* **1086**, 16–28 (2019).
69. Marcou, G., Horvath, D. & Solov, V. Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions. *Mol Inf.*



70. OECD. Principles For The Validation, For Regulatory Purposes, of QSAR models. https://www2.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (2004).
71. Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discov.* **1**, 31–52 (2006).
72. ChemAxon. JChem Base, version 22.19.0 (2022).
73. Ayers, M. ChemSpider: The Free Chemical Database. *Royal Society of Chemistry* https://www.chemspider.com (2023)
74. CAS. SciFinder. https://scifinder.cas.org (2023).
75. OECD, eChemPortal, https://www.echemportal.org/echemportal/.
76. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
77. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).
78. Pedretti, A., Mazzolari, A., Gervasoni, S., Fumagalli, L. & Vistoli, G. The VEGA suite of programs: an versatile platform for cheminformatics and drug design projects. *Bioinformatics.* **37**, 1174–1175 (2021).
79. US EPA. User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool) A Program to Estimate Toxicity from Molecular Structure. https://www.epa.gov/chemical-research/users-guide-test-version-42-toxicity-estimation-software-tool-program-estimate (2016).
80. Mansouri, K., Grulke, C. M., Judson, R. S. & Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics* **10**, 10 (2018).
81. Lin, A. *et al.* Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **13**, 540–554 (2018).
82. Bonachera, F. Isida/fragmentor 2017 user guide. 25.
83. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **34**, 348–356 (2015).
84. Pedregosa, F. *et al* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2825–2830 (2011).
85. Chemical Computing Group ULC. Molecular Operating Environment (MOE). (2022).
86. Liu, F. T., Ting, K. M. & Zhou, Z.-H. in *2008 Eighth IEEE International Conference on Data Mining*. Isolation Forest. 413–422 (2008).
87. Huuskonen, J., Salo, M. & Taskinen, J. Neural Network Modeling for Estimation of the Aqueous Solubility of Structurally Related Drugs. *J. Pharm. Sci.* **86**, 450–454 (1997).
88. Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **41**, 1605–1616 (2001).
89. Liu, R. & So, S.-S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **41**, 1633–1639 (2001).
90. Klamt, A., Eckert, F., Hornig, M., Beck, M. E. & Bürger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **23**, 275–281 (2002).
91. Engkvist, O. & Wrede, P. High-Throughput, In Silico Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **42**, 1247–1249 (2002).
92. Chen, X., Cho, S. J., Li, Y. & Venkatesh, S. Prediction of aqueous solubility of organic compounds using a quantitative structure–property relationship. *J. Pharm. Sci.* **91**, 1838–1852 (2002).
93. Wegner, J. K. & Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **43**, 1077–1084 (2003).
94. Cheng, A. & Merz, K. M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure−Property Relationships. *J. Med. Chem.* **46**, 3572–3580 (2003).
95. Yan, A. & Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds by Topological Descriptors. *QSAR Comb. Sci.* **22**, 821–829 (2003).
96. Lind, P. & Maltseva, T. Support vector machines for the estimation of aqueous solubility. *J. Chem. Inf. Comput. Sci.* **43**, 1855–1859 (2003).
97. Yan, A., Gasteiger, J., Krug, M. & Anzali, S. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput. Aided Mol. Des.* **18**, 75–87 (2004).
98. Hou, T. J., Xia, K. & Zhang, W. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **44**, 266–275 (2004).
99. Fröhlich, H., Wegner, J. K. & Zell, A. Towards Optimal Descriptor Subset Selection with Support Vector Machines in Classification and Regression. *QSAR Comb. Sci.* **23**, 311–318 (2004).
100. Votano, J. R., Parham, M., Hall, L. H., Kier, L. B. & Hall, L. M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodivers.* **1**, 1829–1841 (2004).
101. Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **45**, 30–38 (2005).
102. Catana, C., Gao, H., Orrenius, C. & Stouten, P. F. W. Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J. Chem. Inf. Model.* **45**, 170–176 (2005).
103. Wassvik, C. M., Holmén, A. G., Bergström, C. A. S., Zamora, I. & Artursson, P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur. J. Pharm. Sci.* **29**, 294–305 (2006).
104. Schwaighofer, A. *et al.* Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. *J. Chem. Inf. Model.* **47**, 407–424 (2007).
105. Cheung, M., Johnson, S., Hecht, D. & Fogel, G. B. Quantitative structure-property relationships for drug solubility prediction using evolved neural networks. in *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)* 688–693 (2008). https://doi.org/10.1109/CEC.2008.4630870.
106. Duchowicz, P. R., Talevi, A., Bruno-Blanch, L. E. & Castro, E. A. New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorg. Med. Chem.* **16**, 7944–7955 (2008).
107. Hughes, L. D., Palmer, D. S., Nigsch, F. & Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **48**, 220–232 (2008).
108. Du-Cuny, L., Huwyler, J., Wiese, M. & Kansy, M. Computational aqueous solubility prediction for drug-like compounds in congeneric series. *Eur. J. Med. Chem.* **43**, 501–512 (2008).
109. Obrezanova, O., Gola, J. M. R., Champness, E. J. & Segall, M. D. Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des.* **22**, 431–440 (2008).
110. Duchowicz, P. R. & Castro, E. A. QSPR Studies on Aqueous Solubilities of Drug-Like Compounds. *Int. J. Mol. Sci.* **10**, 2558–2577 (2009).
111. Ghafourian, T. & Bozorgi, A. H. A. Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes. *Eur. J. Pharm. Sci.* **40**, 430–440 (2010).
112. Muratov, E. N. *et al.* New QSPR equations for prediction of aqueous solubility for military compounds. *Chemosphere* **79**, 887–890 (2010).
113. Jain, P. & Yalkowsky, S. H. Prediction of aqueous solubility from SCRATCH. *Int. J. Pharm.* **385**, 1–5 (2010).
114. Eric, S. *et al.* The importance of the accuracy of the experimental data for the prediction of solubility. *J. Serbian Chem. Soc.* **75**, 483–495 (2010).

115. Louis, B., Agrawal, V. K. & Khadikar, P. V. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. *Eur. J. Med. Chem.* **45**, 4018–4025 (2010).
116. Fatemi, M., Heidari, A. & Ghorbanzadeh, M. Prediction of Aqueous Solubility of Drug-Like Compounds by Using an Artificial Neural Network and Least-Squares Support Vector Machine. *Bull. Chem. Soc. Jpn.* **83**, 1338–1345 (2010).
117. Salahinejad, M., Le, T. C. & Winkler, D. A. Aqueous solubility prediction: do crystal lattice interactions help? *Mol. Pharm.* **10**, 2757–2766 (2013).
118. McDonagh, J. L., Nath, N., De Ferrari, L., van Mourik, T. & Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **54**, 844–856 (2014).
119. Kim, S., Jinich, A. & Aspuru-Guzik, A. MultiDK: A Multiple Descriptor Multiple Kernel Approach for Molecular Discovery and Its Application to Organic Flow Battery Electrolytes. *J. Chem. Inf. Model.* **57**, 657–668 (2017).
120. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **57**, 1757–1772 (2017).
121. Cho, H. & Choi, I. S. Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology. *ChemMedChem* **14**, 1604–1609 (2019).
122. Cho, H. & Choi, I. S. Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology. *Chem Med Chem* **14**, 1604 (2019).
123. Deng, T. & Jia, G. Prediction of aqueous solubility of compounds based on neural network. *Mol. Phys.* **118**, e1600754 (2020).
124. Gao, P., Zhang, J., Sun, Y. & Yu, J. Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architectures. *Phys. Chem. Chem. Phys.* **22**, 23766–23772 (2020).
125. Falcón-Cano, G., Molina, C. & Cabrera-Pérez, M. A. ADME prediction with KNIME: In silico aqueous solubility consensus model based on supervised recursive random forest approaches. *ADMET DMPK* **8**, 251–273 (2020).
126. Shen, W. X. *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat Mach Intell* **3**, 334–343 (2021).
127. Tosca, E. M., Bartolucci, R. & Magni, P. Application of Artificial Neural Networks to Predict the Intrinsic Solubility of Drug-Like Molecules. *Pharmaceutics* **13**, 1101 (2021).
128. Wieder, O. *et al.* Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks. *Molecules* **26**, 6185 (2021).
129. Chen, J.-H. & Tseng, Y. J. Different molecular enumeration influences in deep learning: an example using aqueous solubility. *Briefings Bioinf* **22**, bbaa092 (2021).
130. Panapitiya, G. *et al.* Predicting Aqueous Solubility of Organic Molecules Using Deep Learning Models with Varied Molecular Representations. *ACS Omega* **7**, 15695–15710 (2022).
131. Hou, Y., Wang, S., Bai, B., Chan, H. C. S. & Yuan, S. Accurate Physical Property Predictions via Deep Learning. *Molecules* **27**, 1668 (2022).
132. Raevsky, O. A., Grigor'ev, V. Y., Polianczyk, D. E., Raevskaja, O. E. & Dearden, J. C. Calculation of aqueous solubility of crystalline un-ionized organic chemicals and drugs based on structural similarity and physicochemical descriptors. *J Chem Inf Model.* **54**, 683–91, https://doi.org/10.1021/ci400692n (2014).
133. Schaper, K.-J., Kunz, B. & Raevsky, O. Analysis of water solubility data on the basis of HYBOT descriptors. Part 2. *QSAR Comb. Sci.* **22**, 943–958, https://doi.org/10.1002/qsar.200330840 (2003).

Author contributions

P.L. is the main author. Data collection, annotation process supervision, modeling and statistical analysis of results were carried out by P.L., C.M. and G.M. Figures and tables preparation by P.L. and G.M. Kinetic data contributed by S.B. Supervision by C.M., G.M. and A.V. The first version of this article was written by P.L. and G.M.; G.M., D.H., C.M. and A.V. led the subsequent revisions.

Competing interests

C. Minoletti and P. Llompart are Sanofi employees and may hold shares and/or stock options in the company. S. Baybekov, D. Horvath, G. Marcou, and A. Varnek have nothing to disclose.

Additional information

Correspondence and requests for materials should be addressed to G.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Outline

In this study we demonstrate that some widely used models fail to deliver robust predictions when applied to new data. To address these issues, we propose a workflow for curating high-quality aqueous solubility datasets and improving predictive reliability. Our findings underscore the need for rigorous dataset validation and highlight the impact of factors such as interlaboratory variability, ionic states, and data provenance. The curated datasets and trained models resulting from this study are made publicly available to facilitate further improvements in solubility modeling.

This approach was applied in collaboration with Sanofi Frankfurt for the design and selection of soluble phosphines, leading to the successful proposition of novel water-soluble phosphines (**Figure 18**). The collaboration resulted in a publication[147], further validating the effectiveness of our methodology in real-world drug discovery applications.

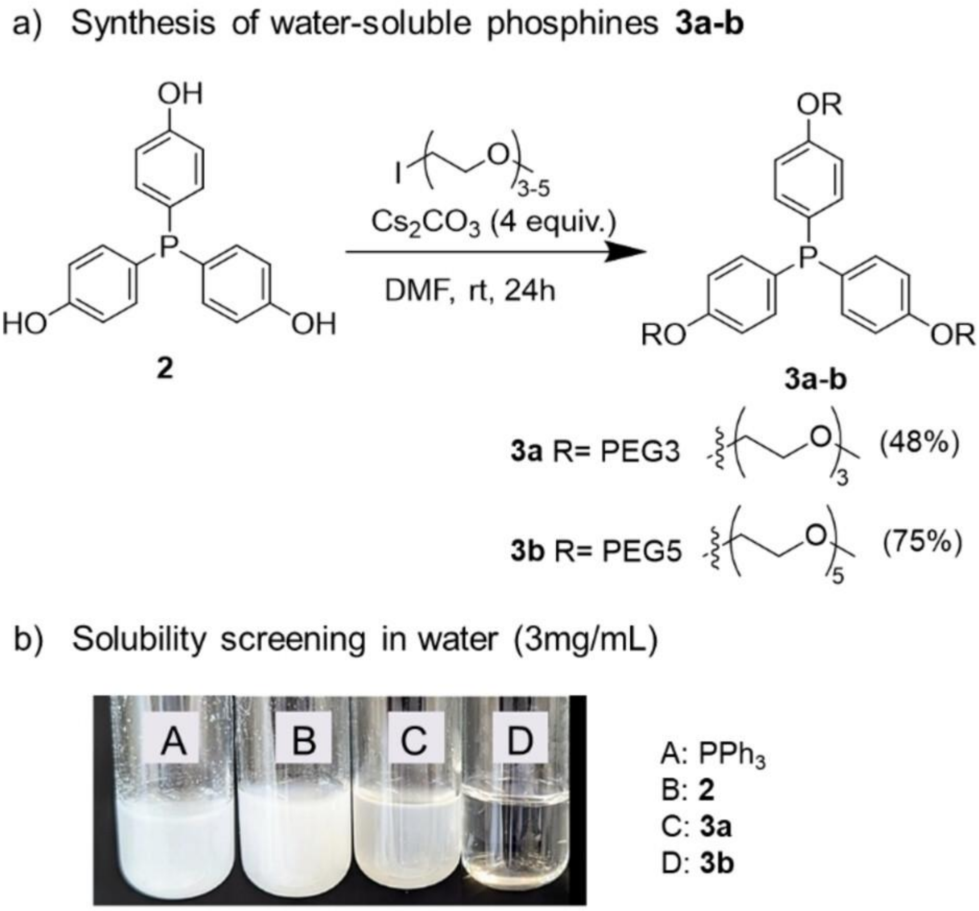


Figure 18: ***Preparation of pegylated phosphines.*** **(a)** Synthesis of water-soluble pegylated phosphines. **(b)** Solubility of different triphenylphosphines in water (3 mg/mL).

4.2. Kinetic Solubility

Introduction

Machine learning has demonstrated potential in predicting solubility, but its application to kinetic solubility remains underexplored. Many existing models focus on thermodynamic solubility, assuming it can serve as a proxy for kinetic solubility, yet recent studies highlight their lack of correlation. Kinetic solubility, often used in early drug discovery, is known for its variability due to differences in experimental setups, solvent residues, and pH control. These factors have contributed to the perception that kinetic solubility is less reproducible, deterring efforts to develop predictive models. In this section, we investigate the acquisition, reproducibility and modelisability of kinetic solubility assays (**Figure 19**). We analyze large kinetic solubility datasets, compare their inter-laboratory consistency, and benchmark machine learning models trained on these datasets.

Main Terminology

Nominal Concentration is the predefined concentration of a compound in solution at which kinetic solubility is measured. It serves as an upper limit for kinetic solubility values in assays.

Precipitation onset is the point at which a compound begins to precipitate out of solution due to exceeding its kinetic solubility threshold, influenced by solvent and pH.

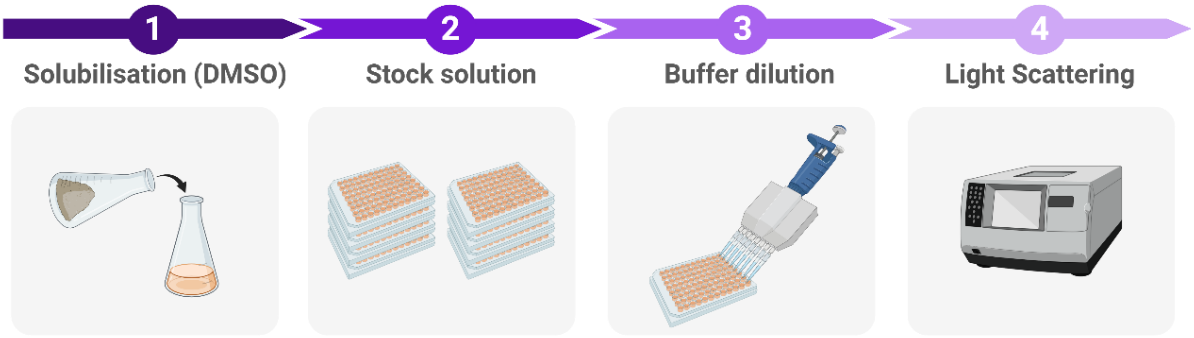


Figure 19: ***Kinetic solubility measurement.*** The process begins with dilution of a compound from an organic stock solution (e.g., DMSO) into a buffer, potentially causing precipitation. After incubation, the solution is filtered or centrifuged, and the dissolved fraction is quantified. This method estimates solubility before equilibrium is reached, making it particularly useful for HTS.

Received: 22 August 2023 | Revised: 25 November 2023 | Accepted: 23 December 2023

DOI: 10.1002/minf.202300216

RESEARCH ARTICLE

molecular informatics

Kinetic solubility: Experimental and machine-learning modeling perspectives

Shamkhal Baybekov[1] | Pierre Llompart[1, 2] | Gilles Marcou[1] | Patrick Gizzi[4] | Jean-Luc Galzi[3, 5] | Pascal Ramos[6] | Olivier Saurel[6] | Claire Bourban[4] | Claire Minoletti[2] | Alexandre Varnek[1]

[1]Laboratoire de Chémoinformatique UMR 7140 CNRS, Institut Le Bel, University of Strasbourg, Strasbourg, France

[2]IDD/CADD, Sanofi, Vitry-Sur-Seine, France

[3]Biotechnologie et signalisation cellulaire UMR 7242 CNRS, École supérieure de biotechnologie de Strasbourg, University of Strasbourg, Illkirch, France

[4]Plateforme de Chimie Biologique Intégrative de Strasbourg UAR 3286 CNRS, University of Strasbourg, Illkirch, France

[5]ChemBioFrance – Chimiothèque Nationale UAR 3035, ENSCM – 240, Montpellier Cedex 5, France

[6]Institut de Pharmacologie et de Biologie Structurale (IPBS), Université de Toulouse, CNRS, Université Toulouse III – Paul Sabatier (UT3), Toulouse, France

Correspondence
Alexandre Varnek, Laboratoire de Chémoinformatique UMR 7140 CNRS, Institut Le Bel, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France.
Email: varnek@unistra.fr

Funding information
French National Research Agency, Grant/Award Number: CSC-IGS ANR-17-EURE-0016

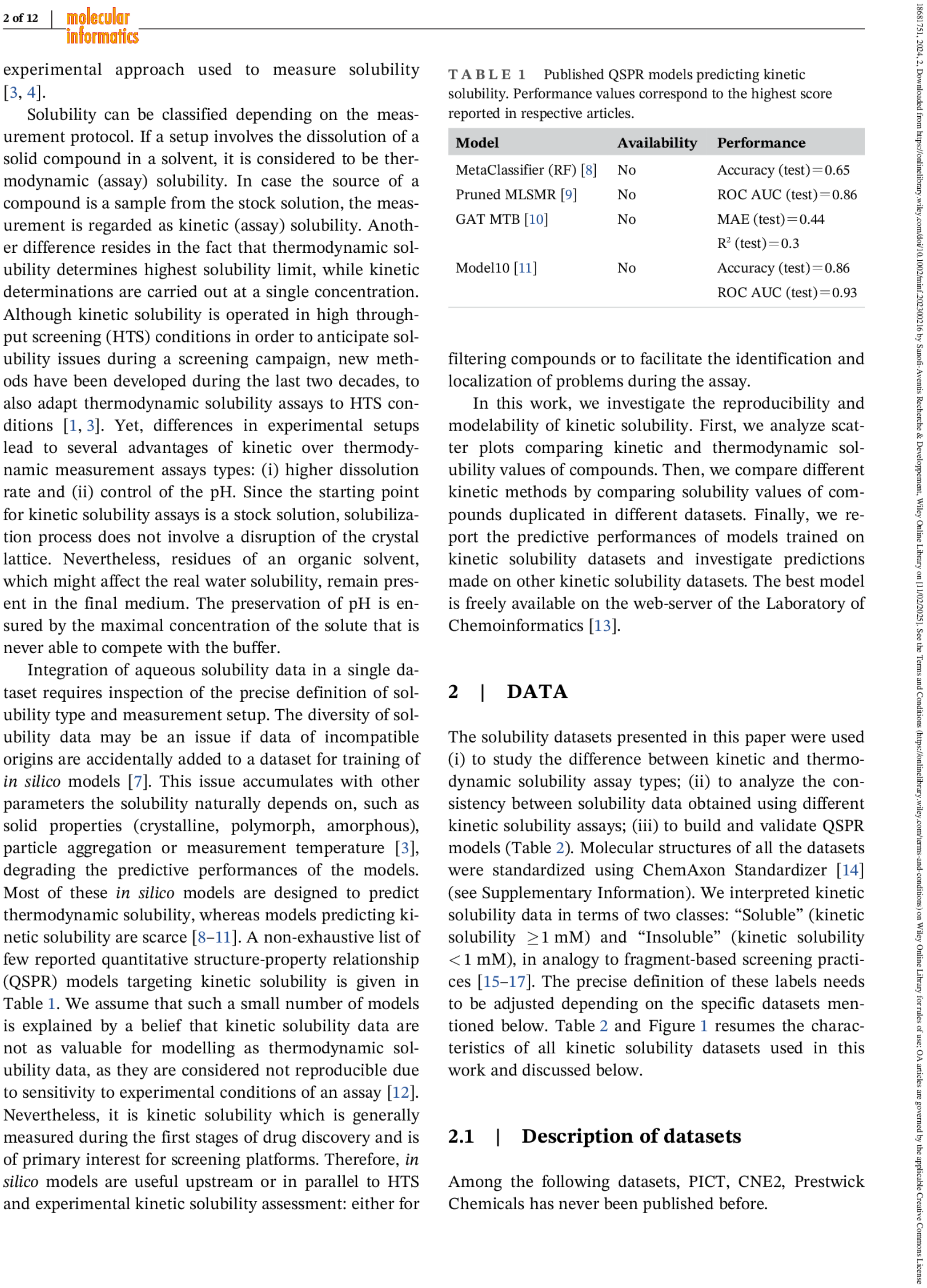
Abstract

Kinetic aqueous or buffer solubility is important parameter measuring suitability of compounds for high throughput assays in early drug discovery while thermodynamic solubility is reserved for later stages of drug discovery and development. Kinetic solubility is also considered to have low inter-laboratory reproducibility because of its sensitivity to protocol parameters [1]. Presumably, this is why little efforts have been put to build QSPR models for kinetic in comparison to thermodynamic aqueous solubility. Here, we investigate the reproducibility and modelability of kinetic solubility assays. We first analyzed the relationship between kinetic and thermodynamic solubility data, and then examined the consistency of data from different kinetic assays. In this contribution, we report differences between kinetic and thermodynamic solubility data that are consistent with those reported by others [1,2] and good agreement between data from different kinetic solubility campaigns in contrast to general expectations. The latter is confirmed by achieving high performing QSPR models trained on merged kinetic solubility datasets. The poor performance of QSPR model trained on thermodynamic solubility when applied to kinetic solubility dataset reinforces the conclusion that kinetic and thermodynamic solubilities do not correlate: one cannot be used as an ersatz for the other. This encourages for building predictive models for kinetic solubility. The kinetic solubility QSPR model developed in this study is freely accessible through the Predictor web service of the Laboratory of Chemoinformatics (https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi).

KEYWORDS
comparison, kinetic solubility, QSPR, thermodynamic solubility

1 | INTRODUCTION

Aqueous solubility is an essential property of a compound to be measured in drug discovery and development [3, 4]. It is a parameter to assess the bioavailability of a compound and it is important to avoid bias on the measurement of a bioactivity, such as a masking effect – i. e. when the saturation of an assay is due to the solubility limit of a compound and not to the biological material tested [5, 6]. Different steps of drug discovery and development focus on different aspects of solubility, which in turn dictates the choice of



experimental approach used to measure solubility [3, 4].

Solubility can be classified depending on the measurement protocol. If a setup involves the dissolution of a solid compound in a solvent, it is considered to be thermodynamic (assay) solubility. In case the source of a compound is a sample from the stock solution, the measurement is regarded as kinetic (assay) solubility. Another difference resides in the fact that thermodynamic solubility determines highest solubility limit, while kinetic determinations are carried out at a single concentration. Although kinetic solubility is operated in high throughput screening (HTS) conditions in order to anticipate solubility issues during a screening campaign, new methods have been developed during the last two decades, to also adapt thermodynamic solubility assays to HTS conditions [1, 3]. Yet, differences in experimental setups lead to several advantages of kinetic over thermodynamic measurement assays types: (i) higher dissolution rate and (ii) control of the pH. Since the starting point for kinetic solubility assays is a stock solution, solubilization process does not involve a disruption of the crystal lattice. Nevertheless, residues of an organic solvent, which might affect the real water solubility, remain present in the final medium. The preservation of pH is ensured by the maximal concentration of the solute that is never able to compete with the buffer.

Integration of aqueous solubility data in a single dataset requires inspection of the precise definition of solubility type and measurement setup. The diversity of solubility data may be an issue if data of incompatible origins are accidentally added to a dataset for training of *in silico* models [7]. This issue accumulates with other parameters the solubility naturally depends on, such as solid properties (crystalline, polymorph, amorphous), particle aggregation or measurement temperature [3], degrading the predictive performances of the models. Most of these *in silico* models are designed to predict thermodynamic solubility, whereas models predicting kinetic solubility are scarce [8–11]. A non-exhaustive list of few reported quantitative structure-property relationship (QSPR) models targeting kinetic solubility is given in Table 1. We assume that such a small number of models is explained by a belief that kinetic solubility data are not as valuable for modelling as thermodynamic solubility data, as they are considered not reproducible due to sensitivity to experimental conditions of an assay [12]. Nevertheless, it is kinetic solubility which is generally measured during the first stages of drug discovery and is of primary interest for screening platforms. Therefore, *in silico* models are useful upstream or in parallel to HTS and experimental kinetic solubility assessment: either for filtering compounds or to facilitate the identification and localization of problems during the assay.

TABLE 1 Published QSPR models predicting kinetic solubility. Performance values correspond to the highest score reported in respective articles.

Model	Availability	Performance
MetaClassifier (RF) [8]	No	Accuracy (test) = 0.65
Pruned MLSMR [9]	No	ROC AUC (test) = 0.86
GAT MTB [10]	No	MAE (test) = 0.44
		R^2 (test) = 0.3
Model10 [11]	No	Accuracy (test) = 0.86
		ROC AUC (test) = 0.93

In this work, we investigate the reproducibility and modelability of kinetic solubility. First, we analyze scatter plots comparing kinetic and thermodynamic solubility values of compounds. Then, we compare different kinetic methods by comparing solubility values of compounds duplicated in different datasets. Finally, we report the predictive performances of models trained on kinetic solubility datasets and investigate predictions made on other kinetic solubility datasets. The best model is freely available on the web-server of the Laboratory of Chemoinformatics [13].

2 | DATA

The solubility datasets presented in this paper were used (i) to study the difference between kinetic and thermodynamic solubility assay types; (ii) to analyze the consistency between solubility data obtained using different kinetic solubility assays; (iii) to build and validate QSPR models (Table 2). Molecular structures of all the datasets were standardized using ChemAxon Standardizer [14] (see Supplementary Information). We interpreted kinetic solubility data in terms of two classes: "Soluble" (kinetic solubility ≥1 mM) and "Insoluble" (kinetic solubility <1 mM), in analogy to fragment-based screening practices [15–17]. The precise definition of these labels needs to be adjusted depending on the specific datasets mentioned below. Table 2 and Figure 1 resumes the characteristics of all kinetic solubility datasets used in this work and discussed below.

2.1 | Description of datasets

Among the following datasets, PICT, CNE2, Prestwick Chemicals has never been published before.

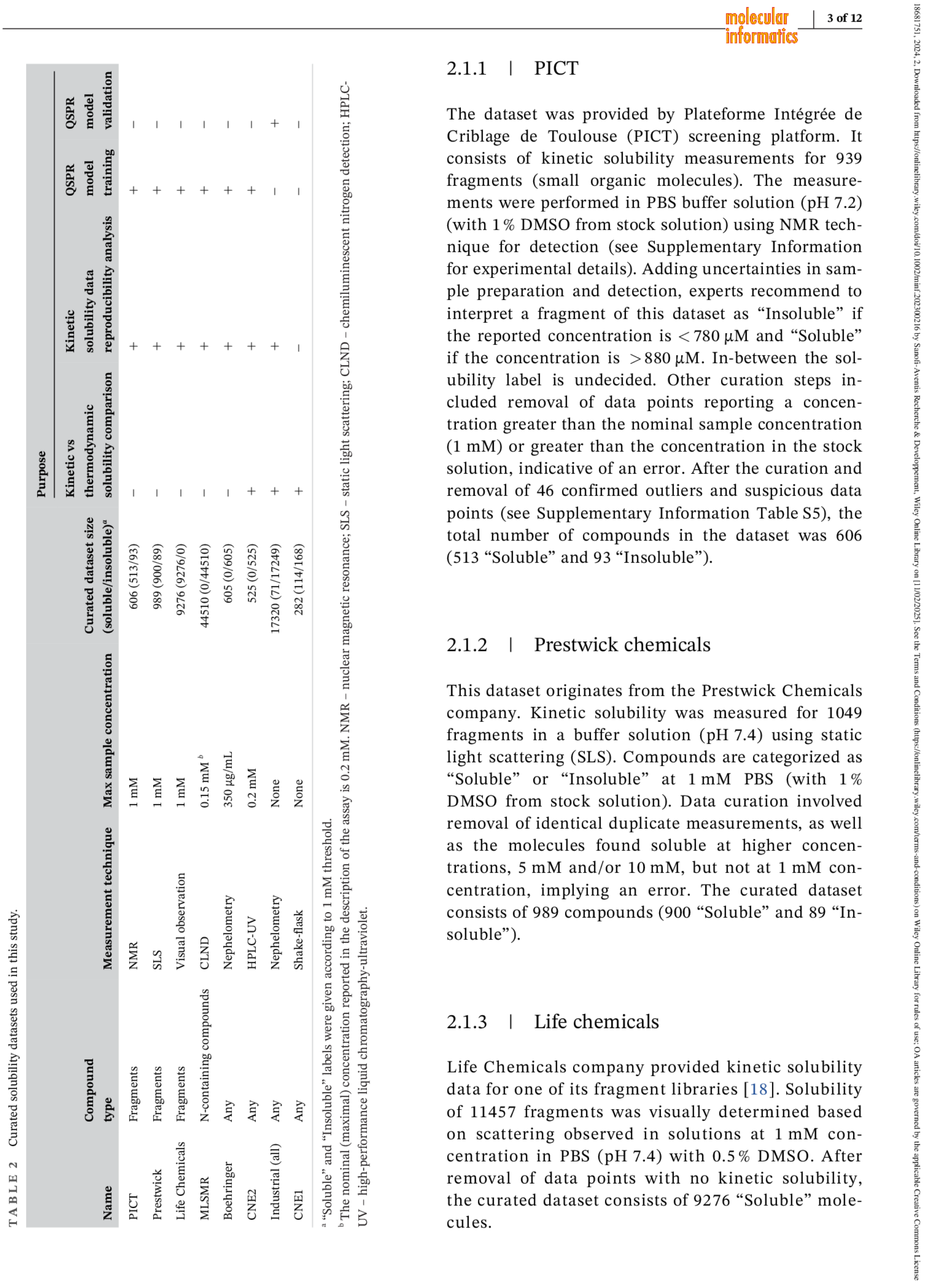


TABLE 2 Curated solubility datasets used in this study.

Name	Compound type	Measurement technique	Max sample concentration	Curated dataset size (soluble/insoluble)[a]	Purpose Kinetic vs thermodynamic solubility comparison	Kinetic solubility data reproducibility analysis	QSPR model training	QSPR model validation
PICT	Fragments	NMR	1 mM	606 (513/93)	–	+	+	–
Prestwick	Fragments	SLS	1 mM	989 (900/89)	–	+	+	–
Life Chemicals	Fragments	Visual observation	1 mM	9276 (9276/0)	–	+	+	–
MLSMR	N-containing compounds	CLND	0.15 mM[b]	44510 (0/44510)	–	+	+	–
Boehringer	Any	Nephelometry	350 μg/mL	605 (0/605)	–	+	+	–
CNE2	Any	HPLC-UV	0.2 mM	525 (0/525)	+	+	+	–
Industrial (all)	Any	Nephelometry	None	17320 (71/17249)	+	+	–	+
CNE1	Any	Shake-flask	None	282 (114/168)	+	–	–	–

[a] "Soluble" and "Insoluble" labels were given according to 1 mM threshold.
[b] The nominal (maximal) concentration reported in the description of the assay is 0.2 mM. NMR – nuclear magnetic resonance; SLS – static light scattering; CLND – chemiluminescent nitrogen detection; HPLC-UV – high-performance liquid chromatography-ultraviolet.

2.1.1 | PICT

The dataset was provided by Plateforme Intégrée de Criblage de Toulouse (PICT) screening platform. It consists of kinetic solubility measurements for 939 fragments (small organic molecules). The measurements were performed in PBS buffer solution (pH 7.2) (with 1 % DMSO from stock solution) using NMR technique for detection (see Supplementary Information for experimental details). Adding uncertainties in sample preparation and detection, experts recommend to interpret a fragment of this dataset as "Insoluble" if the reported concentration is < 780 μM and "Soluble" if the concentration is > 880 μM. In-between the solubility label is undecided. Other curation steps included removal of data points reporting a concentration greater than the nominal sample concentration (1 mM) or greater than the concentration in the stock solution, indicative of an error. After the curation and removal of 46 confirmed outliers and suspicious data points (see Supplementary Information Table S5), the total number of compounds in the dataset was 606 (513 "Soluble" and 93 "Insoluble").

2.1.2 | Prestwick chemicals

This dataset originates from the Prestwick Chemicals company. Kinetic solubility was measured for 1049 fragments in a buffer solution (pH 7.4) using static light scattering (SLS). Compounds are categorized as "Soluble" or "Insoluble" at 1 mM PBS (with 1 % DMSO from stock solution). Data curation involved removal of identical duplicate measurements, as well as the molecules found soluble at higher concentrations, 5 mM and/or 10 mM, but not at 1 mM concentration, implying an error. The curated dataset consists of 989 compounds (900 "Soluble" and 89 "Insoluble").

2.1.3 | Life chemicals

Life Chemicals company provided kinetic solubility data for one of its fragment libraries [18]. Solubility of 11457 fragments was visually determined based on scattering observed in solutions at 1 mM concentration in PBS (pH 7.4) with 0.5 % DMSO. After removal of data points with no kinetic solubility, the curated dataset consists of 9276 "Soluble" molecules.

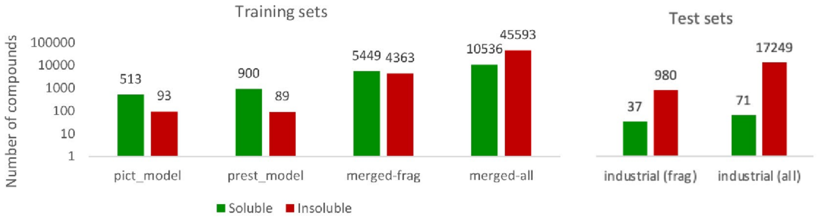


FIGURE 1 Distribution of "Soluble" (green)/"Insoluble" (red) classes in training and test sets. The population axis follows a logarithmic scale.

2.1.4 | MLSMR

The Molecular Libraries Small Molecule Repository (MLSMR) [19] is a collection of small molecules compiled under the initiative of National Institutes of Health (NIH) and screened by Sanford-Burnham Center for Chemical Genomics (SBCCG). To our knowledge, MLSMR is the largest kinetic solubility dataset available in PubChem and it is composed of 57824 data points measured in PBS (pH 7.4) using quantitative chemiluminescent nitrogen detection (CLND). Although, 0.2 mM was reported as the nominal concentration of a sample, a large fraction of the reported concentration (about 31 % of the dataset) is in the range of (0.15; 0.151]. Based on this observation, we assumed 0.15 mM as the actual sample nominal concentration and removed data points which reported concentration greater than or equal to 0.15 mM (13262 data points). Additionally, data curation included removal of duplicate molecules while taking median of their solubility values (mean of standard deviations over the duplicates = 9.85 μM). The resulting curated dataset contained 44510 nitrogen containing compounds which are insoluble at 0.15 mM, and therefore labeled "Insoluble" at 1 mM.

2.1.5 | Boehringer

Boehringer Ingelheim Pharma GmbH & Co. shared a dataset of 789 kinetic solubility measurements [20] performed in PBS (pH 7.4) using nephelometry method. Data points with reported precipitate formation in DMSO stock solution and those for which solubility value was only bounded (relation denoted as ">") were removed. The curated dataset contained 605 compounds that are all "Insoluble" at 1 mM. This dataset was used for QSPR modelling. The full dataset (789 data points) was used to discuss the alignment of solubility values between different kinetic solubility assays.

2.1.6 | CNE1 and CNE2

Chimiothèque Nationale Essentielle (CNE) is a representative collection of physical samples of pure compounds from a larger chemical library of biologically relevant substances and natural extracts called Chimiothèque Nationale [21]. CNE1 is referring to the first generation of this representative collection of 640 compounds, most of which has been depleted. CNE2 is a currently available new representative collection of 1040 compounds. Aqueous solubility of both of these collections have been measured by the "Plateforme de Chimie Biologique Intégrative de Strasbourg" (PCBIS) screening platform. PCBIS has measured thermodynamic solubility for CNE1 collection, whereas CNE2 collection was screened for kinetic solubility. Thermodynamic solubility was measured using shake-flask method, whereas kinetic solubility was measured using HPLC-UV method, at 200 μM nominal concentration (see Supplementary Information for details). Data curation process was identical to Oprisiu [22]. Insoluble compounds which solubility was lower than the limit of detection have been ignored for the discussion. In addition, for CNE2, the following data points were removed:

- entries with reported concentration > 210 μM, implying an experimental error;
- measurements with signs of impurity (multiple peaks in chromatogram);
- compounds with observed precipitation in stock solutions.

The CNE1 contains 282 compounds and the curation step yielded 525 compounds in CNE2, all of which are insoluble based on 1 mM threshold. CNE1 and CNE2 datasets were used to analyze differences between thermodynamic and kinetic solubility assay types, whereas the latter was also used for QSPR model training.

2.1.7 | Industrial data

The kinetic solubility dataset provided by Sanofi contained solubility values of 18407 compounds measured from a 10 mM stock in PBS (pH 7.4) using nephelometry technique. The curation procedure involved duplicate molecule processing by taking median solubility value, and removal of data points in [0.8; 1.2] mM range according to expert opinion. The latter step is related to possible experimental error that could potentially change solubility label based on 1 mM threshold. The curated dataset was composed of 17320 compounds, including 71 "Soluble" and 17249 "Insoluble" compounds. A subset of the curated dataset composed of 1017 fragment-like compounds only consisted of 37 "Soluble" and 980 "Insoluble" compounds. Fragments were defined according to the rule of 3 (Ro3) [23]: calculated logP ≤ 3, molecular weight < 300 g/mol, number of hydrogen bond donors ≤ 3, number of hydrogen bond acceptors ≤ 3.

A subset of compounds for which both thermodynamic and kinetic solubility were measured contained 334 molecules. It was used to investigate the relationship between thermodynamic and kinetic solubility assay types. The whole dataset, "*industrial (all)*", and the fragment-like subset, "*industrial (frag)*", were used as test sets for external validation of the trained QSPR models.

2.2 | Preparation of the merged kinetic solubility training set

In this section, we describe the preparation of the merged dataset comprising data of PICT, Prestwick Chemicals, Life Chemicals, Boehringer, CNE2, and MLSMR. The dataset "*industrial (all)*" and its subset "*industrial (frag)*" containing fragment-like compounds are used as external validation for QSPR models: they have been considered a posteriori, after all model building and validation have been concluded.

We identified duplicated compounds between the different datasets and tried to resolve the conflicting (Tables 3 and 4). PICT and Prestwick Chemicals have 5 compounds in common but the labels are in agreement. The labels of 2 compounds out of 27 in common between PICT and Life Chemicals datasets do not match. These 2 data were ignored because we could not resolve this conflict. There are 4 duplicates between the PICT and MLSMR datasets; labels differed for 3 of them and the discrepancy could not be solved for 1 of them – this data was ignored. For the remaining 2, the "Soluble" label was accepted because the reported concentration in MLSMR was close enough to the nominal concentration to assume that in fact, these compounds were fully dissolved.

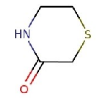
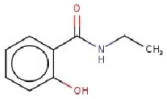
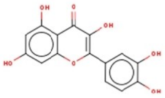
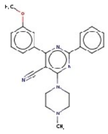
We found 3 CNE2 molecules that had contradicting solubility class labels relative to other datasets (2 molecules between CNE2 and Prestwick Chemicals; 1 molecule between CNE2 and Life Chemicals). The 2 CNE2 molecules had solubility values (179 μM, 180 μM) close enough to the nominal sample concentration (200 μM) to assume that the compounds were in fact fully dissolved, considering measurement uncertainty. For this reason, the labels "Soluble" from both Prestwick Chemicals and Life Chemicals have been accepted. The remaining CNE2 compound had "Insoluble" class label (39 μM solubility value) which contradicted Life Chemicals' "Soluble" label. As we could not resolve this contradiction, the datapoint has not been included in the merged dataset.

The MLSMR had 208 molecules in duplicate with the other datasets. After a thorough analysis, a large population (116 molecules) of data points was in the [140; 150] μM range, which is close enough to the nominal value of 150 μM, to assume that the compounds were in fact fully dissolved. For these 116 MLSMR data points we accepted the labels "Soluble" from the other datasets. We could not resolve the contradicting labels for the remaining 92 MLSMR duplicate measurements and these datapoints were ignored.

TABLE 3 The number of common compounds between each pair of kinetic solubility datasets. The LC and Prestwick datasets are composed of categorical values only, whereas the other datasets contained numerical values.

	Boehringer	LC	MLSMR	PICT	Prestwick	CNE2
LC	0					
MLSMR	12	189				
PICT	0	28	14			
Prestwick	0	39	169	5		
CNE2	0	1	8	1	5	
Industrial (all)	1	19	92	0	11	3

TABLE 4 Comparison of kinetic solubility of compounds common to pairs of datasets. The table is composed of cases when only one compound was in common between a given pair of datasets. The case of "Industrial (all)" vs CNE2 compound is an exception: it is reported separately from the scatter plot presented in Figure 6, because it could not be quantified in CNE2 measurements.

Compound	Dataset A and solubility	Dataset B and solubility	Comment
	Life Chemicals Soluble at 1 mM (−3 log)	CNE2 0.18 mM (−3.74 log)	Difference between log values = 0.74 Good alignment (difference within 1 log unit)
	PICT 0.9 mM (−3.05 log)	CNE2 0.25 mM (−3.61 log)	Difference between log values = 0.56 Good alignment (difference within 1 log unit)
	Industrial (all) 0.001 mM (−6 log)	CNE2 < 10 μM	One can assume good alignment considering the limitations of the reported data in both the industrial dataset and the CNE2.
	Industrial (all) 0.006 mM (−5.22 log)	Boehringer 0.005 mM (−5.3 log)	Difference between log values = 0.08 Good alignment (difference within 1 log unit)

3 | METHOD

3.1 | Molecular descriptors

We used ISIDA substructural molecular fragments (SMF) [24] representing 2D substructures (fragments) of various topologies (sequence of atoms only, sequence of atoms and bonds, atom-centered fragments, triplets) and sizes (see Table S1 in Supplementary Information). The descriptor value is the occurrence of a given fragment in the chemical structure. The minimal length of fragment descriptors was 2 atoms, while the maximal length varied from 2 to 5 atoms. Combination of different topologies and sizes resulted in generation of 112 descriptor sets.

3.2 | Machine learning method

Support Vector Machine (SVM) method was implemented in this study for the generation of kinetic solubility QSPR models and potential outliers' detection. The SVM method offers the advantage of robustness against outliers, thanks to its epsilon-insensitive loss function. The libsvm 3.24 software package [25] was used for training and validation of SVM models. Selection of optimal SVM hyperparameters, SVM kernels and descriptor sets was performed using genetic algorithm (GA) implemented in the libsvm-GA package [26].

Four statistical metrics are used here: sensitivity, specificity, balanced accuracy (BA), Matthew's correlation coefficient (MCC). They are calculated using the equations given below (TP – true positive; TN – true negative; FP – false positive; FN – false negative). In this context, soluble class is regarded as "Positive" class, and insoluble class is regarded as "Negative" class.

$$Sensitivity = TP/(TP + FN) \tag{1}$$

$$Specificity = TN/(TN + FP) \tag{2}$$

$$BA = (Sensitivity + Specificity)/2 \tag{3}$$

$$MCC = (TN \times TP - FN \times FP)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \tag{4}$$

3.3 | Modeling workflow

The modeling workflow of kinetic solubility QSPR models applied in this study can be divided into 3 steps: (1) molecular descriptor calculation; (2) model building and validation using cross-validation; (3) consensus model preparation (Figure 2). ISIDA fragment descriptors were computed for each training set during cross-validation. The hyperparameters of the models were optimized using a GA, with the cross-validation performances as scoring function. The top performing models were included in a consensus model. The selected models were then retrained on the whole dataset and included in the consensus model integrated into the ISIDA Predictor software [27] (available both as desktop software and web service [13]). The ISIDA Predictor software was used to predict kinetic solubility on the industrial data. The reported external performances concern this application of the model.

In addition to the application of QSPR models, the ISIDA Predictor software incorporates an assessment of predicted value confidences. Scoring of prediction confidence is based on the number of applied models and concordance between the predicted labels given by each applied individual model of the consensus model. Each individual model prediction is considered according to the model's applicability domain, defined by fragment control rule [28]. Fragment control states that if a test molecule contains at least one new fragment compared to those observed in the training set, the model is not applied.

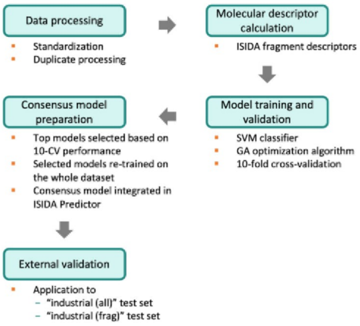

Data processing
Standardization
Duplicate processing
Molecular descriptor calculation
ISIDA fragment descriptors
Model training and validation
SVM classifier
GA optimization algorithm
10-fold cross-validation
Consensus model preparation
Top models selected based on 10-CV performance
Selected models re-trained on the whole dataset
Consensus model integrated in ISIDA Predictor
External validation
Application to
– "industrial (all)" test set
– "industrial (frag)" test set

FIGURE 2 The modelling workflow of kinetic solubility QSPR models. The main steps are pre-processing data, computing molecular descriptor, training and validating individual models and implementation of the consensus model. External validation results from application of the final consensus model to the test sets (*industrial (all)* and *industrial (frag)* datasets). SVM – Support Vector Machine; GA – Genetic Algorithm; 10-CV – 10-fold cross-validation.

The ISIDA Predictor provides 4 labels of prediction confidence: "Low", "Average", "Good", "Optimal". In this paper, for kinetic solubility QSPR models we considered only the test compounds with "Optimal" prediction confidence.

While an ideal classification model would excel at predicting compounds from both classes, in the context of kinetic solubility, the primary goal is to identify and eliminate insoluble molecules. From a statistical perspective, the model should exhibit high Specificity (the ability to predict insoluble molecules accurately) while still maintaining high BA and MCC. Performance metrics for the developed kinetic models are summarized in Table 5.

We also challenged an independent thermodynamic solubility QSPR model to predict the kinetic solubility label using a 1 mM threshold. This QSPR model has been trained on a dataset comprised of 42159 industrial and public solubility data (doi:10.57745/CZVZIA). The model was trained using Chemprop software package [29] that implements a message passing neural network method. The validation performance on a test set of 5728 compounds was RMSE (root mean squared error) = 0.59.

TABLE 5 10-fold cross-validation (10-CV) performance of consensus QSPR models developed in this work.[a]

Model	Training set	# individual models in the consensus model	BA$_{10\text{-}CV}$[b]	Standard deviation (BA$_{10\text{-}CV}$)	MCC$_{10\text{-}CV}$[b]	Standard deviation (MCC$_{10\text{-}CV}$)	Sensitivity (10-CV)	Specificity (10-CV)
prest_model	Prestwick Chemicals	5	0.68	0.09	0.39	0.16	0.96	0.4
pict_model	PICT	3	0.71	0.06	0.46	0.17	0.94	0.48
merged-frag_model	Merged (frag)	7	0.87	0.01	0.75	0.02	0.91	0.84
merged-all_model	Merged (all)	12	0.93	0.004	0.86	0.005	0.88	0.98

[a] Each representing ensemble of individual SVM models built on ISIDA fragment descriptors.
[b] BA – balanced accuracy; MCC – Matthew's correlation coefficient.

4 | RESULTS AND DISCUSSION

4.1 | Kinetic and thermodynamic solubility

Saal and Petereit [2] described three different types of relationship between kinetic and thermodynamic solubility visualized on Figure 3. The first one (Zone A) corresponds to compounds fully dissolved in a kinetic solubility measurement because their thermodynamic solubility is equal to or larger than the nominal of the measure. The second type (Zone B) is typical for the compounds whose kinetic solubility is larger than thermodynamic one. This behavior can be explained by the solid-state form of the precipitate that may differ from a kinetic to a thermodynamic measurement [30]. In kinetic solubility measurements, the solid that forms can be amorphous or a metastable crystal polymorph; thermodynamic measurements start from a crystal that must be solubilized and are expected to let only the lowest soluble solid to form. The measurement can be complicated if the compound leads to polymorphic crystal structures [31]. The third type (Zone C) represents compounds for whom kinetic and thermodynamic solubilities correlate.

In this context two new datasets – Industrial and Chimiothèque Nationale Essentielle version 1 and 2 (CNE1 and CNE2) – have been analyzed. The former dataset shows a rather different pattern (Figure 4) from what is expected (Figure 3). The scattered data points are organized along several horizontal lines, at certain kinetic solubility values. This dataset corresponds to several kinetic solubility determination campaigns carried out at different concentrations. These horizontal lines correspond to the different nominal concentrations of the many nephelometry kinetic measurements aggregated in this dataset. The contributors to this dataset were looking for the nominal concentration at which each compound begins to appear insoluble. To this end, they scanned several of them and reported a concentration that appear to behave as in the zone A exemplified in the Figure 3.

In Figure 5, the solubility values distribution aligns with expectations (Figure 3). While the majority of data points are accumulated at about −3.7 log kinetic solubility, the others are instances of the case when kinetic solubility is greater than or equal to the thermodynamic solubility. Apart from 6 outlying data points, the overall picture resembles the pattern described by Saal and Petereit [2]. The 6 compounds on the lower right hand of the plot, not matching the expectations are disclosed in the Supplementary Information (Table S4). The limit of detection at −3.7 log has been explained by Saal and

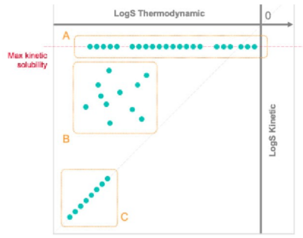


FIGURE 3 Different types of relationship between thermodynamic and kinetic solubility. Zone A: kinetic solubility of compounds is at the nominal concentration; zone B: kinetic solubility greater than thermodynamic solubility; zone C: kinetic solubility equals to thermodynamic solubility.

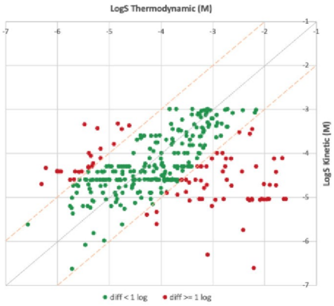


FIGURE 4 Comparison of kinetic and thermodynamic solubility values of the industrial dataset (334 compounds). Green dots represent differences <1 log unit between kinetic and thermodynamic values, red dots > =1 log unit. Orange dashed lines show a 1 log margin.

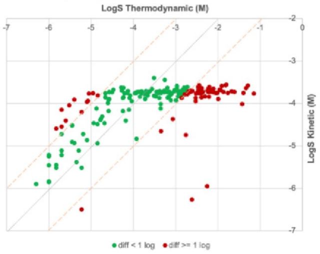


FIGURE 5 Comparison of kinetic and thermodynamic solubility values of Chimiothèque Nationale Essentielle dataset (186 compounds). Green dots represent differences <1 log unit between kinetic and thermodynamic values, red dots > =1 log unit. Orange dashed lines show a 1 log margin.

Petereit [2] as resulting from the nominal concentration and the maximum DMSO concentration allowed in the incubation medium.

The difference between kinetic and thermodynamic solubility measurements can originate from solvent-mediated transformations occurring between different polymorphic forms of the compound [31, 32]. Recrystallization leads to the most stable polymorphic form which is characterized by its lower solubility. Measurement of a compound at any other metastable form results in different concentration (kinetic solubility) as it did not reach equilibrium state with the solution. Equally important factor is the quality of the measured compounds. Compounds with a low purity will lead to stock solutions with concentration errors, followed by calibration errors and finally, measurement errors. Additionally, it is now better understood that "kinetic solubility" does not refer to a kinetic phenomenon, and therefore, this terminology is contested [32].

4.2 | Analysis of available kinetic datasets

This section reports the comparison of different kinetic solubility datasets based on common compounds between each pair of datasets. The findings of this study have been used to build the merged datasets (see section "Preparation of the merged kinetic solubility training set").

In Table 3, a number of common compounds for each pair of kinetic solubility datasets is given. The analysis of common compounds was conducted in two ways: by scatter plots, for datasets containing numerical values; by pairwise comparison of datasets containing categorical values. The cases where there was only one common compound were studied individually.

Scatter plots presented in Figure 6 generally show the good agreement between datasets (within 1 log margin). Vertical alignment of data points observed in scatter plots involving MLSMR data correspond to the upper limit of value set by the nominal sample concentration. For one compound the reported solubility is < 10 μM in CNE2 and 1 μM in the industrial dataset.

The pairwise comparison of the LC/PICT, LC/Prestwick Chemicals and PICT/Prestwick Chemicals dataset pairs shows (Figure 7) consistency of kinetic solubility data: only 2 molecules out of 16 are differently labeled in LC and PICT, in the other dataset all labels are fully consistent. The datasets whose max solubility value is less than 1 mM (MLSMR, CNE2) were not considered during this comparison, since the solubility label between 1 mM and their nominal concentration cannot be decided.

Table 4 consists of cases where only one molecule was common to pairs of datasets, except for a compound common to "Industrial (all)" and CNE2, which was detected below the limit of quantification of the CNE2 measures. Overall, these data confirm the good

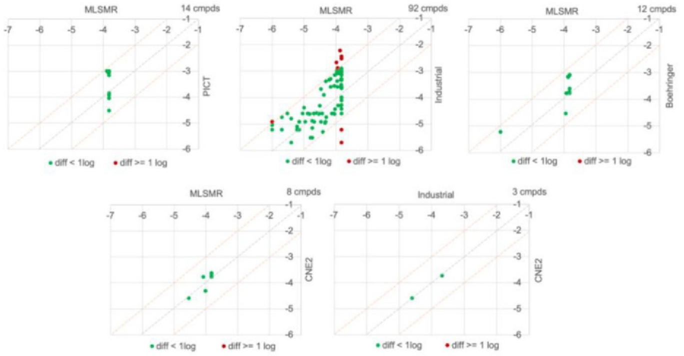
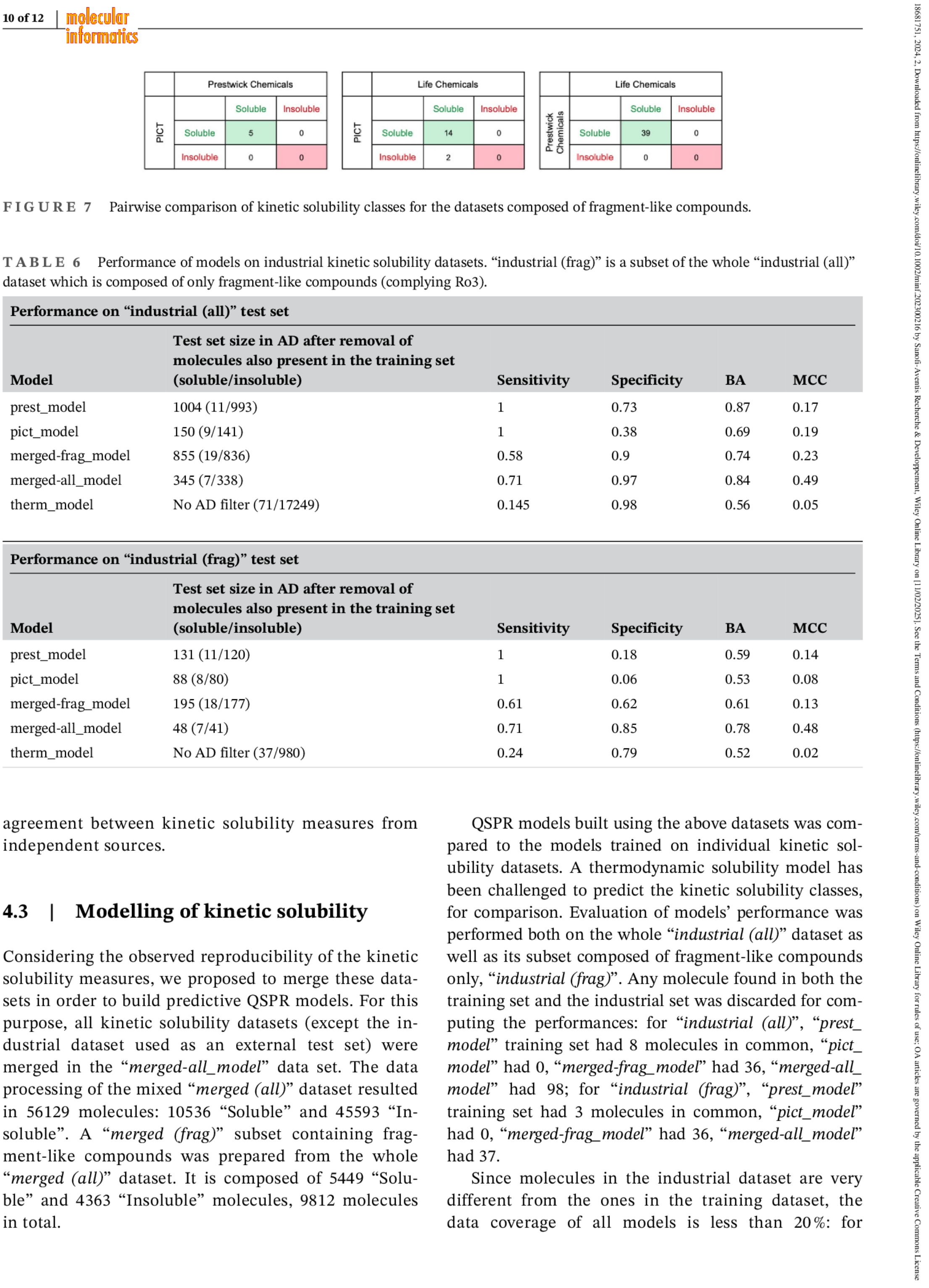


FIGURE 6 Scatter plots comparing kinetic solubility values of dataset pairs. The unit is logS (in molar). The number of common compounds is given at the top right corner of the plot. Green dots represent cases when the absolute difference between kinetic solubility values is less than 1 log unit and red dots indicate when the difference is greater than or equal to 1 log unit. Orange dashed lines shows 1 log margin.



		Prestwick Chemicals	
		Soluble	Insoluble
PICT	Soluble	5	0
	Insoluble	0	0

		Life Chemicals	
		Soluble	Insoluble
PICT	Soluble	14	0
	Insoluble	2	0

		Life Chemicals	
		Soluble	Insoluble
Prestwick Chemicals	Soluble	39	0
	Insoluble	0	0

FIGURE 7 Pairwise comparison of kinetic solubility classes for the datasets composed of fragment-like compounds.

TABLE 6 Performance of models on industrial kinetic solubility datasets. "industrial (frag)" is a subset of the whole "industrial (all)" dataset which is composed of only fragment-like compounds (complying Ro3).

Performance on "industrial (all)" test set					
Model	**Test set size in AD after removal of molecules also present in the training set (soluble/insoluble)**	**Sensitivity**	**Specificity**	**BA**	**MCC**
press_model	1004 (11/993)	1	0.73	0.87	0.17
pict_model	150 (9/141)	1	0.38	0.69	0.19
merged-frag_model	855 (19/836)	0.58	0.9	0.74	0.23
merged-all_model	345 (7/338)	0.71	0.97	0.84	0.49
therm_model	No AD filter (71/17249)	0.145	0.98	0.56	0.05

Performance on "industrial (frag)" test set					
Model	**Test set size in AD after removal of molecules also present in the training set (soluble/insoluble)**	**Sensitivity**	**Specificity**	**BA**	**MCC**
prest_model	131 (11/120)	1	0.18	0.59	0.14
pict_model	88 (8/80)	1	0.06	0.53	0.08
merged-frag_model	195 (18/177)	0.61	0.62	0.61	0.13
merged-all_model	48 (7/41)	0.71	0.85	0.78	0.48
therm_model	No AD filter (37/980)	0.24	0.79	0.52	0.02

agreement between kinetic solubility measures from independent sources.

4.3 | Modelling of kinetic solubility

Considering the observed reproducibility of the kinetic solubility measures, we proposed to merge these datasets in order to build predictive QSPR models. For this purpose, all kinetic solubility datasets (except the industrial dataset used as an external test set) were merged in the "*merged-all_model*" data set. The data processing of the mixed "*merged (all)*" dataset resulted in 56129 molecules: 10536 "Soluble" and 45593 "Insoluble". A "*merged (frag)*" subset containing fragment-like compounds was prepared from the whole "*merged (all)*" dataset. It is composed of 5449 "Soluble" and 4363 "Insoluble" molecules, 9812 molecules in total.

QSPR models built using the above datasets was compared to the models trained on individual kinetic solubility datasets. A thermodynamic solubility model has been challenged to predict the kinetic solubility classes, for comparison. Evaluation of models' performance was performed both on the whole "*industrial (all)*" dataset as well as its subset composed of fragment-like compounds only, "*industrial (frag)*". Any molecule found in both the training set and the industrial set was discarded for computing the performances: for "*industrial (all)*", "*prest_model*" training set had 8 molecules in common, "*pict_model*" had 0, "*merged-frag_model*" had 36, "*merged-all_model*" had 98; for "*industrial (frag)*", "*prest_model*" training set had 3 molecules in common, "*pict_model*" had 0, "*merged-frag_model*" had 36, "*merged-all_model*" had 37.

Since molecules in the industrial dataset are very different from the ones in the training dataset, the data coverage of all models is less than 20%: for

"*industrial (all)*", "*prest_model*" was applied to 1004 molecules with "Optimal" confidence prediction label (5.8 % of the "*industrial (all)*" with no common molecules with the training set of "*prest_model*"), "*pict_model*" to 150 molecules (0.9 %), "*merged-frag_model*" to 855 molecules (4.9 %), "*merged-all_model*" to 345 molecules (2 %); for "*industrial (frag)*", "*prest_model*" was applied to 88 molecules with "Optimal" confidence prediction label (12.9 % of the "*industrial (all)*" with no common molecules with the training set of "*prest_model*"), "*pict_model*" to 131 molecules (8.7 %), "*merged-frag_model*" to 195 molecules (19.9 %), "*merged-all_model*" to 48 molecules (4.9 %).

The results show that models trained on a combination of kinetic solubility datasets ("*merged-all_model*", "*merged-frag_model*") show higher MCC and Specificity values, compared to those trained on individual datasets, both in "*industrial (all)*" and "*industrial (frag)*" test sets (Table 6). When applied to the "*industrial (frag)*" test set, the "*merged-frag-_model*" demonstrates inferior results compared to the "*merged-all_model*". The latter benefits from a more extensive training set, despite the former's specialization, which includes only fragment-like compounds. Moreover, one can see that the ratio of soluble to insoluble molecules in the "*merged-all_model*" (≈0.2) is closer to the ratio in the "*industrial (frag)*" test set (≈0.07), rather than the more equally distributed training set of the "*merged-frag_model*" (≈1.25). Actually, the mismatch of the prior expectation of the other kinetic solubility models ("*prest_model*", "*pict_model*") compared to the actual "Soluble"/"Insoluble" distributions observed in the various dataset can have a negative impact on their performances. This adds to the weaknesses of these models resulting from the relatively small size of their training sets.

For early drug discovery solubility screening campaigns, it is better to identify and remove insoluble compounds. For this reason, it is preferable for a QSPR model to have high predictive rate of insoluble molecules (*Specificity*), while preserving a high BA and MCC. Given that, the "*merged-all_model*" is a better candidate to be used for virtual screening (see Supplementary Information Table S2 for details). The use of a thermodynamic solubility model for such task seems a wrong idea, as illustrated by the performance of a recent predictive QSPR model used for this task (*therm_model*, Table 6).

The benchmarking of existing models that were described in Table 1 and Table S3, is not possible due to unavailability of those models.

5 | CONCLUSIONS

The analysis of kinetic and thermodynamic solubility data confirmed the previously known patterns [2] of relationship between these two solubility types, namely, the three scenarios: (i) upper limit of kinetic solubility constrained by the assay setup, (ii) overestimation of kinetic solubility relative to thermodynamic solubility, (iii) equal kinetic and thermodynamic solubility.

Our analysis also demonstrated that the kinetic solubility data obtained using different measurement protocols are in good agreement with each other, indicating good inter-laboratory reproducibility.

This allowed us to merge the kinetic solubility data into a single dataset on which predictive models were trained. This dataset (doi:10.57745/ZWS0WC) contains exclusive data from Prestwick Chemicals, PICT and CNE2 never reported so far. The modelability of the merged dataset using different detection methods strengthen the conclusion that kinetic solubility data are not as assay-dependent as initially assumed. It should be noted that the model trained on thermodynamic solubility data fails to evaluate kinetic solubility, emphasizing that these are conceptually related but different measurements.

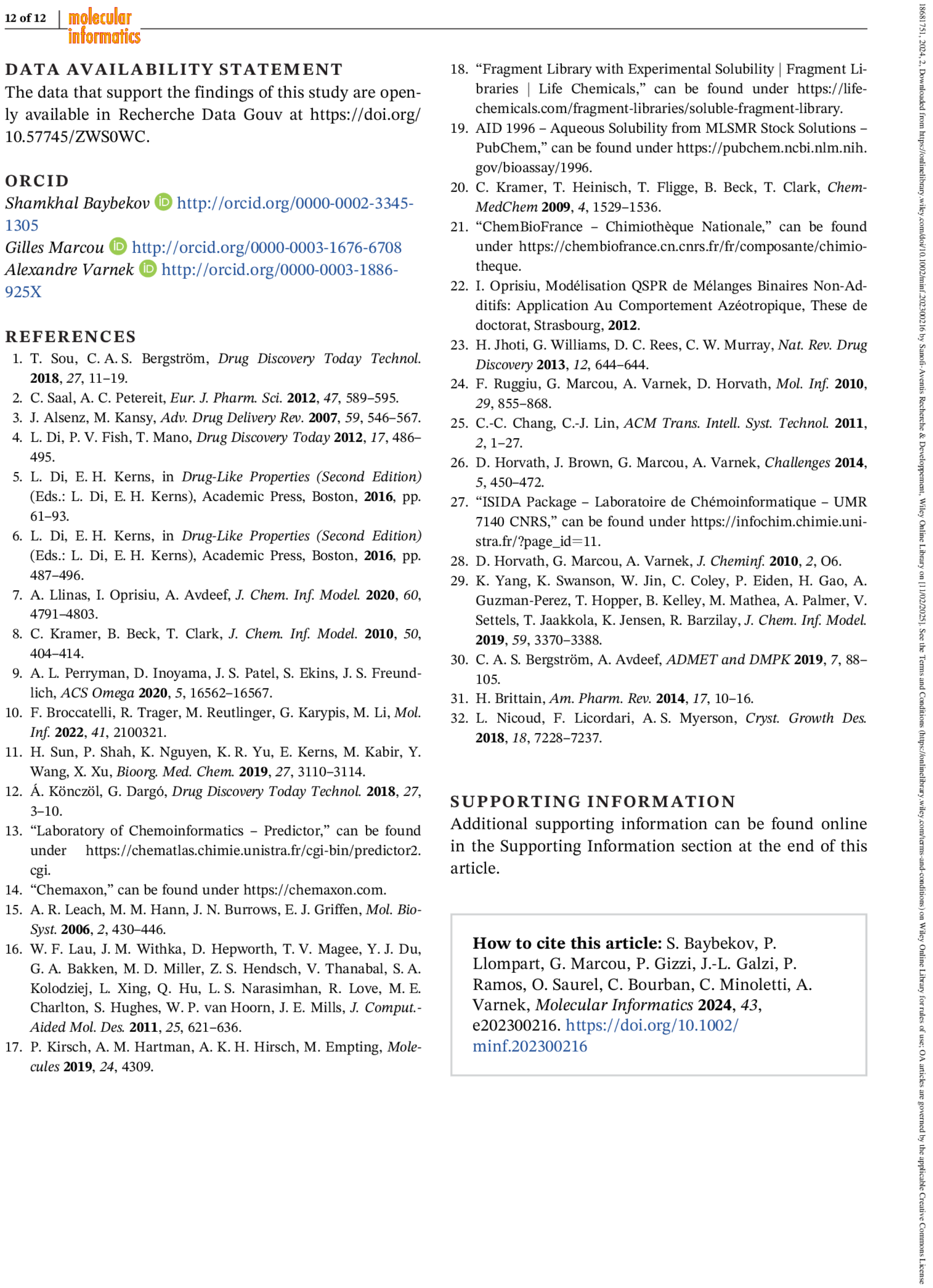
This contribution led to the publicly available QSPR model predicting kinetic solubility freely accessible through the Predictor web service of the Laboratory of Chemoinformatics (https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi). The model can be used for prioritization of screening compounds by preliminary assessing kinetic solubility at pH 7.4 and at 1 mM nominal concentration and a DMSO maximal concentration of 2 % in the incubation medium. It is recommended to consider only "Optimal" predicted values when applying this model.

ACKNOWLEDGMENTS

Shamkhal Baybekov thanks the CSC Graduate School funded by the French National Research Agency (CSC-IGS ANR-17-EURE-0016) for a PhD fellowship. The fragment library and the Bruker Avance III HD 600 MHz NMR spectrometer of the Integrated Screening Platform of Toulouse (PICT) were funded by CNRS, Université Paul Sabatier, Infrastructures en Biologie Santé et Agronomie European Structural Funds, and the Midi-Pyrénées Region.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.



DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Recherche Data Gouv at https://doi.org/10.57745/ZWS0WC.

ORCID

Shamkhal Baybekov http://orcid.org/0000-0002-3345-1305
Gilles Marcou http://orcid.org/0000-0003-1676-6708
Alexandre Varnek http://orcid.org/0000-0003-1886-925X

REFERENCES

1. T. Sou, C. A. S. Bergström, *Drug Discovery Today Technol.* **2018**, *27*, 11–19.
2. C. Saal, A. C. Petereit, *Eur. J. Pharm. Sci.* **2012**, *47*, 589–595.
3. J. Alsenz, M. Kansy, *Adv. Drug Delivery Rev.* **2007**, *59*, 546–567.
4. L. Di, P. V. Fish, T. Mano, *Drug Discovery Today* **2012**, *17*, 486–495.
5. L. Di, E. H. Kerns, in *Drug-Like Properties (Second Edition)* (Eds.: L. Di, E. H. Kerns), Academic Press, Boston, **2016**, pp. 61–93.
6. L. Di, E. H. Kerns, in *Drug-Like Properties (Second Edition)* (Eds.: L. Di, E. H. Kerns), Academic Press, Boston, **2016**, pp. 487–496.
7. A. Llinas, I. Oprisiu, A. Avdeef, *J. Chem. Inf. Model.* **2020**, *60*, 4791–4803.
8. C. Kramer, B. Beck, T. Clark, *J. Chem. Inf. Model.* **2010**, *50*, 404–414.
9. A. L. Perryman, D. Inoyama, J. S. Patel, S. Ekins, J. S. Freundlich, *ACS Omega* **2020**, *5*, 16562–16567.
10. F. Broccatelli, R. Trager, M. Reutlinger, G. Karypis, M. Li, *Mol. Inf.* **2022**, *41*, 2100321.
11. H. Sun, P. Shah, K. Nguyen, K. R. Yu, E. Kerns, M. Kabir, Y. Wang, X. Xu, *Bioorg. Med. Chem.* **2019**, *27*, 3110–3114.
12. Á. Könczöl, G. Dargó, *Drug Discovery Today Technol.* **2018**, *27*, 3–10.
13. "Laboratory of Chemoinformatics – Predictor," can be found under https://chematlas.chimie.unistra.fr/cgi-bin/predictor2.cgi.
14. "Chemaxon," can be found under https://chemaxon.com.
15. A. R. Leach, M. M. Hann, J. N. Burrows, E. J. Griffen, *Mol. BioSyst.* **2006**, *2*, 430–446.
16. W. F. Lau, J. M. Withka, D. Hepworth, T. V. Magee, Y. J. Du, G. A. Bakken, M. D. Miller, Z. S. Hendsch, V. Thanabal, S. A. Kolodziej, L. Xing, Q. Hu, L. S. Narasimhan, R. Love, M. E. Charlton, S. Hughes, W. P. van Hoorn, J. E. Mills, *J. Comput.-Aided Mol. Des.* **2011**, *25*, 621–636.
17. P. Kirsch, A. M. Hartman, A. K. H. Hirsch, M. Empting, *Molecules* **2019**, *24*, 4309.
18. "Fragment Library with Experimental Solubility | Fragment Libraries | Life Chemicals," can be found under https://lifechemicals.com/fragment-libraries/soluble-fragment-library.
19. AID 1996 – Aqueous Solubility from MLSMR Stock Solutions – PubChem," can be found under https://pubchem.ncbi.nlm.nih.gov/bioassay/1996.
20. C. Kramer, T. Heinisch, T. Fligge, B. Beck, T. Clark, *ChemMedChem* **2009**, *4*, 1529–1536.
21. "ChemBioFrance – Chimiothèque Nationale," can be found under https://chembiofrance.cn.cnrs.fr/fr/composante/chimiotheque.
22. I. Oprisiu, Modélisation QSPR de Mélanges Binaires Non-Additifs: Application Au Comportement Azéotropique, These de doctorat, Strasbourg, **2012**.
23. H. Jhoti, G. Williams, D. C. Rees, C. W. Murray, *Nat. Rev. Drug Discovery* **2013**, *12*, 644–644.
24. F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855–868.
25. C.-C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
26. D. Horvath, J. Brown, G. Marcou, A. Varnek, *Challenges* **2014**, *5*, 450–472.
27. "ISIDA Package – Laboratoire de Chémoinformatique – UMR 7140 CNRS," can be found under https://infochim.chimie.unistra.fr/?page_id=11.
28. D. Horvath, G. Marcou, A. Varnek, *J. Cheminf.* **2010**, *2*, O6.
29. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
30. C. A. S. Bergström, A. Avdeef, *ADMET and DMPK* **2019**, *7*, 88–105.
31. H. Brittain, *Am. Pharm. Rev.* **2014**, *17*, 10–16.
32. L. Nicoud, F. Licordari, A. S. Myerson, *Cryst. Growth Des.* **2018**, *18*, 7228–7237.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: S. Baybekov, P. Llompart, G. Marcou, P. Gizzi, J.-L. Galzi, P. Ramos, O. Saurel, C. Bourban, C. Minoletti, A. Varnek, *Molecular Informatics* **2024**, *43*, e202300216. https://doi.org/10.1002/minf.202300216

Outline

Contrary to a common assumption, our findings indicate that kinetic solubility data from different protocols exhibit good agreement, supporting the feasibility of robust predictive modeling. We further show that thermodynamic solubility models fail to generalize to kinetic solubility, reinforcing the necessity for dedicated QSPR models to be used in the preparation of plates for HTS. To address these challenges, we present a workflow for curating high-quality kinetic solubility datasets and training reliable predictive models. The curated datasets and trained models from this study are made publicly available to support further improvements in solubility modeling.

Chapter 5. Modeling of Drug Absorption

5.1. Multi-Task for Permeability Prediction

Introduction

The journey of a drug through the body involves multiple pharmacokinetic phases each determining therapeutic efficacy and safety.

Passing the Absorption Barrier

Absorption, the first step, encompasses the transfer of a xenobiotic from the gastrointestinal tract, primarily the small intestine, into systemic circulation. After the dosage form disintegrates and the active ingredient dissolves in digestive fluids, the compound must cross the epithelial cells either paracellularly (between cells) or transcellularly (through cells) (**Figure 20**). In any case, most drugs must first enter the systemic circulation before reaching their target.

Main Terminology

Xenobiotics are any foreign chemical substance within an organism.

Therapeutic efficacy is the ability of a drug to produce the desired treatment effect.

Recovery a.k.a mass-balance is the efficiency with which a drug is retrieved or detected after a process.

Many small-molecule drugs exert their effect by binding deep within the hydrophobic core of proteins. These buried sites are typically less accessible to water, making them difficult targets for polar compounds. Strong and selective binding at such interfaces often relies on hydrophobic interactions, which favor lipophilic ligands. While medicinal chemistry has emphasized reducing lipophilicity to improve solubility, metabolic stability, and overall pharmacokinetics, many clinically successful drugs still span a wide range of lipophilicity values. For instance, Desflurane binds within intrahelical transmembrane cavities of the GABA$_A$ receptor, and Propofol targets similar sites in the glycine receptor. Flecainide, an antiarrhythmic drug, reaches its site in the Nav1.5 sodium channel through lateral lipid-facing fenestrations.[37] Yet, although increasing lipophilicity may enhance target binding, it often comes at the cost of reduced bioavailability, and higher risk of off-target effects.

To navigate this trade-off, medicinal chemists leverage structural features to control these endpoints. Fluorine or trifluoromethyl groups are often introduced not only to increase stability and fine-tune lipophilicity and binding kinetics. In some cases, drugs exploit internal hydrogen bonds, as intramolecular H-bonding in a hydrophobic pocket can act as a strong, directional anchor that reinforces binding affinity. This strategy is observed in Ivacaftor. Similarly, polar groups like sulfonamides can act as amphiphilic anchors at the protein–membrane interface, as seen with Fasiglifam.[33]

To further characterize membrane permeation, permeability assays such as Caco-2 and PAMPA are commonly used. The Caco-2 assay models intestinal absorption through a monolayer of human epithelial cells, capturing both passive and active transport mechanisms. In contrast, the PAMPA assay focuses exclusively on passive diffusion by measuring a compound's ability to cross an artificial lipid membrane. Comparing permeability between these two methods helps determine whether a molecule primarily relies on passive diffusion or engages in transporter-mediated processes (**Figure 21**).[148] While comparing results from both assays can help identify transport route, such dual profiling is rarely performed in practice due to cost and resource constraints.

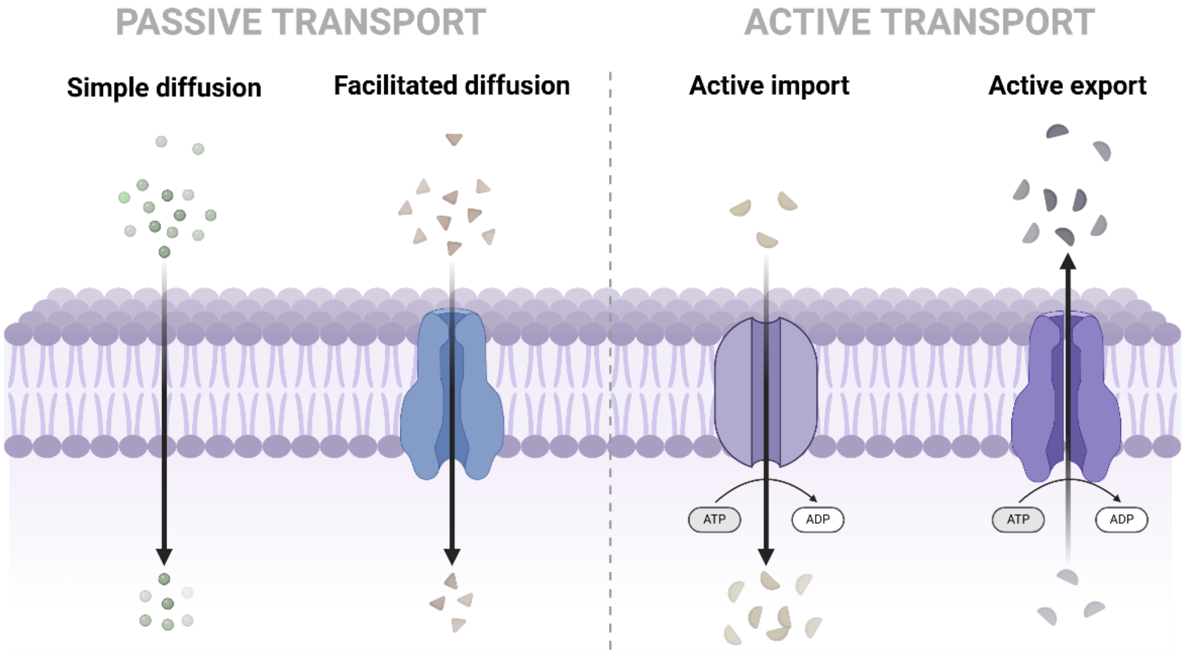


Figure 20: *Mechanisms of membrane permeation.*

Although these assays focus on the membrane transport, oral delivery also depends on overcoming other barriers. To fully grasp a drug's early fate, one must also consider degradation, metabolic transformation, and transporter-mediated efflux. A key limitation to absorption is enzymatic hydrolysis by enzymes such as peptidases and esterases. Interestingly, this liability can be turned into an advantage: prodrugs are often designed with cleavable substituents (e.g., phosphates) that enhance solubility or permeability and are selectively activated by hydrolases or cytochrome P450 enzymes. Even when degradation is avoided and permeability is favorable, a compound may still be actively expelled from enterocytes by efflux transporters. P-glycoprotein (P-gP). This process operates in concert with intestinal CYP enzymes further limiting systemic exposure through what is known as the "intestinal first-pass" effect, a barrier that acts in addition to hepatic first-pass metabolism and can substantially reduce the bioavailable fraction of orally administered drugs.

Distribution in the System

Once in systemic circulation, the distribution phase covers how a xenobiotic moves via the blood and partitions among different tissues. Blood is composed of red blood cells, white blood cells, and plasma. The plasma itself contains about plasma proteins, primarily albumins, α-1-acid glycoproteins (AGP), lipoproteins, and globulins. Albumin generally binds acidic or neutral drugs, while AGP binds basic or neutral compounds. These reversible bindings create an equilibrium between bound (reservoir) and free (active) fractions, with only the free fraction able to transit, exert therapeutic effects, be metabolized, or undergo elimination.

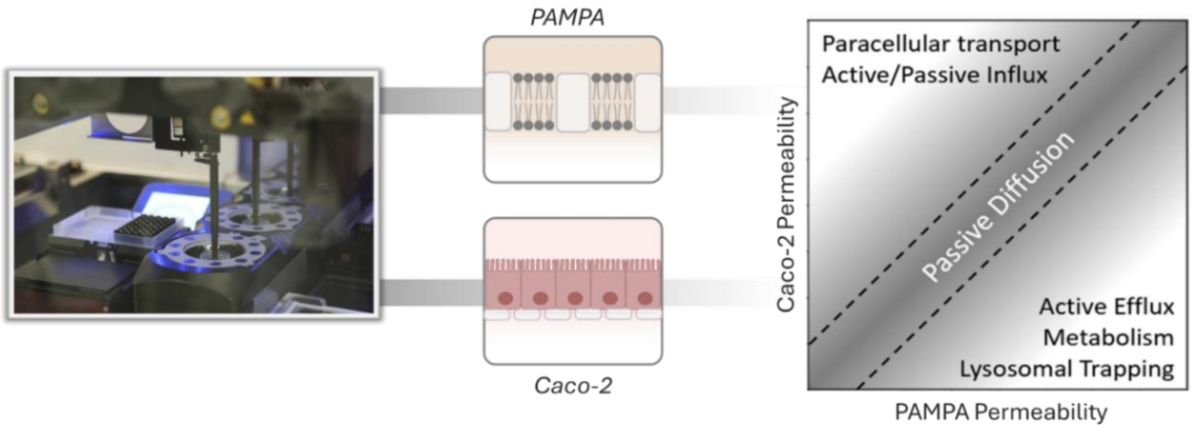


Figure 21: *Comparison of Caco-2 and PAMPA permeability assays.*

Once a drug reaches systemic circulation, its distribution refers to the reversible transfer from blood into various tissues. Initially, the unbound fraction (f_u) of the compound distributes within the extracellular space. Further penetration into cells or deep tissue compartments depends on the molecule's physicochemical properties (e.g., lipophilicity, size, polarity) and the location of its pharmacological target.

For example, lipophilic antibiotics such as Azithromycin show extensive tissue distribution, often accumulating in phagocytic cells and intracellular compartments. While this favors efficacy against intracellular pathogens, excessive sequestration can also limit the free concentration available to bind bacterial ribosomes, highlighting a delicate balance between distribution and target engagement. In contrast, hydrophilic compounds like aminoglycosides are largely restricted to extracellular fluids due to their polarity and poor membrane permeability.[149]

A key pharmacokinetic parameter used to quantify the extent of tissue distribution is the volume of distribution (Vd), defined as:

$$Vd = \frac{Amount\ of\ drug\ in\ the\ body}{Plasma\ concentration}$$

This parameter reflects the apparent, not anatomical, volume into which a drug disperses to yield the observed plasma concentration. High Vd values typically indicate extensive tissue uptake and low plasma levels, often observed in lipophilic or weakly plasma protein-bound compounds. Low Vd values suggest the drug is mostly confined to the bloodstream, commonly due to strong binding to plasma proteins or high polarity.[150] Another fundamental parameter influencing distribution is the fraction unbound in plasma which determines the portion of drug that is free to leave the vascular space, interact with targets, undergo metabolism, or be eliminated. It is calculated as:

$$f_u = \frac{C_{unbound}}{C_{total}}$$

Where the variable,

$C_{unbound}$ is the free (unbound) drug concentration in plasma,

C_{total} is the total drug concentration in plasma (bound + unbound).

Only the unbound drug is pharmacologically active, able to cross membranes, interact with targets, undergo metabolism, or be excreted. Binding to plasma proteins, primarily albumin (for acidic and neutral drugs) and AGP (for basic and some neutral drugs), acts as a dynamic reservoir that reduces the f_u, thereby modulating both distribution and clearance.

A low f_u (e.g., ~1%) indicates that most of the drug is protein-bound, which typically limits distribution into tissues, results in a low Vd, slows clearance, and may prolong the drug's half-life. In contrast, a high f_u (e.g., >10%) means more drug is available in its free form, facilitating broader tissue distribution and often leading to a higher Vd and faster engagement with peripheral compartments.

f_u is measured using equilibrium dialysis, ultrafiltration, or ultracentrifugation which are techniques that estimate the proportion of free versus bound drug in plasma, typically conducted in vitro using human or animal plasma (e.g., rat, dog).

Vd is determined from in vivo pharmacokinetic studies, typically after intravenous dosing to avoid absorption bias. It is calculated from early plasma concentration-time profiles using non-compartmental or model-based methods.

Understanding the relationship between f_u and Vd is essential when predicting tissue exposure, drug efficacy, and the potential for drug–drug interactions. For instance, displacement of a highly protein-bound drug by a co-administered compound can transiently increase f_u, elevate free plasma concentrations, and raise the risk of toxicity.

However, even when f_u is high and Vd suggests favorable distribution, access to certain tissues may still be restricted by biological barriers. The most notable example is the blood–brain barrier (BBB). Formed by tightly connected endothelial cells, the BBB blocks paracellular diffusion and actively limits drug entry through a network of efflux transporters such as P-gP and BCRP. These features mean that only a narrow subset of compounds, typically small, lipophilic, non-ionized molecules that are not efflux substrates, can penetrate the CNS. Consequently, less than 2% of small molecules intended for central nervous system targets successfully achieve therapeutic brain concentrations, making BBB penetration a major challenge in neuropharmacology.[151]

Research Approach

This chapter presents a large-scale analysis and predictive modeling of absorption data from both public and industrial sources, examining relationships between major permeability parameters and unveiling common misconceptions about transport routes. We employ MTL to develop predictive models for absorption and validate their performance across diverse datasets, underscoring the influence of protocol variations on model robustness. Recovery, distribution coefficients, and topological polar surface area emerge as critical factors in Multi-Parameter Optimization (MPO), offering clearer directions for lead selection. We also incorporate Generative Topographic Mapping for chemical space visualization, aiding the identification of absorption hurdles and refining lead optimization strategies.

Garbage in, garbage out: An industrial perspective on drug absorption modeling

P. Llompart[1,2], C. Minoletti[2], G. Marcou[1,*], A. Varnek[1]

[1]Laboratory of Cheminformatics, UMR7140, University of Strasbourg, Strasbourg, France

[2]IDD/CADD, Sanofi, Vitry-Sur-Seine, France

Keywords: drug permeability, QSPR, deep learning, ADMET

Abstract

Lead optimization failures are often linked to poor absorption, compounded by strong efflux transport and low recovery. Here we report a comprehensive analysis and modeling of public and industrial data on adsorption of organic molecules. Comparative analysis of one pharma-industrial chemical space was used to examine the relationship between critical permeability parameters. Our findings highlighted misconceptions in the transport route characterization. We demonstrated the importance of considering recovery, distribution coefficient, and topological polar surface area during Multi-Parameter Optimization. A Multi-Task Learning approach was employed for predictive model development. The models built on the public data were validated on the industrial data, revealing key discrepancies influenced by variation in experimental protocols. Our analysis emphasizes the model building on proprietary data in industrial absorption evaluations, which allows to avoid applicability domain issues and standardized measurement protocols. Finally, the integration of predictive models with Generative Topographic Mapping for chemical space exploration introduces a novel strategy to better understand optimization challenges. This work proposes a visual approach for MPO to improve drug discovery efficiency. The developed public models and curated public datasets are publicly accessible.

Introduction

The exploration of drug permeability constitutes an essential facet of pharmacokinetics and pharmacology, serving as a critical determinant in the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) profile of drug molecules. This parameter underpins several key aspects, including bioavailability and therapeutic efficacy of drugs within the human body[1]. Drug absorption in general refers to the ability of a compound to cross two different membranes: the intestine and the blood-brain barrier (BBB). This process is governed by both the drug's physicochemical attributes and the nature of the biological barriers it encounters. The small intestine, a principal site for oral drug absorption, boasts heterogeneous cellular architecture and an extended surface area due to the presence of folds, villi, and microvilli, factors that significantly influence absorption[2]. In contrast, the blood-brain barrier, which regulates brain uptake, consists of tightly joined endothelial cells that restrict the entry of most substances to maintain CNS homeostasis[3].

Biological membranes act as selective barriers, and their interaction with drugs is influenced by factors including the drug's molecular size, lipophilicity, and hydrogen bonding capacity, as elucidated by Lipinski's Rule of Five[4]. Additional factors such as molecular flexibility and the topological polar surface area are considered critical determinants for BBB absorption[5]. The process is even more complex considering the intricacies of different transport mechanisms. While small molecular entities (SME) permeate via passive diffusion and active transport, macromolecules require specific mechanisms such as pinocytosis and receptor-mediated endocytosis. This study, however, focuses on SME absorption, leaving the endocytosis transport outside of our focus.

In recent years, the focus has shifted towards employing computational methods for a more reliable and systematic approach to drug development. The evolution of *in vitro* and *in silico* models has greatly enhanced the understanding of drug permeability[6,7]. Particularly, machine learning and data analytics have been employed to analyze large datasets, paving the way for the development of quantitative structure–property relationships (QSPR)[8–17]. These advancements, however, are not without their limitations. Discrepancies in datasets across different labs and the lack of consensus on optimal modeling parameters emphasizes the need for continuous refinement in our understanding and prediction of drug permeability[18].

The aim of this study is to enrich our understanding of intestinal absorption data, contrasting between publicly available and industry owned sets. We systematically explore various aspects of drug absorption throughout the chemical space to pinpoint crucial molecular properties and substructures that govern permeation (Figure 1). Our approach involved analyzing in-vitro assays to identify discrepancies in existing permeability models and developing robust *in silico* models. We propose to use Multi-Task Learning (MTL) in this context, based on Graph Neural Network (GNN) and Generative Topographic Mapping (GTM). We report and analyze when this strategy leads to enhanced QSAR models and highlight synergies (and antagonisms) between datasets. To this end we gathered datasets expressing specific aspects of permeation: permeability, efflux ratios, and solubility. Pharmaceutical industry data have been used for validation of models trained on public data. This validation exposed limitations in public models that we analyze as mostly due to experimental variabilities across the data sources. By integrating our findings with GTM, we construct a comprehensive absorption landscape map that elucidates how specific absorption requirements can be achieved and highlights the inherent difficulties in meeting these criteria.

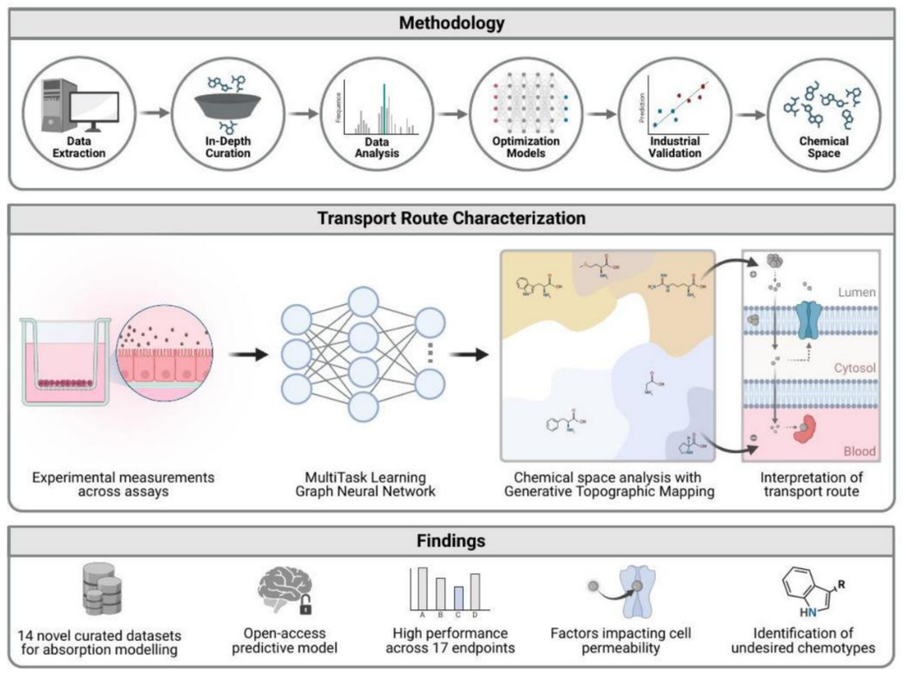

Methodology
Data Extraction
In-Depth Curation
Frequence
Data Analysis
Optimization Models
Prediction
Industrial Validation
Chemical Space
Transport Route Characterization
Lumen
Cytosol
Blood
Experimental measurements across assays
MultiTask Learning Graph Neural Network
Chemical space analysis with Generative Topographic Mapping
Interpretation of transport route
Findings
14 novel curated datasets for absorption modelling
Open-access predictive model
A B C D
High performance across 17 endpoints
Factors impacting cell permeability
R
HN
Identification of undesired chemotypes

Figure 1: Research workflow of the study. *Methodology*. Progressing from Data Extraction to Chemical Space Mapping through Analysis, Curation, and Machine Learning (ML) Integration. *Transport Route Characterization*. The predictive models embedding is analyzed using chemography approaches and interpreted for transport route. *Findings*. To resume our claim regarding permeability, we gathered new datasets, release new predictive models for 17 public endpoints, ensuring biological and chemical interpretability.

Background: Drug permeability

Understanding the mechanism of drug absorption is pivotal for optimizing drug efficacy. The absorption pathways across a cell monolayer include passive diffusion, active transport, paracellular, and transcellular transport, each contributing to the drug's overall permeability. Employing *in vitro* models' assays can provide valuable insights into these processes. However, the experimental conditions must be accounted to avoid misleading analysis interpretations[19].

Routes of passage

Paracellular and transcellular pathways represent the main route of drug absorption. While the transcellular passage is mainly used by lipophilic compounds such as propranolol, the paracellular is more relevant for hydrophilic molecules of low molecular weight (like mannitol). Drug transport is function of the membrane surface area and morphology. The lack of microvilli in cell-based assays reduce the available surface which hinder the paracellular transport of lipophilic compounds, requiring more time to be absorbed[20].

Active transport involves drug transporters. They can be categorized into uptake and efflux types and belong to two primary transporter superfamilies: the ATP-binding cassette (ABC) and the solute carrier (SLC and SLCO) families. These transporters can either facilitate or opposes the flux of molecules resulting of their concentration gradients, thereby influencing the permeability of the biological membranes, and ultimately, the bioavailability. Efflux transporters like P-glycoprotein (P-gP, MDR1 gene), and multidrug resistance proteins (MRP 1–6) can be considered as acting as a barrier. Found in tissues like the small intestine, colon, bile duct, and BBB, the P-gP pump prevents the diffusion of toxic compounds and drugs as, for instance, paclitaxel and etoposide[21,22].

In-vitro assay systems

One of the critical aspects of *in vitro* studies is the choice of the model used to estimate the permeability *in vivo*. Various cell lines such as Caco-2 (Cancer colon 2), MDCK (Madin-Darby Canine Kidney), and artificial membrane models like PAMPA (Parallel Artificial Membrane Permeability Assay) have been employed to mimic biological barriers[23,24]. PAMPA, introduced in 1998, mimics the intestinal epithelium using a hydrophobic filter usually coated with a mixture of lecithin and phospholipids[25]. Unlike cellular models, PAMPA focuses solely on passive diffusion. This model serves as a high through-put primary screening tool in the early stages of drug discovery[26–28].

The relationship between PAMPA permeability and lipophilicity, particularly LogD and LogP, is complex and subject to ongoing debate. While initial studies indicated a linear correlation between LogD and PAMPA, recent research suggests a parabolic relationship with optimal permeability at LogD pH 7.4 PBS ($LogD_{7.4}$) with values of 2.0 to 5.0 log[29]. This contrasts with studies emphasizing a strong link between PAMPA and LogP[30]. The current consensus, as illustrated by Avdeef, identifies PAMPA as a valuable tool for permeability assessment. PAMPA is depicted as correlated to Caco-2 while being much more rapid and cost-effective than Caco-2 and LogP[31]. However, comparing PAMPA or PAMPA-BBB with Caco-2 assays is considered crucial for identifying actively transported compounds.

Caco-2 cells, originally isolated from human colorectal adenocarcinoma, are widely considered the industry standard for studying intestinal drug absorption[23]. These cells spontaneously differentiate into enterocytes and form tight junctions, mimicking the human intestinal epithelial barrier in 21 days[25]. They exhibit both passive and active transport mechanisms[32]. However, certain limitations exist: the absence of villi formation and variable protein expression across laboratories[28,33]. An alternative is the Madin-Darby Canine Kidney (MDCK) epithelial cell lines, derived from canine kidney tissue. They are advantageous over Caco-2 for their shorter culture times of 3-7 days[24,33]. They are used to estimate epithelial transport with higher (MDCKI) or lower (MDCKII) monolayer resistance. However, their non-human origin brings specific challenges: the low activity of transporters expressed by MDCK such as the P-gP pump make them non valid for mechanistic studies[34,35]. Engineered MDCK lines exist to study specific transport mechanisms. MDCK-MDR1 overexpresses P-gP to mimic the BBB efflux, while MDCK-LE suppresses it, aiming at the assessment of passive permeability. MDCK-MDR1 and BBB lipid composition are close, making them good model to study Central Nervous System (CNS) permeability[33,36,37].

Absorption metrics

To sum up, absorption is measured on *in vitro* assays that models, schematically, either the intestinal or the brain epithelium. Absorption is quantified by three parameters that are measured in these assays: the apparent permeability (P_{app}), the efflux ratio (ER), and the recovery (Rec).

The experiment starts by injection of a concentration of drug on the *apical* (or donor) side of the membrane while *basolateral* (or acceptor) compartment drug concentration is null. It ends when the equilibrium concentrations of the drug on both side of the membrane has been reached (Figure 2).

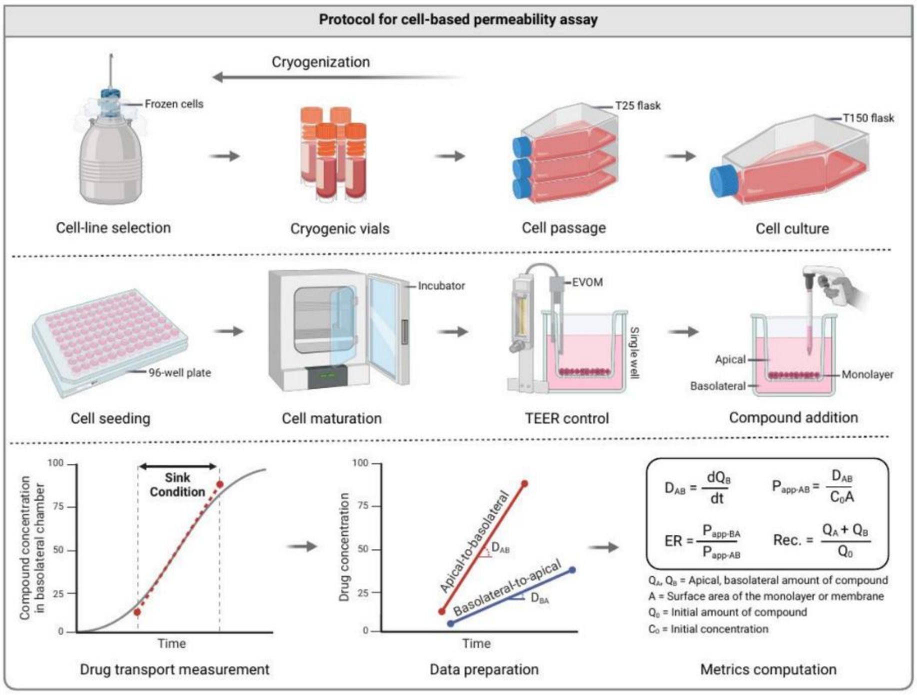

Protocol for cell-based permeability assay
Cryogenization
Frozen cells
Cell-line selection
Cryogenic vials
T25 flask
Cell passage
T150 flask
Cell culture
96-well plate
Cell seeding
Incubator
Cell maturation
EVOM
Single well
TEER control
Apical
Basolateral
Monolayer
Compound addition
Compound concentration in basolateral chamber
100
75
50
25
0
Sink Condition
Time
Drug transport measurement
Drug concentration
100
75
50
25
0
Apical-to-basolateral
D_{AB}
Basolateral-to-apical
D_{BA}
Time
Data preparation
$D_{AB} = \frac{dQ_B}{dt}$
$P_{app\text{-}AB} = \frac{D_{AB}}{C_0 A}$
$ER = \frac{P_{app\text{-}BA}}{P_{app\text{-}AB}}$
$Rec. = \frac{Q_A + Q_B}{Q_0}$
Q_A, Q_B = Apical, basolateral amount of compound
A = Surface area of the monolayer or membrane
Q_0 = Initial amount of compound
C_0 = Initial concentration
Metrics computation

Figure 2: Standard method for the measurement of *in vitro* permeability. The protocol presents the cell acquisition and culture, its differentiation, and the control of the Trans-Epithelial Electrical Resistance (TEER) of the monolayer until the addition of the compound. The readout is the measure of the compound concentration in the acceptor chamber over time. These readouts are interpreted to deduce the endpoint values.

Apparent permeability

The P_{app} quantifies the flux of the drug across the membrane, from the donor (apical) to the acceptor (basolateral) compartment, relative to the initial drug concentration and the surface area of the monolayer or membrane. For the measurement of permeability to be accurate and meaningful, it is essential to maintain sink conditions. These conditions ensure the drug concentration in the acceptor compartment remain negligeable compared to the donor compartment, emulating drug dispersion by blood flow in the body. This prevents drug accumulation on the acceptor side, allowing for uninterrupted transport and reliable measurement. The P_{app} is defined according to equation (1):

$$P_{app} = \frac{dQ}{dt} \cdot \frac{1}{C_0 . A} \qquad (1)$$

Where:

$\frac{dQ}{dt}$ is the quantity of drug transported through the membrane per surface unit of the membrane and time unit at given moment of the experiment,

$C_0$is the initial drug concentration on the apical (or donor) side,

A is the surface area of the monolayer or membrane.

Efflux ratio

The ER is a measure of the activity of efflux transporters, which pump xenobiotics out of cells and maintain a concentration difference between both side of the membrane. It is calculated as the ratio of the apparent permeability in the basolateral-to-apical direction (involving efflux transport) to the apparent permeability in the apical-to-basolateral direction (typically involving passive diffusion and possibly uptake transport). The equation for the efflux ratio is (2):

$$ER = \frac{P_{app,\ B-to-A}}{P_{app,\ A-to-B}} \qquad (2)$$

Where:

$P_{app,\ B-to-A}$ is the apparent permeability in the basolateral-to-apical direction,

$P_{app,\ A-to-B}$ is the apparent permeability in the apical-to-basolateral direction.

An ER>>1 suggests active efflux transport and an ER<<1 an active uptake. An ER~1 suggests a passive transport.

Recovery

Recovery is the amount of test compound that can be accounted for at the end of an experiment. It is calculated as the sum of the amounts of the compound in the apical and basolateral compartments, relative to the initial amount. The equation for recovery is (3):

$$Rec = \frac{Q_A + Q_B}{Q_0} . 100 \qquad (3)$$

Where:

Q_A is the amount of the compound in the apical compartment,

Q_B is the amount in the basolateral compartment,

Q_0 is the initial amount of the compound.

Recovery is used as a quality control of the assay. A Rec between 80% and 100% indicates that the assay performance is satisfactory, and that the compound is stable and not extensively bound to the assay apparatus.

However, several factors can influence the recovery rate. When Rec values are under 80 %, a significant amount of the tested drug is missing from the experiment, either because it has accumulated in the assay membrane or assay apparatus, or because the compound being degraded or metabolized. A Rec over 100 % can sometimes be observed. Such discrepancy can be indicative of metabolic process if readouts are obtained from LC-MS-MS or issue in the analytical part of the assay.[38,39] Researchers must be mindful of the intrinsic limitations of each model and metrics, as variability in permeability results from heterogeneous factors and assay conditions[19,40]. These factors range from laboratory-specific conditions to intrinsic cell line properties[41,42].

Limitations

Early-stage drug candidates are typically engineered with a focus on structure-activity relationship (SAR) targeting potency toward a protein target and a favorable pharmacological profile. This optimization often leads to candidates that are lipophilic and possess poor aqueous solubility, sometimes lower than 0.01 mg/mL.

Stock compound solubility

The low aqueous solubility is a major hurdle when evaluating these candidates in permeability models, especially those involving cell-based assays like Caco-2. These cells are sensitive to typical organic cosolvents like dimethylsulfoxide (DMSO) or propylene glycol (PG) and using them at concentrations above 5% compromises the integrity of the cellular tight junctions, whereas PAMPA assay can be used up to 10% DMSO. This complicates data interpretation: cosolvents may subtly deteriorate the membrane and, as the concentration of the tested compound rises in the comportments of the experiment (basal, apical or the membrane itself), it may de-solubilize and compromise its quantitative detection. This can render the data unreliable and impact mass balance recovery[25].

Non-specific binding

One of the most pervasive issues is non-specific binding to plastic devices and cells. It impacts the drug concentration in both the donor and receiver compartments, undermining the recovery and estimation of permeability. Labware and plastic binding mainly concern highly lipophilic drug candidates. Intracellular binding, or *ion trapping*, affects basic compounds by trapping them in lysosomes. This process is absent from artificial membrane assays[43]. In cell-based assays, compounds can display a high membrane retention and a low solubility, leading to a recovery loss. To counteract this effect, it has been demonstrated that including additives like Bovine Serum Albumin (BSA) or surfactant improves compounds' desorption from the membrane[44,45]. The rational for the presence of BSA in the acceptor compartment is to mimic the in-vivo environment where circulating blood induces sink effect. These conditions reduce the accumulation of lipophilic compounds in the cell monolayer. Lastly, BSA limits the adsorption on plastic surface and filters. Summing up, the addition of BSA improves recovery while maintaining the biological relevance of the experimental model membrane.

pH-partition hypothesis

In-vitro assays typically operate optimally at neutral pH, whereas the lumen to blood exhibits as 6.5/7.4 pH gradient[46]. This difference in pH levels is not trivial; it significantly influences the absorption behavior of both acidic and basic drug compounds. In-vitro permeability assay can either apply iso-pH (a pH of 7.4 in the apical and basolateral sides) or gradient-pH media (apical to basolateral pH from 6.5 to 7.4). The iso-pH setup simplifies the experimental design and models better the ileum absorption but deactivates Di/Tri peptide transporters (PEPT1). The gradient-pH setup is better to simulate general intestinal absorption but introduces 'false efflux'. The ionic charge difference between the basal and apical compartment decreases in bases absorption and increase in acid efflux[47].

Metabolic enzymes

Cell-based models express enzymes found at the brush border of enterocytes such as alkaline phosphatase, sucrase, and amino peptidase[33]. But they also express metabolic enzymes from Phase I (CYP P450) and II (hydrolases, carboxylesterases, and uridine diphosphoglucuronosyl transferases).

These enzymes play a significant role in metabolizing solutes during permeability measurements, impacting the mass balance and thus, the measured permeability[21]. Yet, they can't be considered as a standardized metabolism model as enzyme expression is variable due to dependence upon media pH and number of passages. For instance, PEPT2 is inactive in iso-pH assay and CYP isozymes are under-expressed in Caco-2 cells in comparison to human tissues[28].

Xenobiotic transporters

As compounds are screened using µM range concentration, active efflux transport may not be limited. Whereas *in vivo* the concentration at site can reach mM range, allowing the blockage of transporters and making the efflux an *in vitro* artifact. This concerns high dosage compound with low toxicity. In the other case, the evaluation of efflux is crucial but can be limited[48]. The measurement of ER depends on the quantity of transporters expressed and active by the cells. But this is highly variable as illustrated in inter-laboratory studies comparing Caco-2 permeability data from multiple teams. This contributes to the variability of the efflux measurements and foster standardization approach to mitigate discrepancies in permeability[7,9,29,49]. Efflux activity can be determined by comparing compound permeability with and without specific P-gP inhibitors, such as verapamil or cyclosporine A. This approach seeks to reduce variability of efflux measurements.

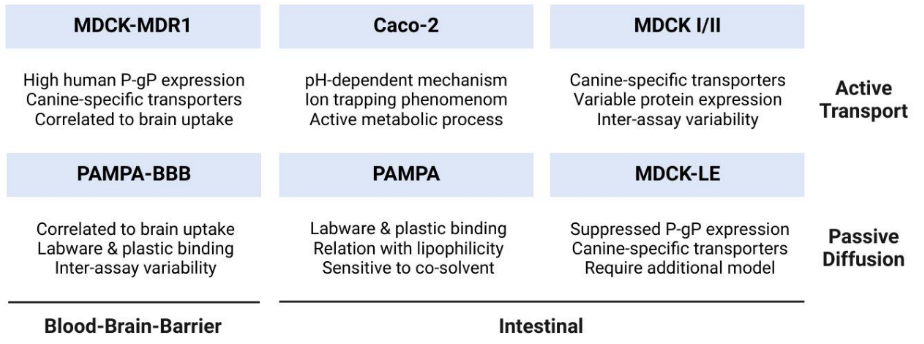
Key concepts of permeability measurement

The above discussion is summarized below. As it is based on an artificial membrane, PAMPA are certainly the more reproducible. But it is also the less biologically relevant assay. On the other hand, Caco-2 and MDCK based assays are impacted by false efflux, variations in the expression of metabolic enzymes and transporters, accounting for reproducibility issues. The permeability measure is influenced also by the solubility of the tested compound and its sensitivity to pH (Figure 3).

Interpretation and sanitization of permeability data must consider:

- **Minimize cell line variability for standardization**: To enhance data reliability, the passage numbers, variability in protein expression and transporter activity should be minimized.
- **Verify recovery to ensure data integrity**: Mass balance and recovery rates in assays can identify issues with compound stability, noise, adsorption, or metabolism.

- **Incorporate transporter inhibitors to evaluate active efflux**: Applying transporter inhibitors can reveal the contribution of active efflux to the overall permeability. It is not recommended to fuse datasets prepared with differences in such treatments.
- **Limit co-solvents for monolayer integrity**: Addition of co-solvent (DMSO) to leverage solubility issues can deteriorate the cell integrity and the permeability measurement. For this reason, data acquisition in the presence of DMSO should be considered separately.
- **Correctly choose the permeability model:** PAMPA mimics an artificial passive diffusion, Caco-2 models intestinal permeability, and MDCK, the passive diffusion or BBB transport. Although they are all labeled as permeability data, they should not be confused nor be fused.
- **Verify the presence of BSA limiting non-specific binding:** Addition of BSA limits the loss of solute through binding to plastic devices, cells, or ion trapping while emulating sink conditions. This is a source of variability observed in permeability values for a given compound.
- **Controlled pH conditions for standardization**: Difference between the apical and basolateral pH influences the absorption process by modulating the activity of xenobiotic transporters. Therefore, the use of a buffer can be another source of variability.
- **Reduced solute concentration to avoid transporter saturation:** *In vitro* cell assays use high compound concentration, saturating transporters and masking the presence of active efflux. Therefore, it is not possible to fuse data prepared using largely differing concentration ranges.



MDCK-MDR1	Caco-2	MDCK I/II	
High human P-gP expression Canine-specific transporters Correlated to brain uptake	pH-dependent mechanism Ion trapping phenomenom Active metabolic process	Canine-specific transporters Variable protein expression Inter-assay variability	Active Transport
PAMPA-BBB	**PAMPA**	**MDCK-LE**	
Correlated to brain uptake Labware & plastic binding Inter-assay variability	Labware & plastic binding Relation with lipophilicity Sensitive to co-solvent	Suppressed P-gP expression Canine-specific transporters Require additional model	Passive Diffusion
Blood-Brain-Barrier	Intestinal		

Figure 3: Factors influencing the reproducibility of permeability measurements in permeability assays.

Considering the impact of all experimental parameters, data curation leads inevitably to fragmentation of an initial dataset of permeability to many related smaller subsets, which are more homogeneous and meaningful both biologically and chemically. This motivates the use of specific machine learning approaches to maximize the benefit of these data, namely, multi-task learning algorithms.

Materials & Methods

Data source

Public experimental values were recovered from three public databases: OChem[50], ChEMBL[51], and BindingDB[52] and used for analysis and validation hereafter. Measurement regarding apparent permeability, recovery, efflux ratio, solubility, hydration free energy, plasma protein binding (PPB), partition coefficient (LogP), distribution coefficient ($LogD_{7.4}$), brain distribution (LogBB), and P-glycoprotein inhibition (pIC50 P-gP) data were collected (Table S1). For data availability reasons, we focused on measures obtained from PAMPA, Caco-2, MDCK, MDCK-LE, and MDCK-MDR1 models. Additional apparent permeability, distribution, and recovery data were sourced from Sanofi, referred to as *Industrial* but not included in the published dataset. Apparent permeability measurements were converted to cm/s and transformed to a base 10 logarithmic scale for modeling purpose[53]. As the efflux ratio spans multiple orders of magnitude, the values were converted to a base 10 logarithmic scale.

Hydration free energy (HFE) quantifies the energy change when a compound transfers from the gas phase to an aqueous environment, indicating its solubilization capacity in water. In the context of this study, HFE is included for two reasons. First, the HFE dataset remains consistent across both academic and industrial segments, allowing for direct comparison of predictive models using identical compounds. This consistency is essential for robustly evaluating model performance across different datasets. Second, HFE functions as a neutral endpoint concerning permeability, providing a neutral reference to examine interactions between tasks in a MTL modeling approach. This inclusion allows assessment of whether MTL models affect performance on an endpoint theoretically independent of permeability, ensuring that any observed synergies or antagonisms are genuine interactions rather than artifacts of the modeling technique. Thus, integrating HFE into the study is crucial for evaluating the hypothesis that MTL models do not inherently improve or degrade performance for endpoints like HFE compared to STL models.

Data curation & standardization

Assay sources were verified by checking methods of measurement (Table S1) to model endpoints with standardized assay across the dataset. Then, for each dataset, structures were standardized using Chemaxon's Jchem (ChemAxon. Jchem Base, version 22.19.0 (2022)) software by removing the salts, removing stereochemistry, removing the aromaticity before recomputing it, ionizing the structure at pH 7.4 and selection of a standard tautomer. After standardization, several chemical structures appeared duplicated because of merging different data sources and ignoring the stereochemistry. In case of duplicated entries, the standard deviation (SDi) (4) and median of the experimental values were computed (Table S2, Figure S1-4). Compounds with a standard deviation exceeding 0.5 log units for apparent permeability and solubility, or 5% for recovery, were excluded. For the remaining compounds, the median value was defined as the response value when considering more than two measurements; otherwise, the worst value among the two measurements was used. The resulting dataset is called AbsSolDB (Figure 4)[17].

$$SDi = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (4)$$

Where:

- x_i denotes each individual observation for given duplicated compound.
- $\bar{x}$ represents the arithmetic mean of duplicated observations.

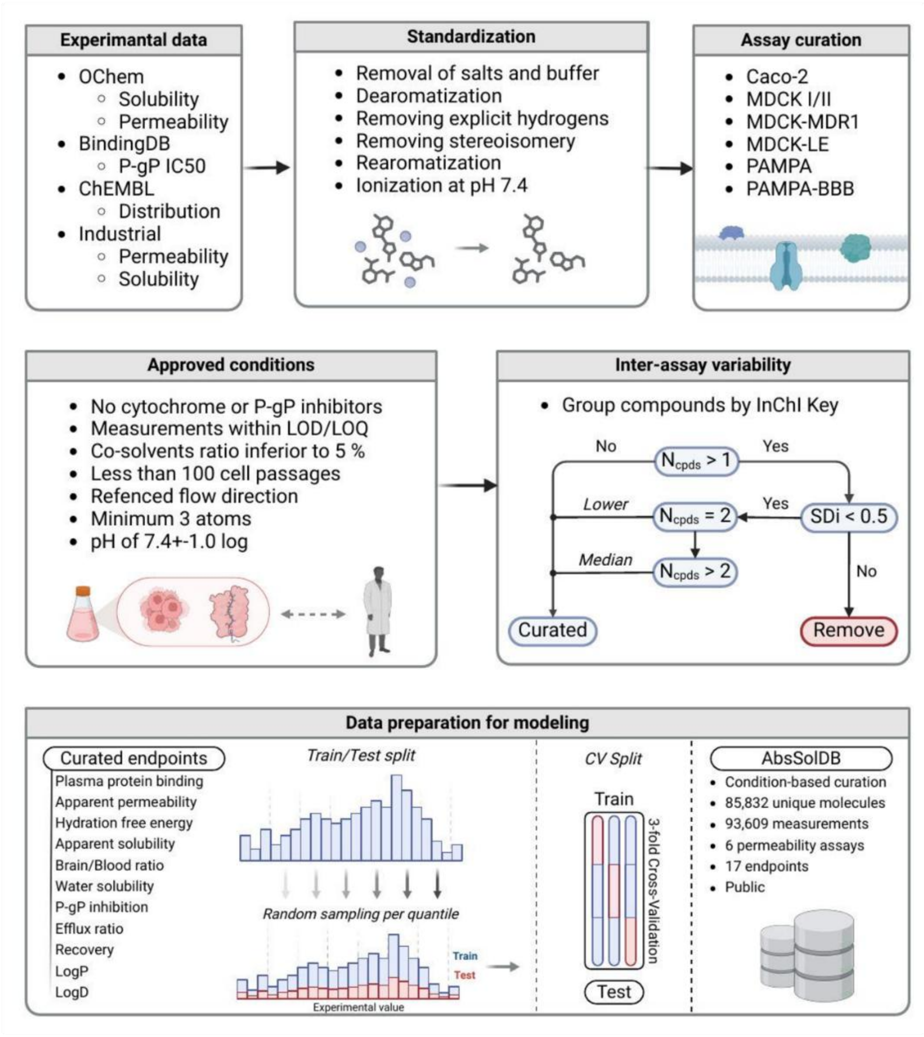


Figure 4: Flowchart describing the guidelines followed from compound standardization to data preparation for modeling. Chemical structures are standardized and ionized using Chemaxon tools. Experimental meta-data are retrieved and used to filter assays by cell-lines, experimental conditions, and presence of inhibitors. When several thermodynamic solubility values are available, an entry is discarded if there is a doubt about which value to keep; otherwise, the median or the smaller value is conserved.

Physico-chemical and structural descriptors

To ensure the performance consistency across tasks in Single Task Learning (STL) models, we evaluated a broad spectrum of descriptor types during model optimization. This approach aimed at exploring diverse molecular representations, thereby better assessing the limit of models. Descriptors' calculation was based only on the 2D structures, justifying that stereoisomers information was ignored. Physicochemical properties, and Morgan fingerprints (ECFP 4, and 6 of 1024 and 2048 bits) were calculated using RDKit[54]. Fragment-based descriptors were computed using ISIDA[55]. Different combination of fragmentation type ('-t 0 -t 3/6/9'), fragment length ('-l2 -u 3/4/5'), and option ('','—UseFormalCharge', '—DoAllWays', '--UseFormalCharge –DoAllWays') were tested. In total, 36 descriptors set were prepared per dataset. Diverse CDK fingerprints were generated (LingoFP, SigFP, GraphFP, ExtFP, PubchemFP, SubFP, AP2DFP, KRFP), and MORDRED 2D descriptors were also prepared. In total, 50 different descriptor sets proposing diverse approaches to represent compounds were obtained.

Modelling approach

Predictive in silico models

Data Partitioning Scheme

The MTL approach involves certain tasks sharing the same compounds, necessitating their simultaneous presence in both training and testing datasets. This requirement, coupled with the sometimes-low effective counts for some tasks, imposes a significant constraint on the composition of cross-validation. Given the sparse data in certain tasks, a conventional random split would not suffice as all tasks were accounted at once. To overcome this, we implemented a quantile-based splitting methodology. This approach was preferred as tasks are processed in parallel even though the datasets are not independent, as they share certain compounds. It's crucial to acknowledge this interconnectedness when composing training and testing datasets to ensure that the model is trained on a representative sample of the data. This technique involves dividing the data distribution of each task into quantiles. We then randomly sample without replacement a proportion of data from each quantile, corresponding to the desired training/testing split ratio. This approach not only maintains the diversity of the dataset but also ensures an equitable representation of data points, particularly tasks with fewer measurements. By splitting all tasks at once, train and test set do not share structures between two different tasks, thus preventing a possible data leakage (Figure 4, S5).

The implementation of this scheme is such that each dataset is partitioned into two subsets: a training subset, constituting 80% of the total data, and a testing subset, comprising the remaining 20%. Within the training subset, we perform further stratification to obtain sets for 3-fold Cross-Validation (CV). Models optimal hyperparameters are optimized during CV by repeatedly training on two sets and testing on one. This approach is crucial for enhancing the stability and generalization of our predictive models[56]. For the validation phase, the model is retrained on the training subset using the optimal hyperparameters, methods, and descriptors identified during CV. This data partitioning scheme is applied to the public and industrial data. The retrained models are then evaluated on the public test set and industrial test set, defined here as an external set. The industrial test is orthogonal to the public test. Different modeling approaches have been applied.

Machine learning algorithms

Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees during training and outputs the average prediction of the individual trees for regression tasks. This approach provides a robust mechanism to handle noisy data and controls overfitting[57].

Support Vector Regression

Support Vector Regression (SVR) applies the principles of Support Vector Machine (SVM) to fit a hyperplane in a high-dimensional space in a way that minimizes the error within a certain threshold. SVR is known for its outlier's robustness, model's stability, and reliable modeling of complex non-linear relationships[58].

Generative Topographic Mapping

Generative Topographic Mapping in an unsupervised dimensionality reduction algorithm. GTM introduces a two-dimensional hypersurface, or manifold, into the initial high-dimensional data space, typically characterized by N dimensions. The objective of GTM is to mold the manifold so that it accurately reflects the distribution of the dataset. This fitting process is achieved through the Expectation Maximization (EM) algorithm, which diligently works to minimize the log-likelihood of the training data. Upon completion of this fitting phase, each data point is projected onto a two-dimensional latent grid composed of K nodes. Within this latent space, data points are represented by vectors of normalized probabilities, a.k.a responsibilities. These responsibilities essentially quantify the association of each data point with the nodes on the manifold's grid[59]. GTM was employed to visualize the chemical space landscape, focusing on two key aspects: the density of data points within this space and the spatial distribution in function of specific properties. In the density landscape, each node cumulates the responsibilities from all projected compounds. This allows the identification of areas with high and low density. GTM property landscape were used as a tool to explore a predefined space, allowing for the analysis of multiple parameters simultaneously[60–62].

Chemical content analysis of the map

Interesting regions of a Generative Topographic Map (GTM) are analyzed using the BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) approach[63]. This method involves decomposing molecules into meaningful substructures based on retrosynthetic principles, which is useful for identifying structural alerts. BRICS fragments are generated by applying rules to break the molecule at specific bonds, primarily involving heteroatoms or functional groups, ensuring the fragments are chemically feasible and relevant.

To analyze GTM regions, we first select a region from the public map and extract public compounds with a responsibility greater than 0.05 to its occupied nodes. For each of these compounds, BRICS fragments are generated, creating a list of structural patterns. These patterns are then counted to determine their frequency within the region. The regions are annotated based on these fragments, arranged in decreasing order of frequency, highlighting common structural features that may impact permeability.

Graph Neural Network

Graph Neural Networks are neural networks designed to work with graph-represented data[64]. Two architectures based on GNN were employed for this study:

- ChemProp focused on learning molecular representations. It uses Message-Passing Neural Networks (MPNN) to aggregate information from the local chemical environment of atoms within a molecule[65].

- AttentiveFP incorporates attention mechanisms to focus on relevant parts of the molecular graph. The aim is to focus on substructures and relations between substructures within a molecule[66].

Parameter optimization and model selection

Hyperparameters tuning was executed using Sequential Model-Based Optimization (SMBO)[67]. SMBO is an application of Bayesian Optimization characterized by its methodical approach to updating the probability model in a sequential manner. Each evaluation of the objective function using a specific set of values informs an update to the model, underpinning the concept that, over time, the model will progressively converge to accurately represent the true objective function optimum.

In this case, the objective function is defined as the Root Mean Square Error (RMSE) of model predictions—over a predefined parameter space (Table S3). For Generative Topographic Mapping, a SMBO was used as Bayesian Optimization method[68]. The free parameters of each machine learning method used, leading to models exhibiting the lowest RMSE on the internal test set, was identified as optimal.

Performances metrics

To assess the performance of our regression models, we employ the coefficient of determination (R2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) per task. R2 (6) represents the goodness-of-fit of a model. It reveals how much of the variance in the dependent variable is captured by the independent variables. It ranges from 0 to 1, where a value closer to 1 indicates better model fit; thus, it is independent of the units in which a given task is expressed.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y}_i)} \qquad (6)$$

Where:

n is the total number of observations,

$\hat{y}_i$ is the predicted value for the i-th observation,

y_i is the actual value for the i-th observation,

$\bar{y}_i$ is the mean of the actual values y.

RMSE (7) quantifies the difference between the predicted and observed values per task, penalizing larger errors more severely by squaring them before averaging. The RMSE is expressed in the units of the given task and is not too sensitive to the composition of the dataset used to compute it - in contrast to R2.

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{7}$$

Where:

n is the total number of observations,

$\hat{y}_i$ is the predicted value for the i-th observation,

y_i is the actual value for the i-th observation.

MAE (8) represents the accuracy of a regression model to a certain task. Compared to RMSE, MAE is less sensitive to outliers or large errors. It is an arithmetic average of absolute errors.

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n} |y_i - \hat{y}_i| \tag{8}$$

Where:

n is the total number of observations,

$\hat{y}_i$ is the predicted value for the i-th observation,

y_i is the actual value for the i-th observation.

MSE (9) measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. It provides a measure of the quality of a predictive model.

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

These metrics are computed for each task, providing a comprehensive measure of the model's performance across various aspects of the data.

Loss function

The MTL models adopt a loss function based on the Mean Squared Error (MSE). Our multi-task loss function is a weighted sum of the MSE for each task (11). We opted for fixed weights per task to compute the multi-task loss function (10). Weights values are normalized to ensure that their sum equals 1. As each task has its own units, the weights represent the different scales covered by these units, maintaining a consistent range of values across all tasks. This ensures a balanced contribution of each task to the overall loss calculation and effectively outputs real values.

The weights are estimated as follows:

$$w_i = \frac{\frac{1}{|\max(values_i) - \min(values_i)|}}{\sum_{j=1}^{n} \frac{1}{|\max(values_j) - \min(values_j)|}} \qquad (10)$$

Where:

w_i is the initial weight for the i-th task,

values$_i$ is the range of experimental values of the i-th task,

n is the total number of tasks.

The multi-task loss function then expresses as:

$$Loss = \sum_{i=1}^{n} MSE_i * w_i \qquad (11)$$

where MSE_i is the Mean Squared Error of the prediction against experimental values for the i-th task.

Ensemble modeling

The consensus prediction was employed to combine the predictions of multiple models into a unified one. The consensus prediction $\langle y \rangle$ is the arithmetic average of predicted values $\hat{y}_i$ by each individual i model. This allow to also compute a standard deviation of the predicted values that is used as a contribution to prediction uncertainty.

Applicability domain

The Applicability Domain (AD) of a predictive model represents a specific area within the Chemical Space (CS) that is well-described by the model's features, effectively marking the boundaries within which the model's predictions are considered reliable. We defined it as a region based on the proximity and similarity of new compounds to those in the training set. We expect to control that predictions are made on compounds for which the model is relevant.

As each task has a specific dataset, we have defined an AD for each task. As MTL models output predictions for all tasks at once, some of them are filtered out, if the query compound does not belong to corresponding applicability domain.

We used One-Class Support Vector Machine (OcSVM)[69,70] to define the applicability domain. OcSVM is used to define a boundary around the training set compounds. In AD determination, OcSVM is used to identify test data points outside that boundary, and thus a prediction about that data point is unreliable. We set the kernel to Radial Basis Function (RBF) with a gamma value of 10^{-5}, which adjusts the expressivity of the kernel to describe the decision boundary. The ν parameter, which balances the model's sensitivity to outliers, was optimized during AD development.

Absorption Score (AS)

Scoring compounds to identify potential leads within a given chemical space is a key strategy in drug development. We introduce an Absorption Score (AS) to reflect the desirability of a compound from an absorption perspective, integrating solubility, efflux, and permeability into a single metric. The AS aims to pinpoint regions within the chemical space that consistently exhibit undesirable absorption characteristics, providing higher confidence in the assessment.

To calculate the AS, solubility, permeability, and efflux measurements are first discretized into "desirable" and "undesirable" groups based on literature-derived thresholds (see Table S2). Each endpoint is analyzed on a GTM of the chemical space, producing class landscape colored with a “desirability” scale from 0 to 1. Merging these landscapes generates consensus desirability landscapes, which combine endpoints into a unified view.

For instance, the consensus solubility landscape aggregates kinetic, apparent, and water solubility data. The consensus score (Sc) (12) for each node in these landscapes is computed as the weighted sum of the individual endpoints, with each endpoint modulated by its respective sum of responsibilities within the node. The formula for the node desirability score is:

$$Sc_{node} = \frac{\sum_{i=1}^{n}(St_{node,i} * Rt_{node,i})}{\sum_{i=1}^{n} Rt_{node,i}} \qquad (12)$$

where $St_{node,i}$ represents the predicted desirability score of the *ith* endpoint on a given node of the map, and $Rt_{node,i}$ denotes the corresponding responsibility. This approach emphasizes the impact of each endpoint based on its relative importance, as indicated by its responsibility on a given region of the chemical space.

Consensus landscapes were computed for solubility (kinetic, apparent buffer, water), efflux ratio (Caco-2, MDCK-MDR1), and apparent permeability (PAMPA, Caco-2, MDCK, MDCK-MDR1, MDCK-LE). These consensus landscapes enhance the reliability of mapping experimental values within certain regions. To facilitate this multiparametric visualization we propose to merge the three consensus landscapes, defining the Absorption Score *(13)*.

The AS provides a nuanced analysis of the chemical space, integrating solubility, permeability, and efflux in one metric. The formula for the AS is:

$$Absorption\ Score = \begin{cases} Sc_{Papp} & if\ Sc_{Papp} > 0.5 \\ \dfrac{Sc_{Efflux} - Sc_{Solubility}}{2} & \end{cases} \qquad (13)$$

The AS ranges from -0.5 to 1. AS from 0.5 to 1 indicates acceptable permeability, for which the impact of any undesirable efflux or solubility issues is not considered. In the other case, if AS is comprised between -0.5 and 0.5, this indicates a suboptimal permeability (Sc_{Papp} < 0.5). This sub-range helps identify whether the permeability issues arise from efflux or solubility:

- **AS tends to -0.5**: Indicates that the compound has desirable solubility properties but undesirable efflux.
- **AS tends to 0.5**: Indicates that the compound has desirable efflux properties but undesirable solubility.
- **AS close to 0**: Suggests that both efflux and solubility are equally problematic, or that other factors are contributing to the poor permeability.

By emphasizing regions where permeability is a limiting factor, the AS facilitates the identification of problematic areas within the chemical space.

Results

Data characterization

To better estimate the relationship between the different endpoints, we conducted dataset analysis focusing on structural redundancy and correlations between tasks (Figure 5).

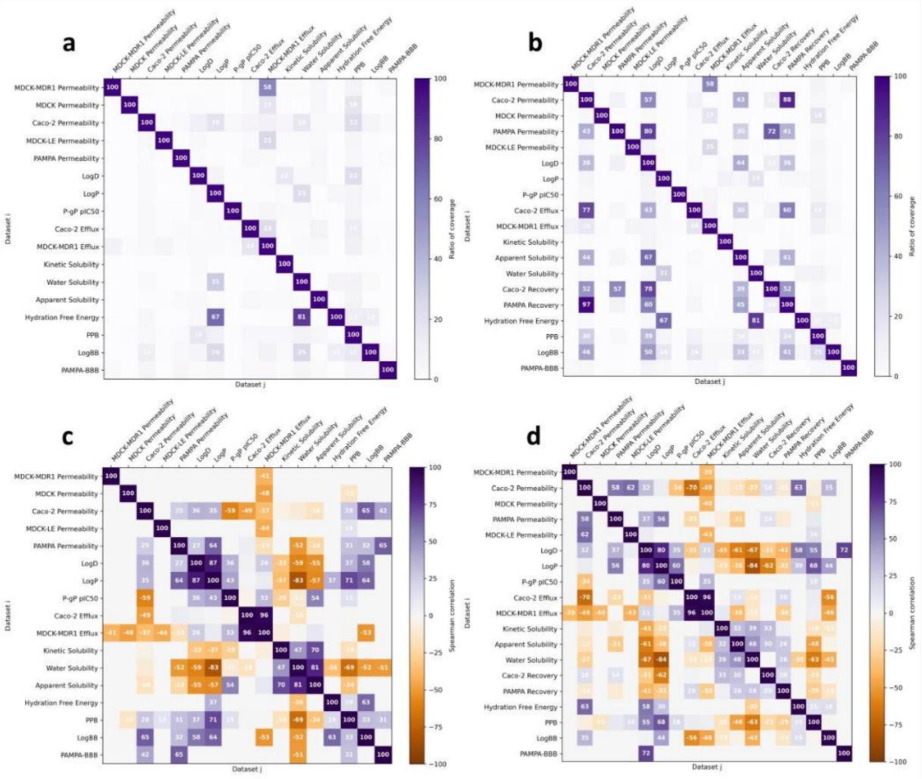


Figure 5: Redundancy and correlation matrices as hierarchical heatmaps. Redundancy matrices are asymmetric. Bottom-left section of the matrices represents the X_{ij} ratio, of chemical compounds from the i^{th} data set within the j^{th} data set. Top-right represents X_{ji} ratio, of chemical compounds from the j^{th} data set within the i^{th} data set. (a) Public. (b) Industrial. Spearman correlation P_{ij} matrices are

symmetric computed using experimental values from shared chemical compounds between the ith dataset with the jth dataset. (c) Public. (d) Industrial.

Public

Public datasets include few shared structures between the endpoints (Figure 5a). Large coverage concerns MDCK-MDR1 permeability with efflux, and hydration free energy with water solubility and LogP. This limited redundancy suggests a diverse yet fragmented chemical space from public data. The correlation analysis (Figure 5c) further elucidates this point. Despite the low coverage, strong correlations are observed, particularly in the inverse relationship between LogP, LogD and solubility, and the positive correlation among MDCK, Caco-2, and PAMPA permeability, which are correlated to efflux.

Industrial

In contrast, industrial data (Figure 5b) exhibit greater coverage. This is particularly evident for LogD$_{7.4}$, apparent solubility, and permeability assays. Weak coverage is observed for LogP, P-gP pIC50, and PPB, aligning with the industry's focus on solubility, permeability, and LogD$_{7.4}$. This illustrates the more systematic research implemented in drug discovery within industry to characterize compounds of interest. The correlation matrix (Figure 5d) shows consistent patterns but with weaker correlations in solubility: differences between the various types of solubilities is more evident.

Comparative insights

Despite their wide-ranging nature, public datasets exhibit significant variations in chemical space across different endpoints, stemming from their compilation from numerous studies. Yet, their strong correlations suggest intrinsic links between the tasks. Industrial data include multiple high coverage and positive/negative correlations between tasks.

Data analysis

To elucidate the links between endpoints, we investigated the dependencies between tasks with strong positive or negative relations (Figure 6).

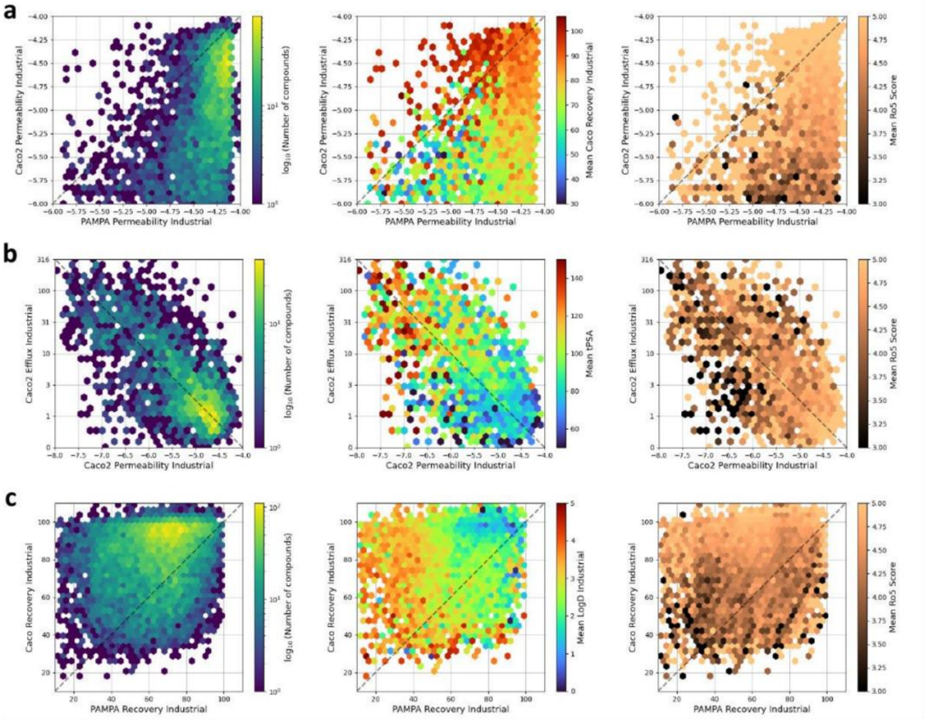


Figure 6: Correlation between absorption endpoints. The color code stands for the density (left column), Lipinsky score (right column) and, in the middle column recovery Caco-2, tPSA and $\text{LogD}_{7.4}$. **(a)** Caco-2 permeability against PAMPA permeability **(b)** Caco-2 efflux against Caco-2 permeability. **(c)** Caco-2 recovery against PAMPA recovery.

Permeability from PAMPA to Caco-2

In comparing apparent permeability between Caco-2 and PAMPA models, it is observed that Caco-2 permeability is generally lower. Literature[71] indicates passive diffusion for compounds adhering to the x=y trend and suggests the involvement of active efflux mechanisms for instances where Caco-2 << PAMPA. Conversely, instances of higher Caco-2 permeability than PAMPA is suggested to be characterized by an uptake process.

Efflux & permeability from Caco-2

Analysis of the Caco-2 apparent permeability against the efflux ratio revealed a strong inverse correlation, emphasizing molecules with a high efflux ratio typically have a topological Polar Surface Area (tPSA) above 120 Å^2. This observation is in line with Lipinski's tPSA criterion to mitigate efflux effects[72].

Recovery from PAMPA to Caco-2

Investigation into the recovery rates between Caco-2 and PAMPA models unveiled that higher recovery in Caco-2 compared to PAMPA often correlates with increased permeability, potentially highlighting the role of metabolism or experimental noise. Specifically, anomalies in Caco-2 recovery exceeding 100% may hint at an overestimation of measured permeability, whereas recoveries below 60% could indicate its underestimation due to non-specific binding or metabolism. The comparative analysis between Caco-2 and PAMPA requires using recovery to elucidate this possible bias.

Recovery & permeability from PAMPA & Caco-2

A focused study on the correlation between recovery and apparent permeability across both Caco-2 and PAMPA models revealed a complex relationship. Optimal Caco-2 was found in compounds with a tPSA below 120 Å^2 and a LogD$_{7.4}$ below 3 log, aligning with Lipinski's rules[4]. The PAMPA model, however, displayed a nuanced relationship where high LogD$_{7.4}$ compounds often showed increased permeability but with potentially reduced recovery due to non-specific binding, especially in compounds exhibiting high lipophilicity (LogD$_{7.4}$ > 3) (Figure S6). This pattern is indicative of potential issues with non-specific binding in the PAMPA assay, particularly for compounds exhibiting high lipophilicity.

Our analysis further reveals that compounds achieving favorable permeability values in PAMPA frequently may display an unfavorable profile when assessed with more biologically relevant assays as exemplified with Caco-2. These observations tend to suggest that PAMPA could be a replacement for LogD$_{7.4}$ measurements and vice versa. Specifically, high P$_{app}$ values in PAMPA tend to favor compounds with elevated LogD$_{7.4}$ values, which, according to Caco-2 data, are often associated with lower recovery rates. LogD$_{7.4}$, known for its high reproducibility, emerges as a crucial indicator of PAMPA recovery in our study. Particularly, we observed that when LogD$_{7.4}$ is equal to or greater than 3, the recovery rates in PAMPA assays often fall below 60%. This finding underlines the enduring relevance of LogD$_{7.4}$ as a reliable measure of hydrophobicity. Coupled with Caco-2 permeability and tPSA data, $\text{LogD}_{7.4}$ forms a robust triad of parameters, facilitating effective multi-parameter optimization in ADMET profiling.

Application of predictive models

Here we examine the efficacy of MTL in comparison to other ML methods. Each model was trained using the same train set and the best parameters identified following CV (Table S4). Their performances were then assessed on both public test sets and external industrial sets. The prediction from the best performing models on the test set are presented (Figure 7). Particularly noteworthy is the consistently superior performance of MTL and RF-STL algorithms on the public test set (Figures S7-9).

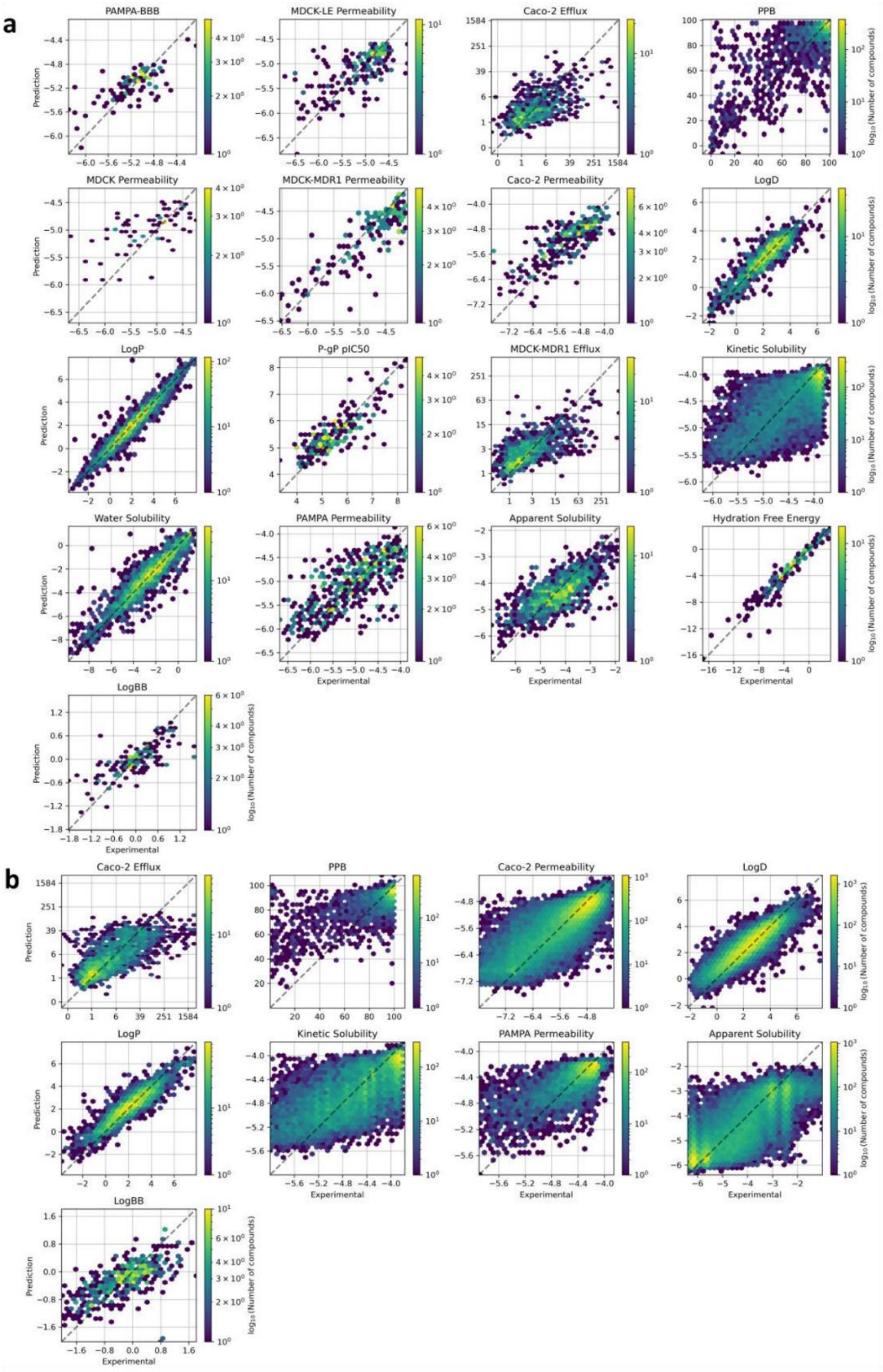

a
PAMPA-BBB
MDCK-LE Permeability
Caco-2 Efflux
PPB
MDCK Permeability
MDCK-MDR1 Permeability
Caco-2 Permeability
LogD
LogP
P-gP pIC50
MDCK-MDR1 Efflux
Kinetic Solubility
Water Solubility
PAMPA Permeability
Apparent Solubility
Hydration Free Energy
LogBB
b
Caco-2 Efflux
PPB
Caco-2 Permeability
LogD
LogP
Kinetic Solubility
PAMPA Permeability
Apparent Solubility
LogBB
Prediction
Experimental
log$_{10}$ (Number of compounds)

Figure 7: Correlation between experimental and predicted properties. (a) Public test (b) Industrial test. The coloration depicts the density of compounds as the base-10 logarithm of the number of unique compounds. MTL: MTL-GNN using ChemProp; GNN: ChemProp STL; ATT: AttentiveFP STL; RF: RandomForest STL; SVR: Support Vector Machine STL; GTM: Generative Topographic Mapping STL.

Analysis of the model performance on the (industrial) external validation set (Figure 7b) revealed good agreement with experimental data overall but failed for certain endpoints: Caco-2 permeability, PAMPA permeability, efflux, and LogBB. We formulated three hypotheses to interpret these inconsistencies. They may stem from (i) chemical space differences between datasets, (ii) experimental condition variability, or (iii) data quality.

Employing an ensemble approach, which involves averaging the predictions of all models for each compound to improve accuracy and robustness, highlighted collective failures in predicting apparent solubility and Caco-2 permeability (Figure S10). Collective failures define the consensus of predictions from multiple models systematically failing, resulting in high variability in consensus prediction errors for these two endpoints. This indicates that compounds were consistently predicted incorrectly across the models for these specific endpoints.

While public and industrial data share the same “apparent solubility” and “Caco-2 permeability” labels, they correspond to different experiments. Industrial data do use BSA and efflux computation is different from public data (Table S1). Apparent buffer solubility from public datasets uses other protocols compared to our industrial data. Using industrial data to develop MTL models (Table S5) showed good performance on an external industrial dataset (Figure 8). Models trained on industrial data appear to be more robust, probably reflecting the more systematic approach in data acquisition. The issue is therefore not much into the quality of the public datasets, but that they do not compare to industrial data.

s

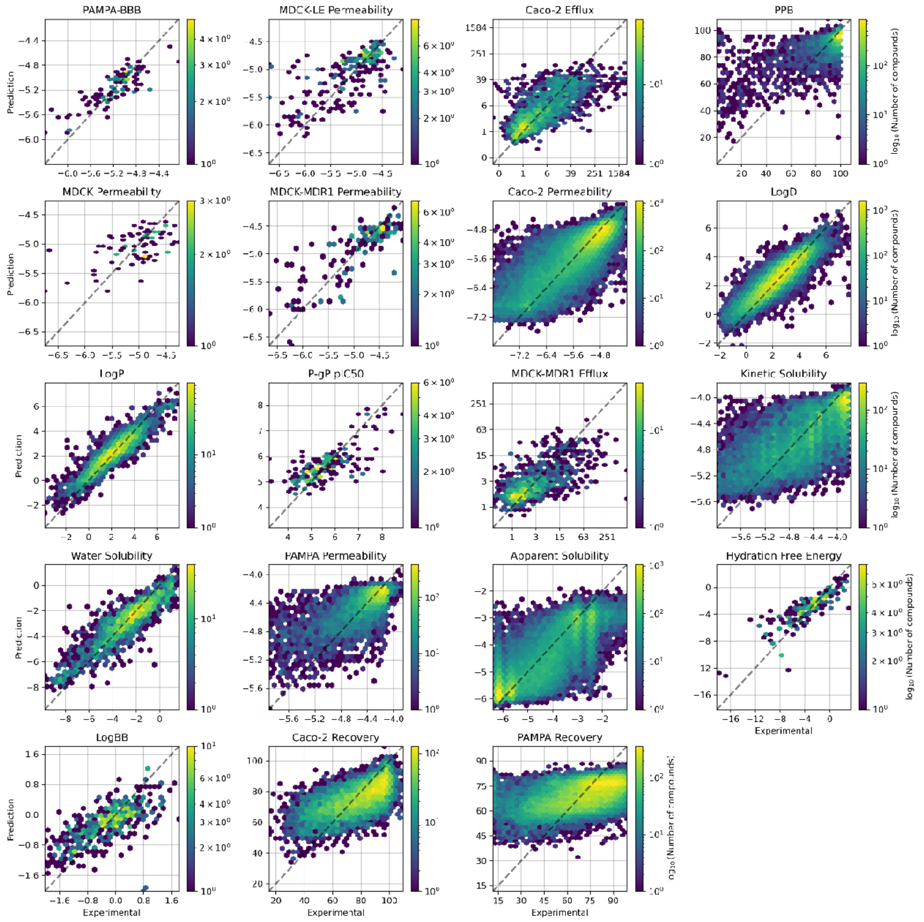


Figure 8: Correlation between experimental and predicted properties from industrial model on industrial test set. The coloration depicts the density of compounds as the base-10 logarithm of the number of unique compounds.

Applicability domain

In this study, we built Applicability Domain (AD) models utilizing latent space representations of compounds, derived from the public MTL-GNN model. This model was selected for its adept representation of the drug absorption space and reliable performances.

For the AD models, the One-Class Support Vector Machine (OcSVM) was employed. These models were trained on latent vectors (LV) representation of the training dataset and tested on the LV representation of the test set. Each endpoint uses a different OcSVM applicability domain, with its own

ν parameter value. The ν parameter measures the partition of data that is out of the AD. By default, it has been set to 0.2. For some endpoints, we observed that the in-AD performances did not change when increasing this value, while for others, a plateau was achieved for larger values. These observations are compiled in the Table 1. The optimized models were subsequently applied considering AD to both public and industrial datasets.

Few endpoints, including PAMPA-BBB, MDCK-LE permeability, PPB, and HFE, were significantly influenced by the AD. The RMSE improvements for PAMPA-BBB and HFE can be attributed to low amounts of data and sparse coverage of the chemical space. Models trained on public data were then applied to industrial data with and without AD. In most cases, the application of AD resulted in an improved RMSE: for apparent solubility, LogD$_{7.4}$, PPB, we observed marginal RMSE reductions of -8%, -10%, and -7% respectively. Yet, the performances of public models on industrial data are disappointing. Using an AD did not solve this issue. Therefore, the generalization of the public models to private data cannot be explained by the content of the public and industrial chemical spaces. We hypothesize that there shall exist some other discrepancies regarding the definition of the endpoints in public databases compared to the industrial data sources.

Table 1: Endpoint Performance of models trained on public data, with and without applicability domain on public and industrial test sets. The outlier sensitivity represents the ν parameter of the OcSVM and the % Out the ratio of compounds defined as Out of Applicability Domain.

Endpoint	**Assay**	**Outlier Sensitivity**	**RMSE Public**			**RMSE Industrial**		
			w/o AD	**w AD**	**% Out**	**w/o AD**	**w AD**	**% Out**
Efflux Ratio	Caco-2	0.2	0.51	0.48	0.12	0.72	0.70	0.05
	MDCK-MDR1	0.2	0.38	0.34	0.13	-	-	-
Apparent Permeability	Caco-2	0.4	0.53	0.51	0.26	0.92	0.92	0.06
	MDCK	0.2	0.49	0.45	0.11	-	-	-
	MDCK-LE	0.5	0.39	0.35	0.51	-	-	-
	MDCK-MDR1	0.2	0.32	0.31	0.17	-	-	-
	PAMPA	0.6	0.47	0.46	0.10	0.72	0.70	0.09
	PAMPA-BBB	0.2	0.29	0.25	0.24	-	-	-
Inhibition	pIC50 P-gP	0.2	0.52	0.51	0.16	-	-	-
Lipophilicity	LogS$_{app}$	0.2	0.62	0.56	0.10	1.14	1.04	0.04
	LogS$_{kin}$	0.2	0.41	0.40	0.11	0.43	0.42	0.11
	LogS$_{w}$	0.4	0.80	0.72	0.12	-	-	-
	LogP	0.4	0.50	0.50	0.13	0.77	0.76	0.09
	LogD$_{7.4}$	0.6	0.60	0.56	0.10	0.98	0.91	0.04
	HFE	0.2	0.96	0.91	0.09	-	-	-
Distribution	LogBB	0.4	0.41	0.37	0.24	0.67	0.64	0.10
	PPB	0.6	16.76	13.45	0.43	14.5	13.5	0.32

Interpreting the absorption chemical space

Absorption is a complex process influenced by numerous factors, including solubility, non-specific binding, efflux, metabolism, and distribution. To computationally dissect this phenomenon, we focused on poorly permeable compounds within the chemical space. We used the GNN-MTL embedding of these compounds to represent the chemical space using a GTM. The number of traits, the RBF width and the regularization were optimized to enhance the GTM's performance in predicting all defined tasks (Table S6). The optimal parameters (number of traits: 80; RBF width: 1.0; regularization: 0.01) were utilized to train a GTM on a 30,000 compounds Frame Set. All compounds were then projected onto this map. The maps were color-coded based on the cumulative sum of responsibilities, a.k.a density landscape (Figures S11 & S12), and the endpoints experimental values of each compound, a.k.a property landscapes by continuous values (Figures S13 & S14) or classes based on expert-defined medicinal chemistry threshold (Figure S15 & S16, Table S2). An additional score was generated by categorizing experimental values into 'good' or 'bad' classes (see ‘Absorption Score (AS), Table S2). This class landscape was superimposed, and scores were summed, using responsibility as a weight for permeability, solubility, and efflux (Figure S17). In this representation, an overall desirability score was plotted (Figure 9), considering all experimental tasks.

This plot highlights numerous clusters of low interest, where multiple endpoints concur on the problematic characteristics of certain subspaces. Numerous clusters with low scores can be observed within the public and industrial landscapes, demonstrating their high resolution (Figure 9b).

Comparing the landscapes of efflux, permeability, and solubility we identified subspaces where poor permeability stemmed either from poor solubility, high efflux, or both (Figure 9c). In this map, red areas indicate poor permeability related to high efflux, blue to poor solubility, and darker spaces indicate both high efflux and low solubility. Since a color on Figure 6c can be obtain only of all needed experimental properties are provided, there are more regions where such color could not be computed. This mapping identifies highly undesirable spaces. It aids in optimizing compounds by signaling potential new solubility limitations when attempting to reduce efflux.

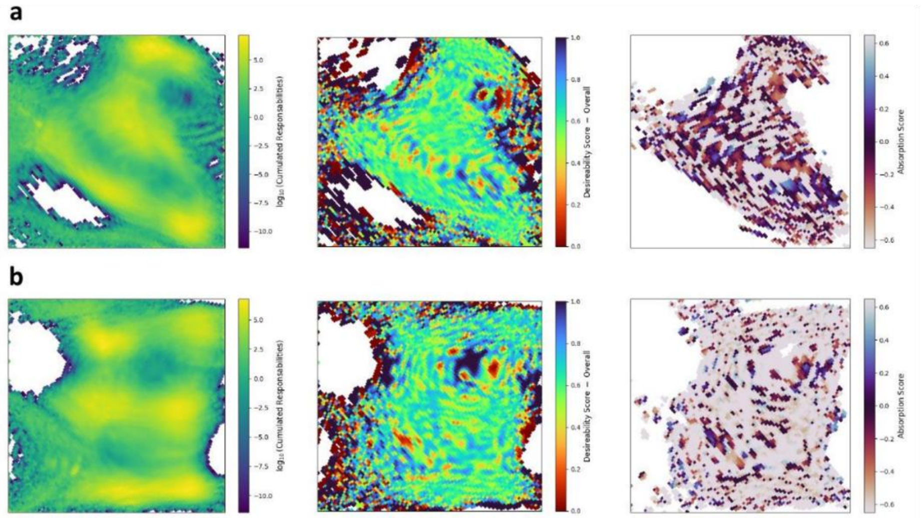


Figure 9: Density and score landscapes of the overall absorption. (a) Public. **(b)** Industrial. The density landscape aggregates projections of all molecules in the datasets. The overall score landscape integrates all endpoints. The Absorption Score landscape, transitioning from red to blue, denotes areas where permeability is influenced by either efflux, solubility, or both. Grey areas are associated to good permeability, unaffected significantly by either efflux or solubility issues.

Identification of undesired fragments

As explained in paragraph “Chemical content analysis of the map”, the public maps were analyzed using BRICS. For solubility issues, fragments like polychlorophenyl, steroids, perfluoro, and alkoxy acyl were identified. For efflux concerns, fragments such as long chains, adamantane, macrocycles, and alkoxy acyl were highlighted (Tables S7 & S8; Figure S18 & S19).

The most frequent BRICS from undesirable regions (Figure 10) revealed compounds like tri-sulfinates, steroids, and macrocycles associated with poor solubility, and benzosulfones, small cyclopeptoids with poor permeability due to high efflux. This analysis suggests chemical modifications to improve the permeability while avoiding introducing new issues.

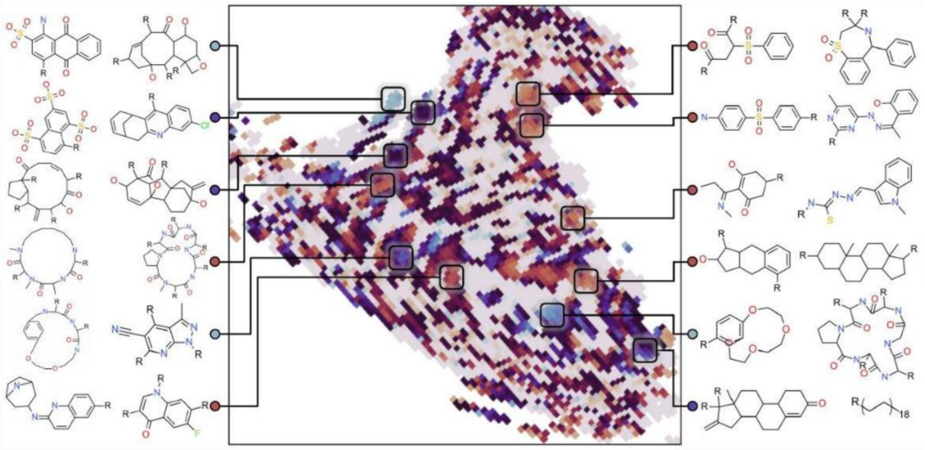


Figure 10: Structural analysis of major BRICS in problematic subspaces. Areas in red and blue represent poor permeability due to efflux or solubility, respectively. Darker areas indicate the co-occurrence of both issues. The chemical substructures are the most frequent BRICS found in the chemical space regions framed by the corresponding boxes.

Case Study

To evaluate the GTM map's capability to accurately differentiate between compounds with desired and undesired properties, particularly within highly similar congeneric series, we projected the series from the study by Degorce *et al.*[73]. Their work focused on developing a series of IRAK4 inhibitors that are orally efficacious, aiming to overcome the challenges of low permeability and high efflux encountered in pyrrolopyrimidine series. Their approach involved optimizing substituents on a pyrrolotriazine core, leading to the identification of compound 30 (Figure 11) resulting from the permeability optimization of this study. This section seeks to evaluate the map's capability to monitor this published optimization procedure and its ability to identify compound 30.

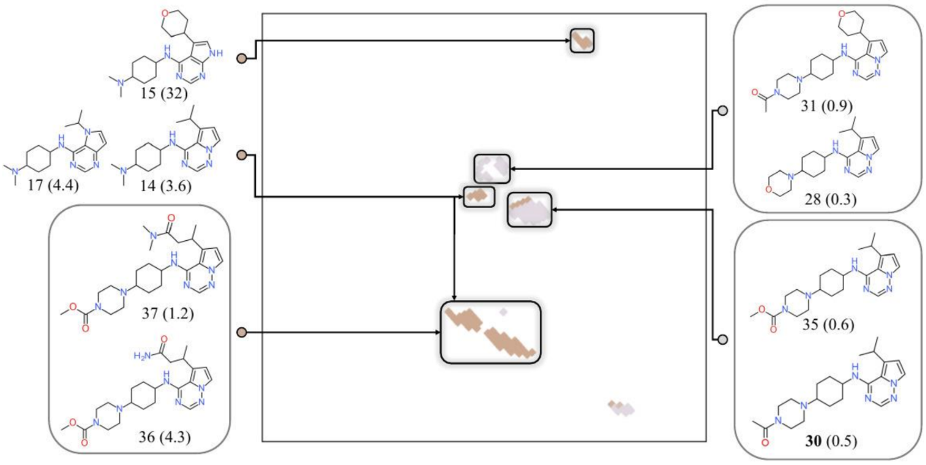


Figure 11: Analysis of the projection of the pyrrolotriazine series. Compounds are annotated with their series ID and in parenthesis, the measured efflux ratio from Caco-2 assay. Areas in light red represent poor permeability due to strong efflux. Light grey colored areas indicate compounds with low efflux and high permeability.

The analysis displays a consensus landscape that groups structurally similar compounds. Despite their similarities, these compounds are divided into distinct subspaces tied to specific structural traits, such as the presence of morpholine and oxane substituted pyrrolotriazine within the upper cluster. This spatial organization not only separates compounds based on efflux and permeability but also places structurally similar compounds in different subspaces of permeability profile.

Analysis of synergetic and antagonistic tasks within MTL

To evaluate the interplay effects between different ADMET endpoints in MTL versus STL, we analyzed the RMSE variations of the public and industrial GNN performance when applied to their respective test set. Our objective was to understand how the tasks in MTL affects predictive performance compared to STL, particularly considering dataset size.

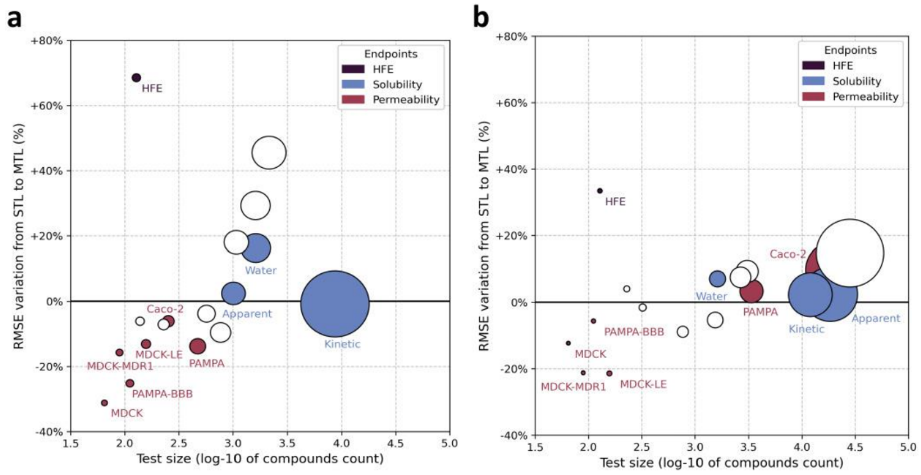


Figure 12: Analysis of synergy and antagonistic effects in GNN MTL versus STL models (a) Public data (b) Industrial data. The size of the point is proportional to the size of the dataset. Points in red represent permeability endpoints, blue represent solubility endpoints, and black represents the HFE, the neutral task. White dots indicate other tasks.

The analysis reveals that for both public and industrial models, the MTL improves RMSE for smaller permeability datasets but degrades performance for unrelated tasks such as HFE, with an RMSE increase exceeding 30% (Figure 12). Larger datasets (>1000 compounds) tend to perform better with STL. Comparing public and industrial datasets, we find that industrial datasets, particularly those related to permeability, show minimal performance differences between STL and MTL due to higher data quality and quantity. These findings indicate that MTL can enhance predictive accuracy for related smaller tasks but may hinder performance for unrelated tasks, emphasizing the need for careful task selection and consideration of dataset characteristics in MTL applications.

Discussion and Conclusions

The present work addressed misconceptions in SME permeability and showcased the efficacy of MTL-GNN models for ADMET optimization through GTM. Traditional optimization processes have largely depended on Caco-2 and PAMPA studies to delineate transport pathways[14,74,75]. Over the year, the development of drug-likeliness filters[4,76,77] has aimed to ease the identification of undesired properties. Our investigation illustrated the discrepancies between public and industrial data. However, we aimed to go beyond and formalize the differences between these data sources. Hence, this work aims to provide a clearer and more comprehensive perspective on permeability challenges. Additionally, it introduces an explainable and predictive approach for SME permeability optimization.

Our comparative analysis of PAMPA and Caco-2 assays for assessing transport routes reveals misconceptions. First, low-to-high permeability compounds in PAMPA/Caco-2 assay comparison may be attributable to labware binding, rather than solely to active efflux mechanisms[74]. Secondly, instances where Caco-2 permeability surpasses PAMPA permeability —previously attributed to uptake—may reflect high recovery rates influenced by metabolic processes or experimental noise. This emphasizes the critical role of metabolic data, acquired through LC-MS/MS, in refining permeability evaluations[28]. Our research elucidates the relationship between $\mathrm{LogD}_{7.4}$ and non-specific binding in PAMPA models, which adversely affects apparent permeability measurements. The impact of $\mathrm{LogD}_{7.4}$ in drug permeability is well-documented[78–80], with poor $\mathrm{LogD}_{7.4}$ values frequently indicative of permeability issues. Yet, we observed compounds with suboptimal $\mathrm{LogD}_{7.4}$ and unreliable PAMPA permeability, as further investigations revealed compromised recovery due to non-specific binding to labware. These observations underscore the limitations of PAMPA assays. Recovery issues lead to underestimated apparent permeability values. The identified optimal recovery range (80% to 95%) is consistent with prior studies[45,81], reinforcing its importance for data analysis and curation.

Moreover, efflux mechanisms markedly affect the permeability of drug-like molecules. We report clear correlation between topological polar surface area and efflux, consistent with previously reported observations[82,83], across various cellular models, establishing tPSA as a consistent efflux determinant.

The application of publicly available models to industrial data highlighted significant limitations, particularly when slight variations in experimental conditions led to model inaccuracies. This is exemplified in the assessment of Caco-2 permeability, where the inclusion of BSA has an impact on the model's performance on low permeability compounds[44,81]. Despite these challenges, our findings suggest that industrial datasets, characterized by consistent parameters and standardized protocols, provide a solid basis for QSPR models with improved performances. All the modeling approaches have been optimized on the public data, and successfully transposed on the industrial data. This demonstrates the advantage of employing public architectural frameworks with industrial data. On the other hand, application of models trained on public data to the industrial setup was in some cases disappointing.

Utilizing the latent vectors from the GNN-MTL models, we achieved high-resolution mapping of the chemical space, integrating it with experimental data for enhanced insight. This approach offers a refined understanding of the chemical factors influencing drug permeability, namely solubility issues and efflux. The ability to derive a consensus experimental landscape for ensemble scoring, defines a novel method to explain ADMET failure.

We expect such analysis to be applicable to the entire ADME-Tox and target inhibition spectrum. Such methods would be beneficial for consensus ADMET profiling and the identification of binding to anti-target families.

Data Availability

The authors declare that the data supporting the findings of this study are available free of charge. The repository features multiple datasets that have been curated for this research.

Code Availability

No custom code has been used.

Author Contributions

PL is the main author. Data collection, annotation process supervision, modeling and statistical analysis of results were carried out by PL, CM and GM. Figures and table preparation by PL and GM. Supervision by CM, GM, and AV. The first version of this article was written by PL and GM; GM, CM and AV led the subsequent revisions.

Notes

C. Minoletti and P. Llompart are Sanofi employees and may hold shares and/or stock options in the company. G. Marcou, and A. Varnek have nothing to disclose.

Abbreviations

ABC: ATP-binding cassette

AD: Applicability Domain

ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity

ATP: Adenosine Triphosphate

BBB: Blood-Brain Barrier

BRICS: Breaking of Retrosynthetically Interesting Chemical Substructures

BSA: Bovine Serum Albumin

Caco-2: Human Colorectal Adenocarcinoma Cell Line

CS: Chemical Space

CV: Cross-Validation

CYP: Cytochrome P450

DMSO: Dimethyl Sulfoxide

ECFP: Extended-Connectivity Fingerprints

EM: Expectation Maximization

ER: Efflux Ratio

GA: Genetic Algorithm

GNN: Graph Neural Network

GTM: Generative Topographic Mapping

HFE: Hydration Free Energy

IC50: Half Maximal Inhibitory Concentration

IRAK4: Interleukin-1 Receptor-Associated Kinase 4

LC-MS-MS: Liquid Chromatography-Tandem Mass Spectrometry

LOD: Limit of Detection

LOQ: Limit of Quantitation

LE: Low Efflux

LV: Latent Vector

MAE: Mean Absolute Error

MDCK: Madin-Darby Canine Kidney

MDR1: Multidrug Resistance Protein 1

MRP: Multidrug Resistance Proteins

MTL: Multi-Task Learning

MPNN: Message Passing Neural Network

MPO: Multi-Parameter Optimization

MSE: Mean Squared Error

PAMPA: Parallel Artificial Membrane Permeability Assay

PEPT1: Peptide Transporter 1

PG: Propylene Glycol

P-gP: P-glycoprotein

PPB: Plasma Protein Binding

OcSVM: One-Class Support Vector Machine

QSPR: Quantitative Structure-Property Relationship

RBF: Radial Basis Function

RF: Random Forest

RMSE: Root Mean Squared Error

SAR: Structure-Activity Relationship

SLC: Solute Carrier

SDi: Inter-laboratory Standard Deviation

SME: Small Molecular Entities

SMBO: Sequential Model-Based Optimization

STL: Single-Task Learning

SVM: Support Vector Machine

TEER: Trans-Epithelial Electrical Resistance

tPSA: topological Polar Surface Area

References

(1) Dahlgren, D.; Lennernäs, H. Intestinal Permeability and Drug Absorption: Predictive Experimental, Computational and In Vivo Approaches. *Pharmaceutics* **2019**, *11* (8), 411. https://doi.org/10.3390/pharmaceutics11080411.

(2) Billat, P.-A.; Roger, E.; Faure, S.; Lagarce, F. Models for Drug Absorption from the Small Intestine: Where Are We and Where Are We Going? *Drug Discovery Today* **2017**, *22* (5), 761–775. https://doi.org/10.1016/j.drudis.2017.01.007.

(3) Abbott, N. J. Blood–Brain Barrier Structure and Function and the Challenges for CNS Drug Delivery. *J Inherit Metab Dis* **2013**, *36* (3), 437–449. https://doi.org/10.1007/s10545-013-9608-0.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews* **2012**, *64*, 4–17. https://doi.org/10.1016/j.addr.2012.09.019.

(5) Ma, X.; Chen, C.; Yang, J. Predictive Model of Blood-Brain Barrier Penetration of Organic Compounds. *Acta Pharmacol Sin* **2005**, *26* (4), 500–512. https://doi.org/10.1111/j.1745-7254.2005.00068.x.

(6) Amidon, G. L.; Lennernäs, H.; Shah, V. P.; Crison, J. R. A Theoretical Basis for a Biopharmaceutic Drug Classification: The Correlation of in Vitro Drug Product Dissolution and in Vivo Bioavailability. *Pharm Res* **1995**, *12* (3), 413–420. https://doi.org/10.1023/A:1016212804288.

(7) Artursson, P.; Karlsson, J. Correlation between Oral Drug Absorption in Humans and Apparent Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cells. *Biochemical and Biophysical Research Communications* **1991**, *175* (3), 880–885. https://doi.org/10.1016/0006-291X(91)91647-U.

(8) *A General Model for Prediction of Caco-2 Cell Permeability - Nordqvist - 2004 - QSAR & Combinatorial Science - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1002/qsar.200330868 (accessed 2024-12-10).

(9) Pham-The, H.; Cabrera-Pérez, M. Á.; Nam, N.-H.; Castillo-Garit, J. A.; Rasulev, B.; Le-Thi-Thu, H.; Casañola-Martin, G. M. In Silico Assessment of ADME Properties: Advances in Caco-2 Cell Monolayer Permeability Modeling. *Current Topics in Medicinal Chemistry* **2018**, *18* (26), 2209–2229. https://doi.org/10.2174/1568026619666181130140350.

(10) Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R. Estimation of ADME Properties in Drug Discovery: Predicting Caco-2 Cell Permeability Using Atom-Based Stochastic and Non-Stochastic Linear Indices. *Journal of Pharmaceutical Sciences* **2008**, *97* (5), 1946–1976. https://doi.org/10.1002/jps.21122.

(11) Paixão, P.; Gouveia, L. F.; Morais, J. A. G. Prediction of the *in Vitro* Permeability Determined in Caco-2 Cells by Using Artificial Neural Networks. *European Journal of Pharmaceutical Sciences* **2010**, *41* (1), 107–117. https://doi.org/10.1016/j.ejps.2010.05.014.

(12) *In Silico Prediction of Intestinal Permeability by Hierarchical Support Vector Regression*. https://www.mdpi.com/1422-0067/21/10/3582 (accessed 2024-12-10).

(13) *QSAR Prediction of Passive Permeability in the LLC-PK1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities - Sherer - 2012 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201100157 (accessed 2024-12-10).

(14) Avdeef, A.; Artursson, P.; Neuhoff, S.; Lazorova, L.; Gråsjö, J.; Tavelin, S. Caco-2 Permeability of Weakly Basic Drugs Predicted with the Double-Sink PAMPA pKaflux<math><mrow Is="true"><mtext Is="true">p</Mtext><msubsup Is="true"><mi Is="true">K</Mi><mtext Is="true">a</Mtext><mrow Is="true"><mtext Is="true">flux</Mtext></Mrow></Msubsup></Mrow></Math> Method. *European Journal of Pharmaceutical Sciences* **2005**, *24* (4), 333–349. https://doi.org/10.1016/j.ejps.2004.11.011.

(15) Press, B. Optimization of the Caco-2 Permeability Assay to Screen Drug Compounds for Intestinal Absorption and Efflux. In *Permeability Barrier: Methods and Protocols*; Turksen, K., Ed.; Humana Press: Totowa, NJ, 2011; pp 139–154. https://doi.org/10.1007/978-1-61779-191-8_9.

(16) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (3), 1253–1268. https://doi.org/10.1021/acs.jcim.8b00785.

(17) *Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches*. https://www.mdpi.com/1999-4923/14/10/1998 (accessed 2024-12-10).

(18) Lee, J. B.; Zgair, A.; Taha, D. A.; Zang, X.; Kagan, L.; Kim, T. H.; Kim, M. G.; Yun, H.; Fischer, P. M.; Gershkovich, P. Quantitative Analysis of Lab-to-Lab Variability in Caco-2 Permeability Assays. *European Journal of Pharmaceutics and Biopharmaceutics* **2017**, *114*, 38–42. https://doi.org/10.1016/j.ejpb.2016.12.027.

(19) Volpe, D. A. Variability in Caco-2 and MDCK Cell-Based Intestinal Permeability Assays. *Journal of Pharmaceutical Sciences* **2008**, *97* (2), 712–725. https://doi.org/10.1002/jps.21010.

(20) Smetanová, L.; Stětinová, V.; Svoboda, Z.; Kvetina, J. Caco-2 Cells, Biopharmaceutics Classification System (BCS) and Biowaiver. *Acta Medica (Hradec Kralove)* **2011**, *54* (1), 3–8.

(21) Gan, L.-S. L.; Thakker, D. R. Applications of the Caco-2 Model in the Design and Development of Orally Active Drugs: Elucidation of Biochemical and Physical Barriers Posed by the Intestinal Epithelium. *Advanced Drug Delivery Reviews* **1997**, *23* (1), 77–98. https://doi.org/10.1016/S0169-409X(96)00427-9.

(22) Di, L.; Kerns, E. H. *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*; Academic Press, 2015.

(23) Hidalgo, I. J.; Raub, T. J.; Borchardt, R. T. Characterization of the Human Colon Carcinoma Cell Line (Caco-2) as a Model System for Intestinal Epithelial Permeability. *Gastroenterology* **1989**, *96* (2, Part 2), 736–749. https://doi.org/10.1016/S0016-5085(89)80072-1.

(24) Irvine, J. D.; Takahashi, L.; Lockhart, K.; Cheong, J.; Tolan, J. W.; Selick, H. E.; Grove, J. R. MDCK (Madin–Darby Canine Kidney) Cells: A Tool for Membrane Permeability Screening. *Journal of Pharmaceutical Sciences* **1999**, *88* (1), 28–33. https://doi.org/10.1021/js9803205.

(25) Balimane, P. V.; Han, Y.-H.; Chong, S. Current Industrial Practices of Assessing Permeability and P-Glycoprotein Interaction. *AAPS J* **2006**, *8* (1), 1. https://doi.org/10.1208/aapsj080101.

(26) Reis, J. M.; Sinko, B.; Serra, C. H. R. Parallel Artificial Membrane Permeability Assay (PAMPA) - Is it Better than Caco-2 for Human Passive Permeability Prediction? *Mini Reviews in Medicinal Chemistry* **2010**, *10* (11), 1071–1076. https://doi.org/10.2174/138955710793177476.

(27) Kansy, M.; Avdeef, A.; Fischer, H. Advances in Screening for Membrane Permeability: High-Resolution PAMPA for Medicinal Chemists. *Drug Discovery Today: Technologies* **2004**, *1* (4), 349–355. https://doi.org/10.1016/j.ddtec.2004.11.013.

(28) van Breemen, R. B.; Li, Y. Caco-2 Cell Permeability Assays to Measure Drug Absorption. *Expert Opinion on Drug Metabolism & Toxicology* **2005**, *1* (2), 175–185. https://doi.org/10.1517/17425255.1.2.175.

(29) Balimane, P. V.; Chong, S.; Morrison, R. A. Current Methodologies Used for Evaluation of Intestinal Permeability and Absorption. *Journal of Pharmacological and Toxicological Methods* **2000**, *44* (1), 301–312. https://doi.org/10.1016/S1056-8719(00)00113-1.

(30) Galinis-Luciani, D.; Nguyen, L.; Yazdanian, M. Is PAMPA a Useful Tool for Discovery? *Journal of Pharmaceutical Sciences* **2007**, *96* (11), 2886–2892. https://doi.org/10.1002/jps.21071.

(31) Avdeef, A. The Rise of PAMPA. *Expert Opinion on Drug Metabolism & Toxicology* **2005**, *1* (2), 325–342. https://doi.org/10.1517/17425255.1.2.325.

(32) Wang, Q.; Strab, R.; Kardos, P.; Ferguson, C.; Li, J.; Owen, A.; Hidalgo, I. J. Application and Limitation of Inhibitors in Drug–Transporter Interactions Studies. *International Journal of Pharmaceutics* **2008**, *356* (1), 12–18. https://doi.org/10.1016/j.ijpharm.2007.12.024.

(33) *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability | Wiley*. Wiley.com. https://www.wiley.com/en-ae/Drug+Bioavailability%3A+Estimation+of+Solubility%2C+Permeability%2C+Absorption+and+Bioavailability-p-9783527605156 (accessed 2024-12-10).

(34) Cole, S.; Bagal, S.; El-Kattan, A.; Fenner, K.; Hay, T.; Kempshall, S.; Lunn, G.; Varma, M.; Stupple, P.; Speed, W. Full Efficacy with No CNS Side-Effects: Unachievable Panacea or Reality? DMPK Considerations in Design of Drugs with Limited Brain Penetration. *Xenobiotica* **2012**, *42* (1), 11–27. https://doi.org/10.3109/00498254.2011.617847.

(35) *Development of a new permeability assay using low-efflux MDCKII cells - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0022354915318542 (accessed 2024-12-10).

(36) Wang, Q.; Rager, J. D.; Weinstein, K.; Kardos, P. S.; Dobson, G. L.; Li, J.; Hidalgo, I. J. Evaluation of the MDR-MDCK Cell Line as a Permeability Screen for the Blood–Brain Barrier. *International Journal of Pharmaceutics* **2005**, *288* (2), 349–359. https://doi.org/10.1016/j.ijpharm.2004.10.007.

(37) Hellinger, É.; Veszelka, S.; Tóth, A. E.; Walter, F.; Kittel, Á.; Bakk, M. L.; Tihanyi, K.; Háda, V.; Nakagawa, S.; Dinh Ha Duy, T.; Niwa, M.; Deli, M. A.; Vastag, M. Comparison of Brain Capillary Endothelial Cell-Based and Epithelial (MDCK-MDR1, Caco-2, and VB-Caco-2) Cell-Based Surrogate Blood–Brain Barrier Penetration Models. *European Journal of Pharmaceutics and Biopharmaceutics* **2012**, *82* (2), 340–351. https://doi.org/10.1016/j.ejpb.2012.07.020.

(38) Wahlang, B.; Pawar, Y. B.; Bansal, A. K. Identification of Permeability-Related Hurdles in Oral Delivery of Curcumin Using the Caco-2 Cell Model. *European Journal of Pharmaceutics and Biopharmaceutics* **2011**, *77* (2), 275–282. https://doi.org/10.1016/j.ejpb.2010.12.006.

(39) Cai, X.; Walker, A.; Cheng, C.; Paiva, A.; Li, Y.; Kolb, J.; Herbst, J.; Shou, W.; Weller, H. Approach to Improve Compound Recovery in a High-Throughput Caco-2 Permeability Assay Supported by Liquid Chromatography–Tandem Mass Spectrometry. *Journal of Pharmaceutical Sciences* **2012**, *101* (8), 2755–2762. https://doi.org/10.1002/jps.23194.

(40) Koljonen, M.; Hakala, K. S.; Ahtola-Sätilä, T.; Laitinen, L.; Kostiainen, R.; Kotiaho, T.; Kaukonen, A. M.; Hirvonen, J. Evaluation of Cocktail Approach to Standardise Caco-2 Permeability Experiments. *European Journal of Pharmaceutics and Biopharmaceutics* **2006**, *64* (3), 379–387. https://doi.org/10.1016/j.ejpb.2006.06.006.

(41) Pires, C. L. Permeability through Caco-2 Cell Monolayers as a Model for BBB: Implementation and Preliminary Evaluation Using Model Compounds. In *Permeability through Caco-2 cell monolayers as a model for BBB: implementation and preliminary evaluation using model compounds*; 2018.

(42) Rubas, W.; Cromwell, M. E. M.; Shahrokh, Z.; Villagran, J.; Nguyen, T.-N.; Wellton, M.; Nguyen, T.-H.; Mrsny, R. J. Flux Measurements across Caco-2 Monolayers May Predict Transport in Human Large Intestinal Tissue. *Journal of Pharmaceutical Sciences* **1996**, *85* (2), 165–169. https://doi.org/10.1021/js950267+.

(43) Bednarczyk, D.; Sanghvi, M. V. The Impact of Assay Recovery on the Apparent Permeability, a Function of Lysosomal Trapping. *Xenobiotica* **2020**, *50* (7), 753–760. https://doi.org/10.1080/00498254.2019.1691284.

(44) Saha, P.; Kou, J. H. Effect of Bovine Serum Albumin on Drug Permeability Estimation across Caco-2 Monolayers. *European Journal of Pharmaceutics and Biopharmaceutics* **2002**, *54* (3), 319–324. https://doi.org/10.1016/S0939-6411(02)00089-9.

(45) Liu, T.; Chang, L.-J.; Uss, A.; Chu, I.; Morrison, R. A.; Wang, L.; Prelusky, D.; Cheng, K.-C.; Li, C. The Impact of Protein on Caco-2 Permeability of Low Mass Balance Compounds for Absorption Projection and Efflux Substrate Identification. *Journal of Pharmaceutical and Biomedical Analysis* **2010**, *51* (5), 1069–1077. https://doi.org/10.1016/j.jpba.2009.12.006.

(46) Neuhoff, S.; Ungell, A.-L.; Zamora, I.; Artursson, P. pH-Dependent Passive and Active Transport of Acidic Drugs across Caco-2 Cell Monolayers. *European Journal of Pharmaceutical Sciences* **2005**, *25* (2), 211–220. https://doi.org/10.1016/j.ejps.2005.02.009.

(47) Miret, S.; Abrahamse, L.; de Groene, E. M. Comparison of in Vitro Models for the Prediction of Compound Absorption across the Human Intestinal Mucosa. *J Biomol Screen* **2004**, *9* (7), 598–606. https://doi.org/10.1177/1087057104267162.

(48) Sun, H.; Chow, E. C.; Liu, S.; Du, Y.; Pang, K. S. The Caco-2 Cell Monolayer: Usefulness and Limitations. *Expert Opinion on Drug Metabolism & Toxicology* **2008**, *4* (4), 395–411. https://doi.org/10.1517/17425255.4.4.395.

(49) Egan, W. J.; Merz, Kenneth M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43* (21), 3867–3877. https://doi.org/10.1021/jm000292e.

(50) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J Comput Aided Mol Des* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.

(51) Wassermann, A. M.; Bajorath, J. BindingDB and ChEMBL: Online Compound Databases for Drug Discovery. *Expert Opinion on Drug Discovery* **2011**, *6* (7), 683–687. https://doi.org/10.1517/17460441.2011.579100.

(52) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Combinatorial Chemistry & High Throughput Screening* **2001**, *4* (8), 719–725. https://doi.org/10.2174/1386207013330670.

(53) Wang, N.-N.; Dong, J.; Deng, Y.-H.; Zhu, M.-F.; Wen, M.; Yao, Z.-J.; Lu, A.-P.; Wang, J.-B.; Cao, D.-S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56* (4), 763–773. https://doi.org/10.1021/acs.jcim.5b00642.

(54) *RDKit*. https://www.rdkit.org/ (accessed 2025-01-02).

(55) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Molecular Informatics* **2010**, *29* (12), 855–868. https://doi.org/10.1002/minf.201000099.

(56) Rosa, G. J. M. The Elements of Statistical Learning: Data Mining, Inference, and Prediction by HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. *Biometrics* **2010**, *66* (4), 1315. https://doi.org/10.1111/j.1541-0420.2010.01516.x.

(57) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(58) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems*; MIT Press, 1996; Vol. 9.

(59) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998**, *21* (1), 203–224. https://doi.org/10.1016/S0925-2312(98)00043-5.

(60) *GTM-Based QSAR Models and Their Applicability Domains - Gaspar - 2015 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201400153 (accessed 2024-12-10).

(61) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Molecular Informatics* **2012**, *31* (3–4), 301–312. https://doi.org/10.1002/minf.201100163.

(62) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **2013**, *53* (12), 3318–3325. https://doi.org/10.1021/ci400423c.

(63) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507. https://doi.org/10.1002/cmdc.200800178.

(64) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* **2009**, *20* (1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.

(65) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17. https://doi.org/10.1021/acs.jcim.3c01250.

(66) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63* (16), 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959.

(67) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discov.* **2015**, *8* (1), 014008. https://doi.org/10.1088/1749-4699/8/1/014008.

(68) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2012; Vol. 25.

(69) Baskin, I. I.; Kireeva, N.; Varnek, A. The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Molecular Informatics* **2010**, *29* (8–9), 581–587. https://doi.org/10.1002/minf.201000063.

(70) Schölkopf, B.; Williamson, R. C.; Smola, A.; Shawe-Taylor, J.; Platt, J. Support Vector Method for Novelty Detection. In *Advances in Neural Information Processing Systems*; MIT Press, 1999; Vol. 12.

(71) Kerns, E. H.; Di, L.; Petusky, S.; Farris, M.; Ley, R.; Jupp, P. Combined Application of Parallel Artificial Membrane Permeability Assay and Caco-2 Permeability Assays in Drug Discovery. *JPharmSci* **2004**, *93* (6), 1440–1453. https://doi.org/10.1002/jps.20075.

(72) Desai, P. V.; Sawada, G. A.; Watson, I. A.; Raub, T. J. Integration of in Silico and in Vitro Tools for Scaffold Optimization during Drug Discovery: Predicting P-Glycoprotein Efflux. *Mol. Pharmaceutics* **2013**, *10* (4), 1249–1261. https://doi.org/10.1021/mp300555n.

(73) Degorce, S. L.; Anjum, R.; Dillman, K. S.; Drew, L.; Groombridge, S. D.; Halsall, C. T.; Lenz, E. M.; Lindsay, N. A.; Mayo, M. F.; Pink, J. H.; Robb, G. R.; Scott, J. S.; Stokes, S.; Xue, Y.

Optimization of Permeability in a Series of Pyrrolotriazine Inhibitors of IRAK4. *Bioorganic & Medicinal Chemistry* **2018**, *26* (4), 913–924. https://doi.org/10.1016/j.bmc.2018.01.008.

(74) Kerns, E. H.; Di, L.; Petusky, S.; Farris, M.; Ley, R.; Jupp, P. Combined Application of Parallel Artificial Membrane Permeability Assay and Caco-2 Permeability Assays in Drug Discovery. *Journal of Pharmaceutical Sciences* **2004**, *93* (6), 1440–1453. https://doi.org/10.1002/jps.20075.

(75) Bermejo, M.; Avdeef, A.; Ruiz, A.; Nalda, R.; Ruell, J. A.; Tsinman, O.; González, I.; Fernández, C.; Sánchez, G.; Garrigues, T. M.; Merino, V. PAMPA—a Drug Absorption in Vitro Model: 7. Comparing Rat in Situ, Caco-2, and PAMPA Permeability of Fluoroquinolones. *European Journal of Pharmaceutical Sciences* **2004**, *21* (4), 429–441. https://doi.org/10.1016/j.ejps.2003.10.009.

(76) Matsson, P.; Doak, B. C.; Over, B.; Kihlberg, J. Cell Permeability beyond the Rule of 5. *Advanced Drug Delivery Reviews* **2016**, *101*, 42–61. https://doi.org/10.1016/j.addr.2016.03.013.

(77) Möbitz, H. Design Principles for Balancing Lipophilicity and Permeability in beyond Rule of 5 Space. *ChemMedChem* **2024**, *19* (5), e202300395. https://doi.org/10.1002/cmdc.202300395.

(78) Waring, M. J. Defining Optimum Lipophilicity and Molecular Weight Ranges for Drug Candidates—Molecular Weight Dependent Lower Log *D* Limits Based on Permeability. *Bioorganic & Medicinal Chemistry Letters* **2009**, *19* (10), 2844–2851. https://doi.org/10.1016/j.bmcl.2009.03.109.

(79) Bhal, S. K.; Kassam, K.; Peirson, I. G.; Pearl, G. M. The Rule of Five Revisited: Applying Log D in Place of Log P in Drug-Likeness Filters. *Mol. Pharmaceutics* **2007**, *4* (4), 556–560. https://doi.org/10.1021/mp0700209.

(80) Rubas, W.; Cromwell, M. E. M. The Effect of Chemical Modifications on Octanol/Water Partition (Log D) and Permeabilities across Caco-2 Monolayers. *Advanced Drug Delivery Reviews* **1997**, *23* (1), 157–162. https://doi.org/10.1016/S0169-409X(96)00433-4.

(81) Krishna, G.; Chen, K.; Lin, C.; Nomeir, A. A. Permeability of Lipophilic Compounds in Drug Discovery Using In-Vitro Human Absorption Model, Caco-2. *International Journal of Pharmaceutics* **2001**, *222* (1), 77–89. https://doi.org/10.1016/S0378-5173(01)00698-6.

(82) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. https://doi.org/10.1021/jm000942e.

(83) Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1585–1600. https://doi.org/10.1021/ci049884m.

Supplementary Information

Garbage in, garbage out: An industrial perspective on drug absorption modeling

P. Llompart[1,2], C. Minoletti[2], G. Marcou[1,*], A. Varnek[1]

[1]Laboratory of Cheminformatics, UMR7140, University of Strasbourg, Strasbourg, France

[2]IDD/CADD, Sanofi, Vitry-Sur-Seine, France

Table S1: Description of the endpoints and their conditions of measurements.

Endpoint	Assay	Source	Conditions	Unit
Efflux Ratio	Caco-2	Industrial	LC-MS/MS Transport Time: 2h TC7 Cell Lines Plate Type: Becton D24 Well 1.0 µM Passage Number: 35-72 BSA A/B: 0.5/5 % Concentration: 1-10 µM	$\log_{10}(P_{app\text{-}elacridar}/P_{app\text{-}wo\text{-}elacridar})$
		Public	Passage Number: 20-100 BSA: 0 % Concentration: 1-10 µM	$\log_{10}(P_{appBA}/P_{appAB})$
	MDCK-MDR1	Public	Passage Number: 20-100 BSA: 0 % Concentration: 1-10 µM	$\log_{10}(P_{appBA}/P_{appAB})$
Apparent Permeability	Caco-2	Industrial	LC-MS/MS Transport Time: 2h TC7 Cell Lines Plate Type: Becton D24 Well 1.0 µM Passage Number: 35-72 BSA A/B: 0.5/5 % Concentration: 1-10 µM	$\log_{10}(P_{app})$
		Public	Passage Number: 35-72 BSA A/B: 0 % Concentration: 1-10 µM	$\log_{10}(P_{app})$
	MDCK MDCK-LE MDCK-MDR1	Public	Passage Number: 35-72 BSA A/B: 0 % Concentration: 1-10 µM	$\log_{10}(P_{app})$
	PAMPA	Industrial	PION double sink LC-MS/MS Transport Time: 2h Membrane: GIT Temperature: 22°C Concentration: 1-10 µM	$\log_{10}(P_{app})$
		Public	pH A/B: 7.4/7.4 Concentration: 1-10 µM	$\log_{10}(P_{app})$
	PAMPA-BBB	Public	Brain lipid extract membrane: PC, PE, PS, PBLE, PVDF DMSO 0-5%	$\log_{10}(P_{app})$
Inhibition	pIC50 P-gP	Public	MDCK-MDR1 cell line Rhodamin 123/Digoxin substrate	$\log_{10}$(IC50)
Physico-Chemical	Apparent Solubility (LogS_{app})	Public	Phosphate buffer 0.1 M. At 25±5°Celsius and pH 7.4±1 log.	$\log_{10}$(M)
		Industrial	Phosphate buffer 0.1 M. At 25±5°Celsius and pH 7.4±1 log.	$\log_{10}$(M)
	Kinetic Solubility (LogS_{kin})	Public	Concentration: 10 mM DMSO Temperature 25°Celsius PBS Buffer pH 7.4	$\log_{10}$(M)
		Industrial	Concentration: 10 mM DMSO Nephelometer (Industrial) Temperature 25°Celsius PBS Buffer pH 7.4±1 log.	$\log_{10}$(M)
	Water Solubility (LogSw)	Industrial	Shake-Flask/Column elution Pure water 25±5°Celsius pH 7±1 log.	$\log_{10}$(M)
	LogP	Public	Temperature: 22-25°C HPLC / ShakeFlask pH 7±1 log Pure Water	$\log_{10}(C_{octanol}/C_{water})$
		Industrial	Temperature: 22-25°C HPLC / ShakeFlask pH 7±1 log Pure Water	$\log_{10}(C_{octanol}/C_{water})$
	$\text{LogD}_{7.4}$	Public	Temperature: 22-25°C	$\log_{10}(C_{octanol}/C_{buffer})$

			RP-HPLC / ShakeFlask pH 7.4 PBS	
		Industrial	Temperature: 22-25°C RP-HPLC / ShakeFlask pH 7.4 PBS	$\log_{10}(C_{octanol}/C_{buffer})$
	Hydration Free Energy (HFE)	Public	Alchemical free energy	$\log_{10}$(kcal/mol)
Distribution	Plasma Protein Binding (PPB)	Public	Rat Matrix: Plasma	$1-(C_{Free}/C_{Total})*100$
		Industrial	Rat Temperature: 37°C Supplier: Harlan Matrix: Plasma Concentration: 5-25 µM LC-MS/MS	$1-(C_{Free}/C_{Total})*100$
	Ratio Brain/Blood (LogBB)	Public	Rat Intravenous administration	AUC_{Brain}/AUC_{Blood}
		Industrial	Rat Intravenous administration	AUC_{Brain}/AUC_{Blood}

Permeability data were excluded if:

- no continuous value was available.
- no information was given about the data sources.
- measurement was done in presence of MDR1 or CYP P450 inhibitors/inducer.
- the flux was measured from compartment B to A for P_{app}.
- the flux was evaluated after more than 100 passage.
- The TEER was lower than 200 ohm.cm2 for Caco-2 and 150 for MDCK

To avoid conserving experimental values measured below or above the limit of quantification.

Table S2: Comparison of unique molecules count before and after curation for each endpoint.

Endpoint	Assay	Source	All	% of public	Threshold	Reference
Efflux Ratio	Caco-2	P	2,355	100	D < 0.3	[1]
		I	5,752	0	D < 0.3	[1]
	MDCK-MDR1	P	3,237	100	D < 0.5	[2]
Apparent Permeability	Caco-2	P	1,228	100	-5.5 < D	Industrial
		I	80,440	0	-5.5 < D	Industrial
	MDCK	P	317	100	-5.0 < D	[3]
	MDCK-LE	P	697	100	-5.0 < D	[4]
	MDCK-MDR1	P	407	100	-5.5 < D	[5]
	PAMPA-BBB	P	559	100	-5.3 < D	[6]
	PAMPA	I	15,463	0	-4.5 < D	Industrial
		P	2,343	100	-4.5 < D	Industrial
Distribution	LogBB	I	1,353	49	0.0 < D	[6]
		P	666	100	0.0 < D	[6]
	PPB	I	13,437	58	90.0 % < D	[7]
		P	7,841	100	90.0 % < D	[7]
Recovery	Caco-2	I	19,067	0	80.0 % < D	Industrial
	PAMPA	I	71,986	0	80.0 % < D	Industrial
Inhibition	pIC50 P-gP	P	1,141	100	D < 6.0	Industrial
Lipophilicity	LogS$_{app}$	I	83,246	0	-4.0 < D	[1]
		P	4,915	100	-4.0 < D	[1]
	LogS$_{kin}$	P	43,386	100	-5.0 < D	Industrial
		I	59,501	73	-5.0 < D	Industrial
	LogS$_{W}$	P	7,957	100	-4.0 < D	[8]
	LogP	P	10,627	100	D < 4.0	[1]
		I	13,060	81	D < 4.0	[1]
	LogD$_{7.4}$	P	5,315	100	D < 4.0	[1]
		I	128,243	4	D < 4.0	[1]
	HFE	P	618	100	D < 4.0	[9]

P : Public

I : Industrial

D : Desired range of measurements in a drug discovery context

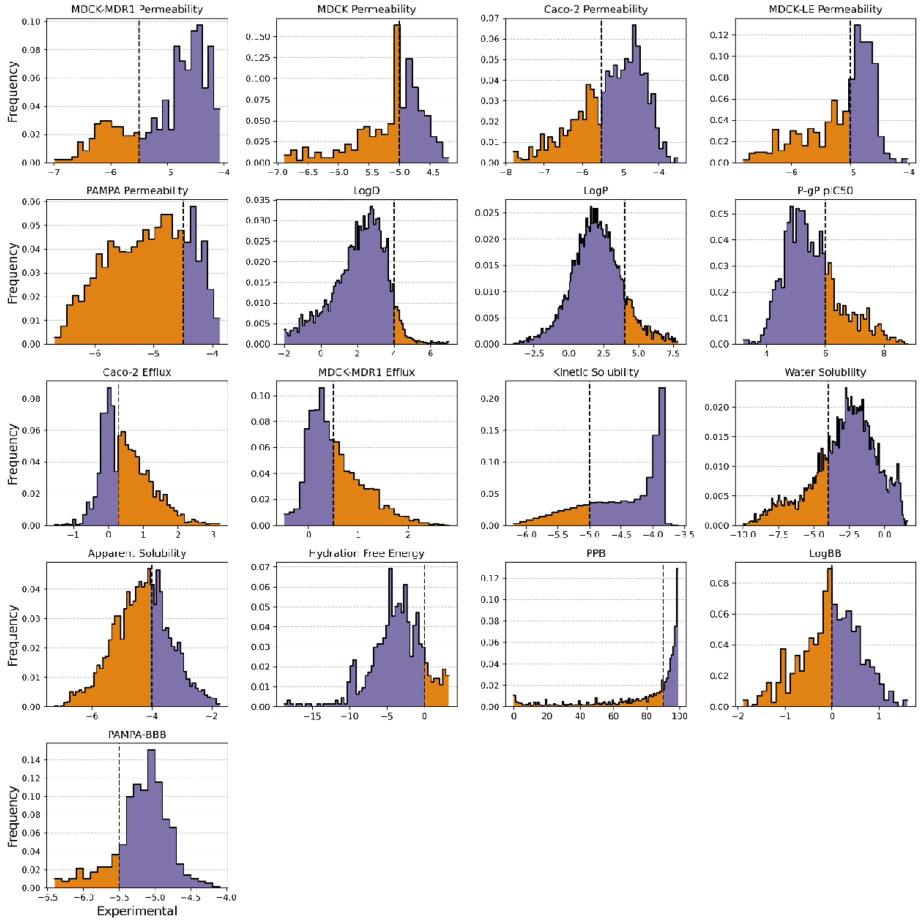


Figure S1: Distribution of the public experimental measurement. Undesired range of properties are colored in orange and defined using medicinal chemistry threshold from the literature or industry. Desirable property range are shown in purple.

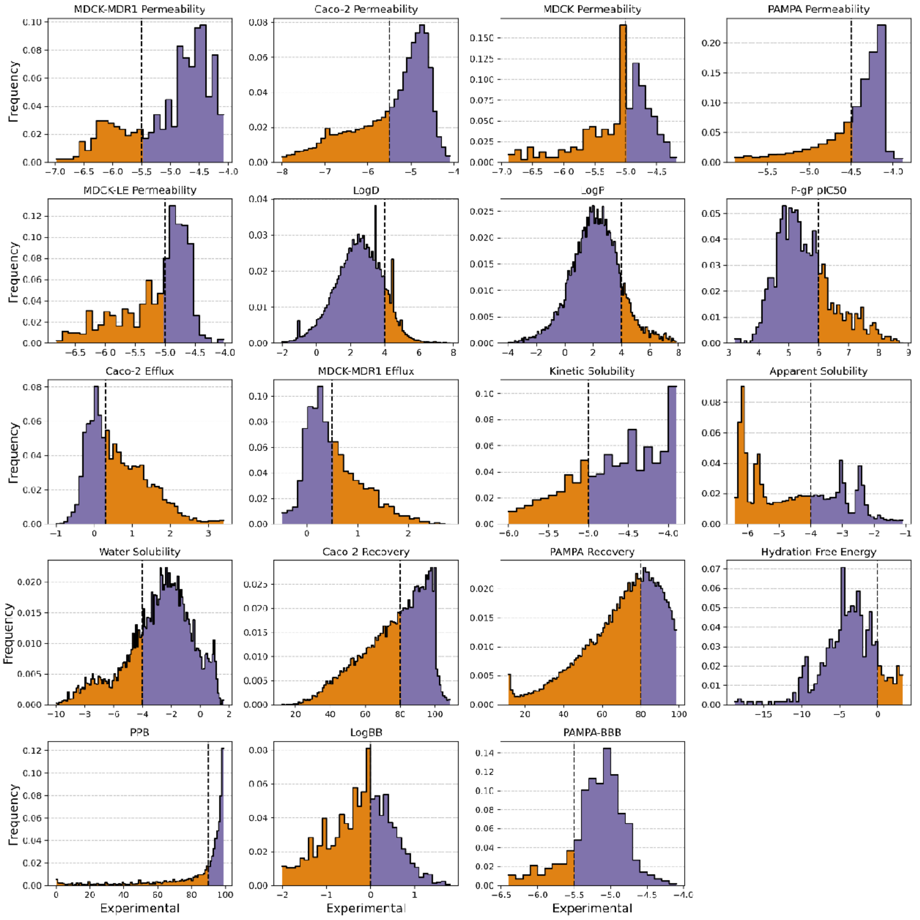


Figure S2: Distribution of the industrial experimental measurement. Undesired range of properties are colored in orange and defined using Medicinal Chemistry threshold from the literature or industry. Desirable property range are shown in purple.

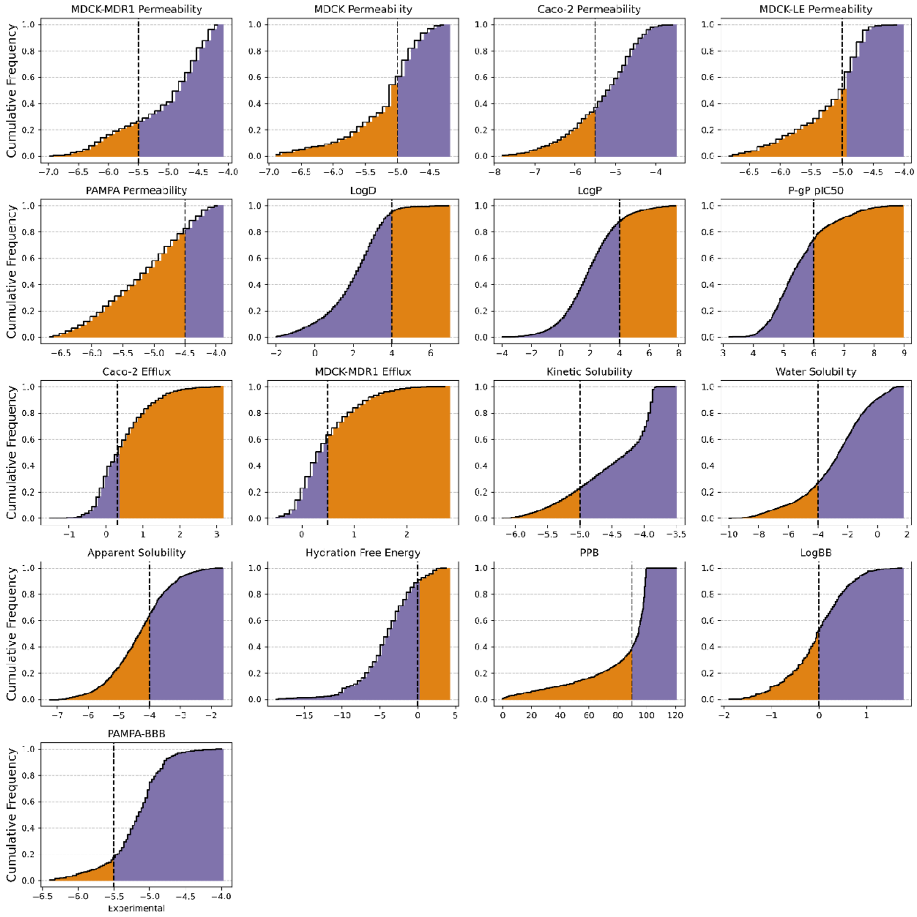


Figure S3: Cumulative distribution of the public experimental measurement. Undesired range of properties are colored in orange and defined using Medicinal Chemistry threshold from the literature or industry. Desirable property range are shown in purple.

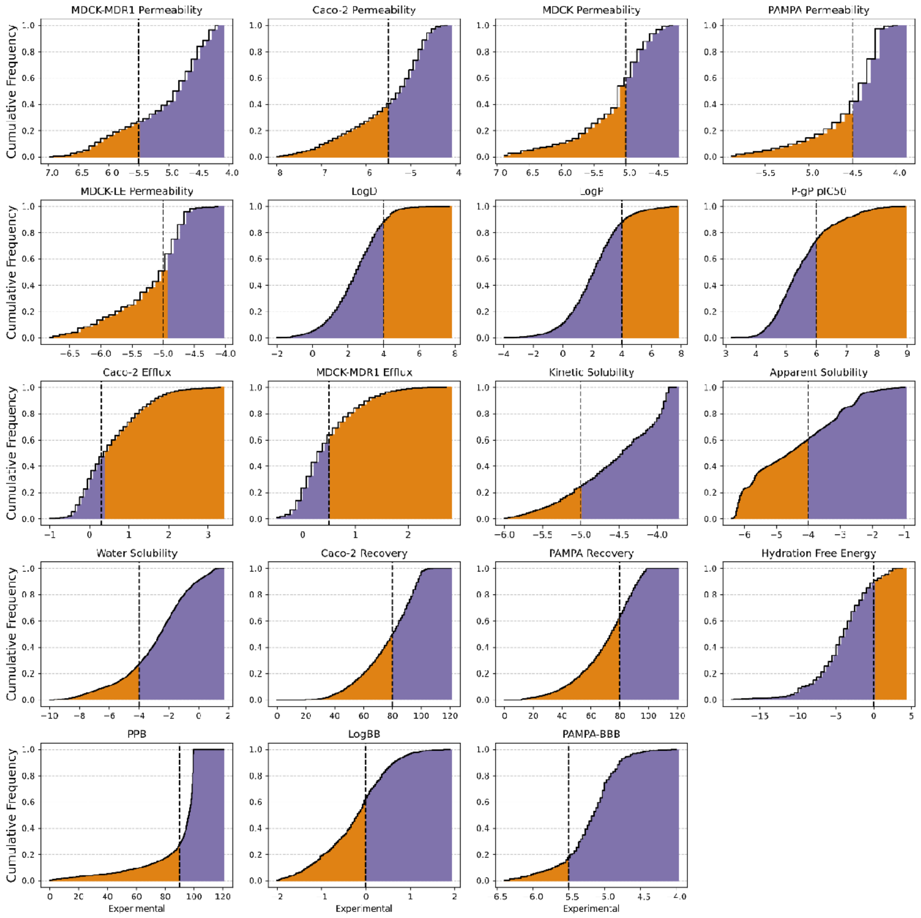


Figure S4: Cumulative distribution of the industrial experimental measurement. Undesired range of properties are colored in orange and defined using Medicinal Chemistry threshold from the literature or industry. Desirable property range are shown in purple.

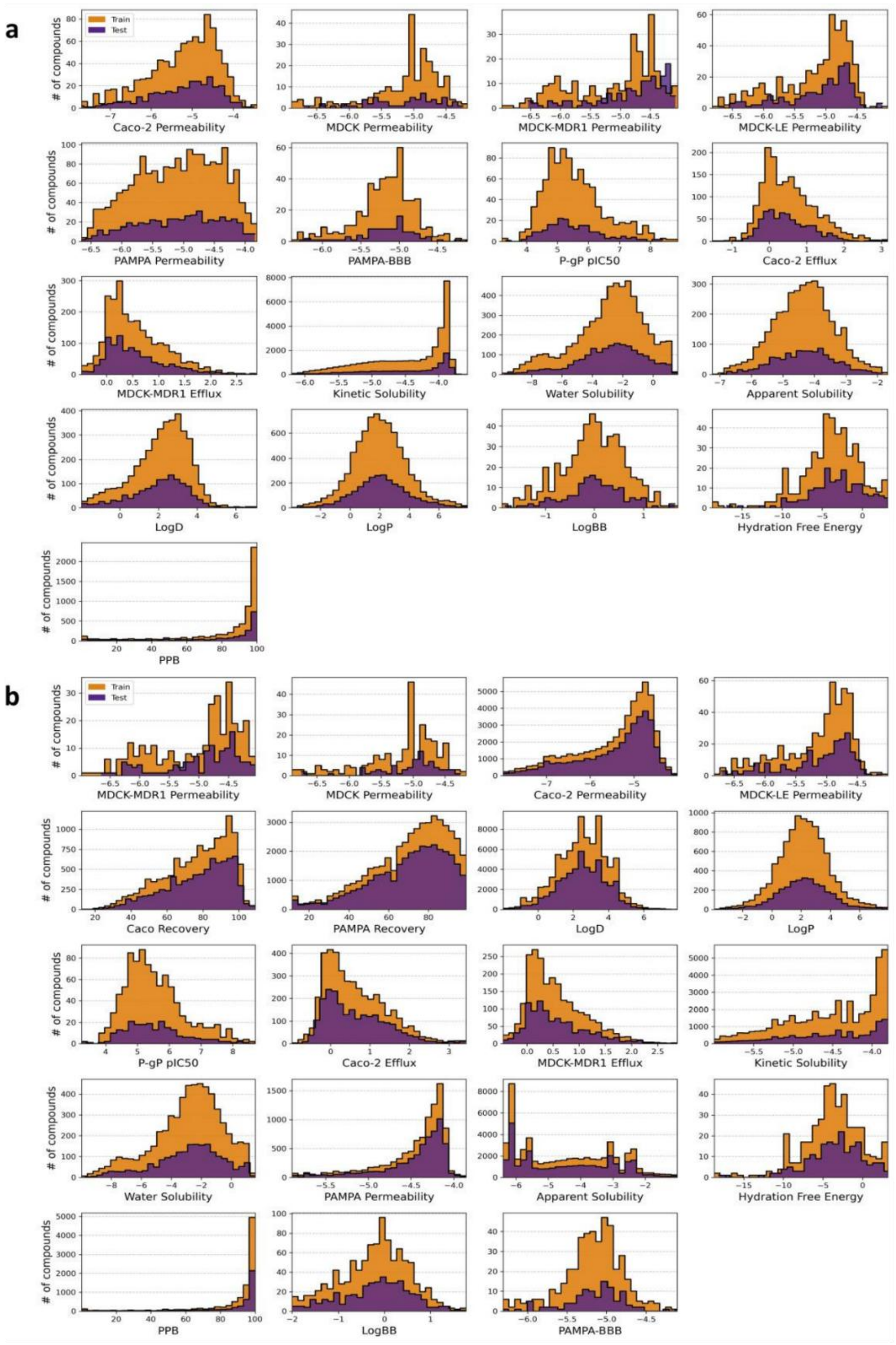

a
Train
Test
# of compounds
Caco-2 Permeability
MDCK Permeability
MDCK-MDR1 Permeability
MDCK-LE Permeability
PAMPA Permeability
PAMPA-BBB
P-gP pIC50
Caco-2 Efflux
MDCK-MDR1 Efflux
Kinetic Solubility
Water Solubility
Apparent Solubility
LogD
LogP
LogBB
Hydration Free Energy
PPB
b
Train
Test
MDCK-MDR1 Permeability
MDCK Permeability
Caco-2 Permeability
MDCK-LE Permeability
Caco Recovery
PAMPA Recovery
LogD
LogP
P-gP pIC50
Caco-2 Efflux
MDCK-MDR1 Efflux
Kinetic Solubility
Water Solubility
PAMPA Permeability
Apparent Solubility
Hydration Free Energy
PPB
LogBB
PAMPA-BBB

Figure S5: Distribution of experimental Measurement after train and test Split. (a) Public (b) Industrial.

Table S3: Hyperparameters search space per public STL model.

Model	Hyperparameter	Possible values
RF	max_depth	[-, 10, 20, 30]
	n_estimators	[100, …, 500]
SVR	C	[0.1, …, 10]
	Epsilon	[0.01, …, 1]
	kernel	[rbf]
GTM	N_nodes	[2000, …, 5000]
	N_rbf	[30, …, 80]
	Width_rbf	[0.001, …, 10]
	Regularization	[0.01, …, 100]
ChemProp	ffn_hidden_size	[200, 300, 400]
	ffn_num_layers	[1, 2, 3]
AttentiveFP	fingerprint_dim	[200, 300, 400]
	radius	[2, 3, 4, 5]
	T	[2, 3, 4, 5]

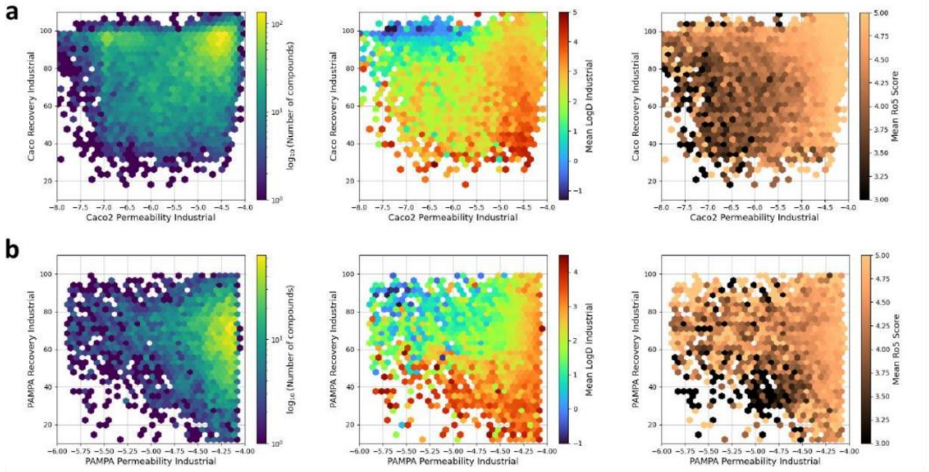


Figure S6: Correlation between permeability and recovery for Caco-2 and PAMPA. Coloration by the base-10 logarithm of the number of compounds, the mean experimental LogD$_{7.4}$, and the mean Lipinsky score per bin. (a) Caco-2 recovery against Caco-2 apparent permeability. (b) PAMPA recovery against PAMPA apparent permeability.

Table S4: Optimal values of hyperparameters per public STL model.

Endpoint	Assay	RF	SVR	GTM	ChemProp	AttentiveFP
Efflux Ratio	Caco-2	9_5_1, -,242	9_5_1, 8.85, 0.38	9_5_2, 4245, 43	200, 2	200, 3, 4
	MDCK-MDR1	3_4_0, 30, 279	3_5_1, 0.9, 0.28	3_4_1, 3700, 62	200, 3	300, 5, 5
Apparent Permeability	Caco-2	9_3_1, 30, 298	9_3_0, 9.50, 0.07	PcFP, 4660, 76	300, 2	200, 5, 2
	MDCK	9_4_3, 30, 356	9_5_0, 1.45, 0.11	9_4_0, 3490, 63	200, 2	400, 3, 4
	MDCK-LE	6_3_0, -, 469	3_4_1, 5.86, 0.02	3_4_0, 3690, 70	400, 1	200, 4, 3
	MDCK-MDR1	6_4_1, -, 413	6_4_1, 5.47, 0.65	6_4_2, 3606, 56	200, 3	200, 4, 5
	PAMPA	9_4_1, -, 477	6_5_1, 4.85, 0.12	6_4_0, 3020, 56	400, 2	200, 3, 4
	PAMPA-BBB	9_3_0, 20, 446	9_3_1, 7.65, 0.18	Ecfp6_2048, 2300, 46	300, 1	300, 4, 3
Distribution	PPB	6_3_3, -, 199	6_3_1, 5.85, 0.57	Rdkit2D, 3660, 39	200, 1	400, 3, 4
	LogBB	9_5_2, 30, 185	9_5_2, 7.4, 0.31	SubFP, 3600, 59	300, 2	200, 4, 4
Inhibition	pIC50 P-gP	6_5_0, 30, 432	9_4_3, 1.77, 0.81	ExtFP, 4180, 37	300, 3	400, 5, 3
Lipophilicity	HFE	9_4_3, 20, 338	9_3_0, 5.65, 0.21	9_3_3, 2700, 56	200, 1	200, 4, 4
	$LogS_{app}$	Rdkit2D, 30, 300	6_5_1, 9.87, 0.18	9_3_2, 2420, 61	200, 3	200, 4, 3
	$LogS_k$	Mordred, -, 323	Rdkit2D, 6.9, 0.40	Rdkit2D, 2295, 64	200, 2	300, 3, 5
	$LogS_w$	Rdkit2D, -, 300	Mordred, 9.5, 0.2	Rdkit2D, 4200, 72	300, 2	200, 3, 4
	LogP	9_5_2, 20, 295	6_5_1, 3.2, 0.25	6_5_2, 4325, 60	200, 2	200, 4, 4
	$LogD_{7.4}$	3_3_1, -, 229	9_5_2, 6.2, 0.6	6_5_1, 3645, 64	300, 3	300, 3, 5

ISIDA descriptors follow a notation such as F_L_O, where:

- F (Fragmentation Type): Defines the method used to generate molecular fragments.
  - 3: Sequences of atoms and bonds + atom count.
  - 6: Atom-centered fragments based on sequences of atoms and bonds + atom count.
  - 9: Atom-centered fragments based on sequences of atoms and bonds of fixed length + atom count.
- L (Path Length): Specifies the range of fragment lengths, measured in atoms.
  - Represents the minimum and maximum number of atoms considered in a fragment.
- O (Fragmentation Options): Additional settings applied to the fragmentation process.
  - 0: None (default).
  - 1: UseFormalCharge.
  - 2: DoAllWays.
  - 3: UseFormalCharge + DoAllWays.

Example:

- 3_4_0: Sequences of atoms and bonds + atom count, with fragments ranging from 2 to 4 atoms in length, and no additional options applied.

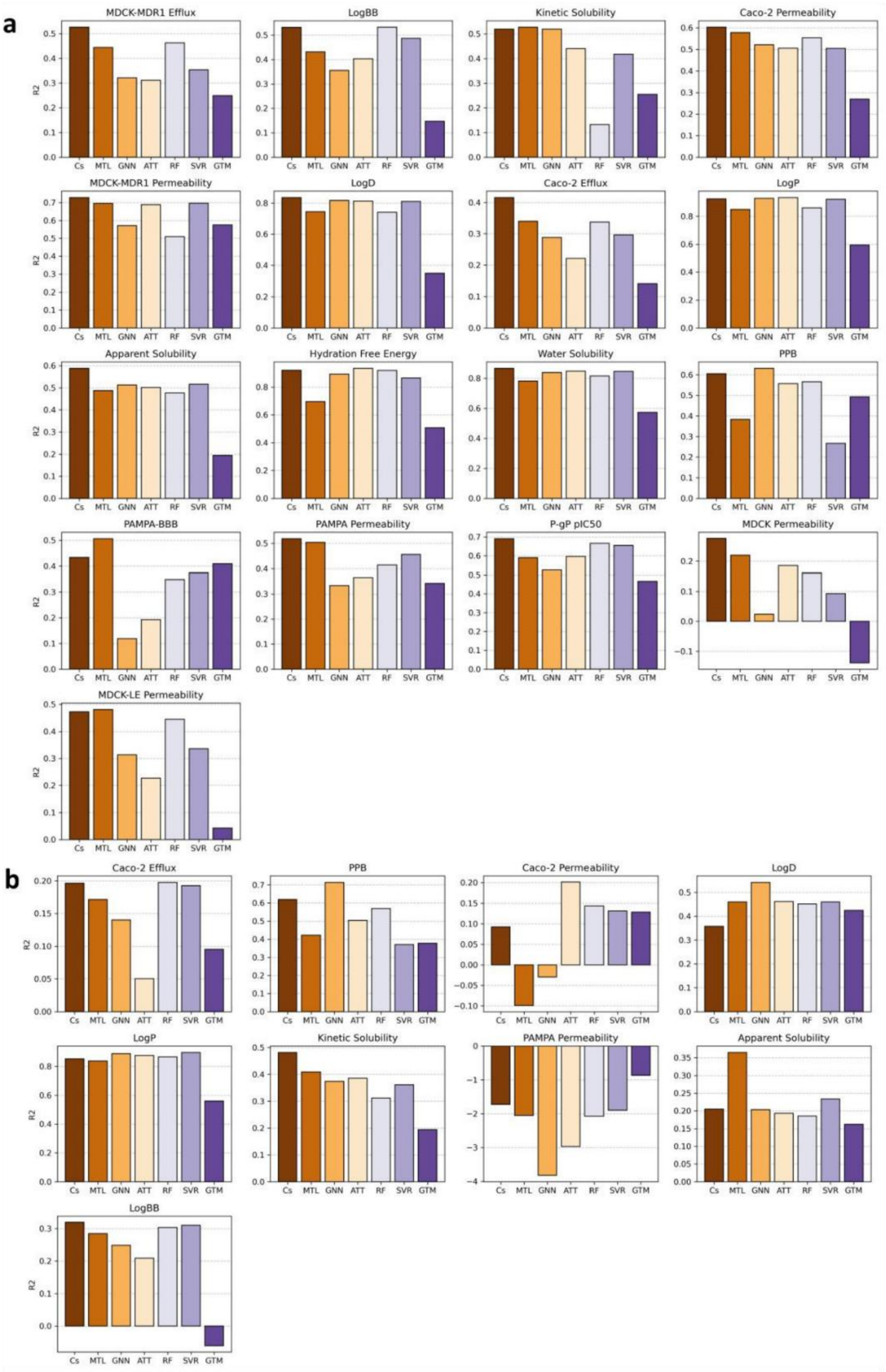

a
MDCK-MDR1 Efflux
LogBB
Kinetic Solubility
Caco-2 Permeability
MDCK-MDR1 Permeability
LogD
Caco-2 Efflux
LogP
Apparent Solubility
Hydration Free Energy
Water Solubility
PPB
PAMPA-BBB
PAMPA Permeability
P-gP pIC50
MDCK Permeability
MDCK-LE Permeability
b
Caco-2 Efflux
PPB
Caco-2 Permeability
LogD
LogP
Kinetic Solubility
PAMPA Permeability
Apparent Solubility
LogBB
R2
Cs
MTL
GNN
ATT
RF
SVR
GTM

Figure S7: Barplots of the R2 performance from the STL and MTL models. The bar color depends on the method. (a) Performances on the Public Test Set. (b) Performances on the Industrial External Set. Cs: Consensus Prediction taking the mean; MTL: MTL-GNN using ChemProp; GNN: ChemProp STL; ATT: AttentiveFP STL; RF: RandomForest STL; SVR: Support Vector Machine STL; GTM: Generative Topographic Mapping STL.

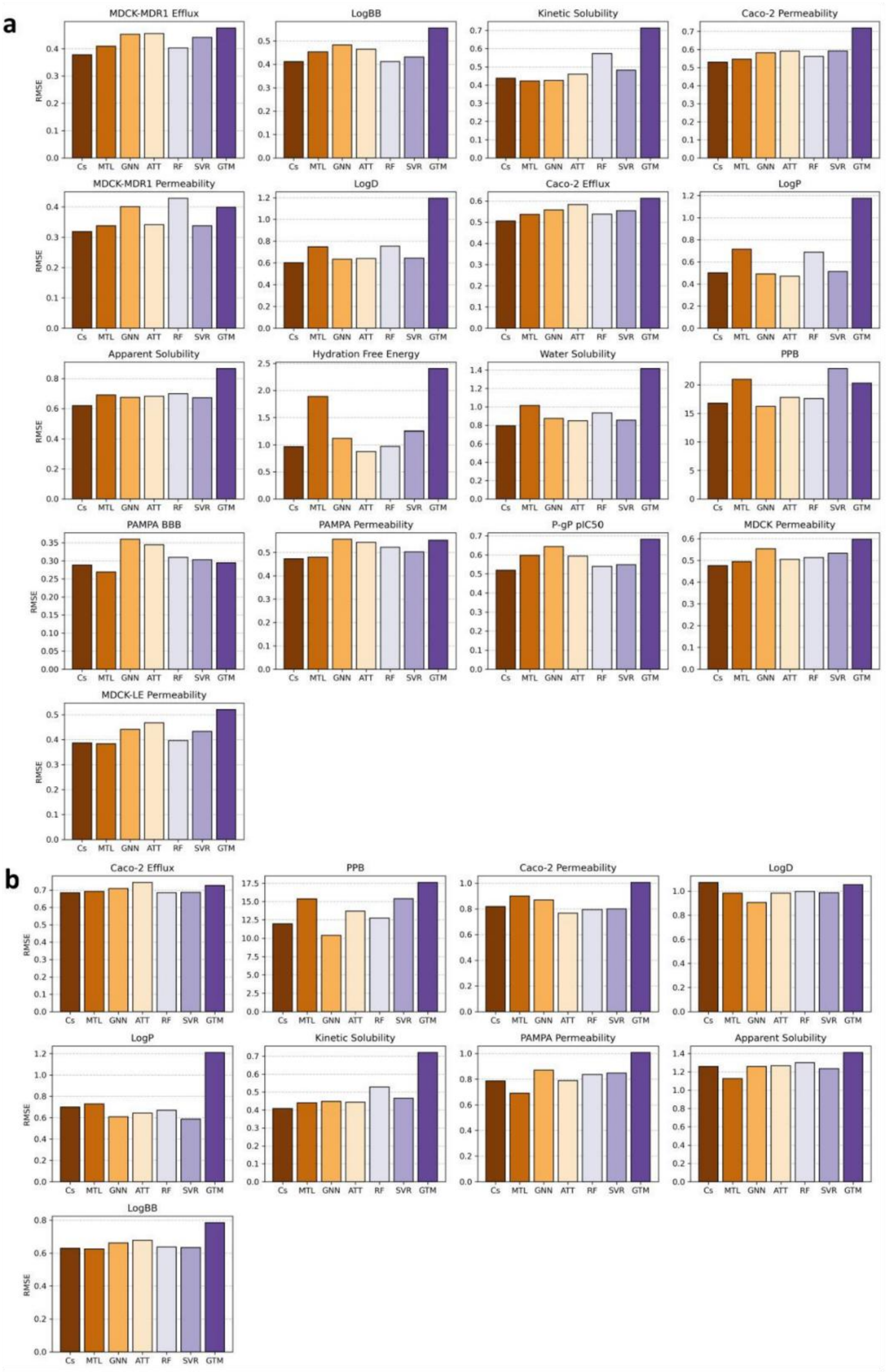

a
MDCK-MDR1 Efflux
LogBB
Kinetic Solubility
Caco-2 Permeability
MDCK-MDR1 Permeability
LogD
Caco-2 Efflux
LogP
Apparent Solubility
Hydration Free Energy
Water Solubility
PPB
PAMPA BBB
PAMPA Permeability
P-gP pIC50
MDCK Permeability
MDCK-LE Permeability
b
Caco-2 Efflux
PPB
Caco-2 Permeability
LogD
LogP
Kinetic Solubility
PAMPA Permeability
Apparent Solubility
LogBB
RMSE
Cs
MTL
GNN
ATT
RF
SVR
GTM

Figure S8: Barplots of the RMSE performance from the STL and MTL models. The bar color depends on the method. (a) Performances on the Public Test Set. (b) Performances on the Industrial External Set. Cs: Consensus Prediction taking the mean; MTL: MTL-GNN using ChemProp; GNN: ChemProp STL; ATT: AttentiveFP STL; RF: RandomForest STL; SVR: Support Vector Machine STL; GTM: Generative Topographic Mapping STL.

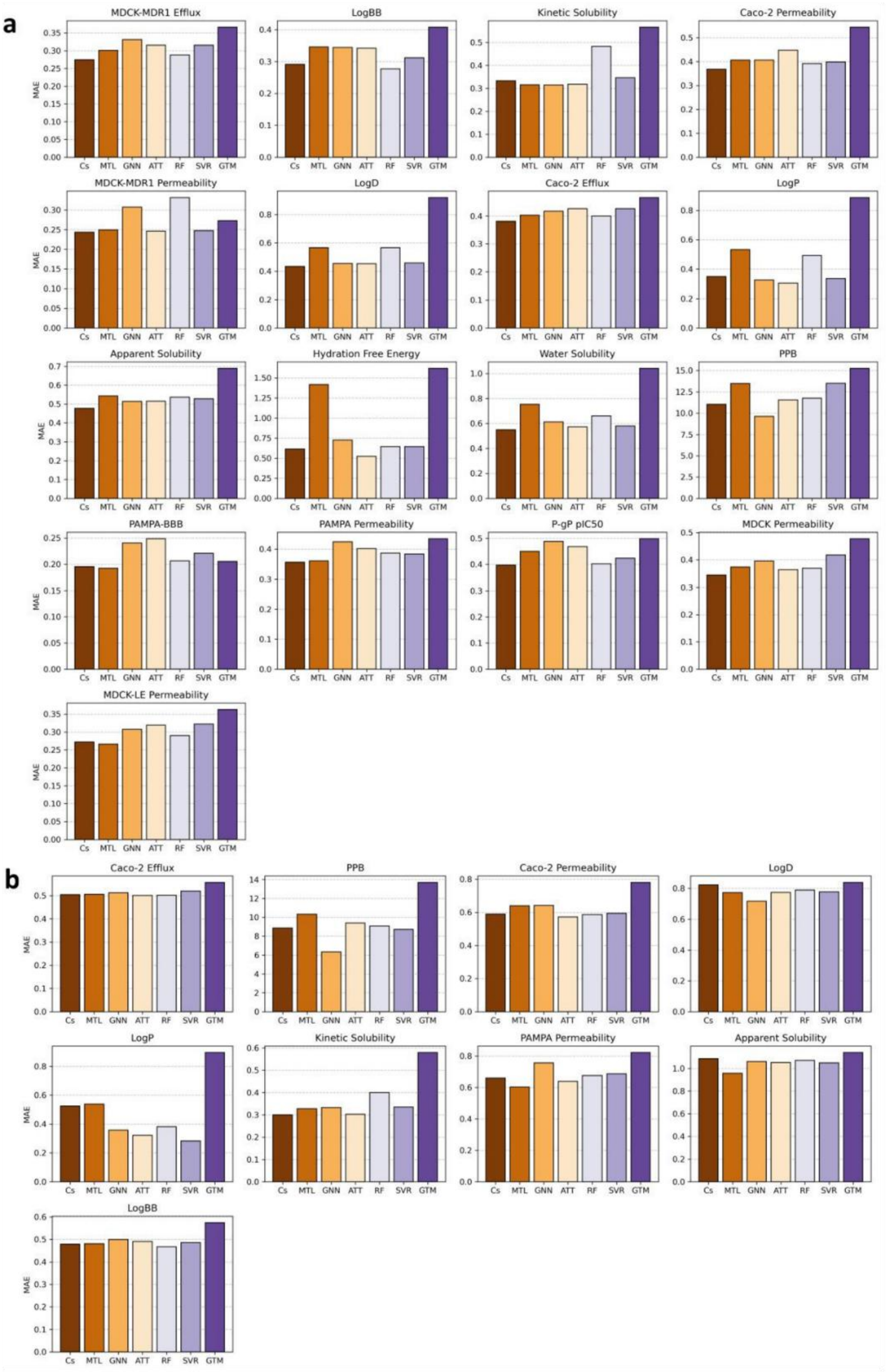

a
MDCK-MDR1 Efflux
LogBB
Kinetic Solubility
Caco-2 Permeability
MDCK-MDR1 Permeability
LogD
Caco-2 Efflux
LogP
Apparent Solubility
Hydration Free Energy
Water Solubility
PPB
PAMPA-BBB
PAMPA Permeability
P-gP pIC50
MDCK Permeability
MDCK-LE Permeability
b
Caco-2 Efflux
PPB
Caco-2 Permeability
LogD
LogP
Kinetic Solubility
PAMPA Permeability
Apparent Solubility
LogBB
MAE
Cs
MTL
GNN
ATT
RF
SVR
GTM

Figure S9: Barplots of the MAE performance from the STL and MTL models. The bar color depends on the method. (a) Performances on the Public Test Set. (b) Performances on the Industrial External Set. Cs: Consensus Prediction taking the mean; MTL: MTL-GNN using ChemProp; GNN: ChemProp STL; ATT: AttentiveFP STL; RF: RandomForest STL; SVR: Support Vector Machine STL; GTM: Generative Topographic Mapping STL.

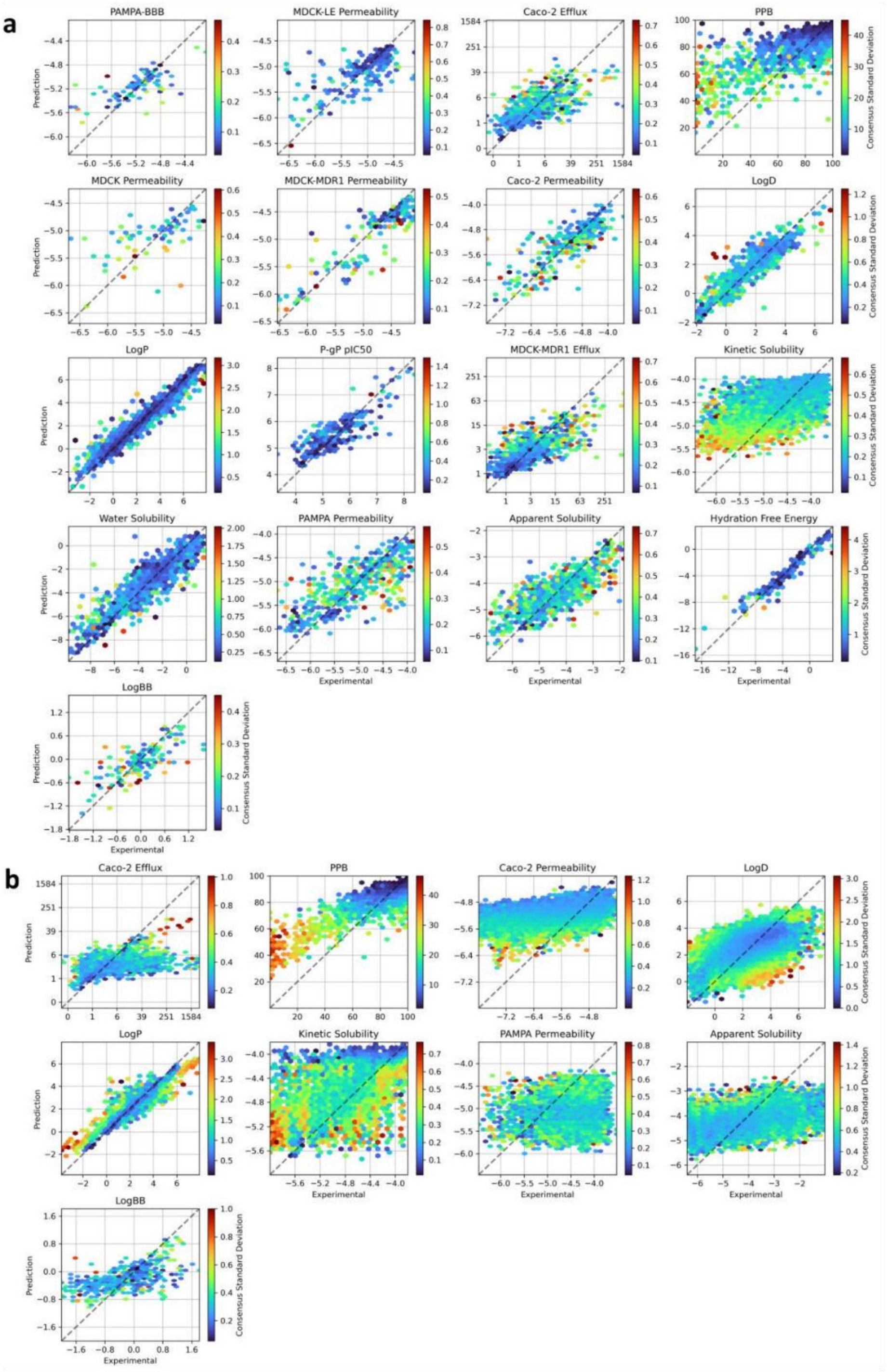

a
PAMPA-BBB
MDCK-LE Permeability
Caco-2 Efflux
PPB
MDCK Permeability
MDCK-MDR1 Permeability
Caco-2 Permeability
LogD
LogP
P-gP pIC50
MDCK-MDR1 Efflux
Kinetic Solubility
Water Solubility
PAMPA Permeability
Apparent Solubility
Hydration Free Energy
LogBB
Prediction
Experimental
Consensus Standard Deviation
b
Caco-2 Efflux
PPB
Caco-2 Permeability
LogD
LogP
Kinetic Solubility
PAMPA Permeability
Apparent Solubility
LogBB
Prediction
Experimental
Consensus Standard Deviation

Figure S10: Correlation between experimental measurement and consensus prediction per endpoint. The bins color depends on the standard deviation of the predictions over all models per compound. (a) Prediction on the Public Test Set. (b) Prediction on the Industrial External Set.

Table S5: **Performances of the public MTL-GNN Model applied on the public test set and the industrial GNN-MTL model applied of the industrial test.** The best performing model per endpoint is highlighted in bold for data shared between industrial and public sets.

Endpoint	Assay	Similar Data	Industrial			Public		
			R2	RMSE	MAE	R2	RMSE	MAE
Efflux Ratio	Caco-2	-	0.54	0.51	0.37	0.34	0.54	0.40
	MDCK-MDR1	Yes	0.44	0.42	0.31	**0.45**	**0.41**	**0.30**
Apparent Permeability	Caco-2	-	0.71	0.46	0.33	0.58	0.55	0.41
	MDCK	Yes	**0.32**	**0.40**	**0.30**	0.22	0.49	0.37
	MDCK-LE	Yes	0.44	0.42	0.29	**0.48**	**0.38**	**0.27**
	MDCK-MDR1	Yes	0.63	0.38	0.28	**0.70**	**0.34**	**0.25**
	PAMPA	-	0.52	0.27	0.19	0.50	0.48	0.36
	PAMPA-BBB	Yes	**0.67**	**0.23**	**0.17**	0.51	0.27	0.19
Recovery	Caco-2	-	0.41	13.67	10.94	-	-	-
	PAMPA	-	0.24	16.94	13.29	-	-	-
Distribution	LogBB	-	0.44	0.56	0.42	0.43	0.45	0.35
	PPB	-	0.46	15.04	8.60	0.38	20.96	13.49
Inhibition	pIC50 P-gP	Yes	0.51	0.67	0.49	**0.59**	**0.68**	**0.45**
Lipophilicity	$LogS_{app}$	-	0.68	0.8	0.6	0.49	0.69	0.54
	$LogS_{kin}$	-	0.48	0.41	0.32	0.53	0.42	0.32
	$LogS_{w}$	Yes	**0.80**	**1.00**	**0.79**	0.78	1.03	0.75
	LogP	-	0.82	0.76	0.56	0.85	0.72	0.53
	$LogD_{7.4}$	-	0.81	0.59	0.44	0.75	0.75	0.57
	HFE	Yes	0.69	1.94	1.38	**0.70**	**1.89**	**1.42**

Table S6: **Performances of the public MTL-GTM Model applied on the public test set.**

Endpoint	Assay	R2	
		MTL-GTM	Best
Efflux Ratio	Caco-2	0.29	0.34
	MDCK-MDR1	0.38	0.45
Apparent Permeability	Caco-2	0.55	0.58
	MDCK	0.19	0.28
	MDCK-LE	0.25	0.48
	MDCK-MDR1	0.57	0.70
	PAMPA	0.41	0.50
	PAMPA-BBB	0.23	0.51
Distribution	LogBB	0.23	0.43
	PPB	0.26	0.38
Inhibition	pIC50 P-gP	0.54	0.59
Lipophilicity	LogS$_{\mathrm{app}}$	0.36	0.49
	LogS$_{\mathrm{kin}}$	0.45	0.53
	LogS$_{\mathrm{w}}$	0.67	0.78
	LogP	0.59	0.85
	LogD$_{7.4}$	0.45	0.75
	HFE	0.53	0.70

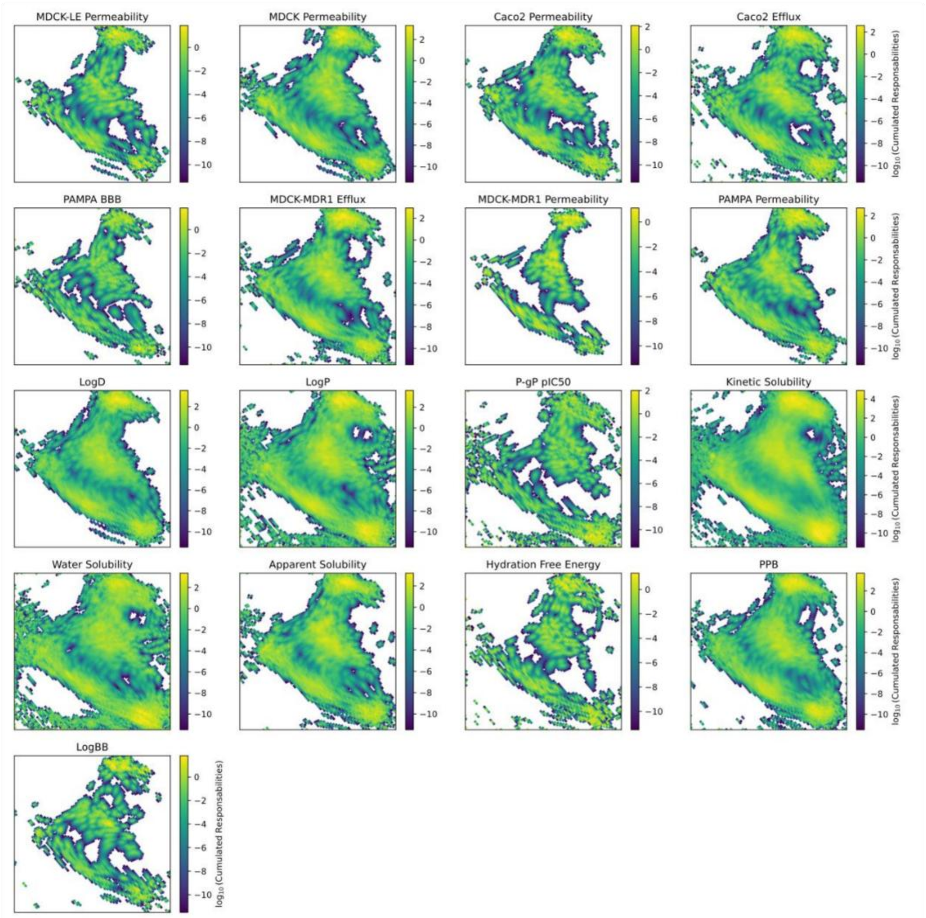


Figure S11: Density landscape of the public library over the public GTM manifold. The quantity of projected compounds is depicted as the base-10 logarithm of the cumulated sum of responsibilities.

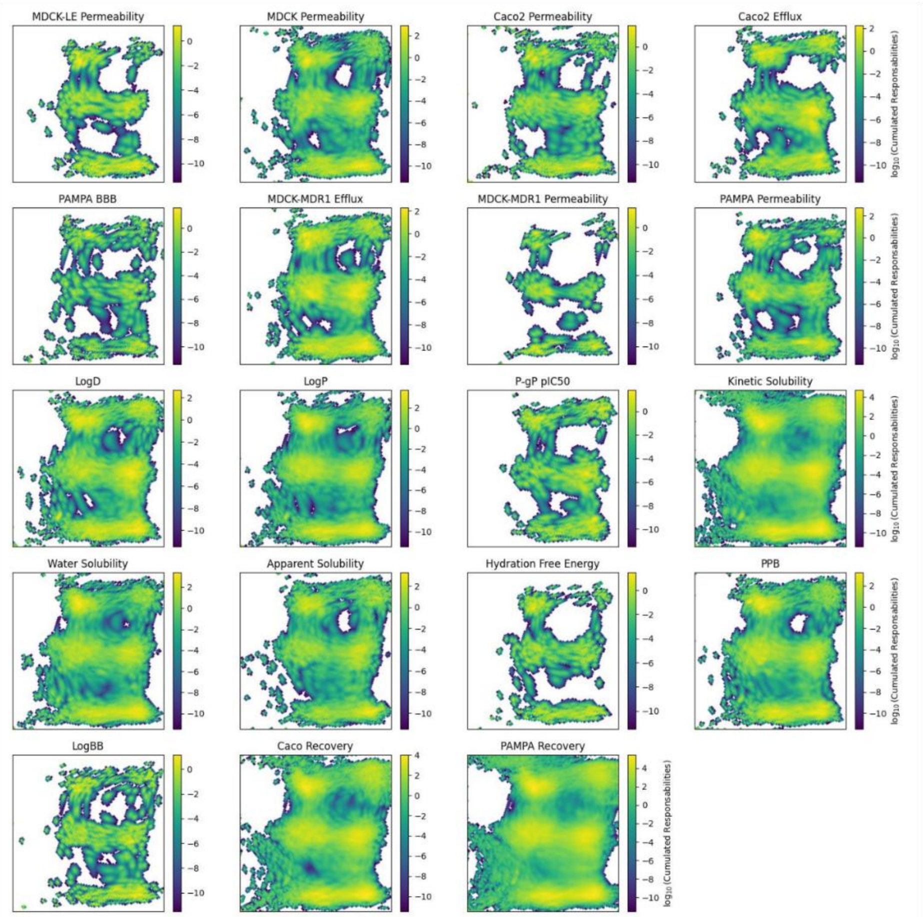


Figure S12: Density landscape of the industrial library over the industrial GTM manifold. The quantity of projected compounds is depicted as the base-10 logarithm of the cumulated sum of responsibilities.

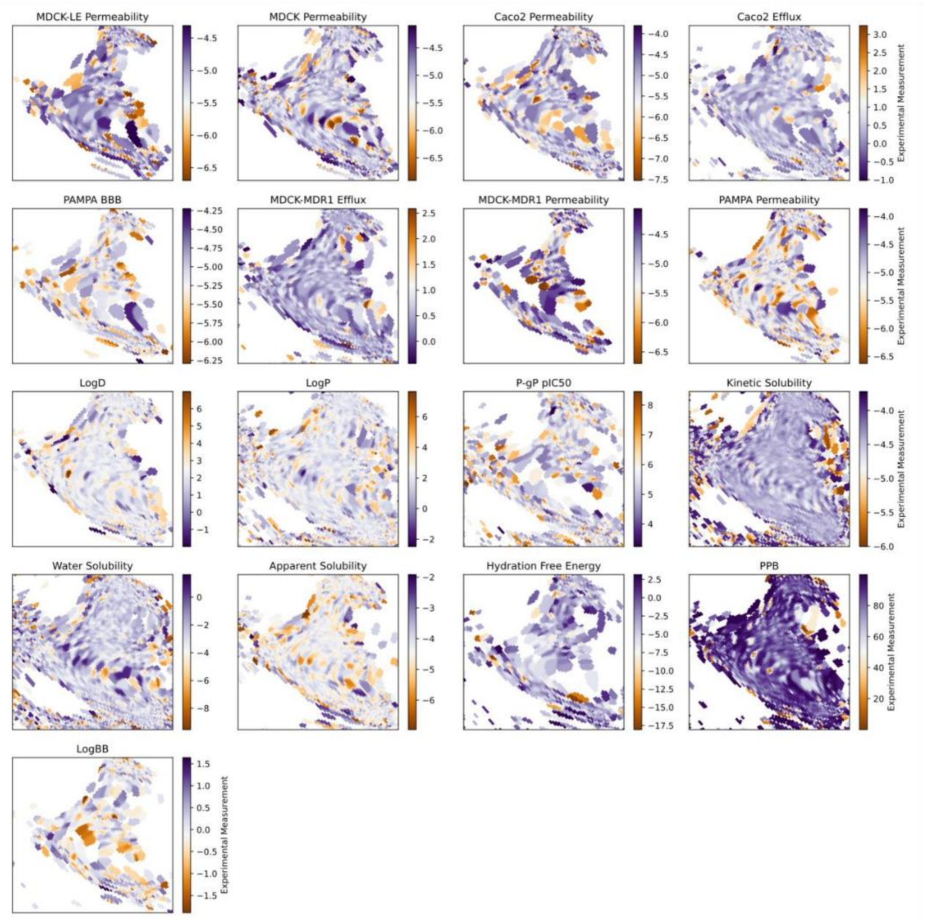


Figure S13: Continuous property landscape of the public library over the public GTM manifold. The responsibility-weighted property depicts the experimental measurement over the map. Undesired ranges are represented in orange and, the inverse in purple.

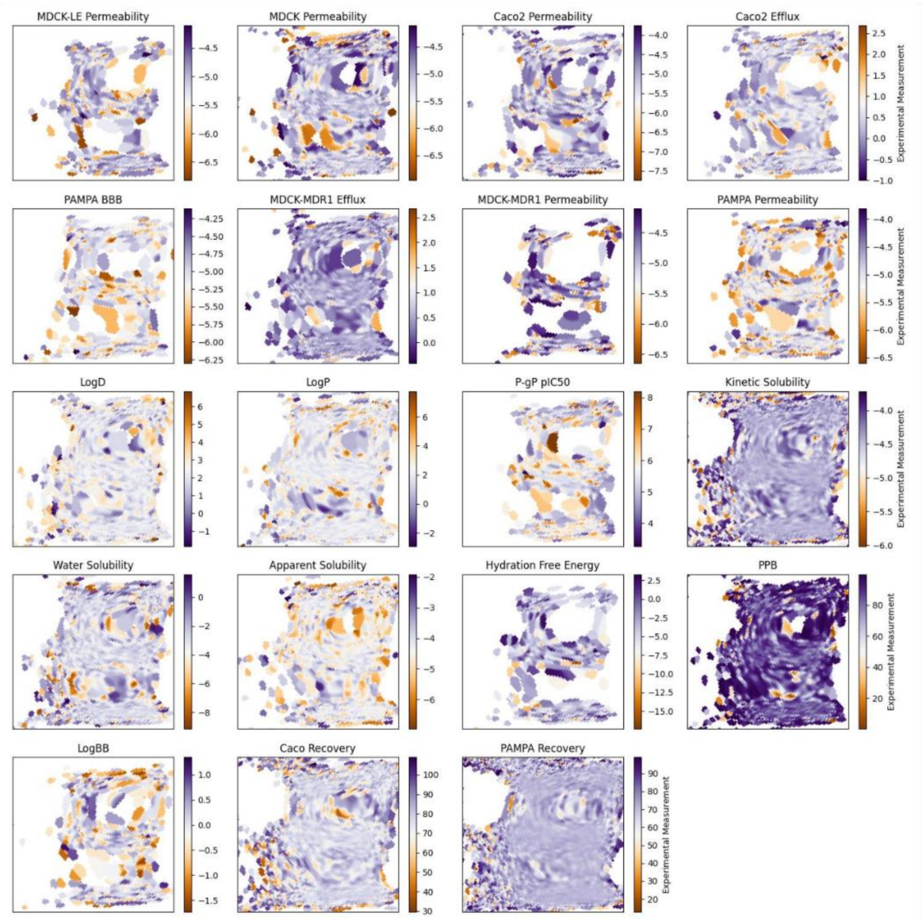


Figure S14: Continuous property landscape of the industrial library over the industrial GTM manifold. The responsibility-weighted property depicts the experimental measurement over the map. Undesired ranges are represented in orange and, the inverse in purple.

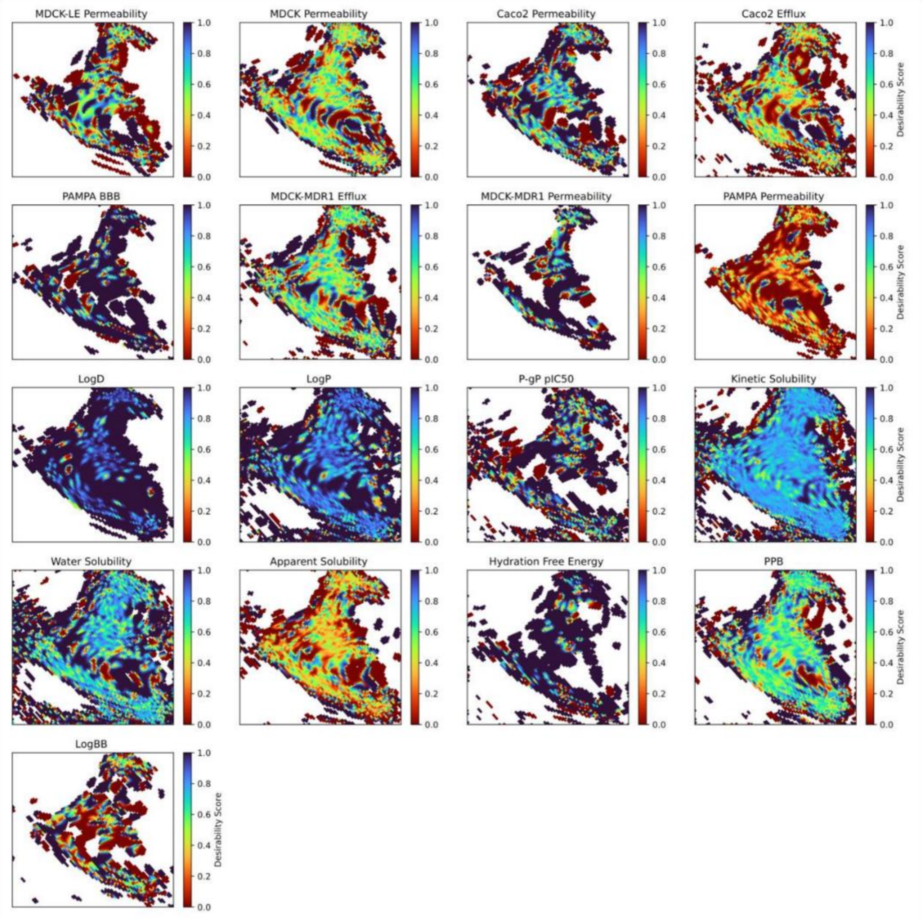


Figure S15: Class property landscape of the public library over the public GTM manifold. The responsibility-weighted class property depicts the desired class over the map. 1 represent the best range of measurement over/below a defined threshold, represented in blue, the inverse in red.

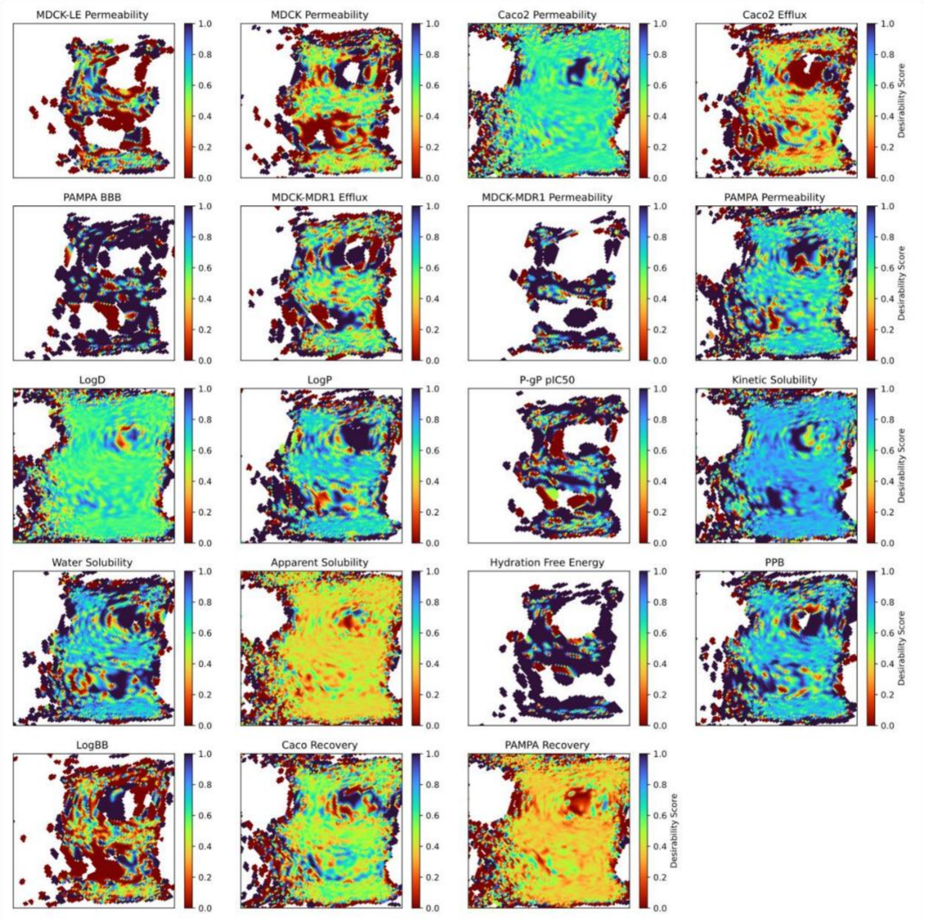


Figure S16: Class property landscape of the industrial library over the industrial GTM manifold. The responsibility-weighted class property depicts the desired class over the map. 1 represent the best range of measurement over/below a defined threshold, represented in blue, the inverse in red.

Figure S17: Score landscapes of the permeability, efflux, and solubility. (a) Public. (b) Industrial. The responsibility-weighted score is obtained by considering the overlay of all landscape associated to a certain endpoint. For instance, MDCK-MDR1 and Caco-2 landscapes are combined to obtain the efflux-specific score landscape. 1 represent the best score, represented in blue, the inverse in red.

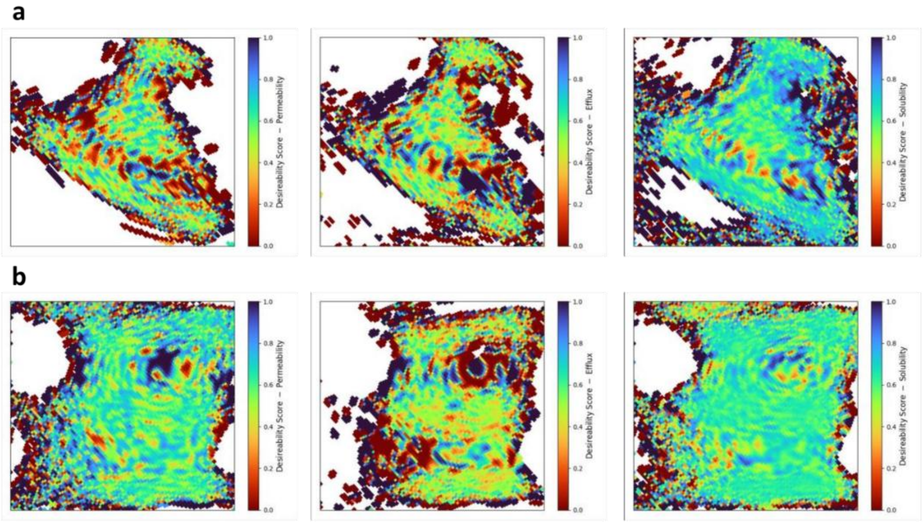


Table S7: Structure and details of the top 30 fragments found in poorly permeable public compounds subject to low solubility. Details present the number of counts per fragment over the full library with the mean absorption score over all compounds sharing this fragment. Score close to 0.5 indicates co-occurrence of permeability and solubility problem, without high efflux.

Fragment SMILES	Details
[*]c1cn2c(C)c([*])nc2c([*])n1	Count: 15 - Score: 0.30
[*]c1ccc2c3c1OC1CCCC4C(C2)N(C)CCC314	Count: 7 - Score: 0.16
[*]c1cc(N)n2nc([*])cc2n1	Count: 21 - Score: 0.15
[*]CC(C)CCC(O)C([*])C	Count: 19 - Score: 0.14
[*]C1CCC2(C)C(=CCC3C2C(O)CC2(C)C([*])C(O)CC32)C1	Count: 19 - Score: 0.14
[*]n1c([*])c(C#N)c(C)c1C	Count: 10 - Score: 0.14
[*]=CC=Nc1cc(C)ccn1	Count: 18 - Score: 0.13
[*]C(C)CC(C)CC(C)C	Count: 11 - Score: 0.13
[*]c1nc(Cl)c(Cl)nc1N	Count: 8 - Score: 0.13
[*]c1cc(Cl)c(Cl)c(Cl)c1[*]	Count: 9 - Score: 0.12
[*]CN1C(=O)COc2c1cc(C)cc2[N+](=O)[O-]	Count: 21 - Score: 0.12
[*]C(N)C(O)=NC1C(=O)N2C1SC(C)(C)C2[*]	Count: 15 - Score: 0.12
[*]c1cccc2nc[nH]c12	Count: 11 - Score: 0.11
[*]CC(C)(C)C(=O)NO	Count: 11 - Score: 0.11
[*]c1cc(Br)c(Br)c(Br)c1	Count: 16 - Score: 0.11
[*]=CCCCCC(=O)O	Count: 17 - Score: 0.10
[*]c1nn2c(O)c([*])c(C)nc2c1[*]	Count: 11 - Score: 0.10
[*]n1c(=O)[nH]c2cc([*])c(F)cc2c1=O	Count: 21 - Score: 0.10
[*]C(=O)NN=Cc1c(Cl)[nH]c2ccccc12	Count: 16 - Score: 0.10
[*]N1CC2CCC1CN2[*]	Count: 17 - Score: 0.10
[*]C(=O)C(N=Nc1ccc([*])cc1[N+](=O)[O-])C(C)=O	Count: 18 - Score: 0.09
[*]c1cc2cccnc2s1	Count: 29 - Score: 0.09
[*]n1c(=O)c([*])c(N)c2ccccc21	Count: 13 - Score: 0.09
[*]c1ccc(Cl)c2c1CC(O)CC2	Count: 18 - Score: 0.09
[*]NC1CCCc2ccc([*])cc21	Count: 12 - Score: 0.09
[*]C(CO)C(C)(C)C	Count: 12 - Score: 0.09
[*]C(=O)C(=COC)c1ccccc1[*]	Count: 18 - Score: 0.09
[*]C1Oc2c(Br)cc(S(F)(F)(F)(F)F)cc2C=C1C(=O)O	Count: 11 - Score: 0.09
[*]C(=O)CC(=O)NNC(=O)CCC	Count: 21 - Score: 0.09
[*]c1nc2cccc([*])c2s1	Count: 19 - Score: 0.09

Table S8: Structure and details of the top 30 fragments found in poorly permeable public compounds subject to high efflux. Details present the number of counts per fragment over the full library with the mean absorption score over all compounds sharing this fragment. Score close to -0.5 indicates co-occurrence of permeability and efflux problem, without poor solubility.

Fragment SMILES	Details
[*]c1cn2c(C)c([*])nc2c([*])n1	Count: 15 - Score: -0.30
[*]c1ccc2c3c1OC1CCCC4C(C2)N(C)CCC314	Count: 7 - Score: -0.16
[*]c1cc(N)n2nc([*])cc2n1	Count: 21 - Score: -0.15
[*]CC(C)CCC(O)C([*])C	Count: 19 - Score: -0.14
[*]C1CCC2(C)C(=CCC3C2C(O)CC2(C)C([*])C(O)CC32)C1	Count: 19 - Score: -0.14
[*]n1c([*])c(C#N)c(C)c1C	Count: 10 - Score: -0.14
[*]=CC=Nc1cc(C)ccn1	Count: 18 - Score: -0.13
[*]C(C)CC(C)CC(C)C	Count: 11 - Score: -0.13
[*]c1nc(Cl)c(Cl)nc1N	Count: 8 - Score: -0.13
[*]c1cc(Cl)c(Cl)c(Cl)c1[*]	Count: 9 - Score: -0.12
[*]CN1C(=O)COc2c1cc(C)cc2[N+](=O)[O-]	Count: 21 - Score: -0.12
[*]C(N)C(O)=NC1C(=O)N2C1SC(C)(C)C2[*]	Count: 15 - Score: -0.12
[*]c1cccc2nc[nH]c12	Count: 11 - Score: -0.11
[*]CC(C)(C)C(=O)NO	Count: 11 - Score: -0.11
[*]c1cc(Br)c(Br)c(Br)c1	Count: 16 - Score: -0.11
[*]=CCCCCC(=O)O	Count: 17 - Score: -0.10
[*]c1nn2c(O)c([*])c(C)nc2c1[*]	Count: 11 - Score: -0.10
[*]n1c(=O)[nH]c2cc([*])c(F)cc2c1=O	Count: 21 - Score: -0.10
[*]C(=O)NN=Cc1c(Cl)[nH]c2ccccc12	Count: 16 - Score: -0.10
[*]N1CC2CCC1CN2[*]	Count: 17 - Score: -0.10
[*]C(=O)C(N=Nc1ccc([*])cc1[N+](=O)[O-])C(C)=O	Count: 18 - Score: -0.09
[*]c1cc2cccnc2s1	Count: 29 - Score: -0.09
[*]n1c(=O)c([*])c(N)c2ccccc21	Count: 13 - Score: -0.09
[*]c1ccc(Cl)c2c1CC(O)CC2	Count: 18 - Score: -0.09
[*]NC1CCCc2ccc([*])cc21	Count: 12 - Score: -0.09
[*]C(CO)C(C)(C)C	Count: 12 - Score: -0.09
[*]C(=O)C(=COC)c1ccccc1[*]	Count: 18 - Score: -0.09
[*]C1Oc2c(Br)cc(S(F)(F)(F)(F)F)cc2C=C1C(=O)O	Count: 11 - Score: -0.09
[*]C(=O)CC(=O)NNC(=O)CCC	Count: 21 - Score: -0.09
[*]c1nc2cccc([*])c2s1	Count: 19 - Score: -0.09

Figure S18: Structure of the top 30 fragments found in poorly permeable public compounds subject to low solubility. Fragments are ordered from worst to acceptable absorption score.

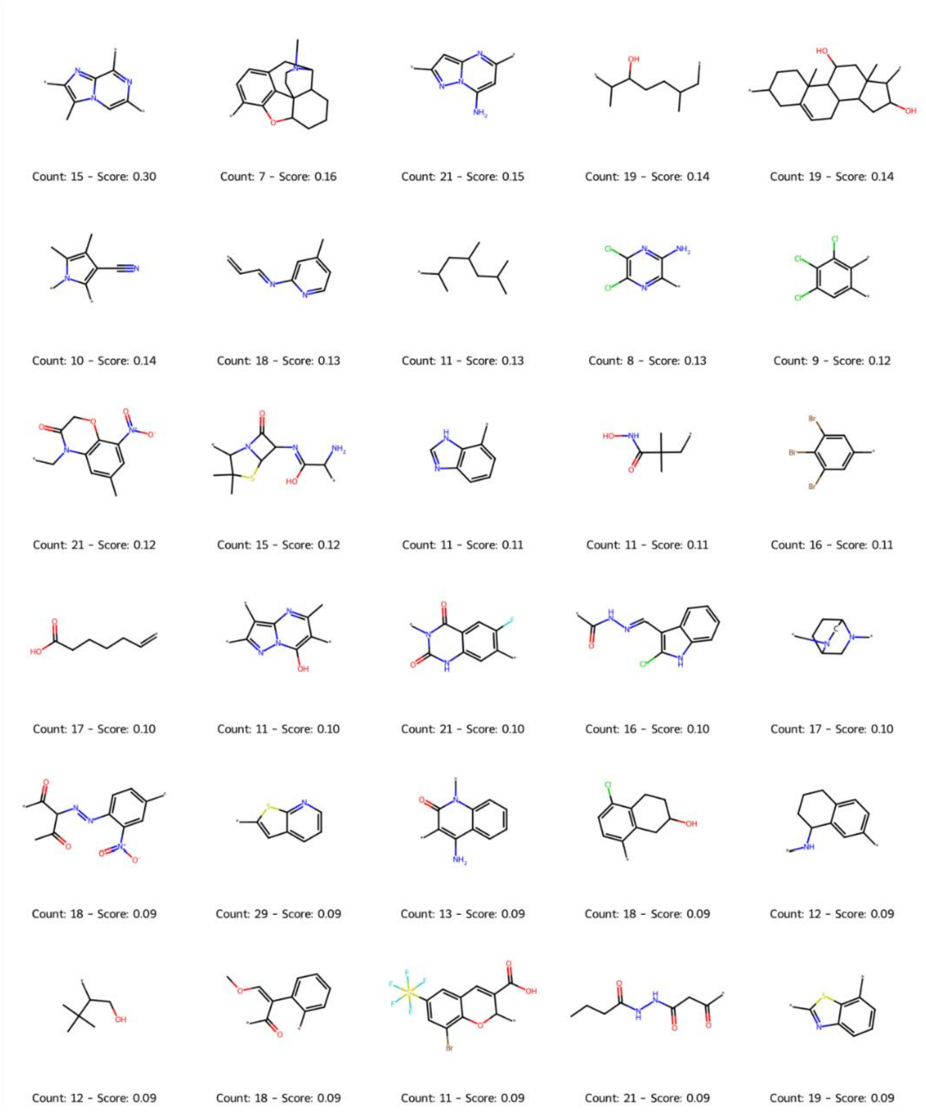
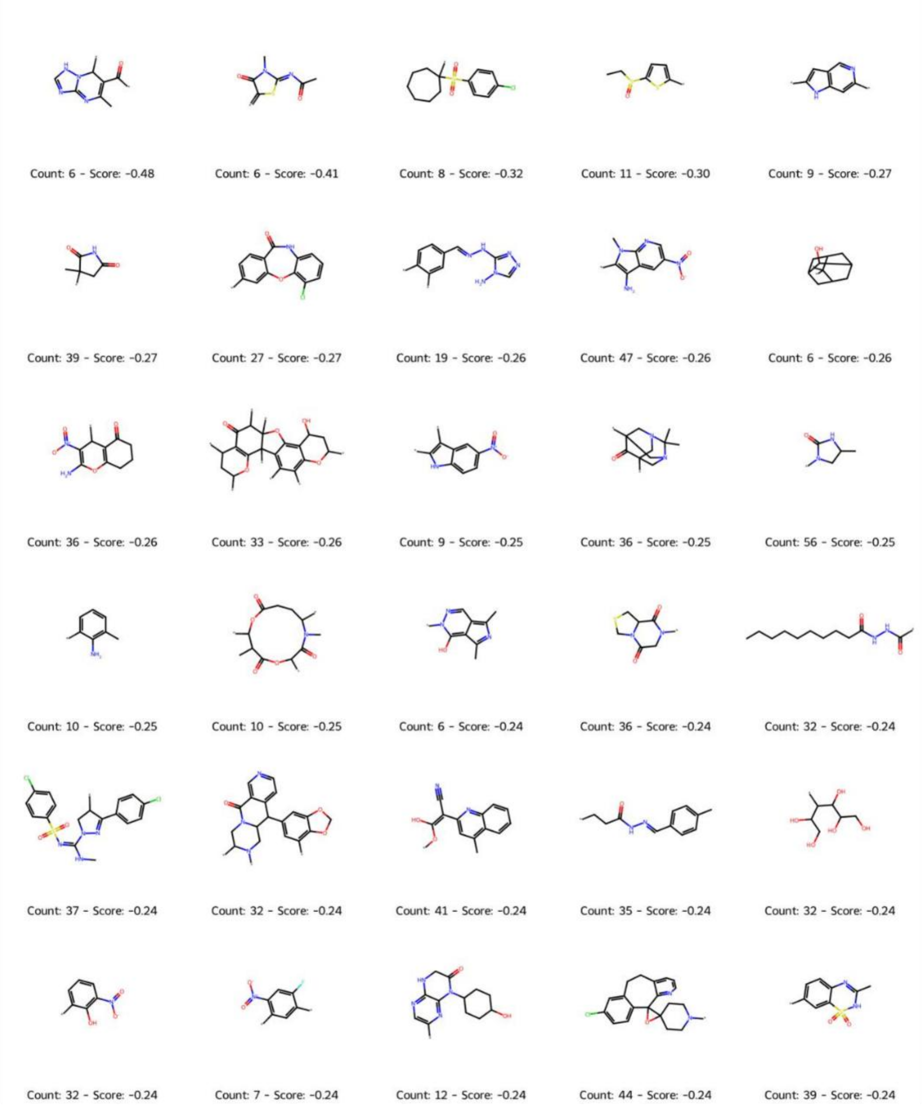


Figure S19: Structure of the top 30 fragments found in poorly permeable public compounds subject to high efflux. Fragments are ordered from worst to acceptable absorption score.



References

(1) Di, L.; Kerns, E. H. *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*; Academic Press, 2015.

(2) Jiang, L.; Kumar, S.; Nuechterlein, M.; Reyes, M.; Tran, D.; Cabebe, C.; Chiang, P.; Reynolds, J.; Carrier, S.; Sun, Y.; Eddershaw, P.; Hay, T.; Chen, W.; Feng, B. Application of a High-Resolution in Vitro Human MDR1-MDCK Assay and in Vivo Studies in Preclinical Species to Improve Prediction of CNS Drug Penetration. *Pharmacology Research & Perspectives* **2022**, *10* (1), e00932. https://doi.org/10.1002/prp2.932.

(3) *Evaluating the Utility of Canine Mdr1 Knockout Madin-Darby Canine Kidney I Cells in Permeability Screening and Efflux Substrate Determination | Molecular Pharmaceutics*. https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.8b00688 (accessed 2025-01-06).

(4) Varma, M. V.; Gardner, I.; Steyn, S. J.; Nkansah, P.; Rotter, C. J.; Whitney-Pickett, C.; Zhang, H.; Di, L.; Cram, M.; Fenner, K. S.; El-Kattan, A. F. pH-Dependent Solubility and Permeability Criteria for Provisional Biopharmaceutics Classification (BCS and BDDCS) in Early Drug Discovery. *Mol. Pharmaceutics* **2012**, *9* (5), 1199–1212. https://doi.org/10.1021/mp2004912.

(5) Wang, Q.; Rager, J. D.; Weinstein, K.; Kardos, P. S.; Dobson, G. L.; Li, J.; Hidalgo, I. J. Evaluation of the MDR-MDCK Cell Line as a Permeability Screen for the Blood–Brain Barrier. *International Journal of Pharmaceutics* **2005**, *288* (2), 349–359. https://doi.org/10.1016/j.ijpharm.2004.10.007.

(6) Mensch, J.; Melis, A.; Mackie, C.; Verreck, G.; Brewster, M. E.; Augustijns, P. Evaluation of Various PAMPA Models to Identify the Most Discriminating Method for the Prediction of BBB Permeability. *European Journal of Pharmaceutics and Biopharmaceutics* **2010**, *74* (3), 495–502. https://doi.org/10.1016/j.ejpb.2010.01.003.

(7) Kratochwil, N. A.; Huber, W.; Müller, F.; Kansy, M.; Gerber, P. R. Predicting Plasma Protein Binding of Drugs: A New Approach. *Biochemical Pharmacology* **2002**, *64* (9), 1355–1374. https://doi.org/10.1016/S0006-2952(02)01074-2.

(8) Venkatraman, V. FP-ADMET: A Compendium of Fingerprint-Based ADMET Prediction Models. *Journal of Cheminformatics* **2021**, *13* (1), 75. https://doi.org/10.1186/s13321-021-00557-5.

(9) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J Comput Aided Mol Des* **2014**, *28* (7), 711–720. https://doi.org/10.1007/s10822-014-9747-x.

Outline

By integrating both public and proprietary data, this work spotlights the importance of tailored predictive tools in industrial drug development to curb applicability domain issues and fortify the reliability of absorption predictions. All developed models and curated datasets are made publicly available to fuel ongoing research and streamline the drug discovery pipeline.

Chapter 6. Large-scale ADMET Profiling

6.1. From ADMET to Bioactivity Prediction

Introduction

After absorption and distribution, some molecules reach their targets unchanged, but many are subject to metabolic transformation, a key determinant of how long it remains in the body, how it is eliminated, and whether it will lead to toxic species. While distribution governs where a drug travels, metabolism determines what the body does to it: whether it is inactivated, bioactivated, or rendered hydrophilic for excretion. These transformations are crucial in clearance, impacting the duration of action and the risk profile.

For orally administered drugs, the first major metabolic hurdle is the “first-pass effect”. Before reaching systemic circulation, a fraction of the absorbed dose may be metabolized in the intestinal wall or liver. Enzymes such as esterases and cytochromes can significantly reduce the concentration of parent drug that enters the bloodstream.

Beyond first-pass metabolism, the drug may continue to undergo biotransformation in the liver, as well as in other compartments such as the intestine, lung, kidney, and even the plasma. These reactions govern not only how the compound is modified, but also how efficiently it can be cleared via hepatic or renal pathways. This phase dictates the pharmacokinetic profile, ultimately influencing dosing and formulation strategies.

Main Terminology

Off-target refers to unintended interactions between a drug and biological targets other than the intended receptor or enzyme, potentially causing adverse effects.

Metabolic oxidation describes a Phase I process, typically involving cytochrome P450 enzymes, where a drug undergoes chemical modification (e.g., addition of oxygen) to increase its polarity.

Metabolic conjugation is a Phase II process in which polar groups (such as glucuronic acid or sulfate) are attached to a drug or its metabolites, enhancing water solubility and promoting excretion.

Metabolism

Metabolic oxidation and metabolic conjugation greatly affect a drug's fate in the body by reducing its bioavailability and potentially generating reactive intermediates. These processes are categorized into Phase I and Phase II reactions.[152] Phase I reactions, such as oxidation, dealkylation, hydroxylation, and deamination, introduce or expose functional groups, increasing a compound's polarity. These transformations are primarily mediated by enzymes such as cytochrome P450 (CYP450), flavin-containing monooxygenases (FMO), and esterases. Many drugs are primarily metabolized by CYP450 isoenzymes, particularly CYP3A4, CYP2C9, CYP2C19, CYP2D6, and CYP1A2. Most of the 200 most frequently prescribed medications in the US rely on these isoforms.[153]

Bioactivation is the enzymatic process where a non-toxic compound is converted into a reactive or toxic metabolite, often by cytochrome P450 enzymes.

Idiosyncratic reaction is an uncommon and unpredictable adverse response to a drug, arising from genetic or immunologic factors, and not related to its primary pharmacological action.

Phase II reactions involve conjugation, where functional groups such as glucuronides, sulfates, or acetyl groups are added to enhance aqueous solubility, facilitating excretion via bile or urine. A compound may serve as a substrate for a metabolic enzyme, be an inhibitor that blocks enzyme function, or act as an inducer that increases enzyme expression, further complicating the metabolism profile (**Figure 22**). Metabolic stability assays assess a compound's susceptibility to be metabolized. One of the most widely used assays is the liver microsomal stability test, where drug depletion over time is measured in microsomal fractions enriched with CYP enzymes. These assays help estimate intrinsic clearance by determining how rapidly a compound is metabolized under controlled conditions. On the other hand, inhibition assays target specific CYP isoforms, measuring whether a molecule inhibits a key metabolic pathway. These experiments are essential in predicting drug-drug interactions, as potent inhibitors can cause the perturbation of expected pharmacokinetics of co-administrated drugs.

Elimination

Elimination (a.k.a. Excretion) proceeds primarily through hepatic and renal routes. In the liver, drug molecules and their metabolites travel to hepatocytes via both the portal vein and the hepatic artery.[154] Enzymes responsible for drug metabolism reside in reticulum endoplasmic and mitochondria of the hepatocytes. Once metabolized, the active compound or its metabolites are secreted into bile canaliculi, which are physically separate from the blood supply. Bile then carries these substances to the intestine, where they can either be excreted with feces or reabsorbed into the bloodstream (enterohepatic recirculation). Any fraction of the drug that remains unmetabolized in the hepatic circulation eventually leaves the liver via the hepatic vein and reenters systemic circulation.

Clearance (CL) is a key parameter of xenobiotics elimination, representing the volume of plasma cleared of a drug per unit time. It provides insight into how efficiently the body eliminates a compound and directly influences drug dosing regimens. A higher clearance value indicates rapid elimination, often requiring more frequent administration, whereas lower clearance values suggest prolonged drug retention in the body.

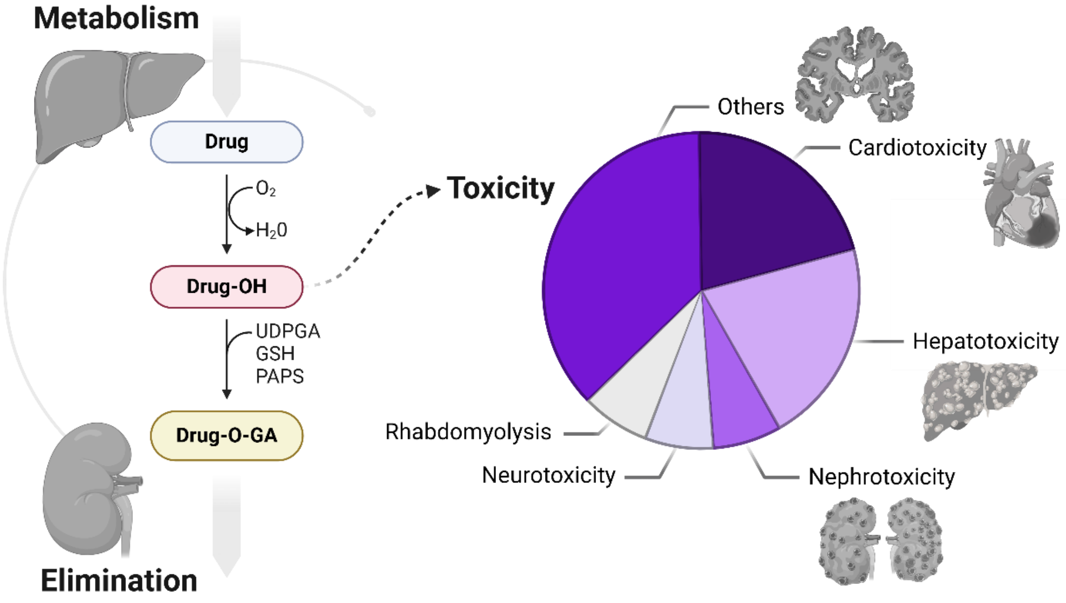


Figure 22: *Drug metabolism and toxicity-related failures in Drug Development.*

Different ways to express clearance exist, with the most fundamental equation being:

$$CL = \frac{Dose}{AUC} \approx CL_{tot} = CL_{hep} + CL_{ren}$$

Where the variable,

$Dose$ is the amount of drug administered (e.g., mg or µmol),

AUC (Area Under the Curve) is the total drug exposure over time (e.g., mg·h/L),

CL_{tot} (Total clearance) is the sum of hepatic (CL_{hep}) and renal clearance (CL_{ren}).

By combining these clearance pathways, a drug's overall elimination efficiency can be evaluated, aiding in the prediction of pharmacokinetics across different patient populations.

Intrinsic clearance (CL_{int}), on the other hand, describes the liver's inherent ability to metabolize a drug independent of hepatic blood flow. It is commonly measured using liver microsomes or hepatocytes in vitro. The relationship between intrinsic clearance and hepatic clearance is described by the well-stirred liver model:

$$CL_{hep} = \frac{f_u * CL_{int} * Q_H}{Q_H + f_u * CL_{int}}$$

Where the variable,

Q_H is the hepatic blood flow (e.g., L/min),

f_u is the fraction of unbound drug in plasma (unitless),

CL_{int} is the intrinsic clearance (e.g., mL/min or L/h, typically normalized per mg of microsomal protein or per million cells in vitro).

This equation differentiates between flow-limited clearance, where drug elimination is governed by hepatic perfusion (high CL_{int}), and capacity-limited clearance, where enzyme activity is the rate-limiting step (low CL_{int}). This distinction is essential for understanding the impact of physiological changes on drug metabolism.

Renal clearance (CLren) quantifies drug elimination through the kidneys. It is influenced by glomerular filtration, active tubular secretion, and passive reabsorption. Renal clearance is determined using the equation:

$$CL_{ren} = f_e * CL$$

where f_e is the fraction of the drug excreted unchanged in urine. This parameter helps assess whether renal elimination is a primary clearance route for a given compound.

Another critical concept is the half-life ($t_{\frac{1}{2}}$, in hours or minutes), which describes how long it takes for the plasma concentration of a drug to decrease by half.

$$t_{\frac{1}{2}} = \frac{0.693 * Vd}{CL}$$

where Vd represents the volume of distribution (e.g., L or mL).

Short half-lives indicate rapid clearance and necessitate frequent dosing, whereas long half-lives suggest prolonged drug activity and extended dosing intervals.

Toxicity

Toxicity can arise from multiple mechanisms, with metabolism and administered dose playing central roles. Off-target toxicity occurs when a drug interacts with unintended molecular targets, such as hERG potassium channels, leading to cardiac arrhythmias, or when it inhibits CYP450 enzymes, directly causing hepatotoxicity through enzyme inhibition or reactive intermediate accumulation.[155,156] Although increased exposure due to drug–drug interactions can indirectly lead to toxicity, intrinsic toxicity generally arises from direct drug or metabolite-driven cellular damage, distinct from simple dose-dependent effects. Excessive dosing or prolonged exposure can therefore precipitate adverse effects.

On-target toxicity arises when a drug engages its intended molecular target in unintended tissues, leading to adverse effects. For example, statins lower cholesterol by inhibiting HMG-CoA reductase in the liver, but can also inhibit the same enzyme in muscle tissue, contributing to myopathy and rhabdomyolysis in susceptible individuals.[157]

Many PIK inhibitors have on-target toxicities due to their role in essential cellular housekeeping functions.[34] Hypersensitivity reactions occur if drugs or their reactive metabolites form covalent bonds with their target, generating haptens that can trigger antibody production and immunological responses.

Drug metabolism can yield reactive intermediates (bioactivation) capable of binding to cellular components or eliciting immune responses. [158] Acetaminophen is a prime example; while the parent drug is safe at therapeutic concentrations, one of its minor metabolites (NAPQI) causes hepatotoxicity. Although drug discovery primarily evaluates parent molecules, identifying potential toxic metabolites remains crucial. Idiosyncratic reactions are especially problematic because they are highly individual (driven by genetic and immunological differences) and rarely detected in early animal models. Often, such reactions come to light only after extensive testing in humans.

Consequently, toxicity evaluations are typically conducted in at least two mammalian species, usually starting with rodents due to their practicality and low cost, and complemented by non-rodents (e.g., dog or pig) when rodents insufficiently reflect human physiology. The route of administration tested usually aligns with intended clinical use, although alternative routes are occasionally employed to circumvent pharmacokinetic limitations (e.g., extensive first-pass metabolism). Modern toxicology studies prioritize maximum tolerated dose, no observed adverse effect level (NOAEL), and exposure margins, rather than routinely determining lethal doses (LD_{50}).[159]

Application of ML

Assessment strategies include in vivo toxicology studies of varying durations, in vitro assays, and computational modeling. QSAR methods correlate molecular descriptors with observed toxicity, in addition to molecular docking preferred to explore possible interactions with toxicologically relevant targets such as hERG channels[160,161] or specific CYP isoforms.[162–165] These in silico approaches, aligned with the European REACH framework, help reduce animal testing by embracing alternative methods and integrating the 3R principle (Reduction, Refinement, Replacement).[166,167] Despite these advances, single-task QSAR models often fail to capture how metabolic transformations or clearance can shift toxicity profiles. This underscores the need to incorporate metabolism-related data into predictive models.[168–170]

In this chapter, we highlight MTL as a solution to broad pharmacokinetic profiling. By simultaneously modeling clearance, half-life, and various toxicity endpoints, MTL takes advantage of correlations among tasks. Recent work suggests that MTL outperforms single-task models for ADMET and potency predictions, paving the way for faster and more reliable drug development. Several open-source web services now leverage early ADMET (eADMET) strategies, but data sources and methods can be redundant or fragmented. To address these gaps, we have built a unified MTL model handling hundreds of continuous tasks in parallel and introduced the OneADMET dataset, a comprehensive curated repository merging diverse ADMET endpoints from public sources (**Figure 23**). This enables a single model to concurrently assess critical pharmacokinetic and toxicological parameters, streamlining early liability detection.

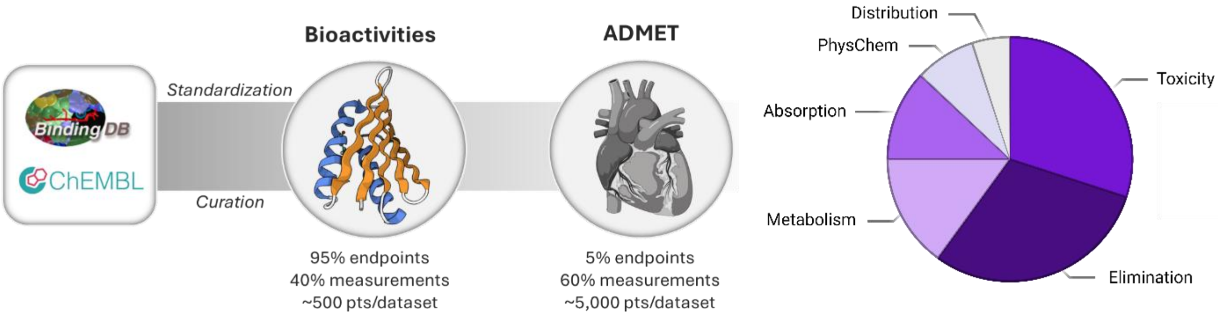


Figure 23: ***Data Integration from ChEMBL and BindingDB.*** A pie chart illustrates the number of tasks per dataset, reflecting dataset diversity and scale.

In Silico eADMET: current situation and novel profilers

Pierre Llompart[1,2], Claire Minoletti[2], Gilles Marcou[1,†], Alexandre Varnek[1]

[1]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg, France

[2]Integrated Drug Discovery, Molecular Design Sciences, Sanofi, Vitry-sur-Seine, France

Keywords : ADMET, QSPR, deep learning, multi-task learning

Abstract

Multi-task learning (MTL) has emerged as a powerful strategy in computational drug discovery, addressing the inherent complexity and variability in pharmacokinetic and pharmacodynamic profiling. This approach promises enhanced predictive performance and generalizability compared to traditional single-task models. Recent studies underscore MTL utility in improving ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) and potency predictions, foundational for efficient drug design. Recent years have seen emerging several exciting open-source web services for early ADMET assessment (eADMET). This contribution aims first to provide a critical review of them regarding data sources, methods, and redundancy between them. Second, building on our observations, we have developed a unified model simultaneously processing hundreds of continuous tasks and we introduce the OneADMET dataset—a comprehensive curated resource from public datasets. Finally, we propose a reference web service for eADMET and bioactivities prediction. Our findings demonstrate that the multi-task model equals or outperforms conventional single-task models in reducing prediction errors. Comparatively, MTL model deployment and maintenance are simpler and computationally more efficient for profiling. This study confirms the robustness of large-scale MTL for detailed pharmacokinetics profiling and contributes to the field with the provision of the OneADMET dataset.

Introduction

Lead optimization is a critical phase of drug discovery. Following hit identification, the aim is to improve the efficacy, selectivity, and pharmacokinetic properties of chemical compounds to be considered as drug candidates. Another aspect is to understand and control their toxicities[1]. Drug candidates are then processed to the next stage: pre-clinical development aiming at improving their safety before testing on humans.

Lead optimization failures related to ADMET properties remain a significant hurdle in drug development. ADMET issues are responsible for approximately 60% of clinical trial failures, leading to substantial delays in delivering new treatments to patients and investment losses, the economic burden of drug development continues to escalate, with the average cost of bringing a new drug to market estimated at over \$2.6 billion[2–4]. Early identification and optimization of ADMET characteristics are therefore key to enhance the success rates of drug. Therefore, new methodologies, data models for eADMET contributes to research and development of new drugs with enhanced productivity[5].

In this contribution, we propose a new Multi-Task Learning model for prediction of 44 eADMET properties and 1,489 bioactivities. The public model and related public dataset called OneADMET are provided open source (Figure 1).

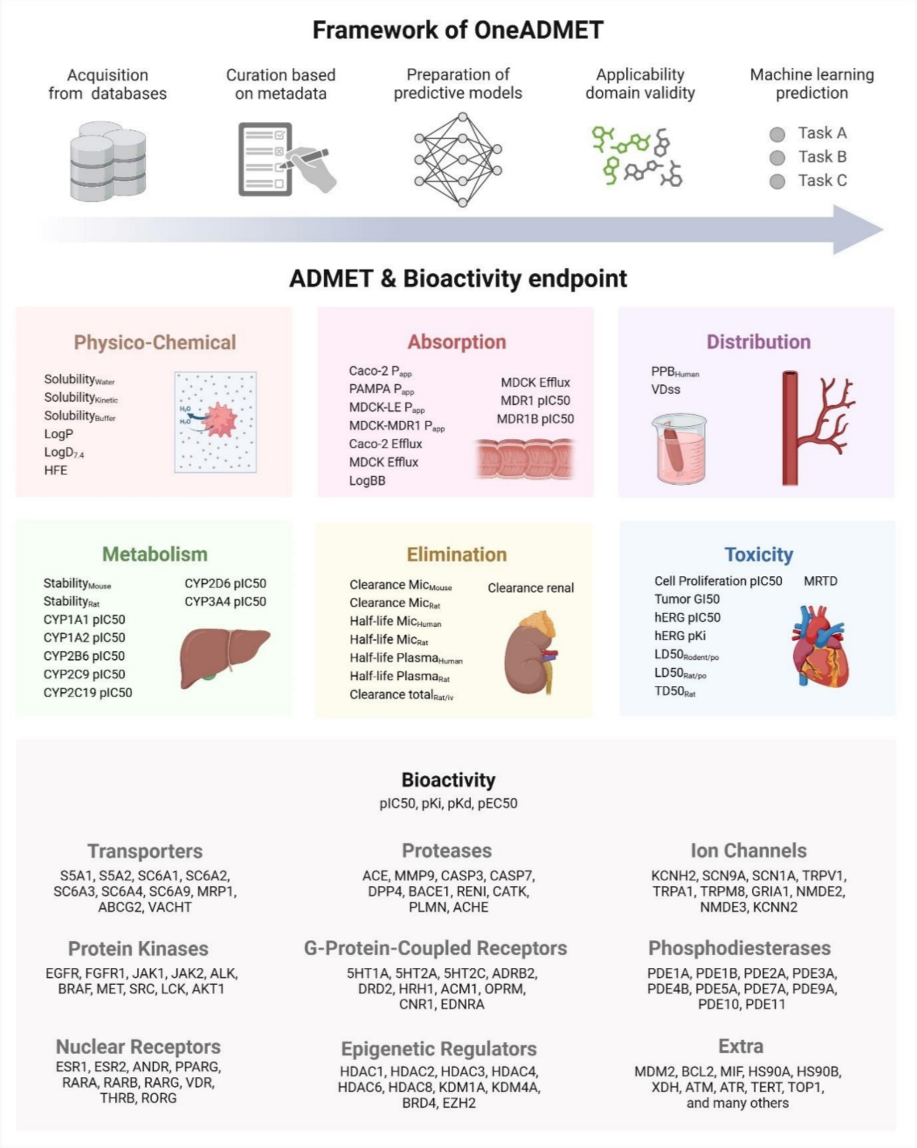


Figure 1: Presentation of the predictive workflow of the webserver, and ADMET to bioactivities endpoints covered.

Multi-Task Learning

MTL consists in optimizing a single model against multiple tasks simultaneously. This approach unifies over several statistical models the choice of common representations and free parameter values. It has been observed that this approach may improve the models generalization for related tasks, although such synergetic behavior between tasks can hardly be anticipated[6,7].

Below, we review the current state of MTL in ADMET modelling and assess available webservers (see Table 1). We focused our attention on the differences in underlying data and models they use. We also report the usage of an applicability domain (AD), defining the chemical space where predictions are reliable, that is crucial to control predictions accuracy. Models without proper AD consideration may yield unreliable predictions for compounds outside their training scope[8].

Emergence of Multi-Task Learning in ADMET modeling

The first reported application of Multi-Task Learning for ADMET was pioneered in 2009 by Varnek et al. They used an Artificial aSsociative Neural Network (ASNN) within an MTL framework to predict tissue-air partition coefficients for human and rat tissues across 11 regression tasks, utilizing a dataset of 648 samples[9]. In 2011, Su et al. advanced the field by employing Max-Margin Conditional Random Fields (MMCRF) to predict bioactivity against cancer cell lines encompassing 4,547 samples and 60 classification tasks[10]. These early studies demonstrated the potential of MTL in cheminformatics and ADMET modeling.

Growth following the Tox21 challenge

The year 2016 marked an acceleration in the development of MTL for ADMET modeling, largely due to the outcomes of the Tox21 Data Challenge. The Tox21 program is a collaborative initiative involving several U.S. federal agencies aimed at developing better toxicity assessment methods[11]. The challenge provided a dataset of 12,707 compounds tested across 12 *in-vitro* assays related to nuclear receptor signaling and stress response pathways. Mayr et al. applied Deep Neural Networks (DNN) within an MTL framework to predict these 12 toxicity-related tasks, achieving top performance on multiple assays and demonstrating significant improvements over STL models[12]. Their approach led to a performance increase of up to 5% compared to some single task models, showcasing the efficacy of MTL in handling complex biological data.

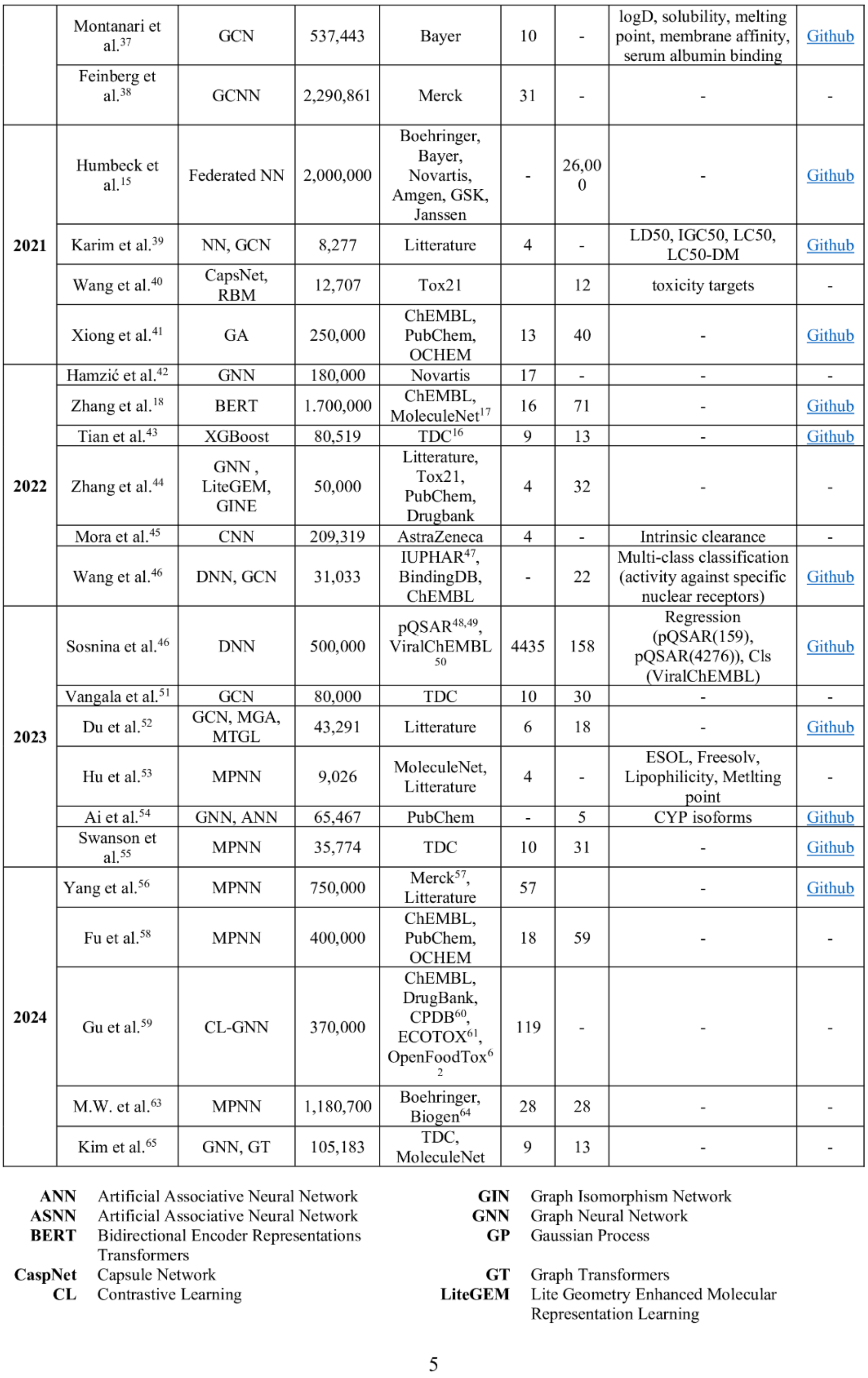
In the subsequent five years, 16 studies explored MTL for ADMET, with seven originating from industrial companies. For instance, Wenzel et al. from Sanofi highlighted the importance of combining tasks from correlated domains when developing average size MTL models of few tasks. They found that adding unrelated tasks could lead to decreased performance, emphasizing the need for careful task selection[13]. Other studies pushed the boundaries of MTL by increasing the number of simultaneous tasks. Zakharov et al. developed a Deep Learning Consensus Architecture (DLCA) trained on 201,599 samples from ChEMBL and Tox21 datasets, covering 820 regression and 12 classification tasks. Their work showcased the scalability of MTL models in handling large and diverse datasets[14]. The MELLODDY (Machine Learning Ledger Orchestration for Drug Discovery) consortium added to MTL the concept of federated learning approach. The consortium including several pharmaceutical companies trained models collaboratively through using arithmetic operations on encrypted data. This initiative encompassed 26,000 classification tasks, including ADMET relevant ones, using data from industrial partners, representing a unique and large-scale application of MTL[15].

In 2022, several authors promoted the Therapeutics Data Commons (TDC) and MoleculeNet to be used as standards for benchmarking along with standardized protocols[16,17]. Since then, at least five studies specifically focused on MTL for ADMET have used these benchmarking datasets in their entirety or in part for model validation. For example, Zhang et al. in 2022 employed a BERT-based model on over 1.7 million samples from ChEMBL and MoleculeNet, covering 71 classification and 16 regression tasks[18]. However, standardized benchmarking datasets are not aiming for data quality and relevance. Walters informally highlighted discrepancies such as duplicate entries with conflicting responses, ionization issues, and standardization problems within these datasets[19]. These issues can lead to perturbed predictions and biased evaluations, potentially compromising model reliability[20].

In response to these challenges, collaborative open-source efforts like the Polaris initiative are emerging to ensure the curation of high-quality, standardized benchmarks[21]. Yet, ongoing work is necessary to address these data quality issues fully, especially since some models trained on problematic datasets are currently deployed on public web servers.

Table 1: Historic of the application of the MTL approaches to the modelling of ADMET properties and bioactivities. The number of tasks and their type of either classification or regression is informed. The number of samples in the dataset used to train the MTL model is added. The quantity is either the exact value or an estimation based on the source of the data in the case of poorly described methods.

Year	Group	Method	# samples	Data source	Reg	Cls	Types	Code
2009	Varnek et al.[9]	ASNN, PLS	648	Katritzky et al.[22]	11	-	tissue-air partition coefficients for human and rat tissues	-
2011	Su et al.[10]	MMCRF	4,547	NCI-Cancer from PCBA[23]	-	60	Bioactivity against cancer cell lines	-
2015	Ramsundar et al.[24]	DNN	37.8M	PCBA, MUV[25], DUD-E[26], Tox21[27]	-	259	-	-
2016	Mayr et al.[12]	DNN	12,707	Tox21	-	12	-	-
2017	Kearnes et al.[28]	DNN, wDNN	280,000	Vertex, Pubchem	-	550	hERG inhibition, solubility, metabolism	-
	Xu et al.[29]	DNN	50,000	Kaggle[30]	-	15	on-target potency, off-target ADME activities	Github
	Ramsundar et al. [31]	Progressive DNN	114,000	Kaggle, Factors, Kinase, UV	316	-	enzymatic inhibition, ADME/Tox	Github
2018	Li et al.[32]	DNN	13,000	PCBA	-	5	inhibition CYP450 isoforms	-
2019	Zakharov et al. [14]	DLCA	201,599	ChEMBL, Tox21	820	12	Regression (ChEMBL), Classification (Tox21)	-
	Wenzel et al.[13]	DNN	81,309	ChEMBL, Sanofi	14	-	-	-
	Liu et al.[33]	DNN, GCNN	250,000	Amgen	-	5	-	-
	Capela et al.[34]	GGRNN, GAIN, GIN	27,057	OChem, ChEMBL, Litterature	6	-	solubility, partition coefficients, boiling point, vapor pressure	Github
2020	Li et al.[35]	DNN-SMOTE	12,707	Tox21	-	12	-	-
	Peng et al.[36]	GNN	12,741	ChEMBL, Tox21, Litterature	4	3	-	-



	Montanari et al.[37]	GCN	537,443	Bayer	10	-	logD, solubility, melting point, membrane affinity, serum albumin binding	Github
	Feinberg et al.[38]	GCNN	2,290,861	Merck	31	-	-	-
2021	Humbeck et al.[15]	Federated NN	2,000,000	Boehringer, Bayer, Novartis, Amgen, GSK, Janssen	-	26,000	-	Github
	Karim et al.[39]	NN, GCN	8,277	Litterature	4	-	LD50, IGC50, LC50, LC50-DM	Github
	Wang et al.[40]	CapsNet, RBM	12,707	Tox21		12	toxicity targets	-
	Xiong et al.[41]	GA	250,000	ChEMBL, PubChem, OCHEM	13	40	-	Github
2022	Hamzić et al.[42]	GNN	180,000	Novartis	17	-	-	-
	Zhang et al.[18]	BERT	1.700,000	ChEMBL, MoleculeNet[17]	16	71	-	Github
	Tian et al.[43]	XGBoost	80,519	TDC[16]	9	13	-	Github
	Zhang et al.[44]	GNN , LiteGEM, GINE	50,000	Litterature, Tox21, PubChem, Drugbank	4	32	-	-
	Mora et al.[45]	CNN	209,319	AstraZeneca	4	-	Intrinsic clearance	-
	Wang et al.[46]	DNN, GCN	31,033	IUPHAR[47], BindingDB, ChEMBL	-	22	Multi-class classification (activity against specific nuclear receptors)	Github
2023	Sosnina et al.[46]	DNN	500,000	pQSAR[48,49], ViralChEMBL[50]	4435	158	Regression (pQSAR(159), pQSAR(4276)), Cls (ViralChEMBL)	Github
	Vangala et al.[51]	GCN	80,000	TDC	10	30	-	-
	Du et al.[52]	GCN, MGA, MTGL	43,291	Litterature	6	18	-	Github
	Hu et al.[53]	MPNN	9,026	MoleculeNet, Litterature	4	-	ESOL, Freesolv, Lipophilicity, Metlting point	-
	Ai et al.[54]	GNN, ANN	65,467	PubChem	-	5	CYP isoforms	Github
	Swanson et al.[55]	MPNN	35,774	TDC	10	31	-	Github
2024	Yang et al.[56]	MPNN	750,000	Merck[57], Litterature	57		-	Github
	Fu et al.[58]	MPNN	400,000	ChEMBL, PubChem, OCHEM	18	59	-	-
	Gu et al.[59]	CL-GNN	370,000	ChEMBL, DrugBank, CPDB[60], ECOTOX[61], OpenFoodTox[62]	119	-	-	-
	M.W. et al.[63]	MPNN	1,180,700	Boehringer, Biogen[64]	28	28	-	-
	Kim et al.[65]	GNN, GT	105,183	TDC, MoleculeNet	9	13	-	-

ANN	Artificial Associative Neural Network	**GIN**	Graph Isomorphism Network
ASNN	Artificial Associative Neural Network	**GNN**	Graph Neural Network
BERT	Bidirectional Encoder Representations Transformers	**GP**	Gaussian Process
CaspNet	Capsule Network	**GT**	Graph Transformers
CL	Contrastive Learning	**LiteGEM**	Lite Geometry Enhanced Molecular Representation Learning

CPDB	Carcinogenic Potency Database
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DLCA	Deep Learning Consensus Architecture
DUD-E	Directory of Useful Decoy - Enhanced
GA	Graph Attention
GAIN	Graph Attention Isomorphism Network
GCN	Graph Convolutional Network
GCNN	Graph Convolutional Neural Network
GGRNN	Gated Recursive Neural Network
GINE	Graph Isomorphism Network with Edge
MGA	Multi-Task Graph Attention
MMCRF	Max-Margin Conditional Random Field
MPNN	Message Passing Neural Network
MTGL	Multi-Task Graph Learning
MUV	Maximum Unbiased Validation
NN	Neural Network
PCBA	PubChem BioAssay
PLS	Partial Least Square
RBM	Restricted Boltzmann Machine
SMOTE	Synthetic Minority Oversampling TEchnique

Webservers for ADMET profilling: Evolution, Challenges, and Futures directions

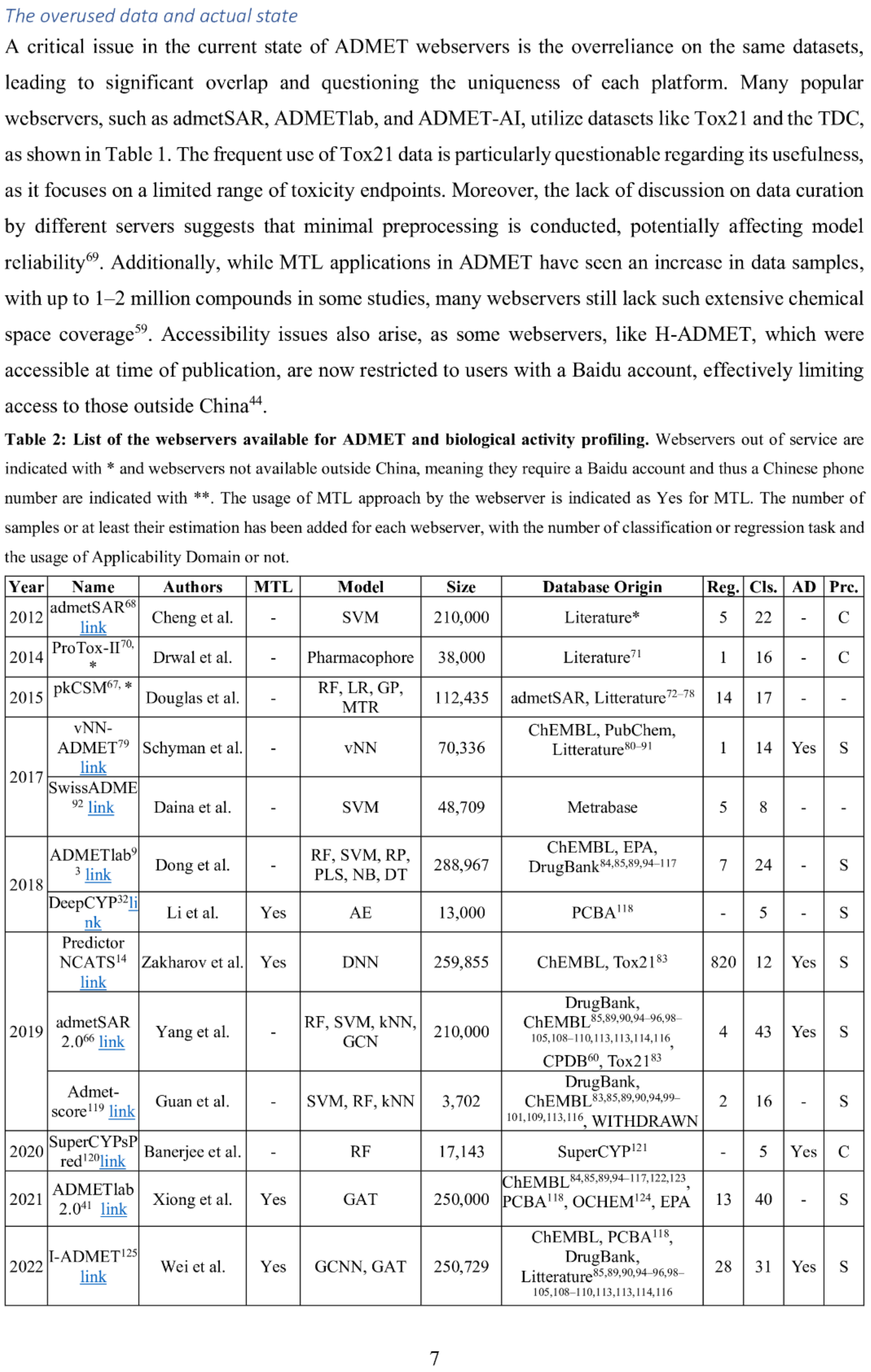
Advances in data science for drug discovery are reflected in the availability of publicly accessible webservers for eADMET predictions. Notable platforms include SwissADME[41], admetSAR[66], ADMETlab[41] and pkCSM[67]. They aim at enabling rapid assessment of compounds across numerous endpoints.

Introduction and early development of ADMET webservers

Over the past 12 years, webservers dedicated to ADMET profiling have been instrumental in drug discovery and toxicity prediction. They implemented early machine learning methods such as Support Vector Machines (SVM) and Random Forests (RF). In 2012, Cheng et al.[68] introduced admetSAR, one of the first webservers in this domain, using SVM on a dataset of approximately 210,000 compounds compiled from literature sources. This platform covered five regression tasks and 22 classification tasks, setting a foundation for subsequent tools. By 2015, pkCSM, developed by Pires et al., integrated multiple machine learning techniques, including RF, Logistic Regression (LR), and Model Tree Regression (MTR), to predict ADMET properties using a dataset of over 112,000 compounds[69]. pkCSM offered predictions for 14 regression tasks and 17 classification tasks, highlighting the versatility of machine learning methods in ADMET profiling.

MTL models-based web services

The landscape of ADMET web services began to shift in 2018 with the introduction of MTL methods. DeepCYP, developed by Li et al. was among the first, using an autoencoders (AE), for predicting cytochrome P450 (CYP) enzyme interactions using a dataset of 13,000 compounds[54]. By 2021, more deep artificial neural network models have been developed: eight out of ten new web services have been proposed based on various MTL architectures. For instance, ADMETlab 2.0 by Xiong et al. utilized Graph Attention Transformer network (GAT) on a dataset of 250,000 compounds from ChEMBL, PubChem, OCHEM, and EPA databases, offering predictions for 13 regression and 40 classification tasks[41]. The average number of training samples also saw a significant increase approaching close to 200,000 compounds covering millions of data points. This expansion reflects the growing availability of chemical data.



The overused data and actual state

A critical issue in the current state of ADMET webservers is the overreliance on the same datasets, leading to significant overlap and questioning the uniqueness of each platform. Many popular webservers, such as admetSAR, ADMETlab, and ADMET-AI, utilize datasets like Tox21 and the TDC, as shown in Table 1. The frequent use of Tox21 data is particularly questionable regarding its usefulness, as it focuses on a limited range of toxicity endpoints. Moreover, the lack of discussion on data curation by different servers suggests that minimal preprocessing is conducted, potentially affecting model reliability[69]. Additionally, while MTL applications in ADMET have seen an increase in data samples, with up to 1–2 million compounds in some studies, many webservers still lack such extensive chemical space coverage[59]. Accessibility issues also arise, as some webservers, like H-ADMET, which were accessible at time of publication, are now restricted to users with a Baidu account, effectively limiting access to those outside China[44].

Table 2: List of the webservers available for ADMET and biological activity profiling. Webservers out of service are indicated with * and webservers not available outside China, meaning they require a Baidu account and thus a Chinese phone number are indicated with **. The usage of MTL approach by the webserver is indicated as Yes for MTL. The number of samples or at least their estimation has been added for each webserver, with the number of classification or regression task and the usage of Applicability Domain or not.

Year	Name	Authors	MTL	Model	Size	Database Origin	Reg.	Cls.	AD	Prc.
2012	admetSAR[68] link	Cheng et al.	-	SVM	210,000	Literature*	5	22	-	C
2014	ProTox-II[70], *	Drwal et al.	-	Pharmacophore	38,000	Literature[71]	1	16	-	C
2015	pkCSM[67], *	Douglas et al.	-	RF, LR, GP, MTR	112,435	admetSAR, Litterature[72–78]	14	17	-	-
2017	vNN-ADMET[79] link	Schyman et al.	-	vNN	70,336	ChEMBL, PubChem, Litterature[80–91]	1	14	Yes	S
	SwissADME[92] link	Daina et al.	-	SVM	48,709	Metrabase	5	8	-	-
2018	ADMETlab[93] link	Dong et al.	-	RF, SVM, RP, PLS, NB, DT	288,967	ChEMBL, EPA, DrugBank[84,85,89,94–117]	7	24	-	S
	DeepCYP[32]link	Li et al.	Yes	AE	13,000	PCBA[118]	-	5	-	S
2019	Predictor NCATS[14] link	Zakharov et al.	Yes	DNN	259,855	ChEMBL, Tox21[83]	820	12	Yes	S
	admetSAR 2.0[66] link	Yang et al.	-	RF, SVM, kNN, GCN	210,000	DrugBank, ChEMBL[85,89,90,94–96,98–105,108–110,113,113,114,116], CPDB[60], Tox21[83]	4	43	Yes	S
	Admet-score[119] link	Guan et al.	-	SVM, RF, kNN	3,702	DrugBank, ChEMBL[83,85,89,90,94,99–101,109,113,116], WITHDRAWN	2	16	-	S
2020	SuperCYPsPred[120]link	Banerjee et al.	-	RF	17,143	SuperCYP[121]	-	5	Yes	C
2021	ADMETlab 2.0[41] link	Xiong et al.	Yes	GAT	250,000	ChEMBL[84,85,89,94–117,122,123], PCBA[118], OCHEM[124], EPA	13	40	-	S
2022	I-ADMET[125] link	Wei et al.	Yes	GCNN, GAT	250,729	ChEMBL, PCBA[118], DrugBank, Litterature[85,89,90,94–96,98–105,108–110,113,113,114,116]	28	31	Yes	S

	toxCSM[126]link	G. C. de Sá et al.	-	RF, XGBoost, AB	133,545	Tox21[83], Litterature[78,81,90,91,114,115,127–139]	5	31	-	-
	ADMETboost[43] link	Tian et al.	Yes	XGBoost	80,519	TDC[17,68,83,94,98,109,112,140,141,141–148,148,149]	8	14	-	-
	H-ADMET[44], **link	Zhang et al.	Yes	GNN, LiteGEM, GINE+	50,000	DrugBank, ChEMBL, CPDB[60], PCBA[118], Littérature[83,95,99,107,114,123,148,150–155], Tox21[83], SuperCYP[121], CYPReact[156]	4	36	-	S
2024	ADMETsar 3.0[59] link	Gu et al.	Yes	CLMGraph	370,000	ChEMBL[107,128,130,131,135,136,152,157–168], DrugBank, TOX, OFT [62]	16	97	-	S
	Deep-PK[169] link	Myung et al.	Yes	MPNN	350,695	ADMETlab 2.0[84,85,89,94–117,122,123], PCBA[118], OCHEM[124], Tox21[83], I-ADMET, toxCSM [78,81,90,91,114,115,127–139], pkCSM[72–78]	24	49	-	S
	Protox 3.0[170] link	Banerjee et al.	-	RF, DNN	263,832	Tox21[83], PCBA[118], CPDB[60], ET[61], OFT[62]	2	60	-	S
	ADMET-AI[55] link	Swanson et al.	Yes	Chemprop-RDKit	80,519	TDC[17,68,83,94,98,109,112,140,141,141–148,148,149]	10	31	-	-
	ADMETlab 3.0[58] link	Yang et al.	Yes	DMPNN	400,000	ChEMBL[171–176], PCBA[118], OCHEM[124]	18	59	-	S

AB	Adaptative Boosting	**kNN**	k-Nearest Neighbors
AE	AutoEncoder	**LiteGEM**	Lite Geometry-Enhanced Molecular
CLMGraph	Contrastive Learning-based Multi-Task Graph	**LR**	Logistic Regression
DNN	Deep Neural Network	**MPNN**	Message-Passing Neural Network
DT	Decision Tree	**MTR**	Model Tree Regression
ET	EcoTox	**NB**	Naïve Bayes
ERT	Extremely Randomized Trees	**OFT**	OpenFoodTox
GAT	Graph Attention Transformer	**PLS**	Partial-Least Square
GCNN	Graph Convolution Neural Network	**RF**	Random Forest
GCN	Graph Convolution Network	**RP**	Recursive Partitioning
GINE+	Graph Isomorphism Network with Edge - Enhanced	**SVM**	Support Vector Machine
GNN	Graph Neural Network	**vNN**	variable Nearest Neighbors
GP	Gaussian Process	**XGBoost**	Extreme Gradient Boosting

*: Out of service, **: Not available outside China, Reg.: Regression endpoints, Cls.: Classification endpoints, Prc.: Preprocessing of the data, C: Curation & Standardization, S: Only standardization.

The overlap of webservers

Although one might assume that each ADMET webserver offers unique functionalities, a detailed analysis reveals substantial overlap in both the datasets utilized and the tasks provided, prompting questions about their overall distinctiveness and utility. Figure 2 showcases the various datasets and tasks offered by these webservers. Data is a scarce resource in ADMET modelling which require experts' knowledge to curate it into modelizable data. Figure 2a follows the publication year of the webservers against the time at which their data were published. Most of the data sources are dating back to 2012, with webservers published in 2024 having most of their data dating back 2012-2015, hence a +10-year gap with the actual state of the public databases. Only ADMETlab 3.0 and admetSAR 3.0 seems to use recent sources, published around 5 years ago. This drawback is accompanied by the lack of preparation and curation (Table 2) of novel data. Figure 2b highlights the proportion of endpoints being cited under the same source. Overall, at least half of the webservers seems to be built on the same datasets, without ongoing any additional curation except a standardization of the SMILES (Table 2). The count is limited to studies citing the source of their data, as certain omitted it. Figure 2c present the most represented tasks from the webservers. The most prevalent tasks involve predicting cytochrome P450 (CYP) inhibitors and substrates, derived from datasets by Veith et al.[109] and Carbon-Mangels et al.[141]. These datasets are part of the TDC and are extensively used across almost all webservers with half of them citing these sources. Following this, the blood-brain barrier (BBB) permeability data from Martins et al.[148], the aqueous solubility (LogS) data from AqSolDb [146], and the human intestinal absorption (HIA) data from Hou et al.[143] are among the most frequently employed. Additionally, P-glycoprotein (P-gP) data from Broccatelli et al.[98] is commonly used. The LD_{50} toxicity data from Zhu et al.[177] and AMES test data from the same study are also prevalent. Furthermore, the Tox21 datasets are utilized by approximately one-third to half of the webservers. This significant overlap in datasets used, shared tasks, lack of curation, age of the datasets indicates that many webservers do not implement specific data curation methods to enhance the quality of their predictive models, thereby affecting their uniqueness and overall effectiveness.

The datasets are used as is, the authors being confident in the results of previously published work and lacking the addition of recent experimental measurements. But aggregation of experimental data is prone to heterogeneity in conditions and numerous studies. The practices and understanding on the data are improving over time. Hence even a legitimate research data source should be re-examined regularly considering these advances and updated where needed.

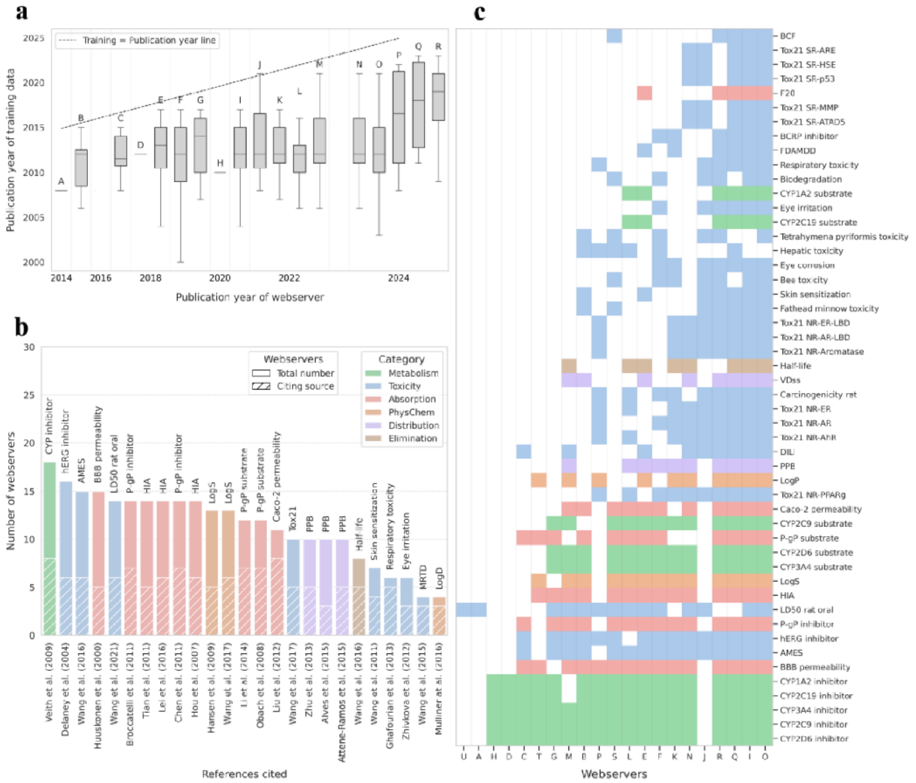


Figure 2: Analysis of the data and endpoints used by webservers. (a) Boxplot of the distribution of training data publication year against the year of the webserver publication. The dashed line highlights the timeline where the publication year of the data equals the publication year of the webserver. Only webservers citing their sources are present and annotated as A: ProTox-II, B: pkCSM, C: vNN-ADMET, D: DeepCYP, E: ADMETlab, F: admetSAR 2.0, G: Admet-score, H: SuperCYPsPred, I: ADMETlab 2.0, J: toxCSM, K: I-ADMET, L: H-ADMET, M: ADMETboost, N: ADMET-AI, O: Deep-PK, P: Protox 3.0, Q: ADMETlab 3.0, R: ADMETsar 3.0, S: admetSAR, T: SwissADME, U: Predictor NCATS. **(b)** Countplot of the most cited reference used by the webservers. The full bar is the number of webservers presenting the endpoint. The hashed bar is the number of webservers citing the reference for the associated endpoint. **(c)** Heatmap of the most popular ADMET endpoints presented by each webserver. Blank spaces mean the endpoint is not delivered by the webserver.

Materials & Methods

Data Collection

We present a dataset which is an extensive collection of endpoints regarding drug discovery endpoints, public experimental values were recovered from three public databases: OChem[124], ChEMBL[178], and BindingDB[179]. These data sources have been selected because of the possibility to source each datapoint, mostly from international reviewed bibliographic sources. Only continuous measurements were recovered as the focus of the work regards regression models. Only datasets containing 30 compounds or more have been included. Experimental units were standardized per endpoint to homogenize the measurements, choosing units that are met most frequently and assuming that they are more consensual. (Table S1).

Metadata curation

A given endpoint is sometime the output of several assays. Although these assays have the same aim, they might differ sufficiently so that the measured output values are not compatible. For this reason, each documented values of our endpoints were reviewed and decisions for merging these sources were taken. The criteria for merging are given Table S2.

Curation and Standardization

The collected datasets went further through several steps of data filtering, curation and standardization. These steps delineate a region of the chemical space that we defined as of interest for drug discovery today. Additionally, it standardizes how compounds and substances of interest are represented. First we filtered out compounds with:

- less than 10 atoms,
- no heteroatom or one carbon atom,
- containing an isotope,
- containing an atom which element is out of this list: C, H, O, N, S, F, Cl, P, Br, I, Si, B.

The standardization procedures included the following steps:

- stripping salt, keeping the largest component,
- removing stereochemistry information – considered as unreliable at this stage,
- removing the aromaticity – it is restored as the last step of the process,
- selecting of a standard tautomer.

After standardization, several chemical structures appeared duplicated because of merging different data sources and ignoring the stereochemistry. In case of duplicated entries, the standard deviation (SDi) *(1)* and median of the experimental values were computed. Compounds with a standard deviation exceeding 0.5 log units, or 5% for endpoints expressed in percent, were excluded[180]. These cutoffs have been chosen to reflect our desired modeling prediction quality, expressed as RMSE. If more than two measurements were available for a non-excluded compound, the median value was taken as the response value. If only two measurements were available for a non-excluded compound, the value reflecting a less favorable outcome (e.g., poor permeability rather than good permeability) was chosen. This approach prioritizes avoiding false positives, focusing instead on false negatives to minimize the risk of pursuing non-viable compounds and save time[180]. Only datasets with at least 30 unique compounds were considered for modeling.

$$SDi = \sqrt{\frac{\sum_{k=1}^{n_i}(x_k - \bar{x})^2}{n-1}} \quad (1)$$

The SDi is the standard deviation of n_i duplicated observations on a given compound, where:

- x_i denotes each individual observation for given duplicated compound.
- $\bar{x}$ represents the arithmetic mean of duplicated observations.

Data Split

For each task, the corresponding dataset is partitioned into two subsets: a training subset, constituting 80% of the total data, and a testing subset, comprising the remaining 20%. Within the training subset, we perform a tuning 3-fold stratified Cross-Validation (CV)[181] for optimization of machine learning method and hyperparameters, as well as optimization of the molecular descriptors sets. Then, a model is retrained on the training subset using the optimal hyperparameters, methods, and descriptors identified during CV. The resulting final models are put for production and used on the test set, using an Applicability Domain (AD). The test set predictions compose an external validation. The public and industrial dataset led to the generation of public and industrial test sets that are orthogonal to each other. The public set is disclosed and used to train the public model, while the industrial set is not disclosed, and only used to validate the public model, and prepare an undisclosed industrial model. External validation performances measures are discussed only if the number of predictions after AD application is larger than 10% of the population of the test set.

To facilitate the analysis, we decided to monitor the presence of the same compounds across different tasks. Due to MTL context, we had to make sure that such compound was found exclusively either in the training set or in the test set. This requirement, coupled with the sometimes-low population datasets of some tasks, imposes an important constraint on the composition of the 3-fold cross-validation folds. We used a single stratified sampling taking advantage of the standardization of the endpoint values. All values were assembled into one dataset that was sampled uniformly per percentile without replacement. All data points associated to a given drawn molecule were located either in a training set or a test set, thus avoiding data leakage.

Molecular Featurization

During this study, we explored diverse molecular representations. Descriptors calculation was based only on the 2D structures, justifying that stereoisomers information was ignored. Descriptors were computed using scikit-fingerprints[182] (Table S3).

Modelling Methods

We used for single task learning, a set of machine learning methods listed in Table S4, both linear and non-linear. Each method uses its own set of hyperparameters that have been optimized using a Bayesian optimization algorithm. The Bayesian optimization was carried over 30 iterations. The optimization

process employed the Tree-structured Parzen Estimator (TPE) algorithm implemented in the Optuna[183] framework. The resolution and the domain of values explored during optimization is provided in the same table. The objective function minimized during optimization root was the averaged mean squared error (RMSE) calculated over the tuning 3-fold cross-validation described earlier.

Hyperparameters Optimization

To achieve optimal model performance, hyperparameter optimization was performed by exploring two main parameter spaces: the method-specific parameter space and the descriptor-specific parameter space. The method-specific space includes parameters inherent to the predictive algorithm, such as the number of estimators in Random Forest or the kernel type in Support Vector Machines. The descriptor-specific space focuses on the parameters impacting the input features such as the number of bits or radius of fingerprint (ECFP), ensuring the best possible representation of molecular properties.

Loss Function of the MTL models

We opted for fixed weights *(2)* per task to compute the multi-task loss function. The MTL models adopt a loss function based on the weighted Mean Squared Error *(3)* (wMSE) defined as a weighted sum of the MSE for each task *(4)*. Weights values are normalized. As each task has its own units, the weights are representing the different scales covered by these units, maintaining a consistent range of values across all tasks. This ensures a balanced contribution of each task to the overall loss calculation and effectively outputs real values.

The weights *(2)* are estimated as follows:

$$w_i = \frac{\frac{1}{|\max(values_i) - \min\,(values_i)|}}{\sum_{j=1}^{n} \frac{1}{\left|\max\left(values_j\right) - \min\,(values_j)\right|}} \qquad (2)$$

Where:

w_i is the initial weight for the i-th task,

values$_i$ is the set of experimental values of the i-th task,

n is the total number of tasks.

MSE *(3)* measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. The MSE of a task is one of the contributing terms of the loss function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (3)$$

Where:

$\hat{y}_i$ is the predicted value for the i-th observation,

y_i is the actual value for the i-th observation.

These metrics are computed for each task, providing a comprehensive measure of the model's performance across different aspects of the data.

The multi-task loss function *(4)* then expresses as:

$$Loss = \sum_{i=1}^{n} MSE_i * w_i \tag{4}$$

where MSE_i is the Mean Squared Error of the prediction against experimental values for the i-th task.

Metrics

To assess the performance of our regression models, we employ the coefficient of determination (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) per task. R^2 *(5)* represents the goodness-of-fit of a model. It reveals how much of the variance in the dependent variable is captured by the independent variables. It ranges from 0 to 1, where a value closer to 1 indicates better model fit; thus, it is independent of the units in which a given task is expressed.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

Where:
n is the total number of observations,
$\hat{y}_i$ is the predicted value for the i-th observation,
y_i is the actual value for the i-th observation,
$\bar{y}_i$ is the mean of the actual values y.

RMSE *(6)* quantifies the difference between the predicted and observed values per task, penalizing larger errors more severely by squaring them before averaging. The RMSE is expressed in the units of the given task and is not too sensitive to the value range of the task, in contrast to R^2.

$$RMSE = \sqrt{MSE} \tag{6}$$

Where:
n is the total number of observations,
$\hat{y}_i$ is the predicted value for the i-th observation,
y_i is the actual value for the i-th observation.

MAE *(7)* represents the accuracy of a regression model to a certain task. Compared to RMSE, MAE is less sensitive to outlying large errors. It is an arithmetic average of errors absolute value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

Where:

n is the total number of observations,

$\hat{y}_i$ is the predicted value for the i-th observation,

y_i is the actual value for the i-th observation.

Applicability Domain

The AD of a predictive model represents the Chemical Space (CS) region that is sampled by the training dataset, marking the boundaries within which the model's predictions are considered reliable. As each task has a specific dataset, we have defined an AD for each task. During this study, we applied the Local Outlier Factor method with standard hyperparameters (n_neighbors=20, contamination=0.2), which are well-suited for handling chemical space variability[184]. LOF is a density-based method that identifies outliers by comparing a sample's local density to that of its neighbors. A score close to 1 indicates a well-represented sample, while higher values suggest outliers. By applying LOF, we identify molecules in underrepresented regions, ensuring predictions remain within a well-defined and reliable space, reducing the risk of extrapolation errors.

Results & Discussion

OneADMET is a meticulously expert curated public dataset designed to support robust and transparent predictive modeling for drug discovery. OneADMET integrates high-quality data from diverse public sources, ensuring consistent data collection protocols. Data for OneADMET were sourced from well-established repositories: OChem and ChEMBL for ADMET-related endpoints, and OChem and BindingDB for biological activity data. OneADMET is the disclosed *public* dataset while the *industrial* dataset, the extended version including the industrial data cannot be disclosed.

The public dataset contains 1,119,719 endpoint values for 738,161 compounds (Figure S1a, S1b). Most compounds adhere to QED drug-likeness standards (Figure 3a). It spans 1,533 endpoints: 44 ADMET endpoints and 1,489 biological activities. Most biological activities are IC50 and Ki measurements and a smaller fraction is expressed as EC50 and Kd data (Figure S1c, S1d). The contrast between the industrial and the OneADMET (public) underscores the focused and systematic nature of industrial measurements (e.g., Caco-2 permeability, LogD7.4), which typically involve 1,000 to 10,000 congeneric compounds, compared to the more diverse but smaller-scale public datasets of 500–1,000 compounds (Figure 3b).

OneADMET is structured into two primary domains: ADMET endpoints and biological activity tasks. While biological activity tasks are more numerous, ADMET tasks involve larger datasets, typically comprising 1,000–10,000 compounds per task, compared to the 100–1,000 compounds per task in biological activity data (Figure 3c, 3d).

We examined the number of endpoints reported per compound. Public compounds are generally measured on a single endpoint (Figure 3e). In contrast, industrial compounds are typically measured on 2–3 endpoints. This is another illustration of the systematic evaluation protocols applied in pharmaceutical pipelines.

We explored relationships between tasks by constructing a correlation matrix based on shared compounds between tasks. Among public datasets from OneADMET, approximately 50% of tasks exhibit positive or negative correlations, forming groups of related activities . These cluster tend to group tasks that are expected and already reported by others as exemplified by the cluster grouping CYP pIC50, lipophilicity, and solubility (Figure S2). Industrial datasets showed stronger degrees of correlation, the consistency of the measurements in the dataset allows to observe relationships that are hardly noticeable in the public dataset. Industrial compounds in an industrial drug development process are investigated for a biological property of interest and quite systematically, for microsomal stability, LogD, and LogP sharing significant data (Figure S3).

Moreover, compound sharing between endpoints revealed limited overlap in public datasets (e.g., CYP and microsomal clearance, Figure S4). In contrast, industrial datasets demonstrated a high degree of overlap in compound measurements across related endpoints, reflecting the systematic approach of industrial drug development processes. For instance, compounds investigated for biological properties were routinely evaluated for complementary ADMET endpoints like microsomal stability, LogD, and LogP, resulting in strong data connections between these endpoints (Figure S5).

We expected that the correlations between tasks would translate into model building synergies for MTL. Highly correlated clusters reflect shared underlying processes or mechanisms, enhancing the potential for MTL to leverage interdependencies between tasks.

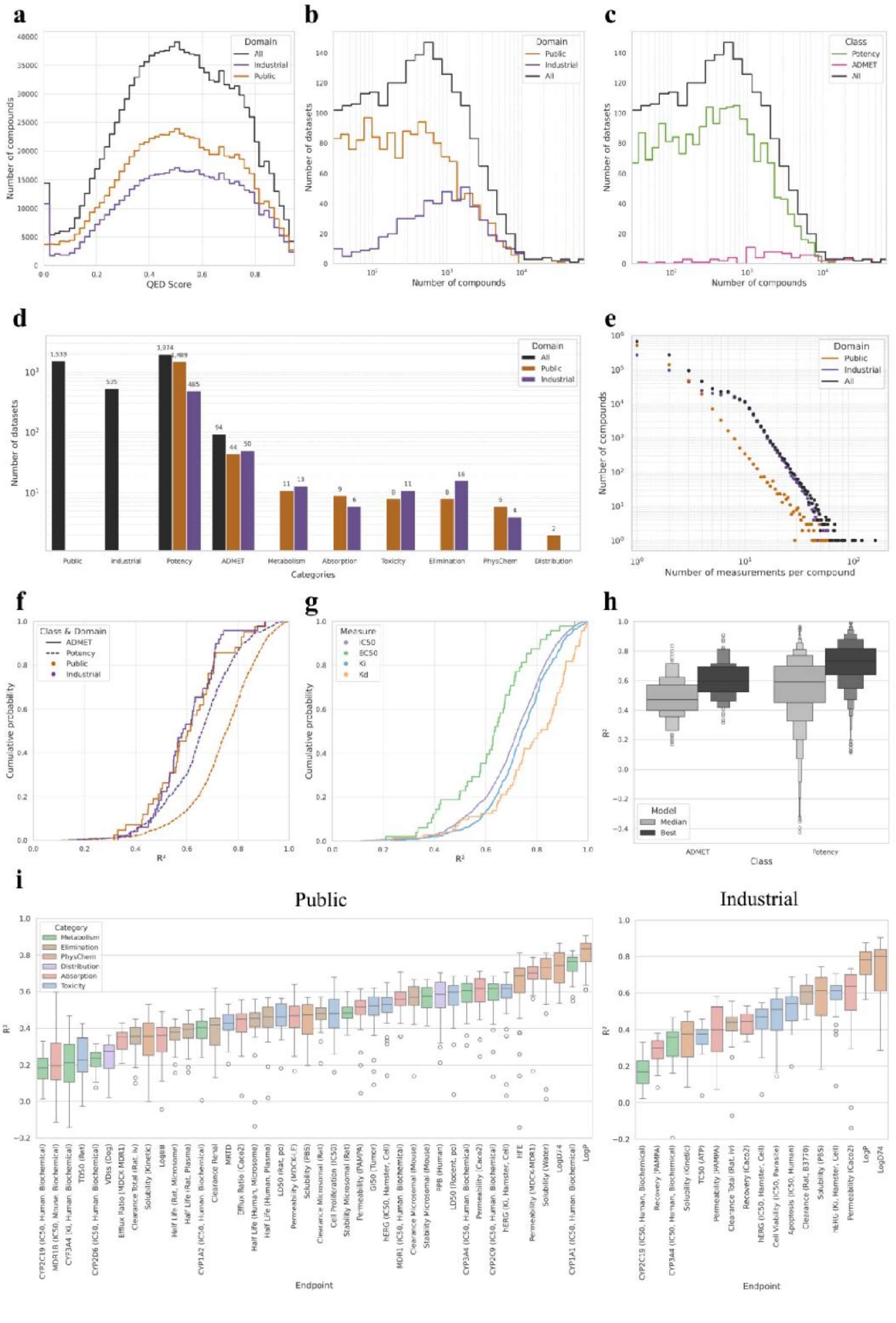


Figure 3: **Presentation of the data of OneADMET from drug-likeliness metrics to overall R^2 performances on the ADMET endpoints.** (**a**) Distribution of the QED of the public (orange), the industrial (purple) and all compounds of the full

set (black). (**b**) Distribution of the size of the datasets in function of their domain as either public, industrial, or all compounds. (**c**) Distribution of the size of the datasets in function of their class as either ADMET or potency a.k.a bioactivities or all. (**d**) Count plot of the number of dataset a.k.a task per domain or/and class. (**e**) Distribution of the number of measurements per compound in function of their domain. (**f**) Cumulative distribution of the median R^2 (Pearson), of individual models selected for the consensus, on the endpoints external test sets in function of their domain or class. (**g**) Cumulative distribution of the median R^2 (Pearson), of individual models selected for the consensus, as a function of the potency task as either EC50, IC50, Ki, or Kd. (**h**) Boxplot of the median R^2 (Pearson), of individual models selected for the consensus for ADMET or potency tasks (**i**) Boxplot distribution of the R^2 (Pearson), of all individual models on the test per ADMET tasks for the public, and for a subset of the ADMET task for the industrial. The color of the box depends on the ADMET category.

Development of ML methods

The public OneADMET dataset was applied to develop predictive models. These models illustrate the modelisability of each task. Models were trained on a training set covering 80% of the data and validated on the remaining 20%. Hyperparameter optimization was conducted as part of the training procedure, based stratified 3-fold cross-validation and 30 iterations of a Bayesian optimization procedure. This approach was conducted systematically and consistently: each model was developed using the same partition of data for training, testing and cross-validation. Both STL (Single-Task Learning) and MTL models were trained on identical data. To summarize, the ChemProp GNN MTL model was optimized and benchmarked against state-of-the-art methods such as Random Forest, XGBoost, STL GNN, Kernel Ridge Regression (KR), K-Nearest Neighbors (KNN), and Support Vector Regression (SVR). Each method was evaluated across an exhaustive combination of descriptors and endpoints, exploring the modelisability of a large space of features.

Performance on ADMET and Biological Activity Endpoints

The performance of predictive models was assessed using R^2 scores per endpoint. The cumulative distribution of the median test performances of models (Figure 3f) was used to compare the performances of the public model applied to the public test and the industrial model applied to the industrial test. First, we observed that the performances are overall similar for ADMET endpoints. Second, biological activity tasks showed generally higher predictive performance, with the median R^2 values averaging at 0.8.

Potency Measures and Endpoint Complexity

Performance was then examined based on potency measures such as IC50, EC50, Kd, and Ki (Figure 3g). IC50 is the concentration of a compound required to inhibit 50% of a target's activity, typically measured in enzyme or cell-based assays, and influenced by substrate concentration and assay conditions. EC50 represents the concentration needed to achieve 50% of the maximal functional response in a cellular system, affected by receptor density, signaling pathways, and assay sensitivity. Kd quantifies the equilibrium dissociation of a ligand from its target, determined in controlled biophysical assays, making it a stable measure of binding affinity. Ki defines the intrinsic binding affinity of an inhibitor, calculated in competitive binding assays and independent of enzyme or substrate concentrations.

EC50 endpoints were the hardest to predict, Kd endpoints the easiest, with IC50 in between. Variability in IC50 values can arise due to differences in assay conditions, including buffer composition, temperature, pH, and experimental setup. Such variability complicates direct comparisons across datasets. EC50 values, dependent on functional responses, are further influenced by receptor density, downstream signaling, and assay sensitivity, leading to higher variability compared to IC50 or Kd. This variability explains the broader error distributions and lower predictive performance observed for EC50 endpoints. In contrast, Kd reflecting binding affinities under controlled conditions, result in more stable and accurate predictions, as reflected in higher R^2 values.

General Predictive Performance

To evaluate the general performance of all predictive models, both median and best-case R^2 scores per endpoint were analyzed (Figure 3h). For ADMET endpoints:

- Median R^2 values were typically in the range of 0.4–0.6.
- Best models showed an increase of +0.1 to +0.2 in R^2 from the median performances to the best model performances, highlighting the importance of adequate hyperparameters optimization.

For biological activity tasks:

- Median R^2 scores ranged from 0.6 to 0.8, with best-case scores often exceeding 0.8.
- Best models showed an increase of +0.05 to 0.1 in R^2.

The median R2 tends to be a bit lowered because of the use of inadequate molecular descriptors. It was consistently observed, as others before use[185], that Estate and MACCS descriptors are generally not the best choice. Also, biological activity endpoints exhibited greater variability in performance, likely due to smaller dataset sizes. Some of these datasets are displaying an obvious SAR on congeneric chemical families – as the pKi TTK (Threonine Tyrosine Kinase) human dataset for instance. On the other hand, other tasks involved a small number of instances that are very diverse so that there sometimes little ground to support a prediction for an external instance, for instance the pIC50 TNR1A (Tumor Necrosis factor Receptor superfamily member 1A) human dataset.

However, we did not investigated outliers in the different datasets. Some errors may remain due to ambiguities in the exact definition of some biological targets can be confusing. For instance, CDK2 activities are often assumed to be acquired on CDK2/cyclinA2-based assays. However, the experimental setup sometimes requires verification to not be confused with an assay involving another cyclin dependent kinase, an affinity measured without cyclin or with another cyclin. An example of a compound tested on CDK2 that can be confused in this way is NU6027 (CHEMBL303948, BDB5566). Other sources of errors that can be cited are: experimental noise, usage of different probes, application to various engineered proteins or cell-lines, compounds stability, ratio of DMSO.

We observed however that biological activity endpoints tend to be more modelisable than ADMET endpoints. This is to confront to a higher experimental noise in ADMET data as observed comparing the SDi for ADMET datasets and biological activity datasets. Yet, optimized models for ADMET endpoints were still able to achieve competitive performance.

Category-specific insights

The R^2 distributions across categories (Figure 3i) reveals that physicochemical endpoints are the easiest to predict, with scores averaging 0.7–0.8. Other ADMET categories, such as metabolism and absorption, exhibited moderate performance (R^2 ~0.5), while toxicity endpoints proved the most challenging, with 20% of tasks achieving $R^2 < 0.4$. Our previous work highlighted similar challenges, particularly in the context of Caco-2 and PAMPA assays. For example, we observed that discrepancies in permeability predictions were often linked to experimental variability, restraining public to be merged with industrial data and making the application of public models to industrial data non-viable.

The measurement of efflux ratios and apparent permeability further exemplifies this disparity. For instance, as highlighted in our previous studies, industrial assays for Caco-2 often incorporate Bovine Serum Albumin (BSA) to enhance absorption modeling, while public datasets omit BSA altogether, relying on simpler protocols. Similarly, PAMPA assays differ in membrane types, pH gradients, and transport conditions, leading to variations in log10 permeability measurements. Boxplots summarizing test performance revealed that industrial and public datasets showed similar distributions, except in the worst-performing cases, where public endpoints for specific toxicity measures had the lowest R^2 values.

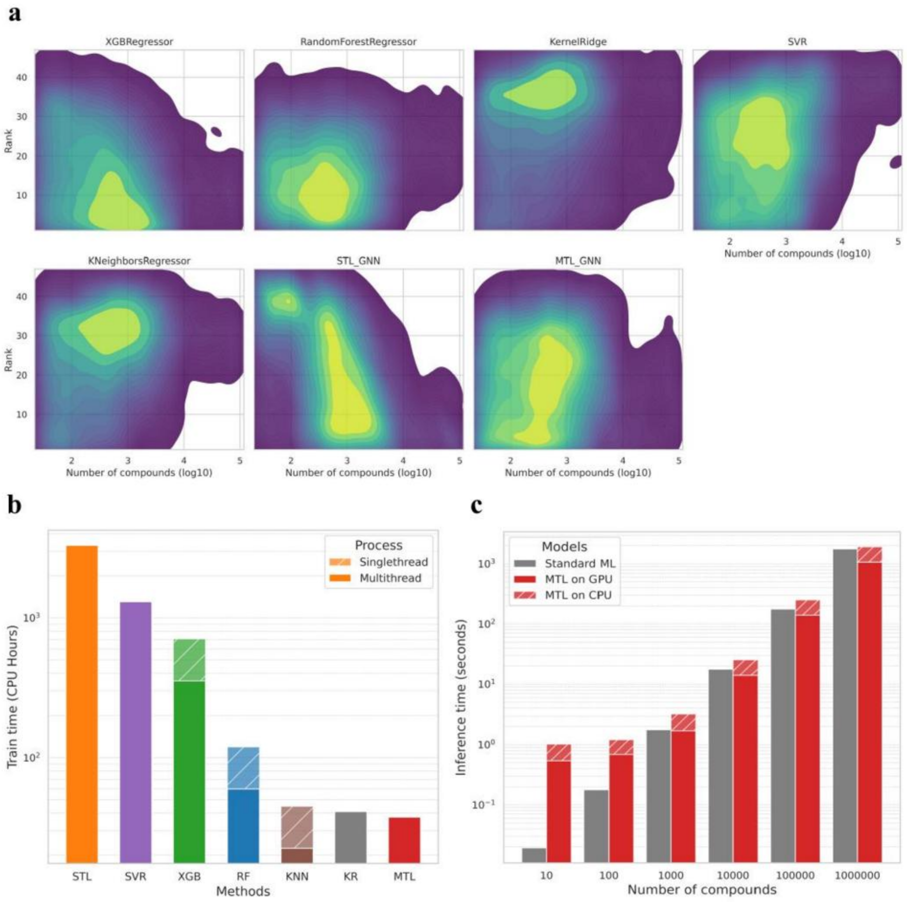


Figure 4: **Analysis of the performance of the predictive models.** (**a**) Density distribution of the rank of the models performances on the test set in function of the logarithm 10 of the size of the dataset. Distributions are colored from purple (low density) to yellow (high density). (MTL: Multi-Task Learning, STL: Single-Task Learning, ML: Machine Learning) (**b**) Barplot of the total time in CPU hours to train and optimized models for all the endpoints, per method. The value is considered using one CPU or GPU for GNN models, on a single threaded process (dashed, not in parallel) or if possible (with RandomForest, XGBoost and k-Nearest Neighbors) on Multi-threaded process (not dashed, in parallel). (**c**) Inference time in seconds in function of the number of compounds by either using standard ML (Random Forest) or the Multi-Task GNN, either on CPU (dashed) or GPU (not dashed).

Method comparisons

To evaluate the impact of descriptors and methods on predictive performance, we conducted a comprehensive comparison across all datasets. The goal was to identify cases where specific descriptors or methods were better suited and to assess the added value of MTL in large-scale applications. For this, we ranked per endpoint all models, each corresponding to the use of one method applied to one set of descriptors, according to their RMSE, assigning the lowest rank to the best-performing combination

(Figure 4a). We also compared methods using R^2 scores to provide a detailed understanding of their efficacy.

A subset of R^2 performances for public *data* is shown in Figures S6, highlighting the close performance of MTL models to STL models. However, STL occasionally exhibited significant performance failures whatever algorithm being used, but not all algorithms are failing for the same tasks. As a result, it is not possible to recommend a solution that would fit all problems optimally, which is another illustration of the No Free Lunch theorem[186].

The relationship between descriptors, methods, and performance was explored using heatmaps of ranks for ADMET and bioactivity endpoints (Figures S7a and S7b). No single method consistently achieved a median rank over all methods below 7, indicating no universally superior combination of method and descriptor. However, XGBoost consistently outperformed RF, which in turn outperformed other methods. For bioactivity endpoints, descriptors like Avalon and ECFP were marginally better, while for ADMET tasks, these descriptors showed more pronounced deviations in performance. Overall, XGBoost or RF paired with Avalon or ECFP descriptors emerged as reliable options for modeling novel datasets, requiring minimal hyperparameter optimization to achieve strong performance with R^2 around 0.6 to 0.9.

We observed that R^2 scores across all endpoints followed a Gaussian distribution when plotted against dataset size (log10) (Figure S7c). The performance of models on larger datasets, particularly ADMET-related endpoints, reflects the nature of these data. ADMET datasets are extensive and span highly diverse chemical spaces, but their cell-based nature introduces significant variability, even with standardized approaches. Even with meticulous curation, conditions can significantly impact model performance, as numerous samples may lack properly annotated metadata. This limitation constrains dataset refinement and is further compounded by inherent assay noise. Factors such as transporter expression, measurement conditions, and compound-specific properties—such as lipophilicity and affinity for laboratory apparatus, contribute to variability, making it challenging to achieve consistently high performance, regardless of dataset size.

We observed a performance peak between $R^2 = 0.6$ and 0.8 for datasets of around 1,000 compounds. Methods such as XGBoost and RF consistently outperformed SVR, KR, and KNN on large datasets. While GNN MTL and STL performed similarly on large datasets, MTL demonstrated superior performance on smaller datasets, outperforming GNN STL, RF, and XGBoost. This observation is particularly interesting as an attractive feature of MTL models is to improve the generalizability of statistical models for datasets that lack data points, using the training on large models to supplement the missing information[187]s. It is nice to observe this statistical behavior in such large MTL model, using a default weighting scheme.

Rank comparison across dataset sizes

To further assess method performance, we analyzed method ranks against dataset sizes (Figure 4a). For small datasets (<500 compounds), STL GNN often underperformed, while MTL GNN maintained in the top ranks. For small datasets, competitive methods are SVR and KNN.

On medium-sized datasets (500–5,000 compounds), the XGBoost and RF appeared more efficient, RF being more efficient for the least populated datasets. However, both STL GNN and MTL GNN appeared also as methods of choice. We found interesting to observe the ranks of the two methods exchanging as the size of the dataset is increasing: MTL for smaller dataset and STL for larger ones.

For large datasets (>5,000 compounds), STL GNN, MTL GNN, and XGBoost emerged as the leading methods. Finally, the MTL GNN and XGboost proved to be useful on a wide range of dataset sizes. On the other hand, the KernelRidge was dominated systematically. In contrast to an SVR, the KernelRidge uses all data points of a dataset to build a model and has a quadratic loss function. It is therefore more susceptible to outlying observation. We hypothesize that it should require higher quality data and a more thorough exploration of the kernel space to make the method shine.

Computational Efficiency in Training and Inference

Modeling also involves considerations of computational time for descriptor computation, training, and inference. Training time was analyzed across dataset sizes (Figure S7d). We performed all calculations on the Amazon Web Service, using multiple g5.4xlarge instances (16 CPUs, 64 RAM, 1 NVIDIA A10G GPU), parallelizing the processes on each, and estimated the CPU threading and GPU time to compute with the timeit module. Therefore, all results reported are only indicative. For MTL we report the time for training the whole model for all endpoints in parallel, while for STL approach we report the time needed to train all models sequentially. The calculation of the molecular descriptors is included in the computation times reported.

Methods like SVR, KNN, and KR showed strong dependence on dataset size, with SVR training times approaching those of STL GNN for large datasets. Among all methods, STL GNN had the slowest training speed, while MTL, RF, and XGBoost were the most scalable. The training time for MTL GNN was comparable to XGBoost. However, the hyperparameters of an MTL methods are optimized once. The method saves a lot of efforts compared to a sequence of STL models where hyperparameters needs to be optimized on each task.

Inference speed was another critical consideration. We report time measurements using a single g5.xlarge instances (4 CPUs, 16 RAM, 1 NVIDIA A10G GPU) machine by either running computing on the GPU, or CPU by single-, or multi-threading. The time required for inference on increasing dataset sizes was compared for standard methods like RF and SVR against MTL GNN on both CPU and GPU (Figure S8). Standardization and descriptors computation emerged as the primary bottleneck. It is observed also that whatever solution used, the computational time is evolving linearly with the size of the dataset to infer. The GNN approaches appear to be more efficient to infer large dataset (>10000 compounds). We hypothesize that it is due to the more efficient calculation of the embedded molecular featurization. MTL GNN can efficiently use GPU and be sensibly faster for the largest datasets.

Applicability Domain

Predictive models are inherently prone to errors, particularly when applied to unseen data that differ from training samples in structure or response values. To mitigate this, predictive models must operate within an AD, a framework designed to identify compounds that fall inside or outside the learning space of the model. For each dataset, an AD was prepared using the Local Outlier Factor algorithm. The AD was designed to identify compounds within the chemical space of the training dataset where predictions are reliable, ensuring that the model does not extrapolate to areas of the chemical space with insufficient representation. To establish the AD for each endpoint, we utilized the best descriptor identified during the benchmark process, ensuring that the AD was tailored to the specific features and characteristics of the dataset. The Local Outlier Factor method was applied using standard hyperparameters. Each endpoint's best single-task learning model was paired with its corresponding LOF model to ensure precise and endpoint-specific AD estimation.

Conclusion

This study presents a comprehensive approach to improving predictive modeling in drug discovery by integrating a unified MTL framework with the extensive OneADMET dataset. With data encompassing over 738,161 compounds and 1,119,719 million measurements across 44 ADMET endpoints and 1,489 biological activity metrics, OneADMET offers a detailed and reliable resource that addresses long-standing data limitations, supporting both academic research and industrial applications[24,188].

A rigorous data standardization process underpins our work, involving careful filtering of duplicates, minimization of experimental noise, and strict quality controls. This foundation accurately reflects the complex chemical and biological realities inherent in drug discovery and builds on earlier studies that advocate for integrative methodologies in handling heterogeneous datasets[189]. Transitioning from STL to MTL offers several advantages:

- STL models require separate hyperparameter optimization for each endpoint, which increases computational complexity as the number of tasks grows. MTL, however, allows for a single optimization step across all tasks, thereby streamlining the process and reducing overhead[6].
- The shared latent representations in MTL facilitate the capture of inter-task dependencies, such as correlations among CYP inhibition, microsomal stability, and lipophilicity. This observation supports previous findings and underscores MTL's potential to improve predictive performance, particularly in scenarios with limited data.
- By consolidating multiple predictive tasks into one model, the MTL approach reduces the complexity involved in deployment and ongoing maintenance compared to managing a suite of STL models[119].
- While MTL may seem more extensive than necessary for a single endpoint, its true value lies in integrating related endpoints to provide richer predictive insights. Importantly, the inference

time remains comparable to that of STL models, as both rely on similar processes such as molecular structure standardization and descriptor computation.

The performance of our models is reflected in best R^2 values of 0.6–0.8 for ADMET tasks and 0.7–0.9 for biological activity tasks, aligning well with state-of-the-art benchmarks[185]. These results suggest that our unified large-scale MTL framework is effective in capturing complex, interrelated biological phenomena while capturing distant tasks.

Nevertheless, challenges remain. Variability in experimental conditions continues to impact model performance, a concern highlighted by several experts[190]. In response, initiatives such as the Polaris project have emerged to enhance data curation and accessibility, helping to address these inconsistencies[21]. Additionally, advances in federated learning[15] and the development of integrative databases offer promising avenues for further improvement.

Looking forward, the OneADMET dataset and our unified MTL framework provide a solid foundation for continued application of large MTL in predictive modeling for drug discovery. Future efforts should focus on ongoing data curation, refining applicability domains, and incorporating new experimental measurements to capture even more complex biological interactions. Expanding this framework to integrate additional chemical and biological information may further enhance predictive performance[191]. In parallel, ongoing improvements in model interpretability, uncertainty estimation, and multi-modal data integration are expected to drive methodological progress, ultimately paving the way for more effective and reliable predictive pipelines.

Data Availability

The authors declare that the data supporting the findings of this study are available free of charge. The repository features multiple datasets that have been curated for this research.

Code Availability

No custom code has been used.

Author Contributions

PL is the main author. Data collection, annotation process supervision, modeling and statistical analysis of results were carried out by PL, CM and GM. Figures and table preparation by PL and GM. Supervision by CM, GM, and AV. The first version of this article was written by PL and GM; GM, CM and AV led the subsequent revisions.

Notes

C. Minoletti and P. Llompart are Sanofi employees and may hold shares and/or stock options in the company. G. Marcou, and A. Varnek have nothing to disclose.

Abbreviations

AD : Applicability Domain

ADMET : Absorption, Distribution, Metabolism, Elimination, and Toxicity

eADMET : early ADMET assessment

AE : AutoEncoder

ANN : Artificial Neural Network

ASNN : Artificial Associative Neural Network

BBB : Blood-Brain Barrier

BERT : Bidirectional Encoder Representations from Transformers

BSA : Bovine Serum Albumin

CapsNet : Capsule Network

CL : Contrastive Learning

CNN : Convolutional Neural Network

CPDB : Carcinogenic Potency Database

CS : Chemical Space

CYP : Cytochrome P450

DL : Deep Learning

DLCA : Deep Learning Consensus Architecture

DNN : Deep Neural Network

DUD-E : Directory of Useful Decoys – Enhanced

EC50 : Half maximal effective concentration

EPA : Environmental Protection Agency

GAT : Graph Attention Transformer

GCN : Graph Convolutional Network

GCNN : Graph Convolutional Neural Network

GGRNN : Gated Graph Recursive Neural Network

GIN : Graph Isomorphism Network

GINE : Graph Isomorphism Network with Edge

GNN : Graph Neural Network

GP : Gaussian Process

GT : Graph Transformer

IC50 : Half maximal inhibitory concentration

IUPHAR : International Union of Basic and Clinical Pharmacology

Kd : Equilibrium dissociation constant

Ki : Inhibition constant

KNN : k-Nearest Neighbors

KR : Kernel Ridge Regression

LD50 : Lethal Dose for 50% of the population

LOF : Local Outlier Factor

LR : Logistic Regression

ML : Machine Learning

MTR : Model Tree Regression

MTL : Multi-Task Learning

MUV : Maximum Unbiased Validation

NB : Naïve Bayes

NCI : National Cancer Institute

OChem : Online Chemical Database

PCBA : PubChem BioAssay

pIC50 : Negative logarithm of the IC50 value

P-gP : P-glycoprotein

PLS : Partial Least Square

QED : Quantitative Estimate of Drug-likeness

RBM : Restricted Boltzmann Machine

RF : Random Forest

RMSE : Root Mean Squared Error

RNN : Recurrent Neural Network

SMILES : Simplified Molecular Input Line Entry System

SMOTE : Synthetic Minority Over-sampling Technique

STL : Single-Task Learning

SVM : Support Vector Machine

SVR : Support Vector Regression

TDC : Therapeutics Data Commons

Tox21 : Toxicology in the 21st Century program

wMSE : Weighted Mean Squared Error

XGBoost : Extreme Gradient Boosting

References

(1) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

(2) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *Journal of Health Economics* **2016**, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

(3) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat Rev Drug Discov* **2004**, *3* (8), 711–716. https://doi.org/10.1038/nrd1470.

(4) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat Rev Drug Discov* **2015**, *14* (7), 475–486. https://doi.org/10.1038/nrd4609.

(5) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat Rev Drug Discov* **2010**, *9* (3), 203–214. https://doi.org/10.1038/nrd3078.

(6) Caruana, R. Multitask Learning. *Machine Learning* **1997**, *28* (1), 41–75. https://doi.org/10.1023/A:1007379606734.

(7) Zhang, Y.; Yang, Q. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering* **2022**, *34* (12), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203.

(8) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29* (6–7), 476–488. https://doi.org/10.1002/minf.201000061.

(9) Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **2009**, *49* (1), 133–144. https://doi.org/10.1021/ci8002914.

(10) Su, H.; Rousu, J. Multi-Task Drug Bioactivity Classification with Graph Labeling Ensembles. In *Pattern Recognition in Bioinformatics*; Loog, M., Wessels, L., Reinders, M. J. T., de Ridder, D., Eds.; Springer: Berlin, Heidelberg, 2011; pp 157–167. https://doi.org/10.1007/978-3-642-24855-9_14.

(11) Capuzzi, S. J.; Politi, R.; Isayev, O.; Farag, S.; Tropsha, A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Front. Environ. Sci.* **2016**, *4*. https://doi.org/10.3389/fenvs.2016.00003.

(12) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*. https://doi.org/10.3389/fenvs.2015.00080.

(13) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (3), 1253–1268. https://doi.org/10.1021/acs.jcim.8b00785.

(14) Zakharov, A. V.; Zhao, T.; Nguyen, D.-T.; Peryea, T.; Sheils, T.; Yasgar, A.; Huang, R.; Southall, N.; Simeonov, A. Novel Consensus Architecture To Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *J. Chem. Inf. Model.* **2019**, *59* (11), 4613–4624. https://doi.org/10.1021/acs.jcim.9b00526.

(15) Humbeck, L.; Morawietz, T.; Sturm, N.; Zalewski, A.; Harnqvist, S.; Heyndrickx, W.; Holmes, M.; Beck, B. Don't Overweight Weights: Evaluation of Weighting Strategies for Multi-Task Bioactivity Classification Models. *Molecules* **2021**, *26* (22), 6959. https://doi.org/10.3390/molecules26226959.

(16) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. arXiv August 28, 2021. https://doi.org/10.48550/arXiv.2102.09548.

(17) *MoleculeNet: a benchmark for molecular machine learning - Chemical Science (RSC Publishing) DOI:10.1039/C7SC02664A*. https://pubs.rsc.org/en/content/articlehtml/2018/sc/c7sc02664a (accessed 2025-01-13).

(18) Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration. *Research* **2022**, *2022*, 0004. https://doi.org/10.34133/research.0004.

(19) *We Need Better Benchmarks for Machine Learning in Drug Discovery*. We Need Better Benchmarks for Machine Learning in Drug Discovery. https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html (accessed 2025-01-13).

(20) Wognum, C.; Ash, J. R.; Aldeghi, M.; Rodríguez-Pérez, R.; Fang, C.; Cheng, A. C.; Price, D. J.; Clevert, D.-A.; Engkvist, O.; Walters, W. P. A Call for an Industry-Led Initiative to Critically Assess Machine Learning for Real-World Drug Discovery. *Nat Mach Intell* **2024**, *6* (10), 1120–1121. https://doi.org/10.1038/s42256-024-00911-w.

(21) *Polaris*. https://polarishub.io (accessed 2025-01-13).

(22) Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E.; Solov'ev, V. P.; Varnek, A. QSAR Modeling of Blood:Air and Tissue:Air Partition Coefficients Using Theoretical Descriptors. *Bioorganic & Medicinal Chemistry* **2005**, *13* (23), 6450–6463. https://doi.org/10.1016/j.bmc.2005.06.066.

(23) *overview of the PubChem BioAssay resource | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/38/suppl_1/D255/3112310 (accessed 2025-01-13).

(24) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. arXiv February 6, 2015. https://doi.org/10.48550/arXiv.1502.02072.

(25) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.

(26) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.

(27). https://tripod.nih.gov/tox21/challenge/data.jsp (Accessed 2025-01-13).

(28) Kearnes, S.; Goldman, B.; Pande, V. Modeling Industrial ADMET Data with Multitask Networks. arXiv January 13, 2017. https://doi.org/10.48550/arXiv.1606.08793.

(29) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57* (10), 2490–2504. https://doi.org/10.1021/acs.jcim.7b00087.

(30) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274. https://doi.org/10.1021/ci500747n.

(31) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076. https://doi.org/10.1021/acs.jcim.7b00146.

(32) Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharmaceutics* **2018**, *15* (10), 4336–4345. https://doi.org/10.1021/acs.molpharmaceut.8b00110.

(33) Liu, K.; Sun, X.; Jia, L.; Ma, J.; Xing, H.; Wu, J.; Gao, H.; Sun, Y.; Boulnois, F.; Fan, J. Chemi-Net: A Molecular Graph Convolutional Network for Accurate Drug Property Prediction. *International Journal of Molecular Sciences* **2019**, *20* (14), 3389. https://doi.org/10.3390/ijms20143389.

(34) Capela, F.; Nouchi, V.; Deursen, R. V.; Tetko, I. V.; Godin, G. Multitask Learning On Graph Neural Networks Applied To Molecular Property Predictions. arXiv October 30, 2019. https://doi.org/10.48550/arXiv.1910.13124.

(35) LI, J.-C. Imbalanced Toxicity Prediction Using Multi-Task Learning and Over-Sampling. In *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*; 2020; pp 1–7. https://doi.org/10.1109/ICMLC51923.2020.9469546.

(36) Peng, Y.; Lin, Y.; Jing, X.-Y.; Zhang, H.; Huang, Y.; Luo, G. S. Enhanced Graph Isomorphism Network for Molecular ADMET Properties Prediction. *IEEE Access* **2020**, *8*, 168344–168360. https://doi.org/10.1109/ACCESS.2020.3022850.

(37) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2020**, *25* (1), 44. https://doi.org/10.3390/molecules25010044.

(38) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63* (16), 8835–8848. https://doi.org/10.1021/acs.jmedchem.9b02187.

(39) Karim, A.; Riahi, V.; Mishra, A.; Newton, M. A. H.; Dehzangi, A.; Balle, T.; Sattar, A. Quantitative Toxicity Prediction via Meta Ensembling of Multitask Deep Learning Models. *ACS Omega* **2021**, *6* (18), 12306–12317. https://doi.org/10.1021/acsomega.1c01247.

(40) Wang, Y.; Wang, B.; Jiang, J.; Guo, J.; Lai, J.; Lian, X.-Y.; Wu, J. Multitask CapsNet: An Imbalanced Data Deep Learning Method for Predicting Toxicants. *ACS Omega* **2021**, *6* (40), 26545–26555. https://doi.org/10.1021/acsomega.1c03842.

(41) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X.; Hou, T.; Cao, D. ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Research* **2021**, *49* (W1), W5–W14. https://doi.org/10.1093/nar/gkab255.

(42) Hamzic, S.; Lewis, R.; Desrayaud, S.; Soylu, C.; Fortunato, M.; Gerebtzoff, G.; Rodríguez-Pérez, R. Predicting In Vivo Compound Brain Penetration Using Multi-Task Graph Neural Networks. *J. Chem. Inf. Model.* **2022**, *62* (13), 3180–3190. https://doi.org/10.1021/acs.jcim.2c00412.

(43) Tian, H.; Ketkar, R.; Tao, P. ADMETboost: A Web Server for Accurate ADMET Prediction. *J Mol Model* **2022**, *28* (12), 408. https://doi.org/10.1007/s00894-022-05373-8.

(44) Zhang, S.; Yan, Z.; Huang, Y.; Liu, L.; He, D.; Wang, W.; Fang, X.; Zhang, X.; Wang, F.; Wu, H.; Wang, H. HelixADMET: A Robust and Endpoint Extensible ADMET System Incorporating Self-Supervised Knowledge Transfer. *Bioinformatics* **2022**, *38* (13), 3444–3453. https://doi.org/10.1093/bioinformatics/btac342.

(45) Martínez Mora, A.; Subramanian, V.; Miljković, F. Multi-Task Convolutional Neural Networks for Predicting in Vitro Clearance Endpoints from Molecular Images. *J Comput Aided Mol Des* **2022**, *36* (6), 443–457. https://doi.org/10.1007/s10822-022-00458-1.

(46) Wang, J.; Lou, C.; Liu, G.; Li, W.; Wu, Z.; Tang, Y. Profiling Prediction of Nuclear Receptor Modulators with Multi-Task Deep Learning Methods: Toward the Virtual Screening. *Briefings in Bioinformatics* **2022**, *23* (5), bbac351. https://doi.org/10.1093/bib/bbac351.

(47) *IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/48/D1/D1006/5613677?login=false (accessed 2025-01-13).

(48) Martin, E. J.; Polyakov, V. R.; Zhu, X.-W.; Tian, L.; Mukherjee, P.; Liu, X. All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, *59* (10), 4450–4459. https://doi.org/10.1021/acs.jcim.9b00375.

(49) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57* (8), 2077–2088. https://doi.org/10.1021/acs.jcim.7b00166.

(50) Nikitina, A. A.; Orlov, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Osolodkin, D. I. Enhanced Taxonomy Annotation of Antiviral Activity Data from ChEMBL. *Database* **2019**, *2019*, bay139. https://doi.org/10.1093/database/bay139.

(51) Vangala, S. R.; Krishnan, S. R.; Bung, N.; Srinivasan, R.; Roy, A. pBRICS: A Novel Fragmentation Method for Explainable Property Prediction of Drug-Like Small Molecules. *J. Chem. Inf. Model.* **2023**, *63* (16), 5066–5076. https://doi.org/10.1021/acs.jcim.3c00689.

(52) Du, B.-X.; Xu, Y.; Yiu, S.-M.; Yu, H.; Shi, J.-Y. ADMET Property Prediction via Multi-Task Graph Learning under Adaptive Auxiliary Task Selection. *iScience* **2023**, *26* (11), 108285. https://doi.org/10.1016/j.isci.2023.108285.

(53) Hu, H.; Bai, Y.; Yuan, Z. Improved Graph-Based Multitask Learning Model with Sparse Sharing for Quantitative Structure–Property Relationship Prediction of Drug Molecules. *AIChE Journal* **2023**, *69* (2), e17968. https://doi.org/10.1002/aic.17968.

(54) Ai, D.; Cai, H.; Wei, J.; Zhao, D.; Chen, Y.; Wang, L. DEEPCYPs: A Deep Learning Platform for Enhanced Cytochrome P450 Activity Prediction. *Front. Pharmacol.* **2023**, *14*. https://doi.org/10.3389/fphar.2023.1099093.

(55) Swanson, K.; Walther, P.; Leitz, J.; Mukherjee, S.; Wu, J. C.; Shivnaraine, R. V.; Zou, J. ADMET-AI: A Machine Learning ADMET Platform for Evaluation of Large-Scale Chemical Libraries. *Bioinformatics* **2024**, *40* (7), btae416. https://doi.org/10.1093/bioinformatics/btae416.

(56) *QComp: A QSAR-Based Data Completion Framework for Drug Discovery*. https://arxiv.org/html/2405.11703v1#bib.bib13 (accessed 2025-01-13).

(57) Yang, B. Iceplussss/QSAR-Complete, 2025. https://github.com/iceplussss/QSAR-Complete (accessed 2025-01-13).

(58) Fu, L.; Shi, S.; Yi, J.; Wang, N.; He, Y.; Wu, Z.; Peng, J.; Deng, Y.; Wang, W.; Wu, C.; Lyu, A.; Zeng, X.; Zhao, W.; Hou, T.; Cao, D. ADMETlab 3.0: An Updated Comprehensive Online ADMET Prediction Platform Enhanced with Broader Coverage, Improved Performance, API Functionality and

Decision Support. *Nucleic Acids Research* **2024**, *52* (W1), W422–W431. https://doi.org/10.1093/nar/gkae236.

(59) Gu, Y.; Yu, Z.; Wang, Y.; Chen, L.; Lou, C.; Yang, C.; Li, W.; Liu, G.; Tang, Y. admetSAR3.0: A Comprehensive Platform for Exploration, Prediction and Optimization of Chemical ADMET Properties. *Nucleic Acids Research* **2024**, *52* (W1), W432–W438. https://doi.org/10.1093/nar/gkae298.

(60) Fitzpatrick, R. B. CPDB: Carcinogenic Potency Database: Roberta Bronson Fitzpatrick, Column Editor. *Medical Reference Services Quarterly* **2008**, *27* (3), 303–311. https://doi.org/10.1080/02763860802198895.

(61) Olker, J. H.; Elonen, C. M.; Pilli, A.; Anderson, A.; Kinziger, B.; Erickson, S.; Skopinski, M.; Pomplun, A.; LaLone, C. A.; Russom, C. L.; Hoff, D. The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment. *Environmental Toxicology and Chemistry* **2022**, *41* (6), 1520–1539. https://doi.org/10.1002/etc.5324.

(62) Dorne, J. L. C. M.; Richardson, J.; Livaniou, A.; Carnesecchi, E.; Ceriani, L.; Baldin, R.; Kovarich, S.; Pavan, M.; Saouter, E.; Biganzoli, F.; Pasinato, L.; Zare Jeddi, M.; Robinson, T. P.; Kass, G. E. N.; Liem, A. K. D.; Toropov, A. A.; Toropova, A. P.; Yang, C.; Tarkhov, A.; Georgiadis, N.; Di Nicola, M. R.; Mostrag, A.; Verhagen, H.; Roncaglioni, A.; Benfenati, E.; Bassan, A. EFSA's OpenFoodTox: An Open Source Toxicological Database on Chemicals in Food and Feed and Its Future Developments. *Environment International* **2021**, *146*, 106293. https://doi.org/10.1016/j.envint.2020.106293.

(63) Walter, M.; Borghardt, J. M.; Humbeck, L.; Skalic, M. Multi-Task ADME/PK Prediction at Industrial Scale: Leveraging Large and Diverse Experimental Datasets. ChemRxiv January 12, 2024. https://doi.org/10.26434/chemrxiv-2024-pf4w9.

(64) *Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.3c00160 (accessed 2025-01-13).

(65) Kim, J.; Chang, W.; Ji, H.; Joung, I. Quantum-Informed Molecular Representation Learning Enhancing ADMET Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (13), 5028–5040. https://doi.org/10.1021/acs.jcim.4c00772.

(66) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. admetSAR 2.0: Web-Service for Prediction and Optimization of Chemical ADMET Properties. *Bioinformatics* **2019**, *35* (6), 1067–1069. https://doi.org/10.1093/bioinformatics/bty707.

(67) *pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.5b00104 (accessed 2025-01-13).

(68) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52* (11), 3099–3105. https://doi.org/10.1021/ci300367a.

(69) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME–Tox Predictions? *Drug Discovery Today* **2006**, *11* (15), 700–707. https://doi.org/10.1016/j.drudis.2006.06.013.

(70) Drwal, M. N.; Banerjee, P.; Dunkel, M.; Wettig, M. R.; Preissner, R. ProTox: A Web Server for the in Silico Prediction of Rodent Oral Toxicity. *Nucleic Acids Research* **2014**, *42* (W1), W53–W58. https://doi.org/10.1093/nar/gku401.

(71) *SuperToxic: a comprehensive database of toxic compounds | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/37/suppl_1/D295/1015670?login=true (accessed 2025-01-14).

(72) Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part II: QSAR Models of Skin Permeability and the Relationships between Skin Permeability and Skin Sensitization. *Toxicology and Applied Pharmacology* **2015**, *284* (2), 273–280. https://doi.org/10.1016/j.taap.2014.12.013.

(73) Berellini, G.; Springer, C.; Waters, N. J.; Lombardo, F. In Silico Prediction of Volume of Distribution in Human Using Linear and Nonlinear Models on a 669 Compound Data Set. *J. Med. Chem.* **2009**, *52* (14), 4488–4495. https://doi.org/10.1021/jm9004658.

(74) Amo, E. M. del; Ghemtio, L.; Xhaard, H.; Yliperttula, M.; Urtti, A.; Kidron, H. Applying Linear and Non-Linear Methods for Parallel Prediction of Volume of Distribution and Fraction of Unbound Drug. *PLOS ONE* **2013**, *8* (10), e74758. https://doi.org/10.1371/journal.pone.0074758.

(75) Suenderhauf, C.; Hammann, F.; Huwyler, J. Computational Prediction of Blood-Brain Barrier Permeability Using Decision Tree Induction. *Molecules* **2012**, *17* (9), 10429–10445. https://doi.org/10.3390/molecules170910429.

(76) Yap, C. W.; Li, Z. R.; Chen, Y. Z. Quantitative Structure–Pharmacokinetic Relationships for Drug Clearance by Using Statistical Learning Methods. *Journal of Molecular Graphics and Modelling* **2006**, *24* (5), 383–395. https://doi.org/10.1016/j.jmgm.2005.10.004.

(77) *Locally Weighted Learning Methods for Predicting Dose-Dependent Toxicity with Application to the Human Maximum Recommended Daily Dose | Chemical Research in Toxicology*. https://pubs.acs.org/doi/full/10.1021/tx300279f (accessed 2025-01-14).

(78) *Modeling Oral Rat Chronic Toxicity | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci8001974 (accessed 2025-01-14).

(79) Schyman, P.; Liu, R.; Desai, V.; Wallqvist, A. vNN Web Server for ADMET Predictions. *Front. Pharmacol.* **2017**, *8*. https://doi.org/10.3389/fphar.2017.00889.

(80) Muehlbacher, M.; Spitzer, G. M.; Liedl, K. R.; Kornhuber, J. Qualitative Prediction of Blood–Brain Barrier Permeability on a Large and Refined Dataset. *J Comput Aided Mol Des* **2011**, *25* (12), 1095–1106. https://doi.org/10.1007/s10822-011-9478-1.

(81) *ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage | Molecular Pharmaceutics*. https://pubs.acs.org/doi/abs/10.1021/mp300023x (accessed 2025-01-14).

(82) *A Generally Applicable Computer Algorithm Based on the Group Additivity Method for the Calculation of Seven Molecular Descriptors: Heat of Combustion, LogPO/W, LogS, Refractivity, Polarizability, Toxicity and LogBB of Organic Compounds; Scope and Limits of Applicability*. https://www.mdpi.com/1420-3049/20/10/18279 (accessed 2025-01-14).

(83) *Profiling of the Tox21 Chemical Collection for Mitochondrial Function to Identify Compounds that Acutely Decrease Mitochondrial Membrane Potential | Environmental Health Perspectives | Vol. 123, No. 1*. https://ehp.niehs.nih.gov/doi/full/10.1289/ehp.1408642 (accessed 2025-01-14).

(84) *Deep Learning for Drug-Induced Liver Injury | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/full/10.1021/acs.jcim.5b00238 (accessed 2025-01-14).

(85) *ADME Evaluation in Drug Discovery. 10. Predictions of P-Glycoprotein Inhibitors Using Recursive Partitioning and Naive Bayesian Classification Techniques | Molecular Pharmaceutics*. https://pubs.acs.org/doi/abs/10.1021/mp100465q (accessed 2025-01-14).

(86) Greene, N.; Fisk, L.; Naven, R. T.; Note, R. R.; Patel, M. L.; Pelletier, D. J. Developing Structure–Activity Relationships for the Prediction of Hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23* (7), 1215–1222. https://doi.org/10.1021/tx1000865.

(87) *Cellular Imaging Predictions of Clinical Drug-Induced Liver Injury | Toxicological Sciences | Oxford Academic*. https://academic.oup.com/toxsci/article/105/1/97/1662976?login=true (accessed 2025-01-14).

(88) *Mixed learning algorithms and features ensemble in hepatotoxicity prediction | Journal of Computer-Aided Molecular Design*. https://link.springer.com/article/10.1007/s10822-011-9468-3 (accessed 2025-01-14).

(89) *ADMET Evaluation in Drug Discovery. 13. Development of in Silico Prediction Models for P-Glycoprotein Substrates | Molecular Pharmaceutics*. https://pubs.acs.org/doi/abs/10.1021/mp400450m (accessed 2025-01-14).

(90) *Benchmark Data Set for in Silico Prediction of Ames Mutagenicity | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci900161g (accessed 2025-01-14).

(91) *Locally Weighted Learning Methods for Predicting Dose-Dependent Toxicity with Application to the Human Maximum Recommended Daily Dose | Chemical Research in Toxicology*. https://pubs.acs.org/doi/full/10.1021/tx300279f (accessed 2025-01-14).

(92) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci Rep* **2017**, *7* (1), 42717. https://doi.org/10.1038/srep42717.

(93) Dong, J.; Wang, N.-N.; Yao, Z.-J.; Zhang, L.; Cheng, Y.; Ouyang, D.; Lu, A.-P.; Cao, D.-S. ADMETlab: A Platform for Systematic ADMET Evaluation Based on a Comprehensively Collected ADMET Database. *Journal of Cheminformatics* **2018**, *10* (1), 29. https://doi.org/10.1186/s13321-018-0283-x.

(94) Wang, N.-N.; Dong, J.; Deng, Y.-H.; Zhu, M.-F.; Wen, M.; Yao, Z.-J.; Lu, A.-P.; Wang, J.-B.; Cao, D.-S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56* (4), 763–773. https://doi.org/10.1021/acs.jcim.5b00642.

(95) *ESOL: Estimating Aqueous Solubility Directly from Molecular Structure | Journal of Chemical Information and Modeling.* https://pubs.acs.org/doi/abs/10.1021/ci034243x (accessed 2025-01-14).

(96) *Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci9901338 (accessed 2025-01-14).

(97) *In silico evaluation of logD7.4 and comparison with other prediction methods - Wang - 2015 - Journal of Chemometrics - Wiley Online Library*. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.2718 (accessed 2025-01-14).

(98) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *J. Med. Chem.* **2011**, *54* (6), 1740–1751. https://doi.org/10.1021/jm101421d.

(99) *P-glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Data set | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci2001583 (accessed 2025-01-14).

(100) *ADME Evaluation in Drug Discovery. 8. The Prediction of Human Intestinal Absorption by a Support Vector Machine | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci7002076 (accessed 2025-01-14).

(101) Wang, N.-N.; Huang, C.; Dong, J.; Yao, Z.-J.; Zhu, M.-F.; Deng, Z.-K.; Lv, B.; Lu, A.-P.; F. Chen, A.; Cao, D.-S. Predicting Human Intestinal Absorption with Modified Random Forest Approach: A Comprehensive Evaluation of Molecular Representation, Unbalanced Data, and Applicability Domain Issues. *RSC Advances* **2017**, *7* (31), 19007–19018. https://doi.org/10.1039/C6RA28442F.

(102) *ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints | Molecular Pharmaceutics*. https://pubs.acs.org/doi/abs/10.1021/mp100444g (accessed 2025-01-14).

(103) *The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding | Pharmaceutical Research*. https://link.springer.com/article/10.1007/s11095-013-1023-6 (accessed 2025-01-14).

(104) Zhivkova, Z.; Doytchinova, I. Quantitative Structure—Plasma Protein Binding Relationships of Acidic Drugs. *Journal of Pharmaceutical Sciences* **2012**, *101* (12), 4627–4641. https://doi.org/10.1002/jps.23303.

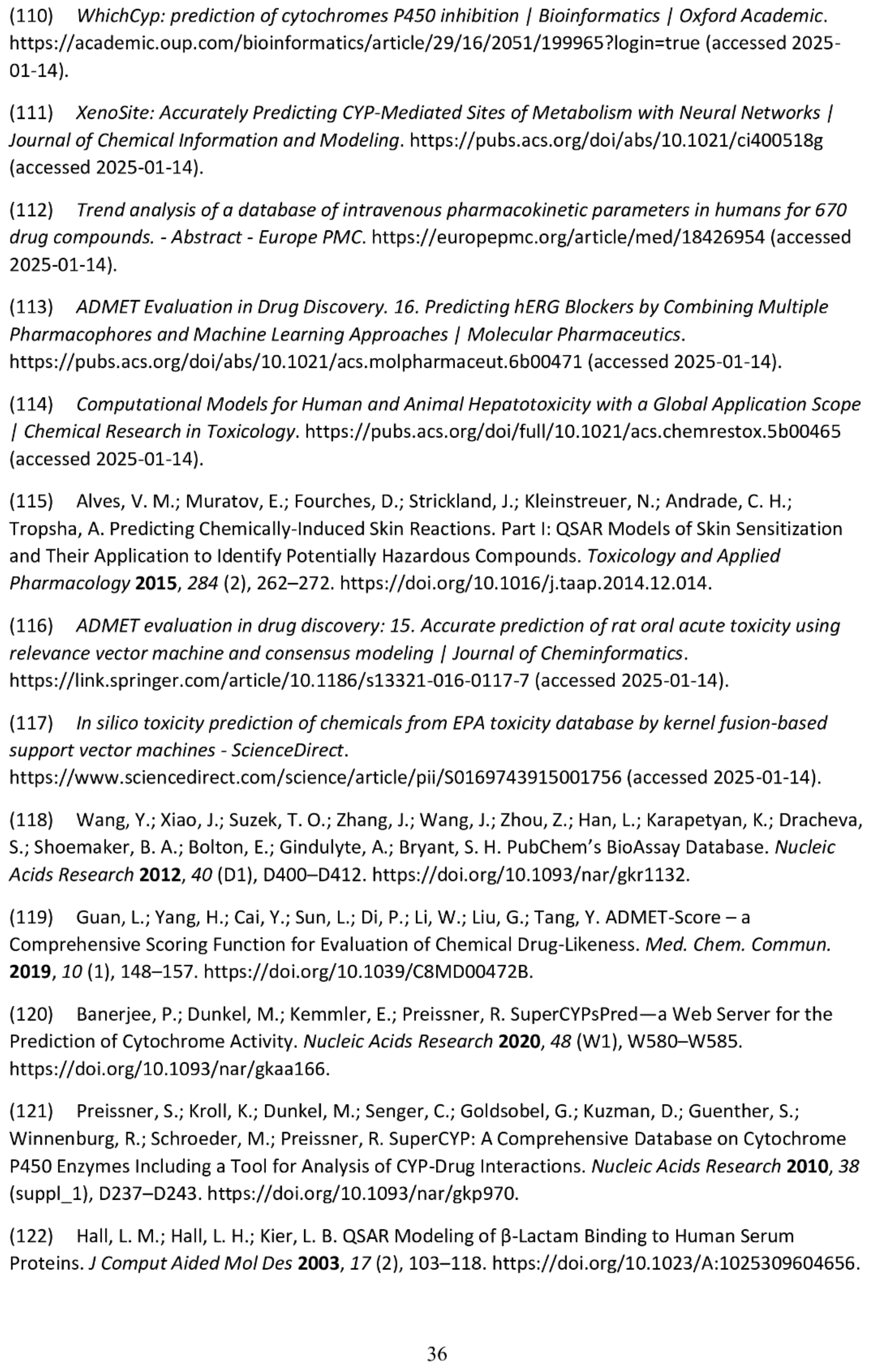
(105) Ghafourian, T.; Amin, Z. QSAR Models for the Prediction of Plasma Protein Binding. *Bioimpacts* **2013**, *3* (1), 21–27. https://doi.org/10.5681/bi.2013.011.

(106) *Effect of Selection of Molecular Descriptors on the Prediction of Blood−Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci050135u (accessed 2025-01-14).

(107) *Estimation of ADME Properties with Substructure Pattern Recognition | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci100104j (accessed 2025-01-14).

(108) *Predicting drug metabolism: experiment and/or computation? | Nature Reviews Drug Discovery*. https://www.nature.com/articles/nrd4581 (accessed 2025-01-14).

(109) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive Characterization of Cytochrome P450 Isozyme Selectivity across Chemical Libraries. *Nat Biotechnol* **2009**, *27* (11), 1050–1055. https://doi.org/10.1038/nbt.1581.



(110) *WhichCyp: prediction of cytochromes P450 inhibition | Bioinformatics | Oxford Academic*. https://academic.oup.com/bioinformatics/article/29/16/2051/199965?login=true (accessed 2025-01-14).

(111) *XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci400518g (accessed 2025-01-14).

(112) *Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. - Abstract - Europe PMC*. https://europepmc.org/article/med/18426954 (accessed 2025-01-14).

(113) *ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches | Molecular Pharmaceutics*. https://pubs.acs.org/doi/abs/10.1021/acs.molpharmaceut.6b00471 (accessed 2025-01-14).

(114) *Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope | Chemical Research in Toxicology*. https://pubs.acs.org/doi/full/10.1021/acs.chemrestox.5b00465 (accessed 2025-01-14).

(115) Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting Chemically-Induced Skin Reactions. Part I: QSAR Models of Skin Sensitization and Their Application to Identify Potentially Hazardous Compounds. *Toxicology and Applied Pharmacology* **2015**, *284* (2), 262–272. https://doi.org/10.1016/j.taap.2014.12.014.

(116) *ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling | Journal of Cheminformatics*. https://link.springer.com/article/10.1186/s13321-016-0117-7 (accessed 2025-01-14).

(117) *In silico toxicity prediction of chemicals from EPA toxicity database by kernel fusion-based support vector machines - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0169743915001756 (accessed 2025-01-14).

(118) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Research* **2012**, *40* (D1), D400–D412. https://doi.org/10.1093/nar/gkr1132.

(119) Guan, L.; Yang, H.; Cai, Y.; Sun, L.; Di, P.; Li, W.; Liu, G.; Tang, Y. ADMET-Score – a Comprehensive Scoring Function for Evaluation of Chemical Drug-Likeness. *Med. Chem. Commun.* **2019**, *10* (1), 148–157. https://doi.org/10.1039/C8MD00472B.

(120) Banerjee, P.; Dunkel, M.; Kemmler, E.; Preissner, R. SuperCYPsPred—a Web Server for the Prediction of Cytochrome Activity. *Nucleic Acids Research* **2020**, *48* (W1), W580–W585. https://doi.org/10.1093/nar/gkaa166.

(121) Preissner, S.; Kroll, K.; Dunkel, M.; Senger, C.; Goldsobel, G.; Kuzman, D.; Guenther, S.; Winnenburg, R.; Schroeder, M.; Preissner, R. SuperCYP: A Comprehensive Database on Cytochrome P450 Enzymes Including a Tool for Analysis of CYP-Drug Interactions. *Nucleic Acids Research* **2010**, *38* (suppl_1), D237–D243. https://doi.org/10.1093/nar/gkp970.

(122) Hall, L. M.; Hall, L. H.; Kier, L. B. QSAR Modeling of β-Lactam Binding to Human Serum Proteins. *J Comput Aided Mol Des* **2003**, *17* (2), 103–118. https://doi.org/10.1023/A:1025309604656.

(123) *In silico Prediction of Chemical Ames Mutagenicity | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci300400a (accessed 2025-01-14).

(124) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J Comput Aided Mol Des* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.

(125) *Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation | Bioinformatics | Oxford Academic*. https://academic.oup.com/bioinformatics/article/38/10/2863/6555376 (accessed 2025-01-13).

(126) de Sá, A. G. C.; Long, Y.; Portelli, S.; Pires, D. E. V.; Ascher, D. B. toxCSM: Comprehensive Prediction of Small Molecule Toxicity Profiles. *Briefings in Bioinformatics* **2022**, *23* (5), bbac337. https://doi.org/10.1093/bib/bbac337.

(127) *Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species | Chemical Research in Toxicology*. https://pubs.acs.org/doi/abs/10.1021/tx900326k (accessed 2025-01-14).

(128) *In silico prediction of chemical respiratory toxicity via machine learning - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S2468111321000037 (accessed 2025-01-14).

(129) *In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.8b00551 (accessed 2025-01-14).

(130) *In silico prediction of chemical genotoxicity using machine learning methods and structural alerts | Toxicology Research | Oxford Academic*. https://academic.oup.com/toxres/article/7/2/211/5559285?login=true (accessed 2025-01-14).

(131) *In Silico Estimation of Chemical Carcinogenicity with Binary and Ternary Classification Methods - Li - 2015 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201400127 (accessed 2025-01-14).

(132) *In silico prediction of terrestrial and aquatic toxicities for organic chemicals*. http://www.nyxxb.cn/en/article/id/20100418 (accessed 2025-01-14).

(133) *In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0045653510013500 (accessed 2025-01-14).

(134) *In Silico Assessment of Chemical Biodegradability | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci200622d (accessed 2025-01-14).

(135) *Quantitative Structure–Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure | Chemical Research in Toxicology*. https://pubs.acs.org/doi/abs/10.1021/tx900189p (accessed 2025-01-14).

(136) *In silico prediction of serious eye irritation or corrosion potential of chemicals - RSC Advances (RSC Publishing) DOI:10.1039/C6RA25267B*. https://pubs.rsc.org/en/content/articlehtml/2017/ra/c6ra25267b (accessed 2025-01-14).

(137) *Development and Comparison of hERG Blocker Classifiers: Assessment on Different Datasets Yields Markedly Different Results - Marchese Robinson - 2011 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201000159 (accessed 2025-01-14).

(138) *In silico prediction of chemical toxicity on avian species using chemical category approaches - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0045653514014003 (accessed 2025-01-14).

(139) *In silico estimation of chemical aquatic toxicity on crustaceans using chemical category methods - Environmental Science: Processes & Impacts (RSC Publishing)*. https://pubs.rsc.org/en/content/articlelanding/2018/em/c8em00220g/unauth (accessed 2025-01-14).

(140) *Mechanistic insights from comparing intrinsic clearance values between human liver microsomes and hepatocytes to guide drug design - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0223523412003959?casa_token=MNBb1xkUT_YAAAAA:3s1kw30Nlv_tSrXYURqsz4j7it5wswWolxgkAPOg_Ts-FqVv3jLnyUFRQu9rNEdyV2FyyXL0VxI (accessed 2025-01-14).

(141) *Selecting Relevant Descriptors for Classification by Bayesian Estimates: A Comparison with Decision Trees and Support Vector Machines Approaches for Disparate Data Sets - Carbon-Mangels - 2011 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201100069 (accessed 2025-01-13).

(142) *Validating ADME QSAR Models Using Marketed Drugs - Vishal Siramshetty, Jordan Williams, Đắc-Trung Nguyễn, Jorge Neyra, Noel Southall, Ewy Mathé, Xin Xu, Pranav Shah, 2021*. https://journals.sagepub.com/doi/full/10.1177/24725552211017520 (accessed 2025-01-14).

(143) *ADME Evaluation in Drug Discovery. 7. Prediction of Oral Absorption by Correlation and Classification | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/10.1021/ci600343x (accessed 2025-01-13).

(144) *Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0731708508001738?via%3Dihub (accessed 2025-01-14).

(145) *Document: Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds (CHEMBL3301361)*. https://www.ebi.ac.uk/explore/document/CHEMBL3301361 (accessed 2025-01-14).

(146) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds. *Sci Data* **2019**, *6* (1), 143. https://doi.org/10.1038/s41597-019-0151-1.

(147) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J Comput Aided Mol Des* **2014**, *28* (7), 711–720. https://doi.org/10.1007/s10822-014-9747-x.

(148) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.* **2012**, *52* (6), 1686–1697. https://doi.org/10.1021/ci300124c.

(149) *In Silico Prediction of Volume of Distribution in Humans. Extensive Data Set and the Exploration of Linear and Nonlinear Methods Coupled with Molecular Interaction Fields Descriptors | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.6b00044?casa_token=SQhDZjizB08AAAAA:zrP4aW5361rH880nm_vVroPqBxjZU8wy0Yb6a0cM7Qvap1aeLQSo0KS3n79wIW0xa64n3Z3V1-lBaQpN (accessed 2025-01-14).

(150) *In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach - Pham The - 2011 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201000118 (accessed 2025-01-14).

(151) *Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0968089612002350 (accessed 2025-01-14).

(152) *Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches | Pharmaceutical Research*. https://link.springer.com/article/10.1007/s11095-013-1222-1 (accessed 2025-01-14).

(153) *Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/ci700096r (accessed 2025-01-14).

(154) *QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure−Information Representation | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/abs/10.1021/jm051245v (accessed 2025-01-14).

(155) *In Silico Prediction of Compounds Binding to Human Plasma Proteins by QSAR Models - Sun - 2018 - ChemMedChem - Wiley Online Library*. https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/cmdc.201700582 (accessed 2025-01-14).

(156) *CypReact: A Software Tool for in Silico Reactant Prediction for Human Cytochrome P450 Enzymes | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.8b00035 (accessed 2025-01-14).

(157) *Chemical rules for optimization of chemical mutagenicity via matched molecular pairs analysis and machine learning methods | Journal of Cheminformatics*. https://link.springer.com/article/10.1186/s13321-023-00707-x (accessed 2025-01-14).

(158) *In silico prediction of drug-induced liver injury with a complementary integration strategy based on hybrid representation - Gu - 2023 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202200284 (accessed 2025-01-14).

(159) *In silico prediction of mitochondrial toxicity of chemicals using machine learning methods - Zhao - 2021 - Journal of Applied Toxicology - Wiley Online Library*. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jat.4141 (accessed 2025-01-14).

(160) *Prediction of the skin sensitising potential and potency of compounds via mechanism-based binary and ternary classification models - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0887233318306337 (accessed 2025-01-14).

(161) *In silico prediction of chemical reproductive toxicity using machine learning - Jiang - 2019 - Journal of Applied Toxicology - Wiley Online Library*. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jat.3772 (accessed 2025-01-14).

(162) *In silico prediction of potential drug-induced nephrotoxicity with machine learning methods - Gong - 2022 - Journal of Applied Toxicology - Wiley Online Library*. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jat.4331 (accessed 2025-01-14).

(163) *In silico prediction of chemical neurotoxicity using machine learning | Toxicology Research | Oxford Academic*. https://academic.oup.com/toxres/article/9/3/164/5826011?login=true (accessed 2025-01-14).

(164) *In silico prediction of hERG potassium channel blockage by chemical category approaches | Toxicology Research | Oxford Academic*. https://academic.oup.com/toxres/article/5/2/570/5568649?login=true (accessed 2025-01-14).

(165) *In silico prediction of chemical aquatic toxicity with chemical category approaches and substructural alerts | Toxicology Research | Oxford Academic*. https://academic.oup.com/toxres/article/4/2/452/5545345?login=true (accessed 2025-01-14).

(166) *In silico prediction of chemical acute contact toxicity on honey bees via machine learning methods - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S088723332100014X (accessed 2025-01-14).

(167) *QSPR model for Caco-2 cell permeability prediction using a combination of HQPSO and dual-RBF neural network - RSC Advances (RSC Publishing) DOI:10.1039/D0RA08209K*. https://pubs.rsc.org/en/content/articlehtml/2020/ra/d0ra08209k (accessed 2025-01-14).

(168) Wang, Z.; Yang, H.; Wu, Z.; Wang, T.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods. *ChemMedChem* **2018**, *13* (20), 2189–2201. https://doi.org/10.1002/cmdc.201800533.

(169) Myung, Y.; de Sá, A. G. C.; Ascher, D. B. Deep-PK: Deep Learning for Small Molecule Pharmacokinetic and Toxicity Prediction. *Nucleic Acids Research* **2024**, *52* (W1), W469–W475. https://doi.org/10.1093/nar/gkae254.

(170) Banerjee, P.; Kemmler, E.; Dunkel, M.; Preissner, R. ProTox 3.0: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Research* **2024**, *52* (W1), W513–W520. https://doi.org/10.1093/nar/gkae303.

(171) *Estimating Screening-Level Organic Chemical Half-Lives in Humans | Environmental Science & Technology*. https://pubs.acs.org/doi/abs/10.1021/es4029414 (accessed 2025-01-14).

(172) *In Silico Prediction of Human Intravenous Pharmacokinetic Parameters with Improved Accuracy | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b00300 (accessed 2025-01-14).

(173) *Improved GNNs for Log D7.4 Prediction by Transferring Knowledge from Low-Fidelity Data | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.2c01564 (accessed 2025-01-14).

(174) *ChemBCPP: A freely available web server for calculating commonly used physicochemical properties - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0169743917304604 (accessed 2025-01-14).

(175) *MF-SuP-pKa: Multi-fidelity modeling with subgraph pooling mechanism for pKa prediction - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S2211383522004622 (accessed 2025-01-14).

(176) *Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism | Journal of Chemical Information and Modeling*. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.2c00038 (accessed 2025-01-14).

(177) Schieferdecker, S.; Rottach, F.; Vock, E. In Silico Prediction of Oral Acute Rodent Toxicity Using Consensus Machine Learning. *J. Chem. Inf. Model.* **2024**, *64* (8), 3114–3122. https://doi.org/10.1021/acs.jcim.4c00056.

(178) Wassermann, A. M.; Bajorath, J. BindingDB and ChEMBL: Online Compound Databases for Drug Discovery. *Expert Opinion on Drug Discovery* **2011**, *6* (7), 683–687. https://doi.org/10.1517/17460441.2011.579100.

(179) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Combinatorial Chemistry & High Throughput Screening* **2001**, *4* (8), 719–725. https://doi.org/10.2174/1386207013330670.

(180) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. Á. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14* (10), 1998. https://doi.org/10.3390/pharmaceutics14101998.

(181) Hastie, T.; Friedman, J.; Tibshirani, R. Model Assessment and Selection. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Friedman, J., Tibshirani, R., Eds.; Springer: New York, NY, 2001; pp 193–224. https://doi.org/10.1007/978-0-387-21606-5_7.

(182) Adamczyk, J.; Ludynia, P. Scikit-Fingerprints: Easy and Efficient Computation of Molecular Fingerprints in Python. arXiv December 12, 2024. https://doi.org/10.48550/arXiv.2407.13291.

(183) *Optuna | Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://dl.acm.org/doi/10.1145/3292500.3330701 (accessed 2025-01-13).

(184) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29* (2), 93–104. https://doi.org/10.1145/335191.335388.

(185) Venkatraman, V. FP-ADMET: A Compendium of Fingerprint-Based ADMET Prediction Models. *Journal of Cheminformatics* **2021**, *13* (1), 75. https://doi.org/10.1186/s13321-021-00557-5.

(186) Adam, S. P.; Alexandropoulos, S.-A. N.; Pardalos, P. M.; Vrahatis, M. N. No Free Lunch Theorem: A Review. In *Approximation and Optimization : Algorithms, Complexity and Applications*; Demetriou, I. C., Pardalos, P. M., Eds.; Springer International Publishing: Cham, 2019; pp 57–82. https://doi.org/10.1007/978-3-030-12767-1_5.

(187) Wang, Y.; Zhao, Z.; Dai, B.; Fifty, C.; Lin, D.; Hong, L.; Wei, L.; Chi, E. H. Can Small Heads Help? Understanding and Improving Multi-Task Generalization. In *Proceedings of the ACM Web Conference 2022*; ACM: Virtual Event, Lyon France, 2022; pp 3009–3019. https://doi.org/10.1145/3485447.3512021.

(188) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*. https://doi.org/10.3389/fenvs.2015.00080.

(189) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23* (6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

(190) Ash, J. R.; Wognum, C.; Rodríguez-Pérez, R.; Aldeghi, M.; Cheng, A. C.; Clevert, D.-A.; Engkvist, O.; Fang, C.; Price, D. J.; Hughes-Oliver, J. M.; Walters, W. P. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. ChemRxiv November 7, 2024. https://doi.org/10.26434/chemrxiv-2024-6dbwv-v2.

(191) Göller, A. H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's *in Silico* ADMET Platform: A Journey of Machine Learning over the Past Two Decades. *Drug Discovery Today* **2020**, *25* (9), 1702–1709. https://doi.org/10.1016/j.drudis.2020.07.001.

Supplementary informations

In Silico eADMET: current situation and novel profilers

Pierre Llompart[1,2], Claire Minoletti[2], Gilles Marcou[1,†], Alexandre Varnek[1]

[1]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg, France

[2]Integrated Drug Discovery, Molecular Design Sciences, Sanofi, Vitry-sur-Seine, France

Table S1: **Standardized units of the modelled ADMET and biological activity endpoints.**

Endpoints	Units
Bioactivities (IC50, EC50, Ki, Kd, GI50, TC50)	-log10(mol/L)
Solubility	log10(mol/L)
Apparent permeability	log10(meter/seconds)
Total/Renal clearance	log10(mL/min/kg)
Microsomal clearance	log10(mL/min/10^6 cells)
Half-life	log10(hours)
Dose (LD50, TD50)	log10(mg/kg)
Volume of Distribution at steady state	log10(L/kg)
Plasma Protein Binding	%
Microsomal stability	%
MRTD	log10(mol/kg)
LogBB	-
Hydration Free Energy	log10(kcal/mol)
Efflux ratio	-
Recovery	%
LogP	-
$LogD_{7.4}$	-

Table S2 : Description of the conditions blueprint used to filter experimental data.

Endpoint	Assay	Conditions
Efflux Ratio	Caco-2	Passage Number: 20-100 BSA: 0 % Concentration: 1-10 µM
	MDCK-MDR1	Passage Number: 20-100 BSA: 0 % Concentration: 1-10 µM
Apparent Permeability	Caco-2	Passage Number: 35-72 BSA A/B: 0 % Concentration: 1-10 µM
	MDCK-LE MDCK-MDR1	Passage Number: 35-72 BSA A/B: 0 % Concentration: 1-10 µM
	PAMPA	pH A/B: 7.4/7.4 Concentration: 1-10 µM
	PAMPA-BBB	Brain lipid extract membrane: PC, PE, PS, PBLE, PVDF DMSO 0-5%
Inhibition	pIC50 P-gP	MDCK-MDR1 cell line Rhodamin 123/Digoxin substrate
Physico-Chemical	Apparent Solubility ($LogS_{app}$)	Phosphate buffer 0.1 M. At 25±5°Celsius and pH 7.4±1 log.
	Kinetic Solubility ($LogS_{kin}$)	Concentration: 10 mM DMSO Temperature 25°Celsius PBS Buffer pH 7.4
	Water Solubility ($LogS_{W}$)	Shake-Flask/Column elution Pure water 25±5°Celsius pH 7±1 log.
	LogP	Temperature: 22-25°C HPLC / ShakeFlask pH 7±1 log Pure Water
	$LogD_{7.4}$	Temperature: 22-25°C RP-HPLC / ShakeFlask pH 7.4 PBS
	Hydration Free Energy (HFE)	Alchemical free energy
Distribution	Plasma Protein Binding (PPB)	Rat Matrix: Plasma
	Ratio Brain/Blood (LogBB)	Rat Intravenous administration
	Volume distribution at steady state (VDss)	Intravenous bolus injection Sampling times over 24-48h LC-MS/MS
Elimination	Half-life plasma	Human or Rat plasma Temperature: 37°C LC-MS/MS
	Half-life microsomal	Human or Rat plasma Temperature: 37°C LC-MS/MS
	Clearance Microsomal	Mouse or Rat Liver microsmoes Cofactor regeneration system (NADPH) Temperature: 37°C Incubation time: 0-60 min Buffer pH: 7.4
	Clearance plasma	Rat or Human LC-MS/MS
	Stability microsomal	Mouse or Rat Liver microsomes

		Compound stability measured under NADPH-dependent oxidation Buffer pH 7.4
	Clearance total	Rat (Sprague Dawley) Intravenous injection LC-MS/MS quantification
Metabolism	CYP1A1	Enzyme activity assay Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS
	CYP1A2	Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Phenacetin
	CYP2B6	Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Bupropion
	CYP2C9	Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Diclofenac
	CYP2C19	Enzyme activity assay Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Omeprazole
	CYP2D6	Enzyme activity assay Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Dextromethorphan
	CYP3A4	Enzyme activity assay Temperature: 37°C Buffer pH: 7.4 Fluorometric or LC-MS/MS Substrate: Midazolam or testosterone
	Stability microsomal	Mouse or Rat Liver microsomes Compound stability measured under NADPH-dependent oxidation Buffer pH 7.4
Toxicity	hERG pIC50	Manual/Automatised patch-clamp or fluorescence-based assays Non engineered hERG for HEK293 and CHO cells Incubation at 37°C Buffer pH 7.4
	Cell proliferation IC50	HEK293 MTT or SRB assay Incubation 24-72h DMSO <= 0.1% Temperature 37°C CO2 5%
	MRTD	Derived from human clinical studies or FDA data; calculated based on the Maximum Recommended Therapeutic Dose (mg/kg/day) normalized to body weight.
	LD50	Acute toxicity test Oral gavage
	GI50	Measured using tumor-derived cell lines; GI50 defined as the concentration inhibiting growth by 50%; assay types include MTT, SRB, or ATP-based assays in 96-well plates; incubation: 24–72h; 37°C; CO_2: 5%.
	TD50	Chronic toxicity study Dose-response over months Wistar rats

Table S3: **Molecular descriptors used for the models' preparation and their optimization domains.**

Descriptors	Hyperparameters	Optimization domain
Avalon	fp_size count	256, 512, 1024, 2048 True, False
ECFP	fp_size radius include_chirality count	1024, 2048, 4096 2, 3 True, False True, False
AtomPairs	fp_size scale_by_hac include_chirality count	1024, 2048, 4096 True, False True, False True, False
PubChem	count	True, False
MACCS	count	True, False
RDKit2D	-	-
EState	type	sum, bit, count

Table S4: **Predictive methods used and their optimization domains.**

Methods	Hyperparameters	Optimization domain
XGBoost	n_estimators learning_rate max_depth subsample	100-300 [0.01-0.3] log = True 3-15 0.6-1.0
K-Nearest Neighbors	n_neighbors weights algorithm	1-30 uniform, distance ball_tree, kd_tree
KernelRidge	alpha kernel gamma	(0.01-10) log = True linear, rbf 0.001-0.1 log = True
Support Vector Machine	C kernel gamma	0.1-5.0 log = True linear, rbf 0.001-0.1 log = True
Random Forest	n_estimators max_depth max_features	100-300 5-20 0.5-1.0
GNN	n_layers layer_size	1, 3 300, 1000, 2500

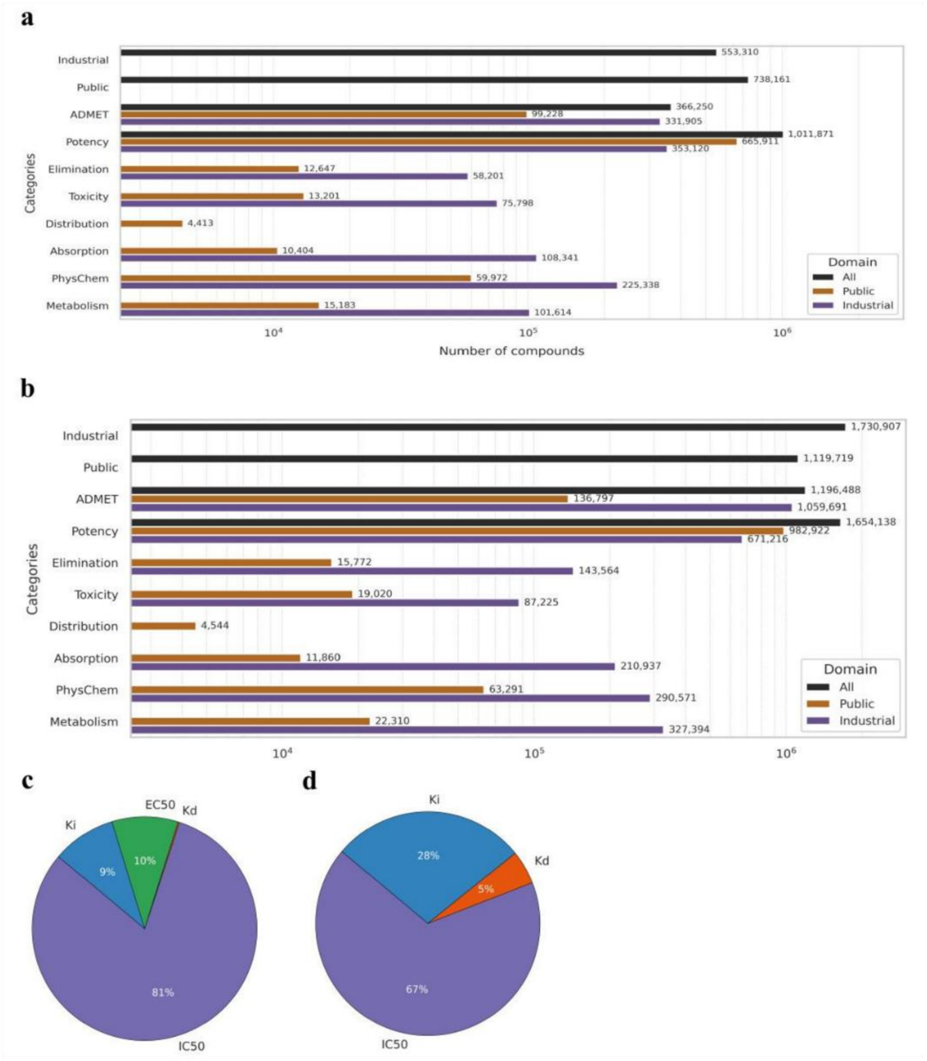


Figure S1: **Presentation of the data of OneADMET and the range of performances.** (**a**) Count plot of the number of unique compounds per domain or/and class. (**b**) Count plot of the number of measurements per domain or/and class. (**c**) Distribution of the industrial potency datasets by bioactivity types considering IC50 (purple), Ki (blue), EC50 (blue), and Kd (orange). (**d**) Distribution of the public potency datasets by bioactivity types considering IC50 (purple), Ki (blue), EC50 (blue), and Kd (orange).

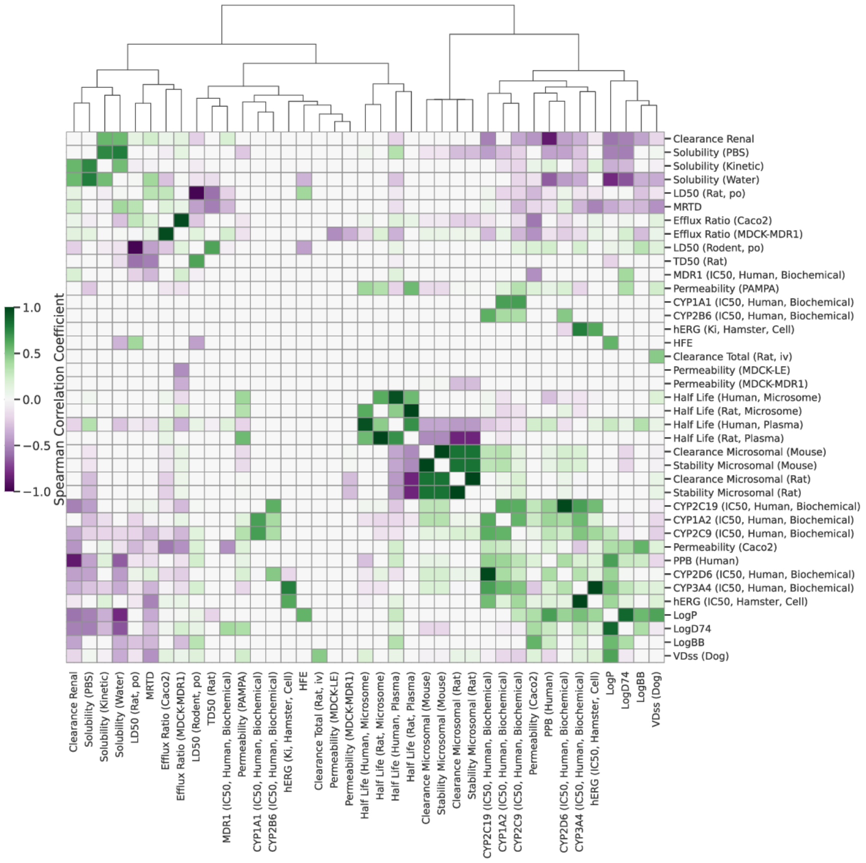


Figure S2: **Hierarchical heatmap of the Spearman correlation between the public ADMET endpoints**. Correlation is only computed if at least 30 compounds are shared between the datasets. Correlations are colored in green and anti-correlations are colored in purple. Datasets sharing less than 30 compounds or with a correlation of 0 are colored in grey.

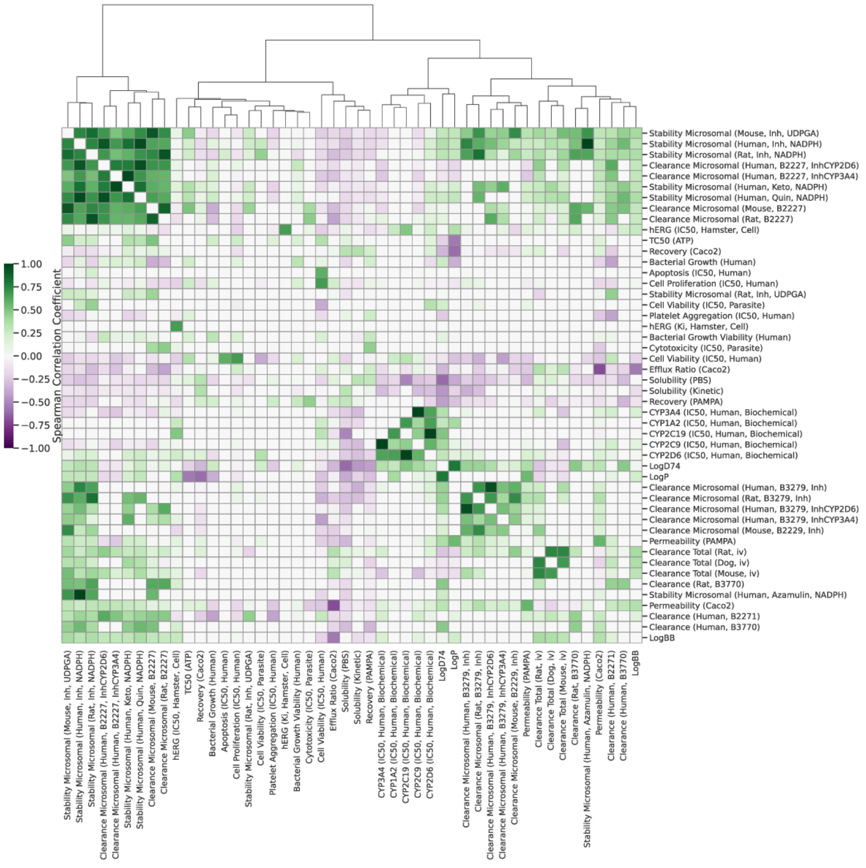


Figure S3: **Hierarchical heatmap of the Spearman correlation between the industrial ADMET endpoints.** Correlation is only computed if at least 30 compounds are shared between the datasets. Correlations are colored in green and anti-correlations are colored in purple. Datasets sharing less than 30 compounds or with a correlation of 0 are colored in grey.

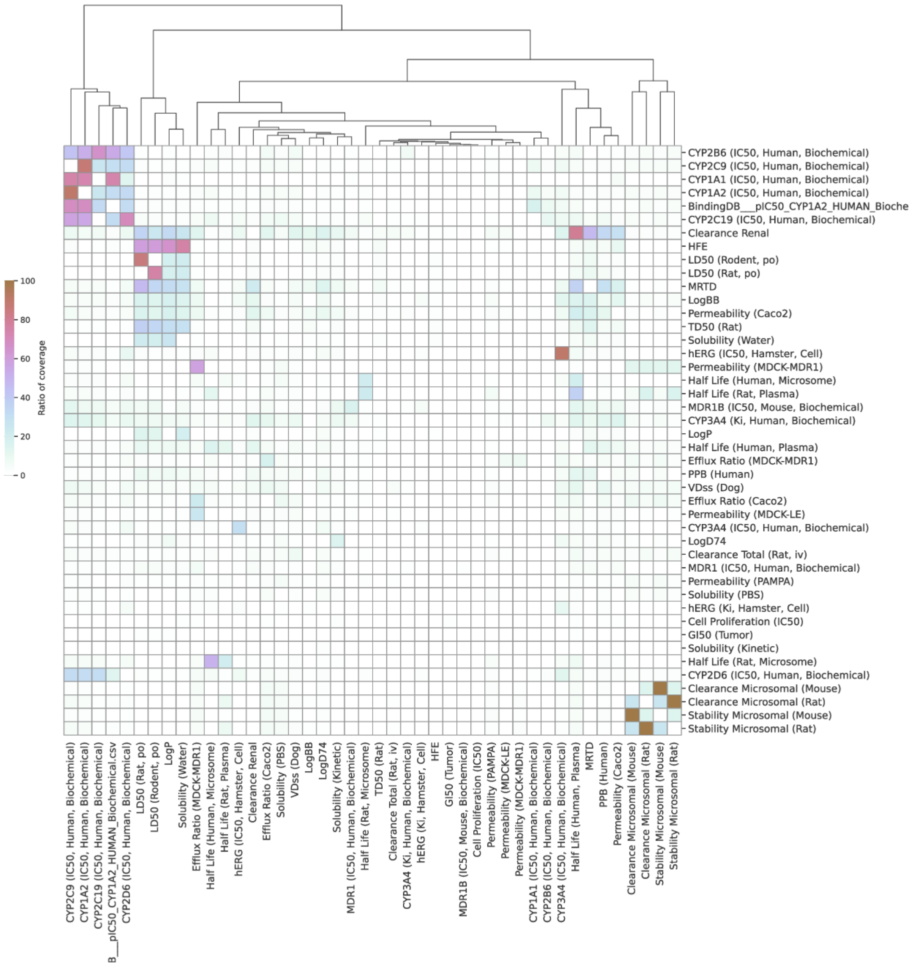


Figure S4: **Hierarchical heatmap of the dataset coverage of compounds between the public ADMET endpoints.** Coverage percentage from dataset A to B is computed as the number of compounds from A found in B over the number of compounds in B. High coverage are colored from pink to brown and low coverage are colored from blue to white, if null. The heatmap present compounds shared from A to B, or B to A, explaining the lack of symmetry.

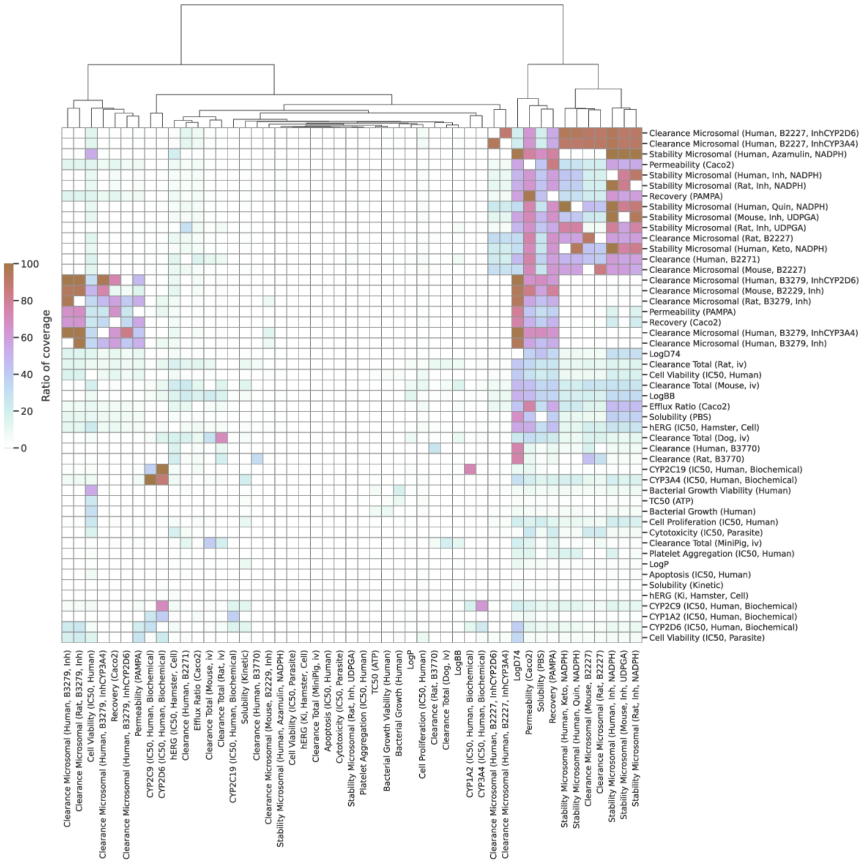
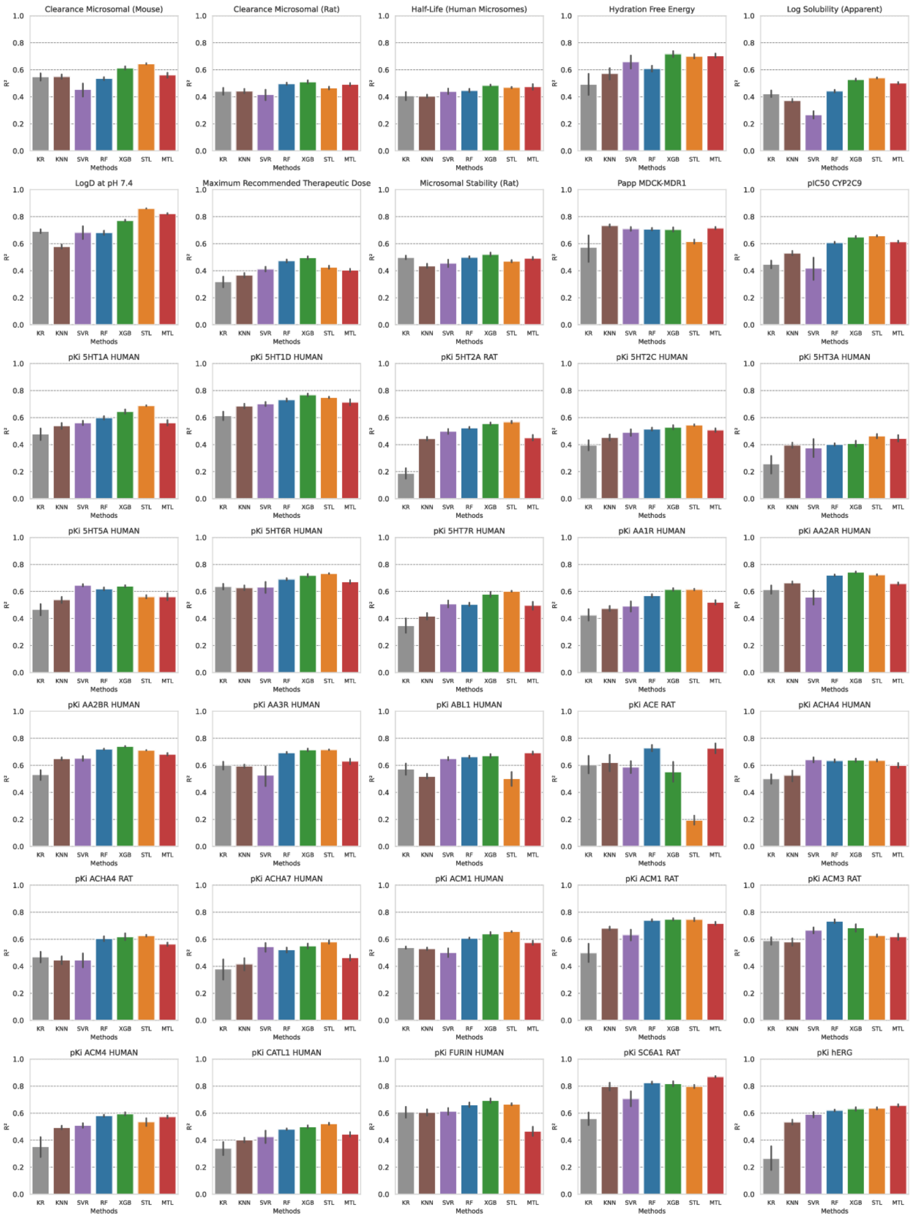


Figure S5: **Hierarchical heatmap of the dataset coverage of compounds between the industrial ADMET endpoints.** Coverage percentage from dataset A to B is computed as the number of compounds from A found in B over the number of compounds in B. High coverage are colored from pink to brown and low coverage are colored from blue to white, if null. The heatmap present compounds shared from A to B, or B to A, explaining the lack of symmetry.


Clearance Microsomal (Mouse)
Clearance Microsomal (Rat)
Half-Life (Human Microsomes)
Hydration Free Energy
Log Solubility (Apparent)
LogD at pH 7.4
Maximum Recommended Therapeutic Dose
Microsomal Stability (Rat)
Papp MDCK-MDR1
pIC50 CYP2C9
pKi 5HT1A HUMAN
pKi 5HT1D HUMAN
pKi 5HT2A RAT
pKi 5HT2C HUMAN
pKi 5HT3A HUMAN
pKi 5HT5A HUMAN
pKi 5HT6R HUMAN
pKi 5HT7R HUMAN
pKi AA1R HUMAN
pKi AA2AR HUMAN
pKi AA2BR HUMAN
pKi AA3R HUMAN
pKi ABL1 HUMAN
pKi ACE RAT
pKi ACHA4 HUMAN
pKi ACHA4 RAT
pKi ACHA7 HUMAN
pKi ACM1 HUMAN
pKi ACM1 RAT
pKi ACM3 RAT
pKi ACM4 HUMAN
pKi CATL1 HUMAN
pKi FURIN HUMAN
pKi SC6A1 RAT
pKi hERG
R²
0.0
0.2
0.4
0.6
0.8
1.0
KR
KNN
SVR
RF
XGB
STL
MTL
Methods

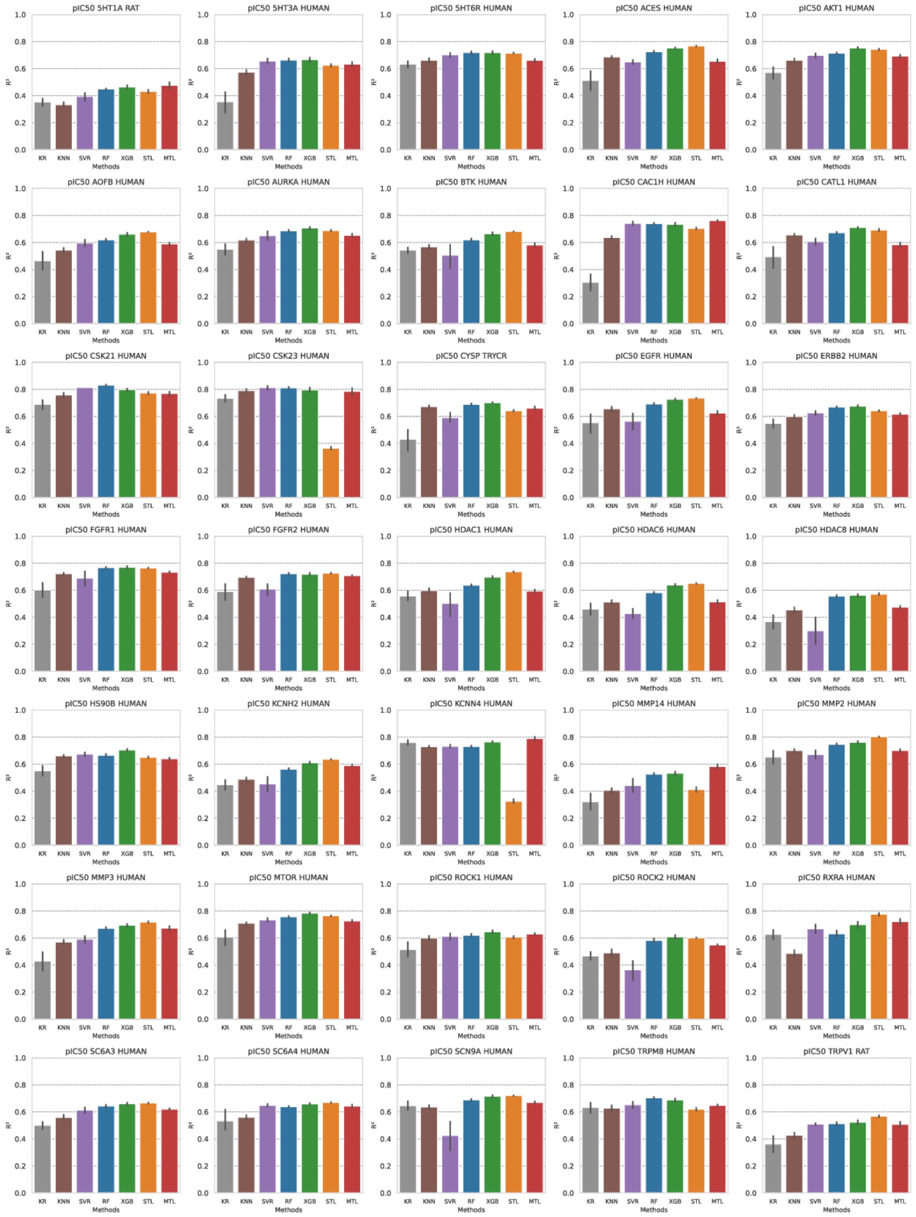


Figure S6: **Barplot of the R^2 performances distribution per method for a subset of 75 public endpoints.** Methods are colored as KR (Kernel Ridge) in grey, KNN (k-nearest neighbors) in brown, SVR (Support Vector Regression) in purple, RF (Random Forest) in blue, XGB (XGBoost) in green, STL (Single-Task Graph Neural Network) in orange, MTL (Multi-Task Graph Neural Network) in red.

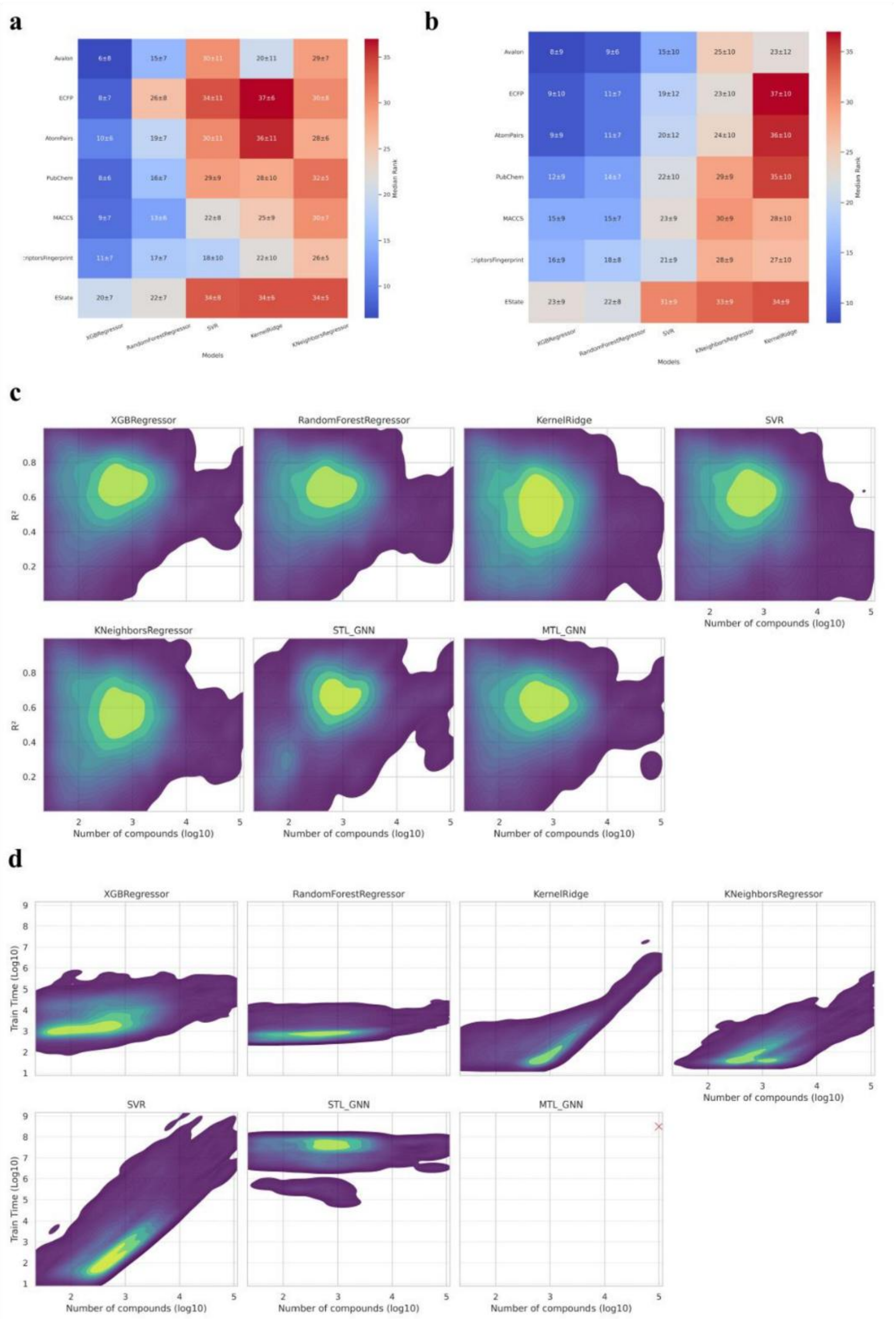


Figure S7: **Comparison the methods per R^2 performances, descriptors, and time to train the models.** (**a**) Density distribution of the performances rank of the models on the test set in function of the logarithm 10 of the size of the dataset.

Distributions are colored from purple (low density) to yellow (high density). (**b**) Heatmap of the mean rank +- the std for the combination of method and descriptors for the ADMETs endpoints. Low ranks are colored in blue and high rank in red. (**c**) Heatmap of the mean rank +- the std for the combination of method and descriptors for the potency endpoints. Low ranks are colored in blue and high rank in red. (**d**) Density distribution of the time to train the models in function of the logarithm 10 of the size of the dataset. Distributions are colored from purple (low density) to yellow (high density).

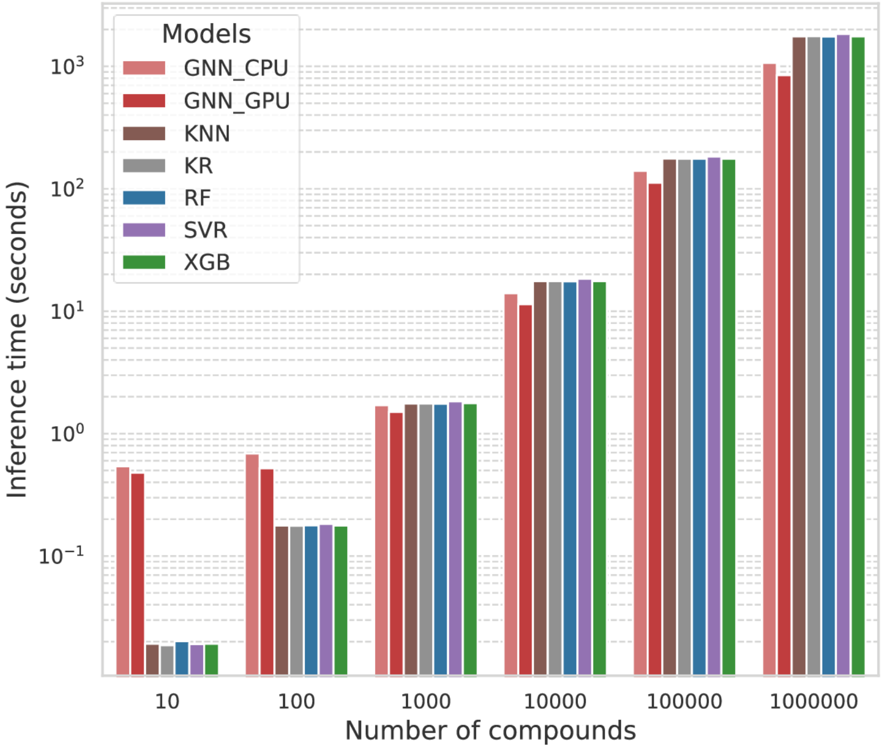


Figure S8: **Inference time of each predictive method against the number of compounds.** The time for inference considers the computation of the descriptors (ECFP2 with 2048 bits in the case of standard ML) and the prediction time.

Outline

In this study we show that this multi-task framework rivals or surpasses conventional single-task models while simplifying large-scale deployment. We also propose a reference web service for ADMET and bioactivity predictions, giving researchers a one-stop solution to flag metabolic liabilities and toxicological red flags before incurring animal studies or clinical trials. Both the OneADMET dataset and the MTL model are released under an open-source license, offering a valuable resource for robust, cost-effective, and ethically responsible drug discovery.

Chapter 7. Collective Intelligence

7.1. Industrial Application

Introduction

In 350 BCE, Aristotle proposed that groups often outperform lone experts when opinions are aggregated. This idea resurfaced in 1907 when Francis Galton observed that crowds at a livestock fair accurately guessed an ox's weight, surpassing individual estimates.[171] In the 1950s, the RAND Corporation formalized similar principles with the Delphi Method, using anonymous expert feedback in iterative rounds to predict events and reach consensus.[172] By 1997, Pierre Lévy had coined the term "collective intelligence" in his book *Collective Intelligence: Mankind's Emerging World in Cyberspace*, stressing how digital platforms could amplify human collaboration.[173]

Advances in technology accelerated the spread of collective intelligence. Wikipedia's founding in 2001 demonstrated that cooperative editing could produce and maintain the world's largest encyclopedia. In 2004, James Surowiecki's *The Wisdom of Crowds* further popularized the concept, using real-world examples to illustrate that organized groups can solve problems more effectively than isolated specialists.[174]

Main Terminology

Wisdom of the crowd refers to how aggregated judgments from diverse, independent contributors can surpass individual expert decisions.

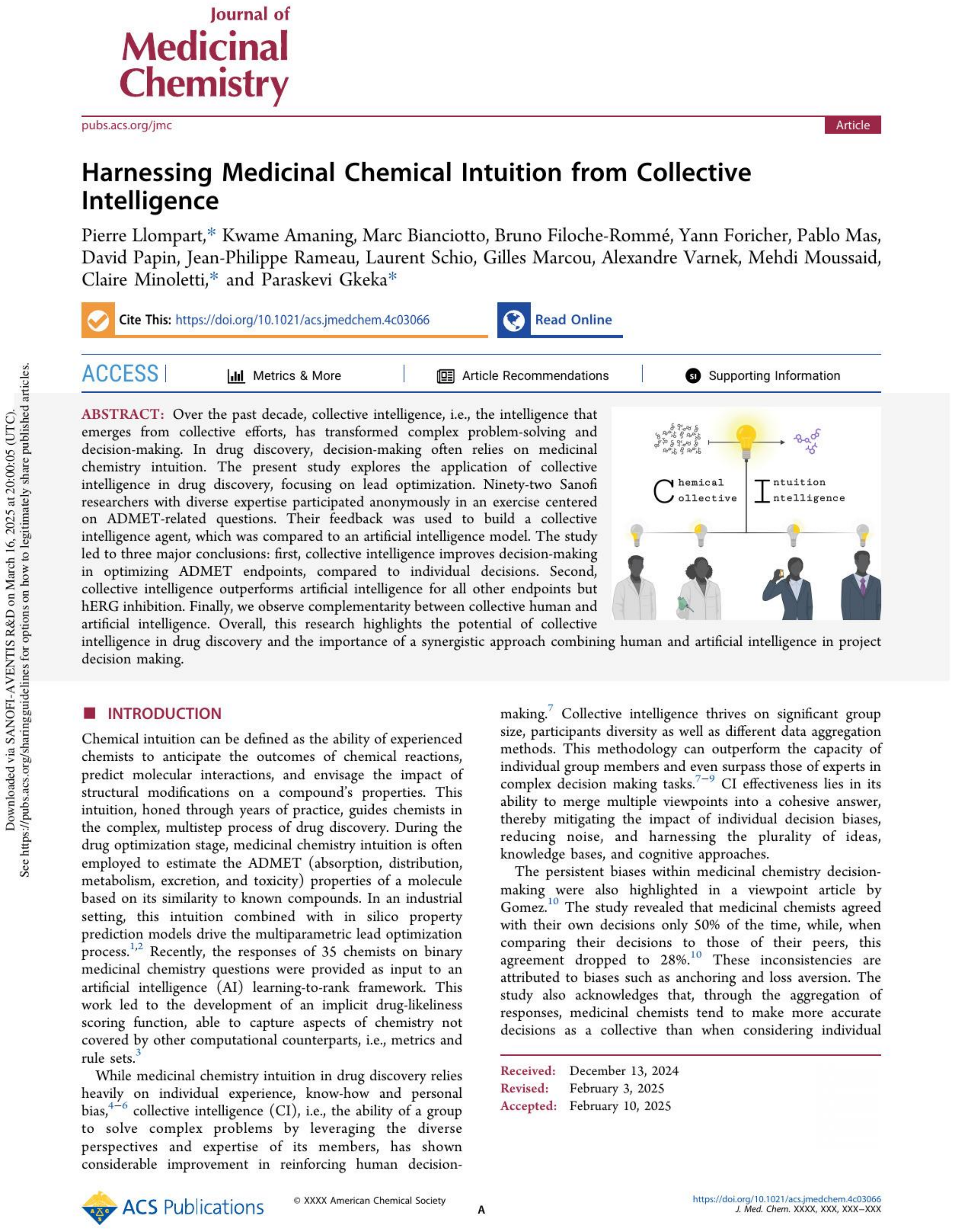
Digital swarm and swarm intelligence describes decentralized and self-organizing group behavior modeled on insect colonies, where simple actions at the individual level collectively solve complex problems.

Peer production is a collaborative model in which volunteers (online) jointly create and refine content or products, typified by open-source software projects.

The Good Judgment Project in the early 2010s used tournaments to show that pooled forecasts often outperformed top analysts in predicting geopolitical events.[175] More recently, "digital swarms" and AI-assisted collaboration platforms have refined how groups share knowledge, distill insights, and reach consensus in real time.

Drug discovery has become a vital testing ground for collective intelligence. Since 2008, the Foldit project developed by the Baker Laboratory has relied on crowdsourced puzzle-solving to unravel complex protein structures, proving that laypeople can sometimes provide breakthroughs that elude experts.[176] Modern initiatives incorporate massive genomic databases, robotic HTS, and global discussion forums where researchers exchange insights on candidate molecules. In emergencies such as new pandemics or target-related health concerns, crowdsourced data analysis and AI-driven approaches have the potential to accelerate drug design and testing. To address this, initiatives such as the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA) and the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) have applied large-scale predictive modeling to assess chemical risks. These projects, led by the U.S. Environmental Protection Agency (EPA), have leveraged global expertise, enlisting 25 international research groups to develop computational models for screening over 55,000 chemical structures.[177]

In this chapter, we explore the intersection of collective intelligence and computational modeling in the context of drug discovery, with a specific focus on lead optimization. Building on prior research demonstrating the effectiveness of aggregated expert input, we conducted a study involving 92 Sanofi researchers from diverse scientific backgrounds. Participants provided anonymous feedback on lead compounds, enabling the construction of a collective intelligence agent whose predictive accuracy was then compared to an artificial intelligence model developed in parallel.



Journal of Medicinal Chemistry

pubs.acs.org/jmc Article

Harnessing Medicinal Chemical Intuition from Collective Intelligence

Pierre Llompart,* Kwame Amaning, Marc Bianciotto, Bruno Filoche-Rommé, Yann Foricher, Pablo Mas, David Papin, Jean-Philippe Rameau, Laurent Schio, Gilles Marcou, Alexandre Varnek, Mehdi Moussaid, Claire Minoletti,* and Paraskevi Gkeka*

Cite This: https://doi.org/10.1021/acs.jmedchem.4c03066 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Over the past decade, collective intelligence, i.e., the intelligence that emerges from collective efforts, has transformed complex problem-solving and decision-making. In drug discovery, decision-making often relies on medicinal chemistry intuition. The present study explores the application of collective intelligence in drug discovery, focusing on lead optimization. Ninety-two Sanofi researchers with diverse expertise participated anonymously in an exercise centered on ADMET-related questions. Their feedback was used to build a collective intelligence agent, which was compared to an artificial intelligence model. The study led to three major conclusions: first, collective intelligence improves decision-making in optimizing ADMET endpoints, compared to individual decisions. Second, collective intelligence outperforms artificial intelligence for all other endpoints but hERG inhibition. Finally, we observe complementarity between collective human and artificial intelligence. Overall, this research highlights the potential of collective intelligence in drug discovery and the importance of a synergistic approach combining human and artificial intelligence in project decision making.

INTRODUCTION

Chemical intuition can be defined as the ability of experienced chemists to anticipate the outcomes of chemical reactions, predict molecular interactions, and envisage the impact of structural modifications on a compound's properties. This intuition, honed through years of practice, guides chemists in the complex, multistep process of drug discovery. During the drug optimization stage, medicinal chemistry intuition is often employed to estimate the ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of a molecule based on its similarity to known compounds. In an industrial setting, this intuition combined with in silico property prediction models drive the multiparametric lead optimization process.[1,2] Recently, the responses of 35 chemists on binary medicinal chemistry questions were provided as input to an artificial intelligence (AI) learning-to-rank framework. This work led to the development of an implicit drug-likeliness scoring function, able to capture aspects of chemistry not covered by other computational counterparts, i.e., metrics and rule sets.[3]

While medicinal chemistry intuition in drug discovery relies heavily on individual experience, know-how and personal bias,[4−6] collective intelligence (CI), i.e., the ability of a group to solve complex problems by leveraging the diverse perspectives and expertise of its members, has shown considerable improvement in reinforcing human decision-making.[7] Collective intelligence thrives on significant group size, participants diversity as well as different data aggregation methods. This methodology can outperform the capacity of individual group members and even surpass those of experts in complex decision making tasks.[7−9] CI effectiveness lies in its ability to merge multiple viewpoints into a cohesive answer, thereby mitigating the impact of individual decision biases, reducing noise, and harnessing the plurality of ideas, knowledge bases, and cognitive approaches.

The persistent biases within medicinal chemistry decision-making were also highlighted in a viewpoint article by Gomez.[10] The study revealed that medicinal chemists agreed with their own decisions only 50% of the time, while, when comparing their decisions to those of their peers, this agreement dropped to 28%.[10] These inconsistencies are attributed to biases such as anchoring and loss aversion. The study also acknowledges that, through the aggregation of responses, medicinal chemists tend to make more accurate decisions as a collective than when considering individual

Received: December 13, 2024
Revised: February 3, 2025
Accepted: February 10, 2025

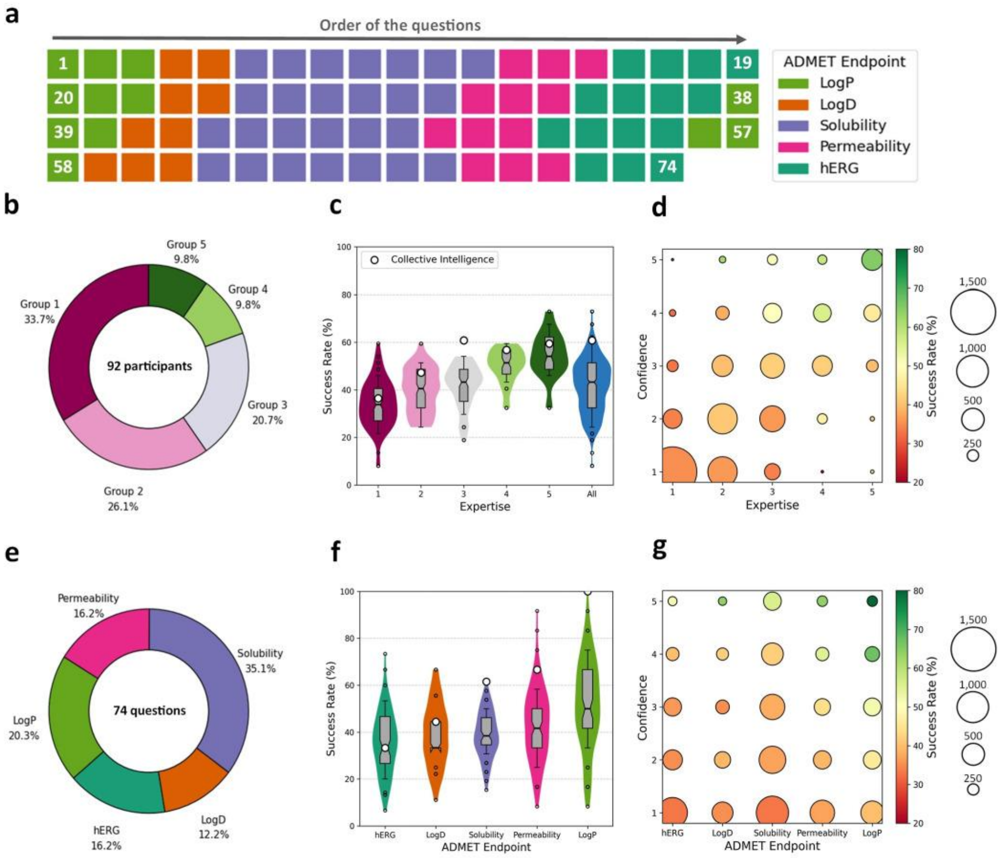
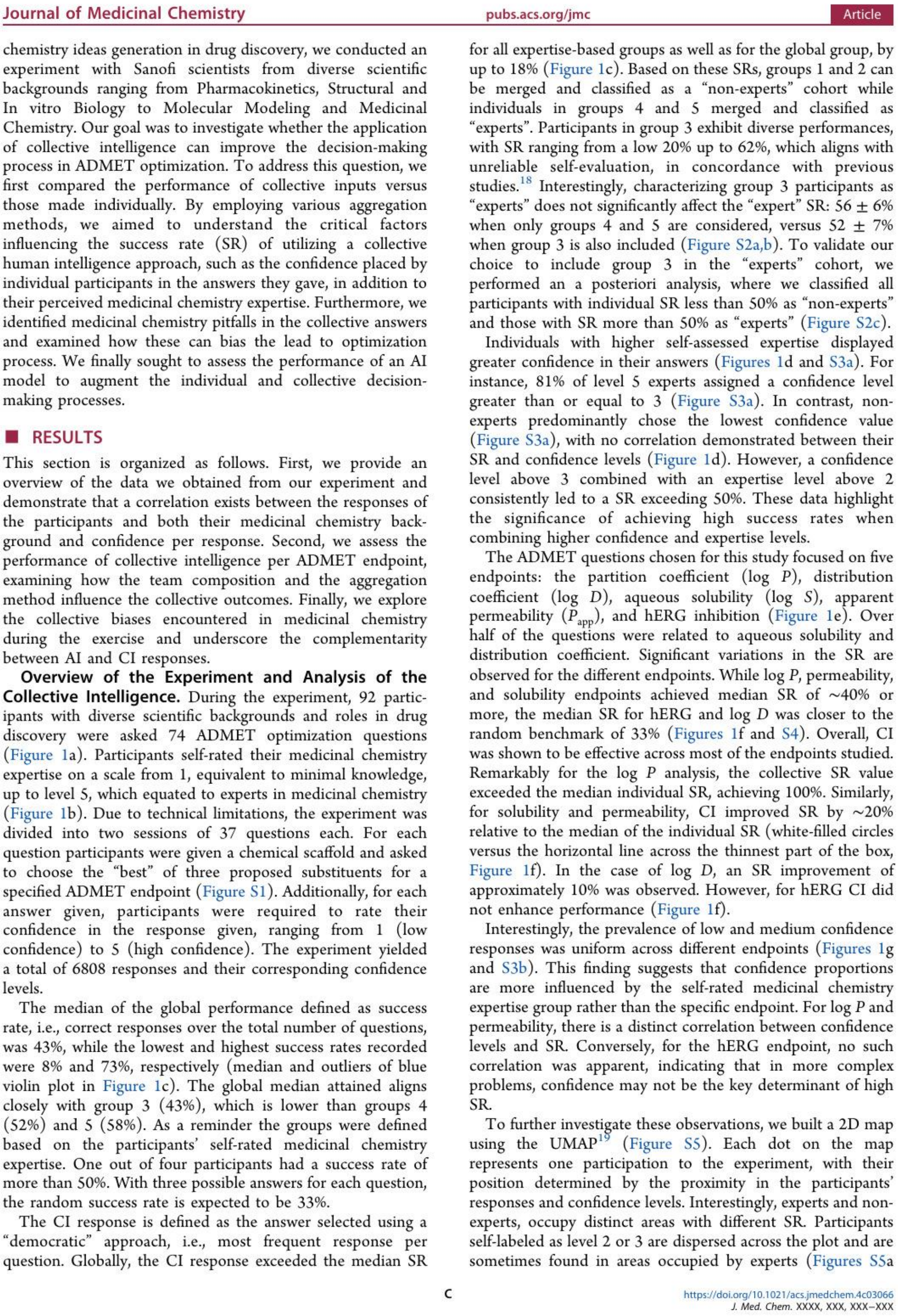


Figure 1. Overview of the collective intelligence experiment and main results. (a) Distribution of the 74 questions in the questionnaire in the form of a waffle plot. Each box corresponds to a question colored by endpoint. The numbers are a guide for the eye that indicate the order of the questions. (b) Participants' partitioning per self-labeled medicinal chemistry expertise, with group 1 corresponding to the least experts and group 5 to the most experts. (c) Violin plots of the SR by expertise level for each group (color code as in 1b) and all participants (blue). The median is shown as a horizontal line across the thinnest part of the boxes. The error bars correspond to the interquartile range. The collective SR (based on the most frequent or "democratic" answers) are shown as white-filled circles, while the outliers are depicted as small circles. (d) Bubble plot of the number of responses (size) and mean SR (color) dependence on medicinal chemistry self-labeling and confidence level per question. The color scale corresponds to the success rate, while the bubble size is indicative of the number of questions for the specific confidence−expertise pair. (e) Endpoint distribution of the 74 questions. (f) Violin plots of the SR for each endpoint (colors as in e). The remaining plot features are as for c. (g) Bubble plot of the number of responses (size) and mean SR (color) dependence on endpoint and confidence level per question. As in subfigure d, the color scale corresponds to the success rate, while the bubble size is indicative of the number of questions for the specific confidence−expertise pair.

responses, reinforcing the value of consolidating diverse inputs for drug discovery projects.

Collective intelligence and chemical intuition have already been combined in the fields of metal−organic frameworks[11] and inorganic chemistry experiments.[12] Nevertheless, the most striking scientific results have been produced for the prediction of biological structures through the Foldit initiative.[13] As of now, Foldit has been applied to other related fields such as small molecule and protein design.[14] Recently, a similar crowdsourcing approach was adopted for RNA design and folding prediction.[15,16] Only a few examples exist where collective intelligence has been applied to the decision making in drug discovery.[3,17] Often drug discovery, and in particular the stage of lead optimization, relies heavily on singular experts or small project teams, however, as illustrated by Hong and Page,[8] groups with diverse perspectives can outperform like-minded experts. In the context of drug design, this diversity, referred to as heterogeneous collective intelligence, could yield a yet more efficient process.

Inspired by the growth of chemistry design environments, such as Torx, LiveDesign, and DesignHub, that provide a robust framework allowing for collaborative feedback and



chemistry ideas generation in drug discovery, we conducted an experiment with Sanofi scientists from diverse scientific backgrounds ranging from Pharmacokinetics, Structural and In vitro Biology to Molecular Modeling and Medicinal Chemistry. Our goal was to investigate whether the application of collective intelligence can improve the decision-making process in ADMET optimization. To address this question, we first compared the performance of collective inputs versus those made individually. By employing various aggregation methods, we aimed to understand the critical factors influencing the success rate (SR) of utilizing a collective human intelligence approach, such as the confidence placed by individual participants in the answers they gave, in addition to their perceived medicinal chemistry expertise. Furthermore, we identified medicinal chemistry pitfalls in the collective answers and examined how these can bias the lead to optimization process. We finally sought to assess the performance of an AI model to augment the individual and collective decision-making processes.

■ RESULTS

This section is organized as follows. First, we provide an overview of the data we obtained from our experiment and demonstrate that a correlation exists between the responses of the participants and both their medicinal chemistry background and confidence per response. Second, we assess the performance of collective intelligence per ADMET endpoint, examining how the team composition and the aggregation method influence the collective outcomes. Finally, we explore the collective biases encountered in medicinal chemistry during the exercise and underscore the complementarity between AI and CI responses.

Overview of the Experiment and Analysis of the Collective Intelligence. During the experiment, 92 participants with diverse scientific backgrounds and roles in drug discovery were asked 74 ADMET optimization questions (Figure 1a). Participants self-rated their medicinal chemistry expertise on a scale from 1, equivalent to minimal knowledge, up to level 5, which equated to experts in medicinal chemistry (Figure 1b). Due to technical limitations, the experiment was divided into two sessions of 37 questions each. For each question participants were given a chemical scaffold and asked to choose the "best" of three proposed substituents for a specified ADMET endpoint (Figure S1). Additionally, for each answer given, participants were required to rate their confidence in the response given, ranging from 1 (low confidence) to 5 (high confidence). The experiment yielded a total of 6808 responses and their corresponding confidence levels.

The median of the global performance defined as success rate, i.e., correct responses over the total number of questions, was 43%, while the lowest and highest success rates recorded were 8% and 73%, respectively (median and outliers of blue violin plot in Figure 1c). The global median attained aligns closely with group 3 (43%), which is lower than groups 4 (52%) and 5 (58%). As a reminder the groups were defined based on the participants' self-rated medicinal chemistry expertise. One out of four participants had a success rate of more than 50%. With three possible answers for each question, the random success rate is expected to be 33%.

The CI response is defined as the answer selected using a "democratic" approach, i.e., most frequent response per question. Globally, the CI response exceeded the median SR for all expertise-based groups as well as for the global group, by up to 18% (Figure 1c). Based on these SRs, groups 1 and 2 can be merged and classified as a "non-experts" cohort while individuals in groups 4 and 5 merged and classified as "experts". Participants in group 3 exhibit diverse performances, with SR ranging from a low 20% up to 62%, which aligns with unreliable self-evaluation, in concordance with previous studies.[18] Interestingly, characterizing group 3 participants as "experts" does not significantly affect the "expert" SR: 56 ± 6% when only groups 4 and 5 are considered, versus 52 ± 7% when group 3 is also included (Figure S2a,b). To validate our choice to include group 3 in the "experts" cohort, we performed an a posteriori analysis, where we classified all participants with individual SR less than 50% as "non-experts" and those with SR more than 50% as "experts" (Figure S2c).

Individuals with higher self-assessed expertise displayed greater confidence in their answers (Figures 1d and S3a). For instance, 81% of level 5 experts assigned a confidence level greater than or equal to 3 (Figure S3a). In contrast, non-experts predominantly chose the lowest confidence value (Figure S3a), with no correlation demonstrated between their SR and confidence levels (Figure 1d). However, a confidence level above 3 combined with an expertise level above 2 consistently led to a SR exceeding 50%. These data highlight the significance of achieving high success rates when combining higher confidence and expertise levels.

The ADMET questions chosen for this study focused on five endpoints: the partition coefficient (log P), distribution coefficient (log D), aqueous solubility (log S), apparent permeability (P_{app}), and hERG inhibition (Figure 1e). Over half of the questions were related to aqueous solubility and distribution coefficient. Significant variations in the SR are observed for the different endpoints. While log P, permeability, and solubility endpoints achieved median SR of ∼40% or more, the median SR for hERG and log D was closer to the random benchmark of 33% (Figures 1f and S4). Overall, CI was shown to be effective across most of the endpoints studied. Remarkably for the log P analysis, the collective SR value exceeded the median individual SR, achieving 100%. Similarly, for solubility and permeability, CI improved SR by ∼20% relative to the median of the individual SR (white-filled circles versus the horizontal line across the thinnest part of the box, Figure 1f). In the case of log D, an SR improvement of approximately 10% was observed. However, for hERG CI did not enhance performance (Figure 1f).

Interestingly, the prevalence of low and medium confidence responses was uniform across different endpoints (Figures 1g and S3b). This finding suggests that confidence proportions are more influenced by the self-rated medicinal chemistry expertise group rather than the specific endpoint. For log P and permeability, there is a distinct correlation between confidence levels and SR. Conversely, for the hERG endpoint, no such correlation was apparent, indicating that in more complex problems, confidence may not be the key determinant of high SR.

To further investigate these observations, we built a 2D map using the UMAP[19] (Figure S5). Each dot on the map represents one participation to the experiment, with their position determined by the proximity in the participants' responses and confidence levels. Interestingly, experts and non-experts, occupy distinct areas with different SR. Participants self-labeled as level 2 or 3 are dispersed across the plot and are sometimes found in areas occupied by experts (Figures S5a

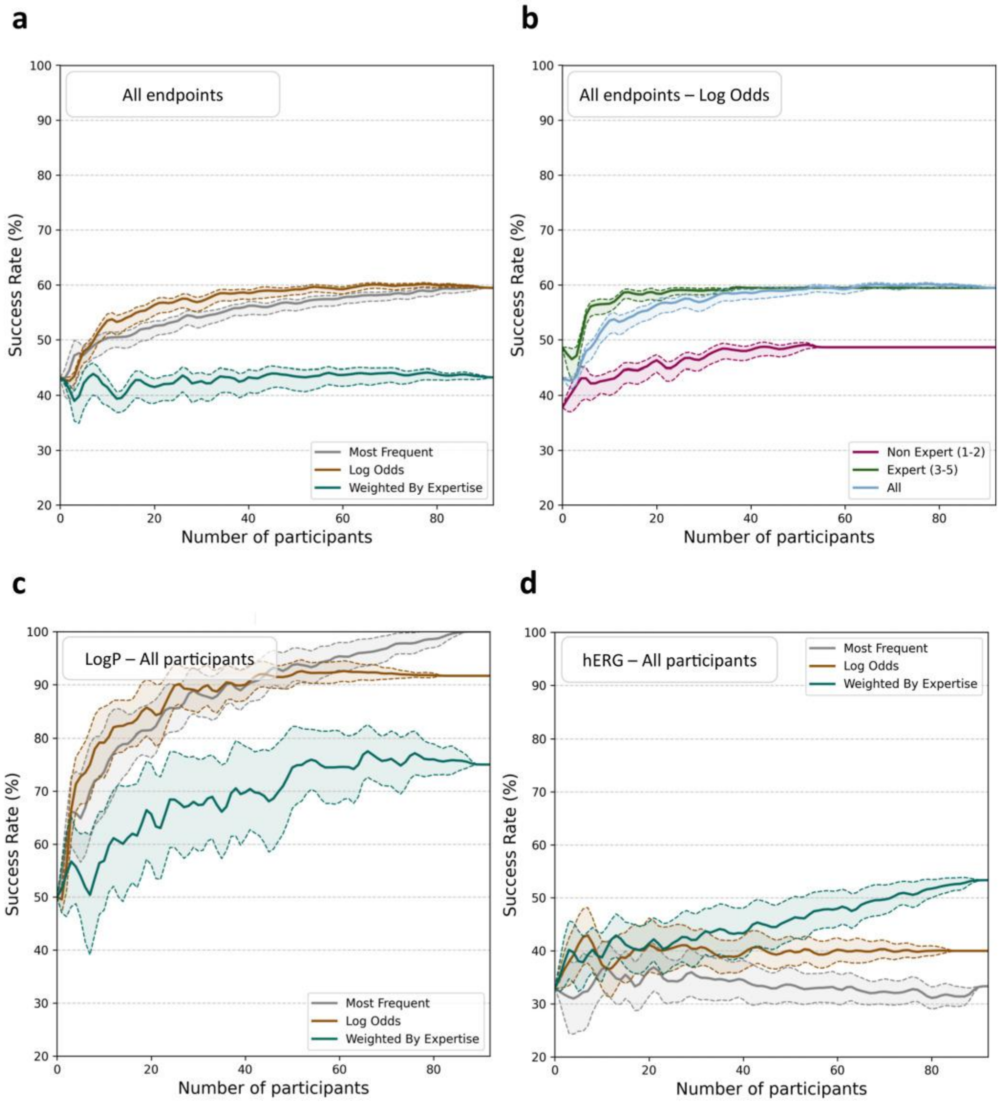


Figure 2. Evolution of the collective success rate. (a) Collective SR for all endpoints using different aggregation methods. The collective answer is either obtained using the "democratic" (most frequent), the confidence-weighted (log odds), or the expertise-weighted aggregation method. (b) Collective SR for all endpoints using log odds for the different participants' groups. (c) Collective SR for log *P* and for all participants. (d) Collective SR for hERG and for all participants.

and S5b). These results suggest significant noise in the self-rated expertise levels. In the second part of the exercise (session two), the participants seem to have understood the difficulty of the questions and adjusted their level of expertise appropriately. This phenomenon is indicated in the graph by more distinct separation between expertise levels (Figure S5c,d).

Overall, our analyses show that the primary determinant correlated to increased SR is the level of confidence individuals place in judging if they have correctly responded to each question.

Collective Performance Dynamics and the Effect of Aggregation Methods. This section of the study is focused on understanding how different aggregation methods affect the

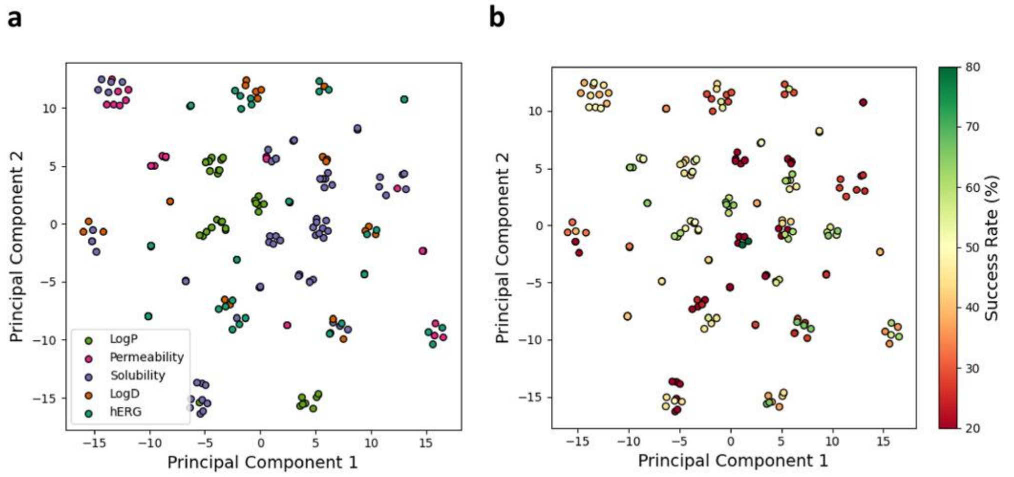


Figure 3. t-SNE map of the CI library. Each point represents a unique compound colored by (a) ADMET endpoint and (b) the success rate of the related question using the CI "most frequent" aggregation method.

SR of CI in drug design, with a particular focus on levels of confidence and expertise, as well as participants population, i.e., self-rated level in medicinal chemistry. To this end, the sample size was iteratively increased from 1 to 92 (the total number of study participants) for each analysis conducted. For each iteration, all unique combinations of individuals without permutations, i.e., without the same participation being accounted for more than once, were analyzed to determine a distinct collective SR distribution. The evolution of this SR was analyzed based on the ADMET endpoint, expertise group, or aggregation method (see Figure 2 and Supporting Information Figures S6−S12).

The collective responses for this study were obtained using six different aggregation methods: the "democratic" approach (most frequent response), log confidence weighting (log odds), fuzzy logic aggregation, confidence weighting, expertise weighting, and coweighting by expertise and confidence. For the questions corresponding to all endpoints together and using the democratic approach, the SR increased by 15% when the analysis was carried out going from a single participant to 20. A SR value converges to 60% when the responses of all 92 participants were considered (Figure 2a). Weighting by expertise had zero effect to the collective intelligence performance, while log confidence weighting achieved a 5% increase compared to the most frequent SR, with only 15 participants. This effect is noticeable for smaller teams, but it becomes less pronounced and diminishes as the group size increases. These results indicate log confidence-based aggregation to be an effective aggregation method, enabling high collective SRs with smaller teams.

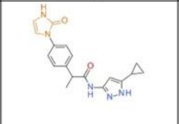
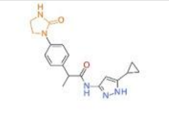
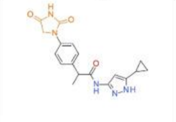
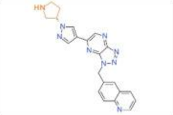
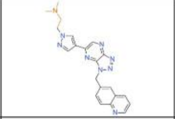
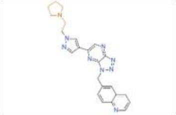
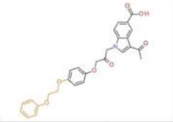
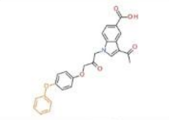
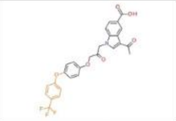
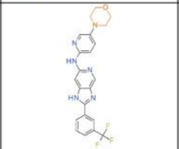
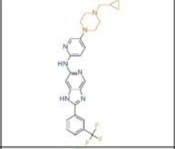
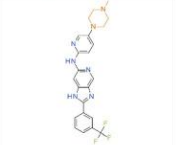
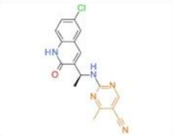
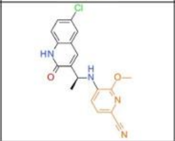
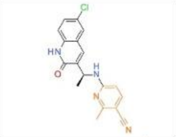
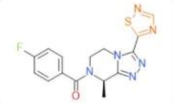
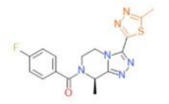
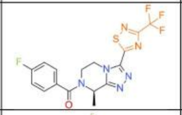
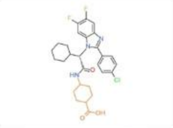
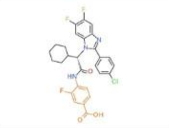
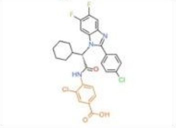
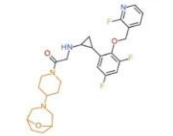
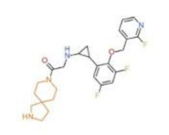
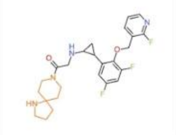
The evolution of CI SR was also analyzed across expertise groups (Figures 2b and S6). The non-expert group requires over 40 participants to reach a SR of approximately 50%, whereas the expert and mixed groups surpass 55% SR with only 10 (experts) and 15 (mixed) participants, respectively. This suggests that an effective CI team for drug design should ideally include some experts and consist of a minimum of 15 members. Notably, the CI SR difference between an all-expert team and a mixed team was minimal, with the mixed team requiring ∼5 more participants to achieve comparable results using the log odds aggregation method (Figure S6).

The collective performance dynamics were also analyzed for each endpoint independently (Figures 2c,d and S6−S11). Using mixed-expertise teams, an 80% SR can be achieved with just ten participants for log *P* (Figure 2c). This trend also held for permeability and solubility (Figures S8 and S9). However, for hERG, over 70 participants were needed to exceed a 50% SR. A noticeable result is that for hERG and only for the groups that include experts, the best performing aggregation method is the one that accounts for the self-rated expertise level in medicinal chemistry, i.e., groups 1 to 5 defined above (Figures 2d, S10, and S12). For log *D* values, the influence of this expertise level was less pronounced (Figures S11 and S12).

We compared the collective SR to the individual mean SR across various endpoints (Figure S13). Simulating groups of 2 to 30 participants with an increment of one participant, we found that CI SR is correlated to the mean individual SR (Figure S13). More specifically, a 5−10% boost in the mean individual success rate leads to a 50% improvement in collective success rates for log *P*, permeability, and log *D*, and a 30% increase for solubility and hERG.

Medicinal Chemistry Pitfalls. This section aims to uncover potential biases in collective decision-making for optimization tasks. A 2D t-SNE[20] map was built to represent the chemical space, with each point corresponding to a compound (Figure 3). The set of all the unique compounds used in the questionnaire is termed as "CI library". This set, consisting of 193 compounds, was projected on the t-SNE map (Figure 3). The map reveals that no endpoint, among the five studied herein, occupies specific areas, indicating that, despite the relatively small size of our data set, the compounds used in this study are diverse. Some compounds overlap between endpoints as they were derived from the same research work, albeit they are not structurally identical (Figure S3). When examining the mapping of the CI SR (using the "most frequent" aggregation method), the performance appears

Table 1. Selection of Challenging Cases from the CI Questionnaire[a]

	ID	SR	Most given	Correct	Third Option	Ref.
Permeability (increase)	**1**	0.20				20
	2	0.26				21
Solubility (increase)	**3**	0.09				22
	4	0.12				23
LogD (decrease)	**5**	0.19				24
	6	0.24				25
hERG (decrease)	**7**	0.19				26
	8	0.26				27

[a]The table presents the collectively worst predicted examples, i.e., worst success rate per ADMET endpoint.

dispersed for all five endpoints, indicating that there are no apparent "easy-to-predict" properties for specific scaffolds or chemical families (Figure S14).

A breakdown of the CI "most frequent" answers with regards to the self-labeled medicinal chemistry groups and endpoints demonstrates no apparent trend between the frequency of an answer per group and its correctness (Figure S15). As expected, for log *P*, the "most frequent" answer corresponds to the correct answer, in agreement with what we presented previously. Despite this, the number of participants per group giving this answer is neither stable nor follows a trend. This is true for all endpoints, where the "most frequent" answer is not necessarily correct, and the number of participants per group does not seem to follow a rule, e.g., one would expect that when the "most frequent" is wrong, the number of experts (groups 4 or 5) giving this answer would decrease compared to the correct answers (Figure S15, wrong answers are annotated with an "x"). This analysis demonstrates that the frequency of an answer, albeit a key factor for the success of the collective

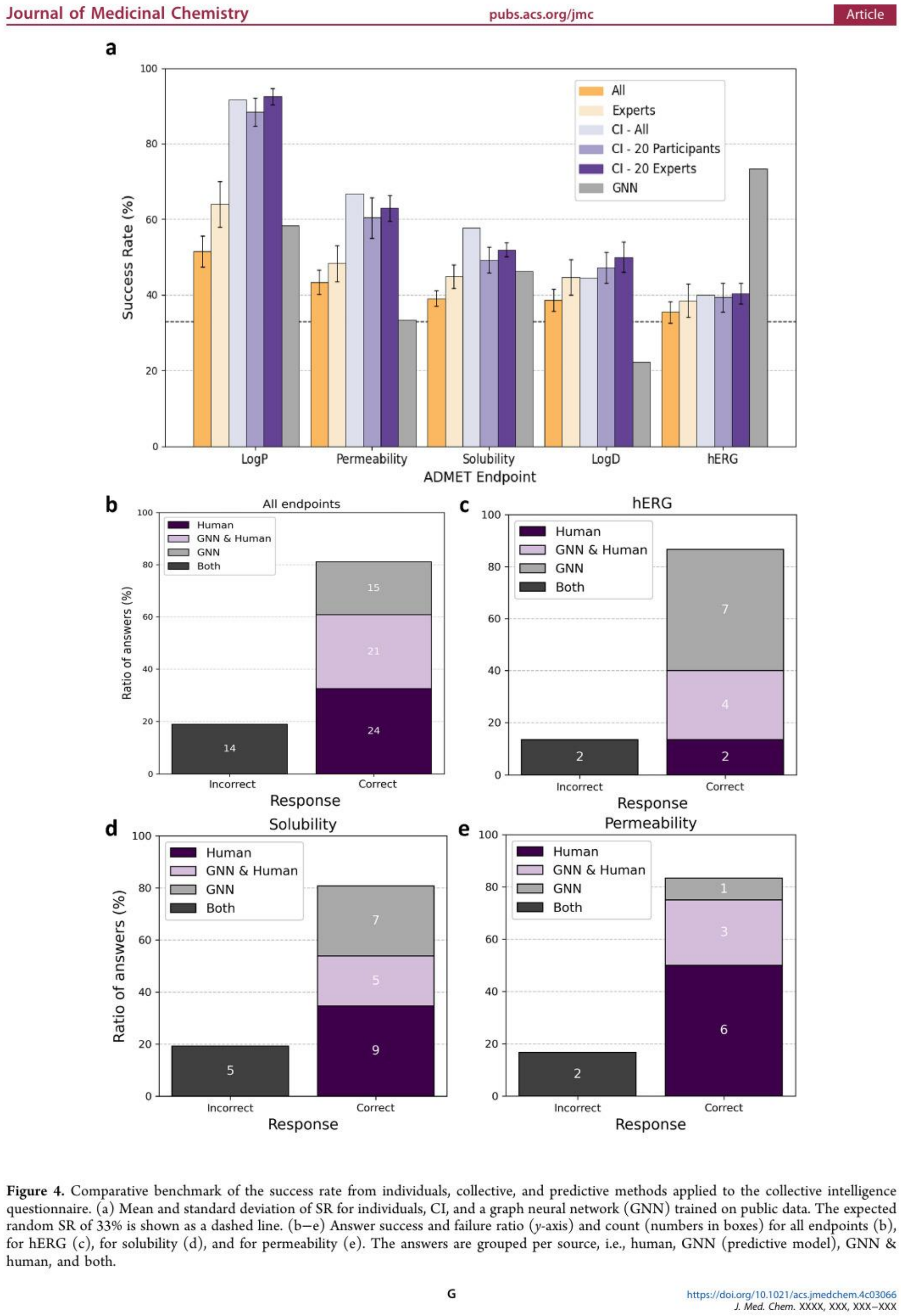
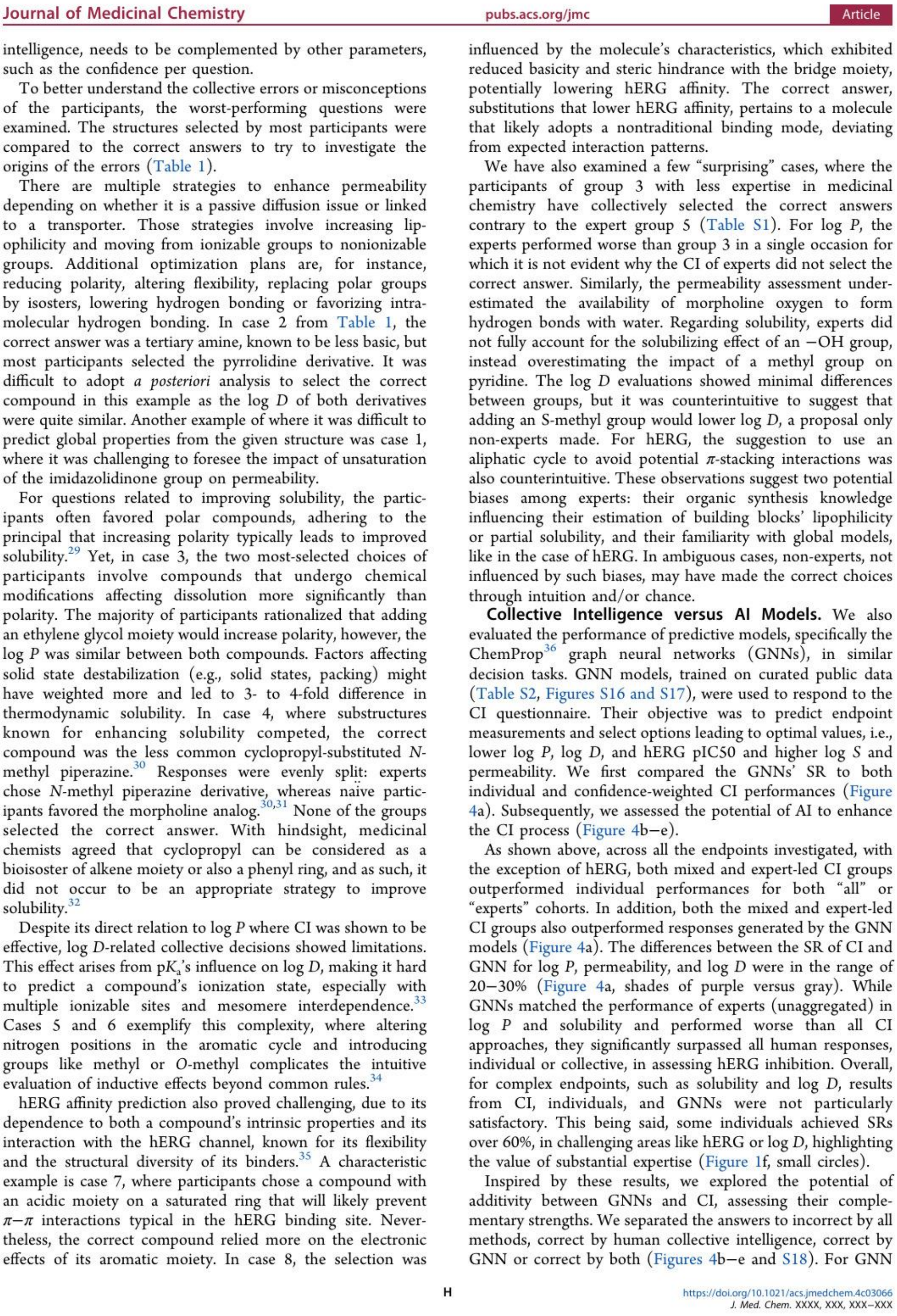


Figure 4. Comparative benchmark of the success rate from individuals, collective, and predictive methods applied to the collective intelligence questionnaire. (a) Mean and standard deviation of SR for individuals, CI, and a graph neural network (GNN) trained on public data. The expected random SR of 33% is shown as a dashed line. (b−e) Answer success and failure ratio (*y*-axis) and count (numbers in boxes) for all endpoints (b), for hERG (c), for solubility (d), and for permeability (e). The answers are grouped per source, i.e., human, GNN (predictive model), GNN & human, and both.

https://doi.org/10.1021/acs.jmedchem.4c03066



intelligence, needs to be complemented by other parameters, such as the confidence per question.

To better understand the collective errors or misconceptions of the participants, the worst-performing questions were examined. The structures selected by most participants were compared to the correct answers to try to investigate the origins of the errors (Table 1).

There are multiple strategies to enhance permeability depending on whether it is a passive diffusion issue or linked to a transporter. Those strategies involve increasing lipophilicity and moving from ionizable groups to nonionizable groups. Additional optimization plans are, for instance, reducing polarity, altering flexibility, replacing polar groups by isosters, lowering hydrogen bonding or favorizing intramolecular hydrogen bonding. In case 2 from Table 1, the correct answer was a tertiary amine, known to be less basic, but most participants selected the pyrrolidine derivative. It was difficult to adopt *a posteriori* analysis to select the correct compound in this example as the log *D* of both derivatives were quite similar. Another example of where it was difficult to predict global properties from the given structure was case 1, where it was challenging to foresee the impact of unsaturation of the imidazolidinone group on permeability.

For questions related to improving solubility, the participants often favored polar compounds, adhering to the principal that increasing polarity typically leads to improved solubility.[29] Yet, in case 3, the two most-selected choices of participants involve compounds that undergo chemical modifications affecting dissolution more significantly than polarity. The majority of participants rationalized that adding an ethylene glycol moiety would increase polarity, however, the log *P* was similar between both compounds. Factors affecting solid state destabilization (e.g., solid states, packing) might have weighted more and led to 3- to 4-fold difference in thermodynamic solubility. In case 4, where substructures known for enhancing solubility competed, the correct compound was the less common cyclopropyl-substituted *N*-methyl piperazine.[30] Responses were evenly split: experts chose *N*-methyl piperazine derivative, whereas naïve participants favored the morpholine analog.[30,31] None of the groups selected the correct answer. With hindsight, medicinal chemists agreed that cyclopropyl can be considered as a bioisoster of alkene moiety or also a phenyl ring, and as such, it did not occur to be an appropriate strategy to improve solubility.[32]

Despite its direct relation to log *P* where CI was shown to be effective, log *D*-related collective decisions showed limitations. This effect arises from pK_a's influence on log *D*, making it hard to predict a compound's ionization state, especially with multiple ionizable sites and mesomere interdependence.[33] Cases 5 and 6 exemplify this complexity, where altering nitrogen positions in the aromatic cycle and introducing groups like methyl or *O*-methyl complicates the intuitive evaluation of inductive effects beyond common rules.[34]

hERG affinity prediction also proved challenging, due to its dependence to both a compound's intrinsic properties and its interaction with the hERG channel, known for its flexibility and the structural diversity of its binders.[35] A characteristic example is case 7, where participants chose a compound with an acidic moiety on a saturated ring that will likely prevent $\pi-\pi$ interactions typical in the hERG binding site. Nevertheless, the correct compound relied more on the electronic effects of its aromatic moiety. In case 8, the selection was influenced by the molecule's characteristics, which exhibited reduced basicity and steric hindrance with the bridge moiety, potentially lowering hERG affinity. The correct answer, substitutions that lower hERG affinity, pertains to a molecule that likely adopts a nontraditional binding mode, deviating from expected interaction patterns.

We have also examined a few "surprising" cases, where the participants of group 3 with less expertise in medicinal chemistry have collectively selected the correct answers contrary to the expert group 5 (Table S1). For log *P*, the experts performed worse than group 3 in a single occasion for which it is not evident why the CI of experts did not select the correct answer. Similarly, the permeability assessment underestimated the availability of morpholine oxygen to form hydrogen bonds with water. Regarding solubility, experts did not fully account for the solubilizing effect of an −OH group, instead overestimating the impact of a methyl group on pyridine. The log *D* evaluations showed minimal differences between groups, but it was counterintuitive to suggest that adding an S-methyl group would lower log *D*, a proposal only non-experts made. For hERG, the suggestion to use an aliphatic cycle to avoid potential π-stacking interactions was also counterintuitive. These observations suggest two potential biases among experts: their organic synthesis knowledge influencing their estimation of building blocks' lipophilicity or partial solubility, and their familiarity with global models, like in the case of hERG. In ambiguous cases, non-experts, not influenced by such biases, may have made the correct choices through intuition and/or chance.

Collective Intelligence versus AI Models. We also evaluated the performance of predictive models, specifically the ChemProp[36] graph neural networks (GNNs), in similar decision tasks. GNN models, trained on curated public data (Table S2, Figures S16 and S17), were used to respond to the CI questionnaire. Their objective was to predict endpoint measurements and select options leading to optimal values, i.e., lower log *P*, log *D*, and hERG pIC50 and higher log *S* and permeability. We first compared the GNNs' SR to both individual and confidence-weighted CI performances (Figure 4a). Subsequently, we assessed the potential of AI to enhance the CI process (Figure 4b−e).

As shown above, across all the endpoints investigated, with the exception of hERG, both mixed and expert-led CI groups outperformed individual performances for both "all" or "experts" cohorts. In addition, both the mixed and expert-led CI groups also outperformed responses generated by the GNN models (Figure 4a). The differences between the SR of CI and GNN for log *P*, permeability, and log *D* were in the range of 20−30% (Figure 4a, shades of purple versus gray). While GNNs matched the performance of experts (unaggregated) in log *P* and solubility and performed worse than all CI approaches, they significantly surpassed all human responses, individual or collective, in assessing hERG inhibition. Overall, for complex endpoints, such as solubility and log *D*, results from CI, individuals, and GNNs were not particularly satisfactory. This being said, some individuals achieved SRs over 60%, in challenging areas like hERG or log *D*, highlighting the value of substantial expertise (Figure 1f, small circles).

Inspired by these results, we explored the potential of additivity between GNNs and CI, assessing their complementary strengths. We separated the answers to incorrect by all methods, correct by human collective intelligence, correct by GNN or correct by both (Figures 4b−e and S18). For GNN

the answer from the unique models was taken, whereas for CI we used the log odds method applied to answers from the full cohort. Analysis showed that GNNs provided correct answers for 20% of the questions where CI failed, while CI succeeded in 32% of cases where GNN struggled. If one was able to combine GNN and CI correct answers, the overall performance would improve from 60% for the collective intelligence group alone to 81% with the addition of GNN over all endpoints. The complementarity between GNNs and CI was particularly evident for solubility and hERG, where GNN would contribute to a 27−47% increase in the SR over the value identified using CI alone. Overall, a potential synergy between GNN and our collective intelligence methods would lead to an impressive SR of 87%, 81%, and 83% for hERG, solubility, and permeability, respectively. For log *P*, CI performs already exceptionally well while for log *D* the more challenging questions were missed by both CI and the GNN model (Figure S18).

■ DISCUSSION

Lead optimization campaigns are driven, or are at least, greatly influenced by the medicinal chemistry intuition of the project chemist leader(s). This medicinal chemistry intuition is inextricably linked to individual drug-likeness standards that depend on the chemist's experience, "know-how", and bias.[5,37,38] Thus, characterizing a clear drug-likeness signal from medicinal chemistry intuition is a challenge.

We have presented an innovative approach to accelerate the lead optimization process that combines notions from the field of collective intelligence, medicinal chemistry, and machine learning. The responses of 92 participants to 74 medicinal chemistry multiple-choice questions offered insights on the influence of expertise and confidence on the application of collective intelligence in drug discovery. It is important to emphasize that both medicinal chemistry expertise and confidence per question were by design included in our questionnaire, in order to be used as parameters in the analysis, and in particular in the data aggregation process.

Through ADMET optimization tasks, we observed varying success rates over self-labeled non-experts and experts in medicinal chemistry. A classification of participants based on success rate (SR) revealed the superiority of teams composed of individuals with varying levels of expertise over those that lacked such variation, in agreement with previous works in the cognitive science field.[8] Moreover, the self-assessment of expertise in medicinal chemistry served as a proxy for subjective confidence, which has been shown to correlate with decision-making accuracy and can effectively enhance group performance through confidence-weighted aggregation.[39] By capturing participants' subjective confidence, we accounted for factors beyond experience that contribute to expertise, such as individual aptitude and continuous learning.[40]

The variability in participants' confidence levels reflects a realistic self-assessment of their knowledge and uncertainty, consistent with the self-consistency model of subjective confidence.[41] Incorporating these confidence levels into our aggregation methods enhances the accuracy of collective judgments, as supported by prior research.[42,43] This approach effectively leverages individual differences in expertise and confidence, improving group decision-making accuracy in complex tasks.

A significant correlation between confidence levels per answer and expertise was observed, with higher confidence generally aligning with higher expertise levels. In parallel, we demonstrated that the primary determinant of SR is the level of confidence per answer expressed by participants, illustrating its importance in decision-making. Another noticeable result was the lack of correlation between SR and intermediate expertise levels. This could be due to a combination of factors: over- or under-confidence among participants, varying experience levels relative to a specific endpoint, and the inherent difficulty of the questions asked.

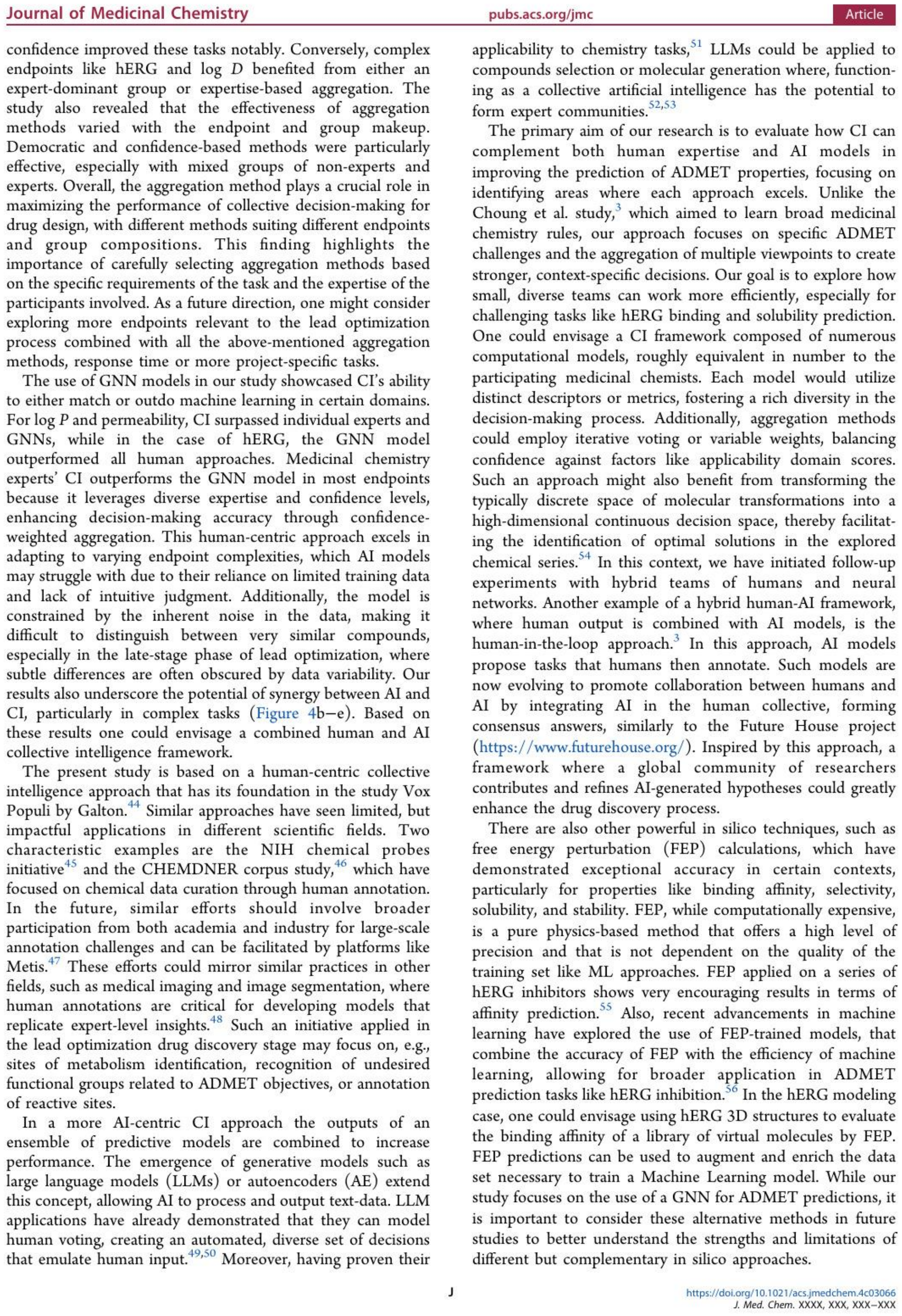
Performance varied across log *P*, log *D*, permeability, solubility, and hERG inhibition endpoints. Aggregation methods that account for collective intelligence significantly enhanced the success rates for tasks such as log *P*, permeability, and solubility, indicating the value of using such methods to address these endpoints at a project level. For example, for log *P*, we observe an average performance of 75% with only 10 participants from diverse backgrounds with the log odds method that uses the information on confidence per response to aggregate the data (Figure 2c). In practice, this method and other examples presented in the Results section (Figures 2c and S12) demonstrate the added value of collective decision-making related to at least certain ADMET endpoints.

Our results indicate that collective intelligence (CI) methods excel for endpoints involving better understood phenomena, such as log *P*, solubility, and permeability, but are less effective in complex areas such as hERG and log *D*. This observation is consistent with Condorcet's Jury Theorem, which suggests that when individual decision-makers have a probability of making a correct decision slightly above random chance (33%), the likelihood of a correct collective decision increases with the number of participants (Figure 2c). Conversely, for more complex endpoints such as hERG and log *D* (Figure 2d), where individual accuracy falls below random chance, increasing the number of participants may paradoxically reduce the likelihood of a correct outcome.

Challenging endpoints, like hERG and log *D*, may require more expert input, detailed structural information, expert-focused aggregation strategies or different presentation of chemical structures, possibly including 3D information. It is likely that the format of the questionnaire itself plays a role in the participants' cognitive process, e.g., a chemist would probably respond differently if the whole molecule was shown rather than only the substitution. Additionally, the brief period for responses may have limited the comprehensive evaluation of complex chemical phenomena, such as tautomeric changes and inductive or mesomeric effects, crucial for efficient optimization. A direction that could explain such cognitive phenomena and possibly further improve data aggregation is the inclusion of a timestamp per question, i.e., response time. This timestamp could help elaborate the discrepancies between intermediate medicinal chemistry levels and SR, easily identify the most challenging questions, and further improve the overall SR by using for example a combination of confidence and timestamp per question as aggregation method. Unfortunately, with the given setup we employed herein (see Methods), it was not possible to register the specific information.

The ensemble of our results demonstrates the need for tailored collective decision-making approaches in drug design, considering the varying complexities of different endpoints, the expertise of project members, and the "expert"-"non-expert" ratio. We found that reweighting responses based on

https://doi.org/10.1021/acs.jmedchem.4c03066



confidence improved these tasks notably. Conversely, complex endpoints like hERG and log *D* benefited from either an expert-dominant group or expertise-based aggregation. The study also revealed that the effectiveness of aggregation methods varied with the endpoint and group makeup. Democratic and confidence-based methods were particularly effective, especially with mixed groups of non-experts and experts. Overall, the aggregation method plays a crucial role in maximizing the performance of collective decision-making for drug design, with different methods suiting different endpoints and group compositions. This finding highlights the importance of carefully selecting aggregation methods based on the specific requirements of the task and the expertise of the participants involved. As a future direction, one might consider exploring more endpoints relevant to the lead optimization process combined with all the above-mentioned aggregation methods, response time or more project-specific tasks.

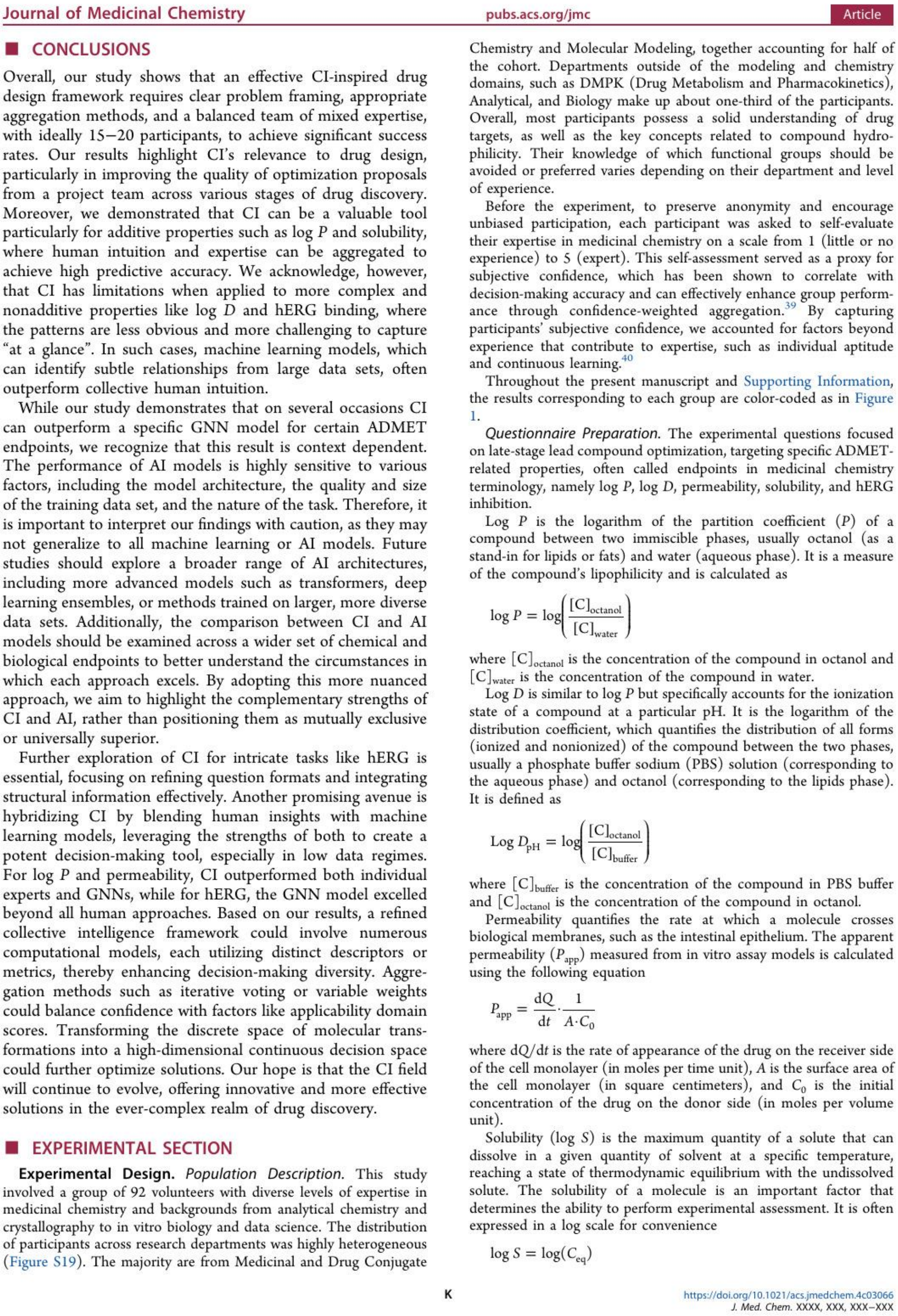
The use of GNN models in our study showcased CI's ability to either match or outdo machine learning in certain domains. For log *P* and permeability, CI surpassed individual experts and GNNs, while in the case of hERG, the GNN model outperformed all human approaches. Medicinal chemistry experts' CI outperforms the GNN model in most endpoints because it leverages diverse expertise and confidence levels, enhancing decision-making accuracy through confidence-weighted aggregation. This human-centric approach excels in adapting to varying endpoint complexities, which AI models may struggle with due to their reliance on limited training data and lack of intuitive judgment. Additionally, the model is constrained by the inherent noise in the data, making it difficult to distinguish between very similar compounds, especially in the late-stage phase of lead optimization, where subtle differences are often obscured by data variability. Our results also underscore the potential of synergy between AI and CI, particularly in complex tasks (Figure 4b−e). Based on these results one could envisage a combined human and AI collective intelligence framework.

The present study is based on a human-centric collective intelligence approach that has its foundation in the study Vox Populi by Galton.[44] Similar approaches have seen limited, but impactful applications in different scientific fields. Two characteristic examples are the NIH chemical probes initiative[45] and the CHEMDNER corpus study,[46] which have focused on chemical data curation through human annotation. In the future, similar efforts should involve broader participation from both academia and industry for large-scale annotation challenges and can be facilitated by platforms like Metis.[47] These efforts could mirror similar practices in other fields, such as medical imaging and image segmentation, where human annotations are critical for developing models that replicate expert-level insights.[48] Such an initiative applied in the lead optimization drug discovery stage may focus on, e.g., sites of metabolism identification, recognition of undesired functional groups related to ADMET objectives, or annotation of reactive sites.

In a more AI-centric CI approach the outputs of an ensemble of predictive models are combined to increase performance. The emergence of generative models such as large language models (LLMs) or autoencoders (AE) extend this concept, allowing AI to process and output text-data. LLM applications have already demonstrated that they can model human voting, creating an automated, diverse set of decisions that emulate human input.[49,50] Moreover, having proven their applicability to chemistry tasks,[51] LLMs could be applied to compounds selection or molecular generation where, functioning as a collective artificial intelligence has the potential to form expert communities.[52,53]

The primary aim of our research is to evaluate how CI can complement both human expertise and AI models in improving the prediction of ADMET properties, focusing on identifying areas where each approach excels. Unlike the Choung et al. study,[3] which aimed to learn broad medicinal chemistry rules, our approach focuses on specific ADMET challenges and the aggregation of multiple viewpoints to create stronger, context-specific decisions. Our goal is to explore how small, diverse teams can work more efficiently, especially for challenging tasks like hERG binding and solubility prediction. One could envisage a CI framework composed of numerous computational models, roughly equivalent in number to the participating medicinal chemists. Each model would utilize distinct descriptors or metrics, fostering a rich diversity in the decision-making process. Additionally, aggregation methods could employ iterative voting or variable weights, balancing confidence against factors like applicability domain scores. Such an approach might also benefit from transforming the typically discrete space of molecular transformations into a high-dimensional continuous decision space, thereby facilitating the identification of optimal solutions in the explored chemical series.[54] In this context, we have initiated follow-up experiments with hybrid teams of humans and neural networks. Another example of a hybrid human-AI framework, where human output is combined with AI models, is the human-in-the-loop approach.[3] In this approach, AI models propose tasks that humans then annotate. Such models are now evolving to promote collaboration between humans and AI by integrating AI in the human collective, forming consensus answers, similarly to the Future House project (https://www.futurehouse.org/). Inspired by this approach, a framework where a global community of researchers contributes and refines AI-generated hypotheses could greatly enhance the drug discovery process.

There are also other powerful in silico techniques, such as free energy perturbation (FEP) calculations, which have demonstrated exceptional accuracy in certain contexts, particularly for properties like binding affinity, selectivity, solubility, and stability. FEP, while computationally expensive, is a pure physics-based method that offers a high level of precision and that is not dependent on the quality of the training set like ML approaches. FEP applied on a series of hERG inhibitors shows very encouraging results in terms of affinity prediction.[55] Also, recent advancements in machine learning have explored the use of FEP-trained models, that combine the accuracy of FEP with the efficiency of machine learning, allowing for broader application in ADMET prediction tasks like hERG inhibition.[56] In the hERG modeling case, one could envisage using hERG 3D structures to evaluate the binding affinity of a library of virtual molecules by FEP. FEP predictions can be used to augment and enrich the data set necessary to train a Machine Learning model. While our study focuses on the use of a GNN for ADMET predictions, it is important to consider these alternative methods in future studies to better understand the strengths and limitations of different but complementary in silico approaches.



CONCLUSIONS

Overall, our study shows that an effective CI-inspired drug design framework requires clear problem framing, appropriate aggregation methods, and a balanced team of mixed expertise, with ideally 15−20 participants, to achieve significant success rates. Our results highlight CI's relevance to drug design, particularly in improving the quality of optimization proposals from a project team across various stages of drug discovery. Moreover, we demonstrated that CI can be a valuable tool particularly for additive properties such as log P and solubility, where human intuition and expertise can be aggregated to achieve high predictive accuracy. We acknowledge, however, that CI has limitations when applied to more complex and nonadditive properties like log D and hERG binding, where the patterns are less obvious and more challenging to capture "at a glance". In such cases, machine learning models, which can identify subtle relationships from large data sets, often outperform collective human intuition.

While our study demonstrates that on several occasions CI can outperform a specific GNN model for certain ADMET endpoints, we recognize that this result is context dependent. The performance of AI models is highly sensitive to various factors, including the model architecture, the quality and size of the training data set, and the nature of the task. Therefore, it is important to interpret our findings with caution, as they may not generalize to all machine learning or AI models. Future studies should explore a broader range of AI architectures, including more advanced models such as transformers, deep learning ensembles, or methods trained on larger, more diverse data sets. Additionally, the comparison between CI and AI models should be examined across a wider set of chemical and biological endpoints to better understand the circumstances in which each approach excels. By adopting this more nuanced approach, we aim to highlight the complementary strengths of CI and AI, rather than positioning them as mutually exclusive or universally superior.

Further exploration of CI for intricate tasks like hERG is essential, focusing on refining question formats and integrating structural information effectively. Another promising avenue is hybridizing CI by blending human insights with machine learning models, leveraging the strengths of both to create a potent decision-making tool, especially in low data regimes. For log P and permeability, CI outperformed both individual experts and GNNs, while for hERG, the GNN model excelled beyond all human approaches. Based on our results, a refined collective intelligence framework could involve numerous computational models, each utilizing distinct descriptors or metrics, thereby enhancing decision-making diversity. Aggregation methods such as iterative voting or variable weights could balance confidence with factors like applicability domain scores. Transforming the discrete space of molecular transformations into a high-dimensional continuous decision space could further optimize solutions. Our hope is that the CI field will continue to evolve, offering innovative and more effective solutions in the ever-complex realm of drug discovery.

EXPERIMENTAL SECTION

Experimental Design. *Population Description.* This study involved a group of 92 volunteers with diverse levels of expertise in medicinal chemistry and backgrounds from analytical chemistry and crystallography to in vitro biology and data science. The distribution of participants across research departments was highly heterogeneous (Figure S19). The majority are from Medicinal and Drug Conjugate Chemistry and Molecular Modeling, together accounting for half of the cohort. Departments outside of the modeling and chemistry domains, such as DMPK (Drug Metabolism and Pharmacokinetics), Analytical, and Biology make up about one-third of the participants. Overall, most participants possess a solid understanding of drug targets, as well as the key concepts related to compound hydrophilicity. Their knowledge of which functional groups should be avoided or preferred varies depending on their department and level of experience.

Before the experiment, to preserve anonymity and encourage unbiased participation, each participant was asked to self-evaluate their expertise in medicinal chemistry on a scale from 1 (little or no experience) to 5 (expert). This self-assessment served as a proxy for subjective confidence, which has been shown to correlate with decision-making accuracy and can effectively enhance group performance through confidence-weighted aggregation.[39] By capturing participants' subjective confidence, we accounted for factors beyond experience that contribute to expertise, such as individual aptitude and continuous learning.[40]

Throughout the present manuscript and Supporting Information, the results corresponding to each group are color-coded as in Figure 1.

Questionnaire Preparation. The experimental questions focused on late-stage lead compound optimization, targeting specific ADMET-related properties, often called endpoints in medicinal chemistry terminology, namely log P, log D, permeability, solubility, and hERG inhibition.

Log P is the logarithm of the partition coefficient (P) of a compound between two immiscible phases, usually octanol (as a stand-in for lipids or fats) and water (aqueous phase). It is a measure of the compound's lipophilicity and is calculated as

$$\log P = \log\left(\frac{[\mathrm{C}]_{\mathrm{octanol}}}{[\mathrm{C}]_{\mathrm{water}}}\right)$$

where $[\mathrm{C}]_{\mathrm{octanol}}$ is the concentration of the compound in octanol and $[\mathrm{C}]_{\mathrm{water}}$ is the concentration of the compound in water.

Log D is similar to log P but specifically accounts for the ionization state of a compound at a particular pH. It is the logarithm of the distribution coefficient, which quantifies the distribution of all forms (ionized and nonionized) of the compound between the two phases, usually a phosphate buffer sodium (PBS) solution (corresponding to the aqueous phase) and octanol (corresponding to the lipids phase). It is defined as

$$\mathrm{Log}\ D_{\mathrm{pH}} = \log\left(\frac{[\mathrm{C}]_{\mathrm{octanol}}}{[\mathrm{C}]_{\mathrm{buffer}}}\right)$$

where $[\mathrm{C}]_{\mathrm{buffer}}$ is the concentration of the compound in PBS buffer and $[\mathrm{C}]_{\mathrm{octanol}}$ is the concentration of the compound in octanol.

Permeability quantifies the rate at which a molecule crosses biological membranes, such as the intestinal epithelium. The apparent permeability (P_{app}) measured from in vitro assay models is calculated using the following equation

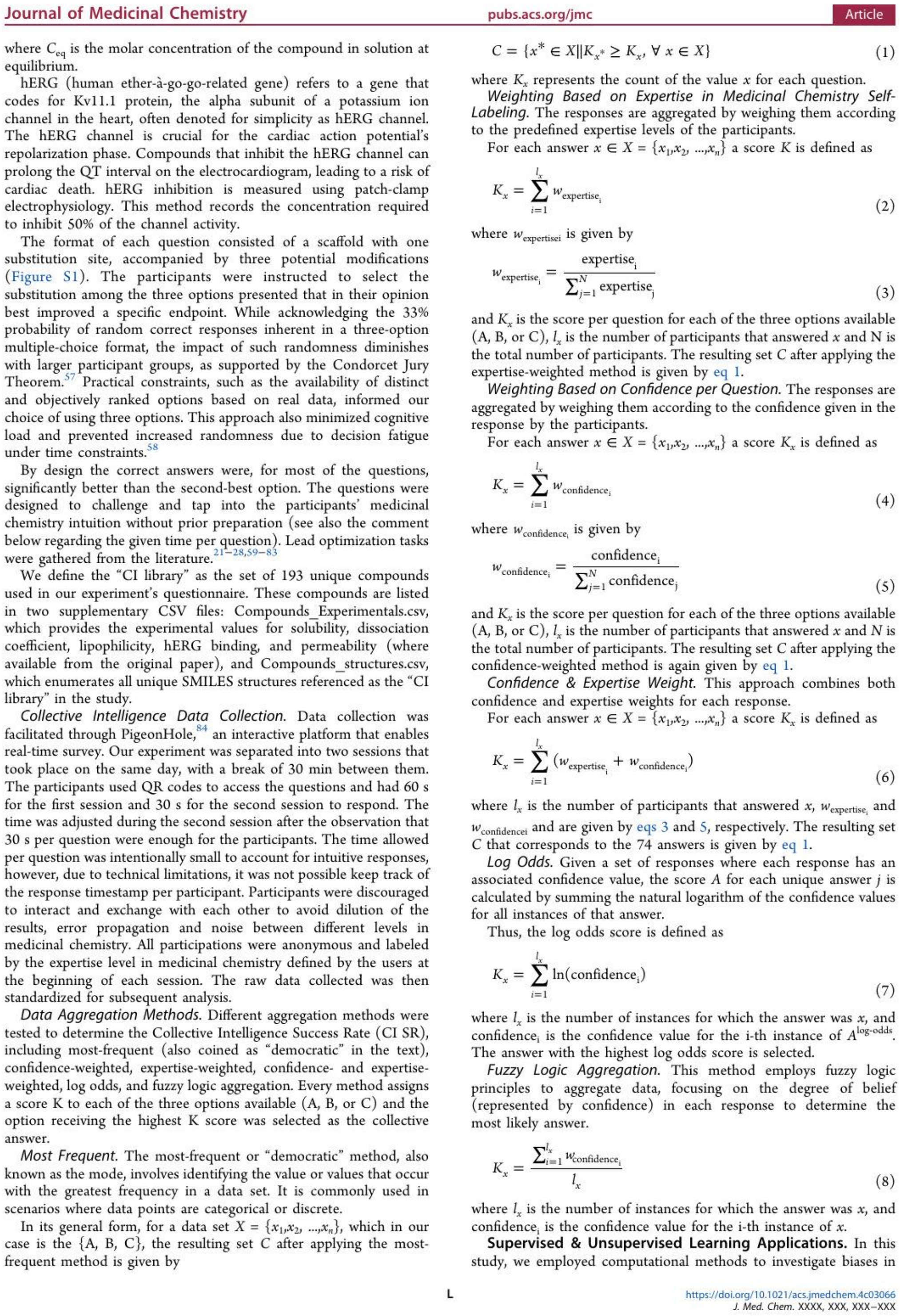
$$P_{\mathrm{app}} = \frac{\mathrm{d}Q}{\mathrm{d}t}\cdot\frac{1}{A\cdot C_0}$$

where $\mathrm{d}Q/\mathrm{d}t$ is the rate of appearance of the drug on the receiver side of the cell monolayer (in moles per time unit), A is the surface area of the cell monolayer (in square centimeters), and C_0 is the initial concentration of the drug on the donor side (in moles per volume unit).

Solubility (log S) is the maximum quantity of a solute that can dissolve in a given quantity of solvent at a specific temperature, reaching a state of thermodynamic equilibrium with the undissolved solute. The solubility of a molecule is an important factor that determines the ability to perform experimental assessment. It is often expressed in a log scale for convenience

$$\log S = \log(C_{\mathrm{eq}})$$

https://doi.org/10.1021/acs.jmedchem.4c03066



where C_{eq} is the molar concentration of the compound in solution at equilibrium.

hERG (human ether-à-go-go-related gene) refers to a gene that codes for Kv11.1 protein, the alpha subunit of a potassium ion channel in the heart, often denoted for simplicity as hERG channel. The hERG channel is crucial for the cardiac action potential's repolarization phase. Compounds that inhibit the hERG channel can prolong the QT interval on the electrocardiogram, leading to a risk of cardiac death. hERG inhibition is measured using patch-clamp electrophysiology. This method records the concentration required to inhibit 50% of the channel activity.

The format of each question consisted of a scaffold with one substitution site, accompanied by three potential modifications (Figure S1). The participants were instructed to select the substitution among the three options presented that in their opinion best improved a specific endpoint. While acknowledging the 33% probability of random correct responses inherent in a three-option multiple-choice format, the impact of such randomness diminishes with larger participant groups, as supported by the Condorcet Jury Theorem.[57] Practical constraints, such as the availability of distinct and objectively ranked options based on real data, informed our choice of using three options. This approach also minimized cognitive load and prevented increased randomness due to decision fatigue under time constraints.[58]

By design the correct answers were, for most of the questions, significantly better than the second-best option. The questions were designed to challenge and tap into the participants' medicinal chemistry intuition without prior preparation (see also the comment below regarding the given time per question). Lead optimization tasks were gathered from the literature.[21−28,59−83]

We define the "CI library" as the set of 193 unique compounds used in our experiment's questionnaire. These compounds are listed in two supplementary CSV files: Compounds_Experimentals.csv, which provides the experimental values for solubility, dissociation coefficient, lipophilicity, hERG binding, and permeability (where available from the original paper), and Compounds_structures.csv, which enumerates all unique SMILES structures referenced as the "CI library" in the study.

Collective Intelligence Data Collection. Data collection was facilitated through PigeonHole,[84] an interactive platform that enables real-time survey. Our experiment was separated into two sessions that took place on the same day, with a break of 30 min between them. The participants used QR codes to access the questions and had 60 s for the first session and 30 s for the second session to respond. The time was adjusted during the second session after the observation that 30 s per question were enough for the participants. The time allowed per question was intentionally small to account for intuitive responses, however, due to technical limitations, it was not possible keep track of the response timestamp per participant. Participants were discouraged to interact and exchange with each other to avoid dilution of the results, error propagation and noise between different levels in medicinal chemistry. All participations were anonymous and labeled by the expertise level in medicinal chemistry defined by the users at the beginning of each session. The raw data collected was then standardized for subsequent analysis.

Data Aggregation Methods. Different aggregation methods were tested to determine the Collective Intelligence Success Rate (CI SR), including most-frequent (also coined as "democratic" in the text), confidence-weighted, expertise-weighted, confidence- and expertise-weighted, log odds, and fuzzy logic aggregation. Every method assigns a score K to each of the three options available (A, B, or C) and the option receiving the highest K score was selected as the collective answer.

Most Frequent. The most-frequent or "democratic" method, also known as the mode, involves identifying the value or values that occur with the greatest frequency in a data set. It is commonly used in scenarios where data points are categorical or discrete.

In its general form, for a data set $X = \{x_1, x_2, ..., x_n\}$, which in our case is the {A, B, C}, the resulting set C after applying the most-frequent method is given by

$$C = \{x^* \in X \| K_{x^*} \geq K_x, \forall\, x \in X\} \tag{1}$$

where K_x represents the count of the value x for each question.

Weighting Based on Expertise in Medicinal Chemistry Self-Labeling. The responses are aggregated by weighing them according to the predefined expertise levels of the participants.

For each answer $x \in X = \{x_1, x_2, ..., x_n\}$ a score K is defined as

$$K_x = \sum_{i=1}^{l_x} w_{\text{expertise}_i} \tag{2}$$

where $w_{\text{expertisei}}$ is given by

$$w_{\text{expertise}_i} = \frac{\text{expertise}_i}{\sum_{j=1}^{N} \text{expertise}_j} \tag{3}$$

and K_x is the score per question for each of the three options available (A, B, or C), l_x is the number of participants that answered x and N is the total number of participants. The resulting set C after applying the expertise-weighted method is given by eq 1.

Weighting Based on Confidence per Question. The responses are aggregated by weighing them according to the confidence given in the response by the participants.

For each answer $x \in X = \{x_1, x_2, ..., x_n\}$ a score K_x is defined as

$$K_x = \sum_{i=1}^{l_x} w_{\text{confidence}_i} \tag{4}$$

where $w_{\text{confidence}_i}$ is given by

$$w_{\text{confidence}_i} = \frac{\text{confidence}_i}{\sum_{j=1}^{N} \text{confidence}_j} \tag{5}$$

and K_x is the score per question for each of the three options available (A, B, or C), l_x is the number of participants that answered x and N is the total number of participants. The resulting set C after applying the confidence-weighted method is again given by eq 1.

Confidence & Expertise Weight. This approach combines both confidence and expertise weights for each response.

For each answer $x \in X = \{x_1, x_2, ..., x_n\}$ a score K_x is defined as

$$K_x = \sum_{i=1}^{l_x} (w_{\text{expertise}_i} + w_{\text{confidence}_i}) \tag{6}$$

where l_x is the number of participants that answered x, $w_{\text{expertise}_i}$ and $w_{\text{confidencei}}$ and are given by eqs 3 and 5, respectively. The resulting set C that corresponds to the 74 answers is given by eq 1.

Log Odds. Given a set of responses where each response has an associated confidence value, the score A for each unique answer j is calculated by summing the natural logarithm of the confidence values for all instances of that answer.

Thus, the log odds score is defined as

$$K_x = \sum_{i=1}^{l_x} \ln(\text{confidence}_i) \tag{7}$$

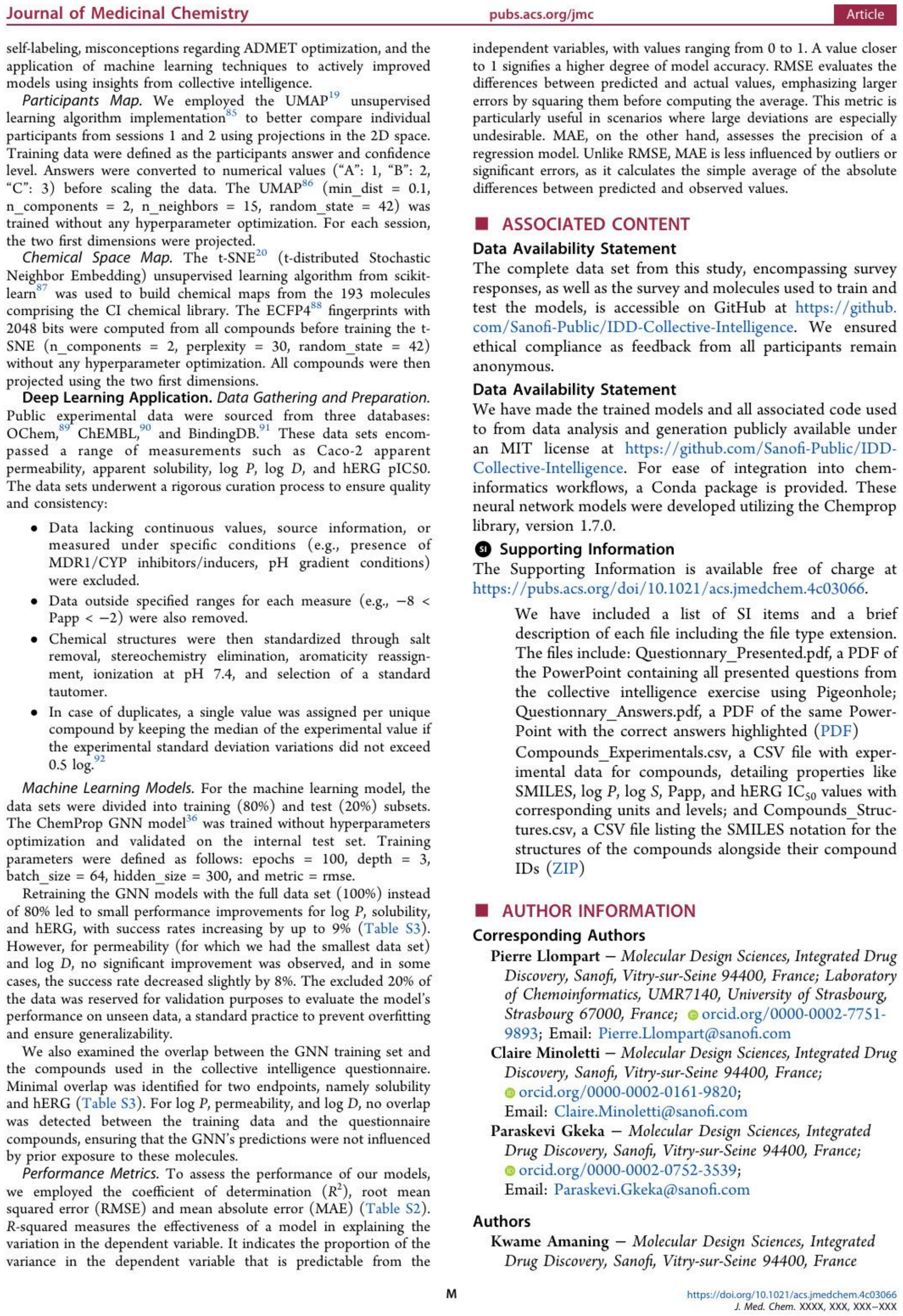
where l_x is the number of instances for which the answer was x, and confidence_i is the confidence value for the i-th instance of $A^{\text{log-odds}}$. The answer with the highest log odds score is selected.

Fuzzy Logic Aggregation. This method employs fuzzy logic principles to aggregate data, focusing on the degree of belief (represented by confidence) in each response to determine the most likely answer.

$$K_x = \frac{\sum_{i=1}^{l_x} w_{\text{confidence}_i}}{l_x} \tag{8}$$

where l_x is the number of instances for which the answer was x, and confidence_i is the confidence value for the i-th instance of x.

Supervised & Unsupervised Learning Applications. In this study, we employed computational methods to investigate biases in



self-labeling, misconceptions regarding ADMET optimization, and the application of machine learning techniques to actively improved models using insights from collective intelligence.

Participants Map. We employed the UMAP[19] unsupervised learning algorithm implementation[85] to better compare individual participants from sessions 1 and 2 using projections in the 2D space. Training data were defined as the participants answer and confidence level. Answers were converted to numerical values ("A": 1, "B": 2, "C": 3) before scaling the data. The UMAP[86] (min_dist = 0.1, n_components = 2, n_neighbors = 15, random_state = 42) was trained without any hyperparameter optimization. For each session, the two first dimensions were projected.

Chemical Space Map. The t-SNE[20] (t-distributed Stochastic Neighbor Embedding) unsupervised learning algorithm from scikit-learn[87] was used to build chemical maps from the 193 molecules comprising the CI chemical library. The ECFP4[88] fingerprints with 2048 bits were computed from all compounds before training the t-SNE (n_components = 2, perplexity = 30, random_state = 42) without any hyperparameter optimization. All compounds were then projected using the two first dimensions.

Deep Learning Application. *Data Gathering and Preparation.* Public experimental data were sourced from three databases: OChem,[89] ChEMBL,[90] and BindingDB.[91] These data sets encompassed a range of measurements such as Caco-2 apparent permeability, apparent solubility, log *P*, log *D*, and hERG pIC50. The data sets underwent a rigorous curation process to ensure quality and consistency:

- Data lacking continuous values, source information, or measured under specific conditions (e.g., presence of MDR1/CYP inhibitors/inducers, pH gradient conditions) were excluded.
- Data outside specified ranges for each measure (e.g., $-8 <$ Papp < -2) were also removed.
- Chemical structures were then standardized through salt removal, stereochemistry elimination, aromaticity reassignment, ionization at pH 7.4, and selection of a standard tautomer.
- In case of duplicates, a single value was assigned per unique compound by keeping the median of the experimental value if the experimental standard deviation variations did not exceed 0.5 log.[92]

Machine Learning Models. For the machine learning model, the data sets were divided into training (80%) and test (20%) subsets. The ChemProp GNN model[36] was trained without hyperparameters optimization and validated on the internal test set. Training parameters were defined as follows: epochs = 100, depth = 3, batch_size = 64, hidden_size = 300, and metric = rmse.

Retraining the GNN models with the full data set (100%) instead of 80% led to small performance improvements for log *P*, solubility, and hERG, with success rates increasing by up to 9% (Table S3). However, for permeability (for which we had the smallest data set) and log *D*, no significant improvement was observed, and in some cases, the success rate decreased slightly by 8%. The excluded 20% of the data was reserved for validation purposes to evaluate the model's performance on unseen data, a standard practice to prevent overfitting and ensure generalizability.

We also examined the overlap between the GNN training set and the compounds used in the collective intelligence questionnaire. Minimal overlap was identified for two endpoints, namely solubility and hERG (Table S3). For log *P*, permeability, and log *D*, no overlap was detected between the training data and the questionnaire compounds, ensuring that the GNN's predictions were not influenced by prior exposure to these molecules.

Performance Metrics. To assess the performance of our models, we employed the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE) (Table S2). *R*-squared measures the effectiveness of a model in explaining the variation in the dependent variable. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, with values ranging from 0 to 1. A value closer to 1 signifies a higher degree of model accuracy. RMSE evaluates the differences between predicted and actual values, emphasizing larger errors by squaring them before computing the average. This metric is particularly useful in scenarios where large deviations are especially undesirable. MAE, on the other hand, assesses the precision of a regression model. Unlike RMSE, MAE is less influenced by outliers or significant errors, as it calculates the simple average of the absolute differences between predicted and observed values.

■ ASSOCIATED CONTENT

Data Availability Statement

The complete data set from this study, encompassing survey responses, as well as the survey and molecules used to train and test the models, is accessible on GitHub at https://github.com/Sanofi-Public/IDD-Collective-Intelligence. We ensured ethical compliance as feedback from all participants remain anonymous.

Data Availability Statement

We have made the trained models and all associated code used to from data analysis and generation publicly available under an MIT license at https://github.com/Sanofi-Public/IDD-Collective-Intelligence. For ease of integration into chem-informatics workflows, a Conda package is provided. These neural network models were developed utilizing the Chemprop library, version 1.7.0.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.4c03066.

> We have included a list of SI items and a brief description of each file including the file type extension. The files include: Questionnary_Presented.pdf, a PDF of the PowerPoint containing all presented questions from the collective intelligence exercise using Pigeonhole; Questionnary_Answers.pdf, a PDF of the same PowerPoint with the correct answers highlighted (PDF)
>
> Compounds_Experimentals.csv, a CSV file with experimental data for compounds, detailing properties like SMILES, log *P*, log *S*, Papp, and hERG IC_{50} values with corresponding units and levels; and Compounds_Structures.csv, a CSV file listing the SMILES notation for the structures of the compounds alongside their compound IDs (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

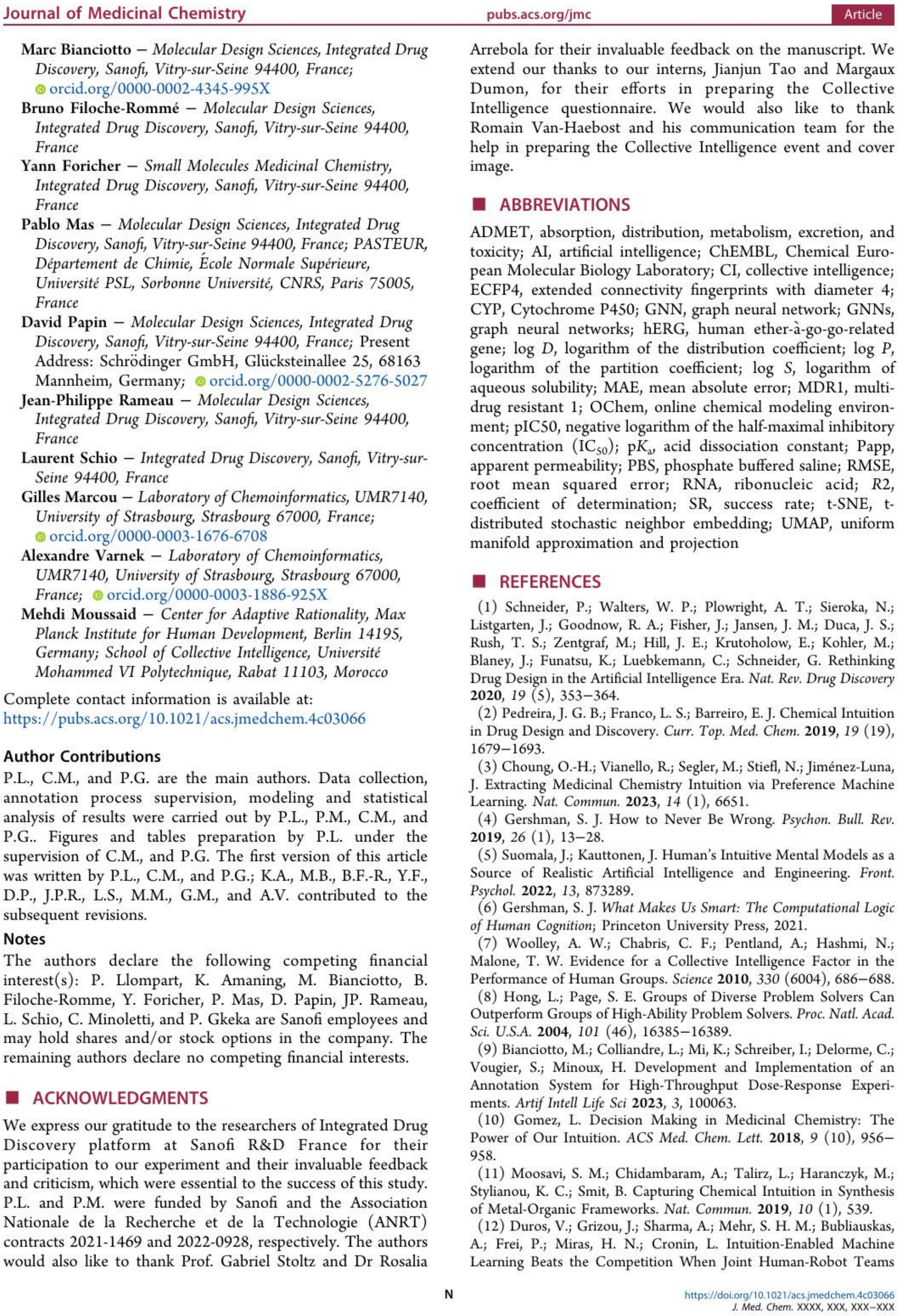
Pierre Llompart − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France; Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg 67000, France;* orcid.org/0000-0002-7751-9893; Email: Pierre.Llompart@sanofi.com

Claire Minoletti − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France;* orcid.org/0000-0002-0161-9820; Email: Claire.Minoletti@sanofi.com

Paraskevi Gkeka − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France;* orcid.org/0000-0002-0752-3539; Email: Paraskevi.Gkeka@sanofi.com

Authors

Kwame Amaning − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France*



Marc Bianciotto − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France;* orcid.org/0000-0002-4345-995X

Bruno Filoche-Rommé − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France*

Yann Foricher − *Small Molecules Medicinal Chemistry, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France*

Pablo Mas − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France; PASTEUR, Département de Chimie, École Normale Supérieure, Université PSL, Sorbonne Université, CNRS, Paris 75005, France*

David Papin − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France;* Present Address: Schrödinger GmbH, Glücksteinallee 25, 68163 Mannheim, Germany; orcid.org/0000-0002-5276-5027

Jean-Philippe Rameau − *Molecular Design Sciences, Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France*

Laurent Schio − *Integrated Drug Discovery, Sanofi, Vitry-sur-Seine 94400, France*

Gilles Marcou − *Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg 67000, France;* orcid.org/0000-0003-1676-6708

Alexandre Varnek − *Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, Strasbourg 67000, France;* orcid.org/0000-0003-1886-925X

Mehdi Moussaid − *Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin 14195, Germany; School of Collective Intelligence, Université Mohammed VI Polytechnique, Rabat 11103, Morocco*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jmedchem.4c03066

Author Contributions

P.L., C.M., and P.G. are the main authors. Data collection, annotation process supervision, modeling and statistical analysis of results were carried out by P.L., P.M., C.M., and P.G.. Figures and tables preparation by P.L. under the supervision of C.M., and P.G. The first version of this article was written by P.L., C.M., and P.G.; K.A., M.B., B.F.-R., Y.F., D.P., J.P.R., L.S., M.M., G.M., and A.V. contributed to the subsequent revisions.

Notes

The authors declare the following competing financial interest(s): P. Llompart, K. Amaning, M. Bianciotto, B. Filoche-Romme, Y. Foricher, P. Mas, D. Papin, JP. Rameau, L. Schio, C. Minoletti, and P. Gkeka are Sanofi employees and may hold shares and/or stock options in the company. The remaining authors declare no competing financial interests.

■ ACKNOWLEDGMENTS

We express our gratitude to the researchers of Integrated Drug Discovery platform at Sanofi R&D France for their participation to our experiment and their invaluable feedback and criticism, which were essential to the success of this study. P.L. and P.M. were funded by Sanofi and the Association Nationale de la Recherche et de la Technologie (ANRT) contracts 2021-1469 and 2022-0928, respectively. The authors would also like to thank Prof. Gabriel Stoltz and Dr Rosalia Arrebola for their invaluable feedback on the manuscript. We extend our thanks to our interns, Jianjun Tao and Margaux Dumon, for their efforts in preparing the Collective Intelligence questionnaire. We would also like to thank Romain Van-Haebost and his communication team for the help in preparing the Collective Intelligence event and cover image.

■ ABBREVIATIONS

ADMET, absorption, distribution, metabolism, excretion, and toxicity; AI, artificial intelligence; ChEMBL, Chemical European Molecular Biology Laboratory; CI, collective intelligence; ECFP4, extended connectivity fingerprints with diameter 4; CYP, Cytochrome P450; GNN, graph neural network; GNNs, graph neural networks; hERG, human ether-à-go-go-related gene; log *D*, logarithm of the distribution coefficient; log *P*, logarithm of the partition coefficient; log *S*, logarithm of aqueous solubility; MAE, mean absolute error; MDR1, multidrug resistant 1; OChem, online chemical modeling environment; pIC50, negative logarithm of the half-maximal inhibitory concentration (IC_{50}); pK_a, acid dissociation constant; Papp, apparent permeability; PBS, phosphate buffered saline; RMSE, root mean squared error; RNA, ribonucleic acid; *R*2, coefficient of determination; SR, success rate; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection

■ REFERENCES

(1) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discovery* **2020**, *19* (5), 353−364.

(2) Pedreira, J. G. B.; Franco, L. S.; Barreiro, E. J. Chemical Intuition in Drug Design and Discovery. *Curr. Top. Med. Chem.* **2019**, *19* (19), 1679−1693.

(3) Choung, O.-H.; Vianello, R.; Segler, M.; Stiefl, N.; Jiménez-Luna, J. Extracting Medicinal Chemistry Intuition via Preference Machine Learning. *Nat. Commun.* **2023**, *14* (1), 6651.

(4) Gershman, S. J. How to Never Be Wrong. *Psychon. Bull. Rev.* **2019**, *26* (1), 13−28.

(5) Suomala, J.; Kauttonen, J. Human's Intuitive Mental Models as a Source of Realistic Artificial Intelligence and Engineering. *Front. Psychol.* **2022**, *13*, 873289.

(6) Gershman, S. J. *What Makes Us Smart: The Computational Logic of Human Cognition*; Princeton University Press, 2021.

(7) Woolley, A. W.; Chabris, C. F.; Pentland, A.; Hashmi, N.; Malone, T. W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **2010**, *330* (6004), 686−688.

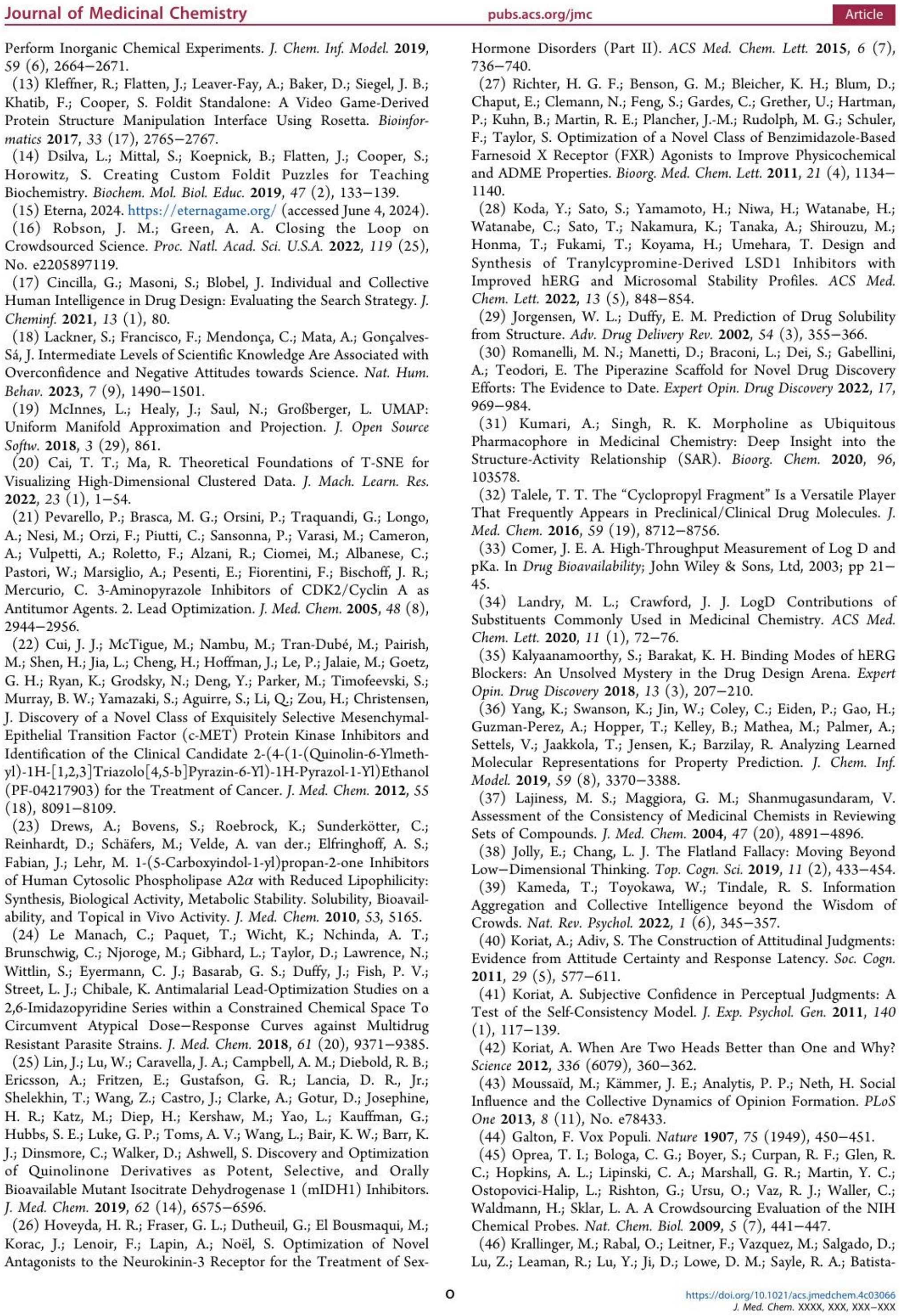
(8) Hong, L.; Page, S. E. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (46), 16385−16389.

(9) Bianciotto, M.; Colliandre, L.; Mi, K.; Schreiber, I.; Delorme, C.; Vougier, S.; Minoux, H. Development and Implementation of an Annotation System for High-Throughput Dose-Response Experiments. *Artif Intell Life Sci* **2023**, *3*, 100063.

(10) Gomez, L. Decision Making in Medicinal Chemistry: The Power of Our Intuition. *ACS Med. Chem. Lett.* **2018**, *9* (10), 956−958.

(11) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing Chemical Intuition in Synthesis of Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10* (1), 539.

(12) Duros, V.; Grizou, J.; Sharma, A.; Mehr, S. H. M.; Bubliauskas, A.; Frei, P.; Miras, H. N.; Cronin, L. Intuition-Enabled Machine Learning Beats the Competition When Joint Human-Robot Teams



Journal of Medicinal Chemistry pubs.acs.org/jmc Article

Perform Inorganic Chemical Experiments. *J. Chem. Inf. Model.* **2019**, *59* (6), 2664−2671.

(13) Kleffner, R.; Flatten, J.; Leaver-Fay, A.; Baker, D.; Siegel, J. B.; Khatib, F.; Cooper, S. Foldit Standalone: A Video Game-Derived Protein Structure Manipulation Interface Using Rosetta. *Bioinformatics* **2017**, *33* (17), 2765−2767.

(14) Dsilva, L.; Mittal, S.; Koepnick, B.; Flatten, J.; Cooper, S.; Horowitz, S. Creating Custom Foldit Puzzles for Teaching Biochemistry. *Biochem. Mol. Biol. Educ.* **2019**, *47* (2), 133−139.

(15) Eterna, 2024. https://eternagame.org/ (accessed June 4, 2024).

(16) Robson, J. M.; Green, A. A. Closing the Loop on Crowdsourced Science. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (25), No. e2205897119.

(17) Cincilla, G.; Masoni, S.; Blobel, J. Individual and Collective Human Intelligence in Drug Design: Evaluating the Search Strategy. *J. Cheminf.* **2021**, *13* (1), 80.

(18) Lackner, S.; Francisco, F.; Mendonça, C.; Mata, A.; Gonçalves-Sá, J. Intermediate Levels of Scientific Knowledge Are Associated with Overconfidence and Negative Attitudes towards Science. *Nat. Hum. Behav.* **2023**, *7* (9), 1490−1501.

(19) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861.

(20) Cai, T. T.; Ma, R. Theoretical Foundations of T-SNE for Visualizing High-Dimensional Clustered Data. *J. Mach. Learn. Res.* **2022**, *23* (1), 1−54.

(21) Pevarello, P.; Brasca, M. G.; Orsini, P.; Traquandi, G.; Longo, A.; Nesi, M.; Orzi, F.; Piutti, C.; Sansonna, P.; Varasi, M.; Cameron, A.; Vulpetti, A.; Roletto, F.; Alzani, R.; Ciomei, M.; Albanese, C.; Pastori, W.; Marsiglio, A.; Pesenti, E.; Fiorentini, F.; Bischoff, J. R.; Mercurio, C. 3-Aminopyrazole Inhibitors of CDK2/Cyclin A as Antitumor Agents. 2. Lead Optimization. *J. Med. Chem.* **2005**, *48* (8), 2944−2956.

(22) Cui, J. J.; McTigue, M.; Nambu, M.; Tran-Dubé, M.; Pairish, M.; Shen, H.; Jia, L.; Cheng, H.; Hoffman, J.; Le, P.; Jalaie, M.; Goetz, G. H.; Ryan, K.; Grodsky, N.; Deng, Y.; Parker, M.; Timofeevski, S.; Murray, B. W.; Yamazaki, S.; Aguirre, S.; Li, Q.; Zou, H.; Christensen, J. Discovery of a Novel Class of Exquisitely Selective Mesenchymal-Epithelial Transition Factor (c-MET) Protein Kinase Inhibitors and Identification of the Clinical Candidate 2-(4-(1-(Quinolin-6-Ylmethyl)-1H-[1,2,3]Triazolo[4,5-b]Pyrazin-6-Yl)-1H-Pyrazol-1-Yl)Ethanol (PF-04217903) for the Treatment of Cancer. *J. Med. Chem.* **2012**, *55* (18), 8091−8109.

(23) Drews, A.; Bovens, S.; Roebrock, K.; Sunderkötter, C.; Reinhardt, D.; Schäfers, M.; Velde, A. van der.; Elfringhoff, A. S.; Fabian, J.; Lehr, M. 1-(5-Carboxyindol-1-yl)propan-2-one Inhibitors of Human Cytosolic Phospholipase A2α with Reduced Lipophilicity: Synthesis, Biological Activity, Metabolic Stability. Solubility, Bioavailability, and Topical in Vivo Activity. *J. Med. Chem.* **2010**, *53*, 5165.

(24) Le Manach, C.; Paquet, T.; Wicht, K.; Nchinda, A. T.; Brunschwig, C.; Njoroge, M.; Gibhard, L.; Taylor, D.; Lawrence, N.; Wittlin, S.; Eyermann, C. J.; Basarab, G. S.; Duffy, J.; Fish, P. V.; Street, L. J.; Chibale, K. Antimalarial Lead-Optimization Studies on a 2,6-Imidazopyridine Series within a Constrained Chemical Space To Circumvent Atypical Dose−Response Curves against Multidrug Resistant Parasite Strains. *J. Med. Chem.* **2018**, *61* (20), 9371−9385.

(25) Lin, J.; Lu, W.; Caravella, J. A.; Campbell, A. M.; Diebold, R. B.; Ericsson, A.; Fritzen, E.; Gustafson, G. R.; Lancia, D. R., Jr.; Shelekhin, T.; Wang, Z.; Castro, J.; Clarke, A.; Gotur, D.; Josephine, H. R.; Katz, M.; Diep, H.; Kershaw, M.; Yao, L.; Kauffman, G.; Hubbs, S. E.; Luke, G. P.; Toms, A. V.; Wang, L.; Bair, K. W.; Barr, K. J.; Dinsmore, C.; Walker, D.; Ashwell, S. Discovery and Optimization of Quinolinone Derivatives as Potent, Selective, and Orally Bioavailable Mutant Isocitrate Dehydrogenase 1 (mIDH1) Inhibitors. *J. Med. Chem.* **2019**, *62* (14), 6575−6596.

(26) Hoveyda, H. R.; Fraser, G. L.; Dutheuil, G.; El Bousmaqui, M.; Korac, J.; Lenoir, F.; Lapin, A.; Noël, S. Optimization of Novel Antagonists to the Neurokinin-3 Receptor for the Treatment of Sex-Hormone Disorders (Part II). *ACS Med. Chem. Lett.* **2015**, *6* (7), 736−740.

(27) Richter, H. G. F.; Benson, G. M.; Bleicher, K. H.; Blum, D.; Chaput, E.; Clemann, N.; Feng, S.; Gardes, C.; Grether, U.; Hartman, P.; Kuhn, B.; Martin, R. E.; Plancher, J.-M.; Rudolph, M. G.; Schuler, F.; Taylor, S. Optimization of a Novel Class of Benzimidazole-Based Farnesoid X Receptor (FXR) Agonists to Improve Physicochemical and ADME Properties. *Bioorg. Med. Chem. Lett.* **2011**, *21* (4), 1134−1140.

(28) Koda, Y.; Sato, S.; Yamamoto, H.; Niwa, H.; Watanabe, H.; Watanabe, C.; Sato, T.; Nakamura, K.; Tanaka, A.; Shirouzu, M.; Honma, T.; Fukami, T.; Koyama, H.; Umehara, T. Design and Synthesis of Tranylcypromine-Derived LSD1 Inhibitors with Improved hERG and Microsomal Stability Profiles. *ACS Med. Chem. Lett.* **2022**, *13* (5), 848−854.

(29) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 355−366.

(30) Romanelli, M. N.; Manetti, D.; Braconi, L.; Dei, S.; Gabellini, A.; Teodori, E. The Piperazine Scaffold for Novel Drug Discovery Efforts: The Evidence to Date. *Expert Opin. Drug Discovery* **2022**, *17*, 969−984.

(31) Kumari, A.; Singh, R. K. Morpholine as Ubiquitous Pharmacophore in Medicinal Chemistry: Deep Insight into the Structure-Activity Relationship (SAR). *Bioorg. Chem.* **2020**, *96*, 103578.

(32) Talele, T. T. The "Cyclopropyl Fragment" Is a Versatile Player That Frequently Appears in Preclinical/Clinical Drug Molecules. *J. Med. Chem.* **2016**, *59* (19), 8712−8756.

(33) Comer, J. E. A. High-Throughput Measurement of Log D and pKa. In *Drug Bioavailability*; John Wiley & Sons, Ltd, 2003; pp 21−45.

(34) Landry, M. L.; Crawford, J. J. LogD Contributions of Substituents Commonly Used in Medicinal Chemistry. *ACS Med. Chem. Lett.* **2020**, *11* (1), 72−76.

(35) Kalyaanamoorthy, S.; Barakat, K. H. Binding Modes of hERG Blockers: An Unsolved Mystery in the Drug Design Arena. *Expert Opin. Drug Discovery* **2018**, *13* (3), 207−210.

(36) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370−3388.

(37) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47* (20), 4891−4896.

(38) Jolly, E.; Chang, L. J. The Flatland Fallacy: Moving Beyond Low−Dimensional Thinking. *Top. Cogn. Sci.* **2019**, *11* (2), 433−454.

(39) Kameda, T.; Toyokawa, W.; Tindale, R. S. Information Aggregation and Collective Intelligence beyond the Wisdom of Crowds. *Nat. Rev. Psychol.* **2022**, *1* (6), 345−357.

(40) Koriat, A.; Adiv, S. The Construction of Attitudinal Judgments: Evidence from Attitude Certainty and Response Latency. *Soc. Cogn.* **2011**, *29* (5), 577−611.

(41) Koriat, A. Subjective Confidence in Perceptual Judgments: A Test of the Self-Consistency Model. *J. Exp. Psychol. Gen.* **2011**, *140* (1), 117−139.

(42) Koriat, A. When Are Two Heads Better than One and Why? *Science* **2012**, *336* (6079), 360−362.

(43) Moussaïd, M.; Kämmer, J. E.; Analytis, P. P.; Neth, H. Social Influence and the Collective Dynamics of Opinion Formation. *PLoS One* **2013**, *8* (11), No. e78433.

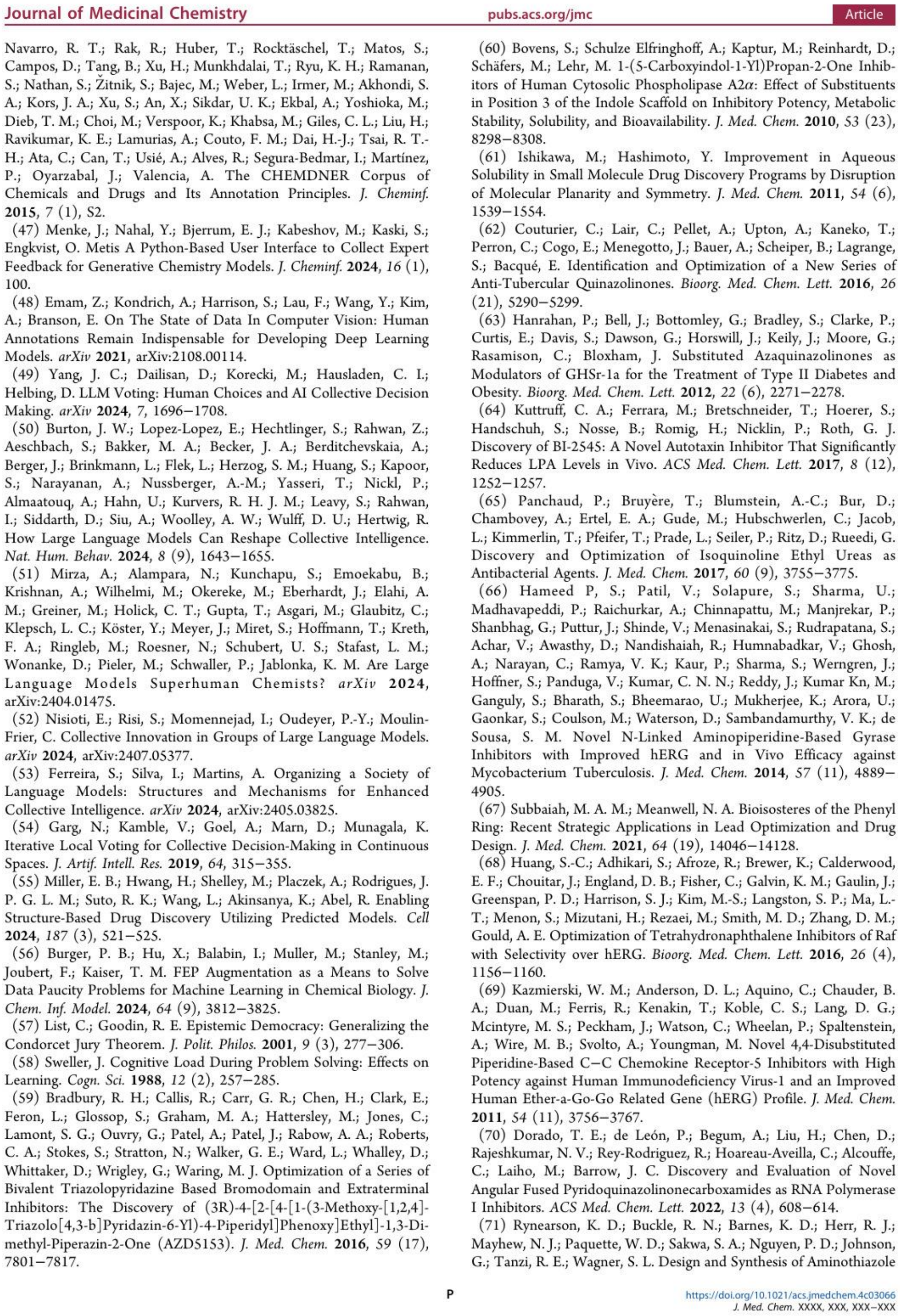
(44) Galton, F. Vox Populi. *Nature* **1907**, *75* (1949), 450−451.

(45) Oprea, T. I.; Bologa, C. G.; Boyer, S.; Curpan, R. F.; Glen, R. C.; Hopkins, A. L.; Lipinski, C. A.; Marshall, G. R.; Martin, Y. C.; Ostopovici-Halip, L.; Rishton, G.; Ursu, O.; Vaz, R. J.; Waller, C.; Waldmann, H.; Sklar, L. A. A Crowdsourcing Evaluation of the NIH Chemical Probes. *Nat. Chem. Biol.* **2009**, *5* (7), 441−447.

(46) Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; Sayle, R. A.; Batista-

o

https://doi.org/10.1021/acs.jmedchem.4c03066
J. Med. Chem. XXXX, XXX, XXX−XXX



Navarro, R. T.; Rak, R.; Huber, T.; Rocktäschel, T.; Matos, S.; Campos, D.; Tang, B.; Xu, H.; Munkhdalai, T.; Ryu, K. H.; Ramanan, S.; Nathan, S.; Žitnik, S.; Bajec, M.; Weber, L.; Irmer, M.; Akhondi, S. A.; Kors, J. A.; Xu, S.; An, X.; Sikdar, U. K.; Ekbal, A.; Yoshioka, M.; Dieb, T. M.; Choi, M.; Verspoor, K.; Khabsa, M.; Giles, C. L.; Liu, H.; Ravikumar, K. E.; Lamurias, A.; Couto, F. M.; Dai, H.-J.; Tsai, R. T.-H.; Ata, C.; Can, T.; Usié, A.; Alves, R.; Segura-Bedmar, I.; Martínez, P.; Oyarzabal, J.; Valencia, A. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *J. Cheminf.* **2015**, *7* (1), S2.

(47) Menke, J.; Nahal, Y.; Bjerrum, E. J.; Kabeshov, M.; Kaski, S.; Engkvist, O. Metis A Python-Based User Interface to Collect Expert Feedback for Generative Chemistry Models. *J. Cheminf.* **2024**, *16* (1), 100.

(48) Emam, Z.; Kondrich, A.; Harrison, S.; Lau, F.; Wang, Y.; Kim, A.; Branson, E. On The State of Data In Computer Vision: Human Annotations Remain Indispensable for Developing Deep Learning Models. *arXiv* **2021**, arXiv:2108.00114.

(49) Yang, J. C.; Dailisan, D.; Korecki, M.; Hausladen, C. I.; Helbing, D. LLM Voting: Human Choices and AI Collective Decision Making. *arXiv* **2024**, *7*, 1696−1708.

(50) Burton, J. W.; Lopez-Lopez, E.; Hechtlinger, S.; Rahwan, Z.; Aeschbach, S.; Bakker, M. A.; Becker, J. A.; Berditchevskaia, A.; Berger, J.; Brinkmann, L.; Flek, L.; Herzog, S. M.; Huang, S.; Kapoor, S.; Narayanan, A.; Nussberger, A.-M.; Yasseri, T.; Nickl, P.; Almaatouq, A.; Hahn, U.; Kurvers, R. H. J. M.; Leavy, S.; Rahwan, I.; Siddarth, D.; Siu, A.; Woolley, A. W.; Wulff, D. U.; Hertwig, R. How Large Language Models Can Reshape Collective Intelligence. *Nat. Hum. Behav.* **2024**, *8* (9), 1643−1655.

(51) Mirza, A.; Alampara, N.; Kunchapu, S.; Emoekabu, B.; Krishnan, A.; Wilhelmi, M.; Okereke, M.; Eberhardt, J.; Elahi, A. M.; Greiner, M.; Holick, C. T.; Gupta, T.; Asgari, M.; Glaubitz, C.; Klepsch, L. C.; Köster, Y.; Meyer, J.; Miret, S.; Hoffmann, T.; Kreth, F. A.; Ringleb, M.; Roesner, N.; Schubert, U. S.; Stafast, L. M.; Wonanke, D.; Pieler, M.; Schwaller, P.; Jablonka, K. M. Are Large Language Models Superhuman Chemists? *arXiv* **2024**, arXiv:2404.01475.

(52) Nisioti, E.; Risi, S.; Momennejad, I.; Oudeyer, P.-Y.; Moulin-Frier, C. Collective Innovation in Groups of Large Language Models. *arXiv* **2024**, arXiv:2407.05377.

(53) Ferreira, S.; Silva, I.; Martins, A. Organizing a Society of Language Models: Structures and Mechanisms for Enhanced Collective Intelligence. *arXiv* **2024**, arXiv:2405.03825.

(54) Garg, N.; Kamble, V.; Goel, A.; Marn, D.; Munagala, K. Iterative Local Voting for Collective Decision-Making in Continuous Spaces. *J. Artif. Intell. Res.* **2019**, *64*, 315−355.

(55) Miller, E. B.; Hwang, H.; Shelley, M.; Placzek, A.; Rodrigues, J. P. G. L. M.; Suto, R. K.; Wang, L.; Akinsanya, K.; Abel, R. Enabling Structure-Based Drug Discovery Utilizing Predicted Models. *Cell* **2024**, *187* (3), 521−525.

(56) Burger, P. B.; Hu, X.; Balabin, I.; Muller, M.; Stanley, M.; Joubert, F.; Kaiser, T. M. FEP Augmentation as a Means to Solve Data Paucity Problems for Machine Learning in Chemical Biology. *J. Chem. Inf. Model.* **2024**, *64* (9), 3812−3825.

(57) List, C.; Goodin, R. E. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *J. Polit. Philos.* **2001**, *9* (3), 277−306.

(58) Sweller, J. Cognitive Load During Problem Solving: Effects on Learning. *Cogn. Sci.* **1988**, *12* (2), 257−285.

(59) Bradbury, R. H.; Callis, R.; Carr, G. R.; Chen, H.; Clark, E.; Feron, L.; Glossop, S.; Graham, M. A.; Hattersley, M.; Jones, C.; Lamont, S. G.; Ouvry, G.; Patel, A.; Patel, J.; Rabow, A. A.; Roberts, C. A.; Stokes, S.; Stratton, N.; Walker, G. E.; Ward, L.; Whalley, D.; Whittaker, D.; Wrigley, G.; Waring, M. J. Optimization of a Series of Bivalent Triazolopyridazine Based Bromodomain and Extraterminal Inhibitors: The Discovery of (3R)-4-[2-[4-[1-(3-Methoxy-[1,2,4]-Triazolo[4,3-b]Pyridazin-6-Yl)-4-Piperidyl]Phenoxy]Ethyl]-1,3-Dimethyl-Piperazin-2-One (AZD5153). *J. Med. Chem.* **2016**, *59* (17), 7801−7817.

(60) Bovens, S.; Schulze Elfringhoff, A.; Kaptur, M.; Reinhardt, D.; Schäfers, M.; Lehr, M. 1-(5-Carboxyindol-1-Yl)Propan-2-One Inhibitors of Human Cytosolic Phospholipase A2α: Effect of Substituents in Position 3 of the Indole Scaffold on Inhibitory Potency, Metabolic Stability, Solubility, and Bioavailability. *J. Med. Chem.* **2010**, *53* (23), 8298−8308.

(61) Ishikawa, M.; Hashimoto, Y. Improvement in Aqueous Solubility in Small Molecule Drug Discovery Programs by Disruption of Molecular Planarity and Symmetry. *J. Med. Chem.* **2011**, *54* (6), 1539−1554.

(62) Couturier, C.; Lair, C.; Pellet, A.; Upton, A.; Kaneko, T.; Perron, C.; Cogo, E.; Menegotto, J.; Bauer, A.; Scheiper, B.; Lagrange, S.; Bacqué, E. Identification and Optimization of a New Series of Anti-Tubercular Quinazolinones. *Bioorg. Med. Chem. Lett.* **2016**, *26* (21), 5290−5299.

(63) Hanrahan, P.; Bell, J.; Bottomley, G.; Bradley, S.; Clarke, P.; Curtis, E.; Davis, S.; Dawson, G.; Horswill, J.; Keily, J.; Moore, G.; Rasamison, C.; Bloxham, J. Substituted Azaquinazolinones as Modulators of GHSr-1a for the Treatment of Type II Diabetes and Obesity. *Bioorg. Med. Chem. Lett.* **2012**, *22* (6), 2271−2278.

(64) Kuttruff, C. A.; Ferrara, M.; Bretschneider, T.; Hoerer, S.; Handschuh, S.; Nosse, B.; Romig, H.; Nicklin, P.; Roth, G. J. Discovery of BI-2545: A Novel Autotaxin Inhibitor That Significantly Reduces LPA Levels in Vivo. *ACS Med. Chem. Lett.* **2017**, *8* (12), 1252−1257.

(65) Panchaud, P.; Bruyère, T.; Blumstein, A.-C.; Bur, D.; Chambovey, A.; Ertel, E. A.; Gude, M.; Hubschwerlen, C.; Jacob, L.; Kimmerlin, T.; Pfeifer, T.; Prade, L.; Seiler, P.; Ritz, D.; Rueedi, G. Discovery and Optimization of Isoquinoline Ethyl Ureas as Antibacterial Agents. *J. Med. Chem.* **2017**, *60* (9), 3755−3775.

(66) Hameed P, S.; Patil, V.; Solapure, S.; Sharma, U.; Madhavapeddi, P.; Raichurkar, A.; Chinnapattu, M.; Manjrekar, P.; Shanbhag, G.; Puttur, J.; Shinde, V.; Menasinakai, S.; Rudrapatana, S.; Achar, V.; Awasthy, D.; Nandishaiah, R.; Humnabadkar, V.; Ghosh, A.; Narayan, C.; Ramya, V. K.; Kaur, P.; Sharma, S.; Werngren, J.; Hoffner, S.; Panduga, V.; Kumar, C. N. N.; Reddy, J.; Kumar Kn, M.; Ganguly, S.; Bharath, S.; Bheemarao, U.; Mukherjee, K.; Arora, U.; Gaonkar, S.; Coulson, M.; Waterson, D.; Sambandamurthy, V. K.; de Sousa, S. M. Novel N-Linked Aminopiperidine-Based Gyrase Inhibitors with Improved hERG and in Vivo Efficacy against Mycobacterium Tuberculosis. *J. Med. Chem.* **2014**, *57* (11), 4889−4905.

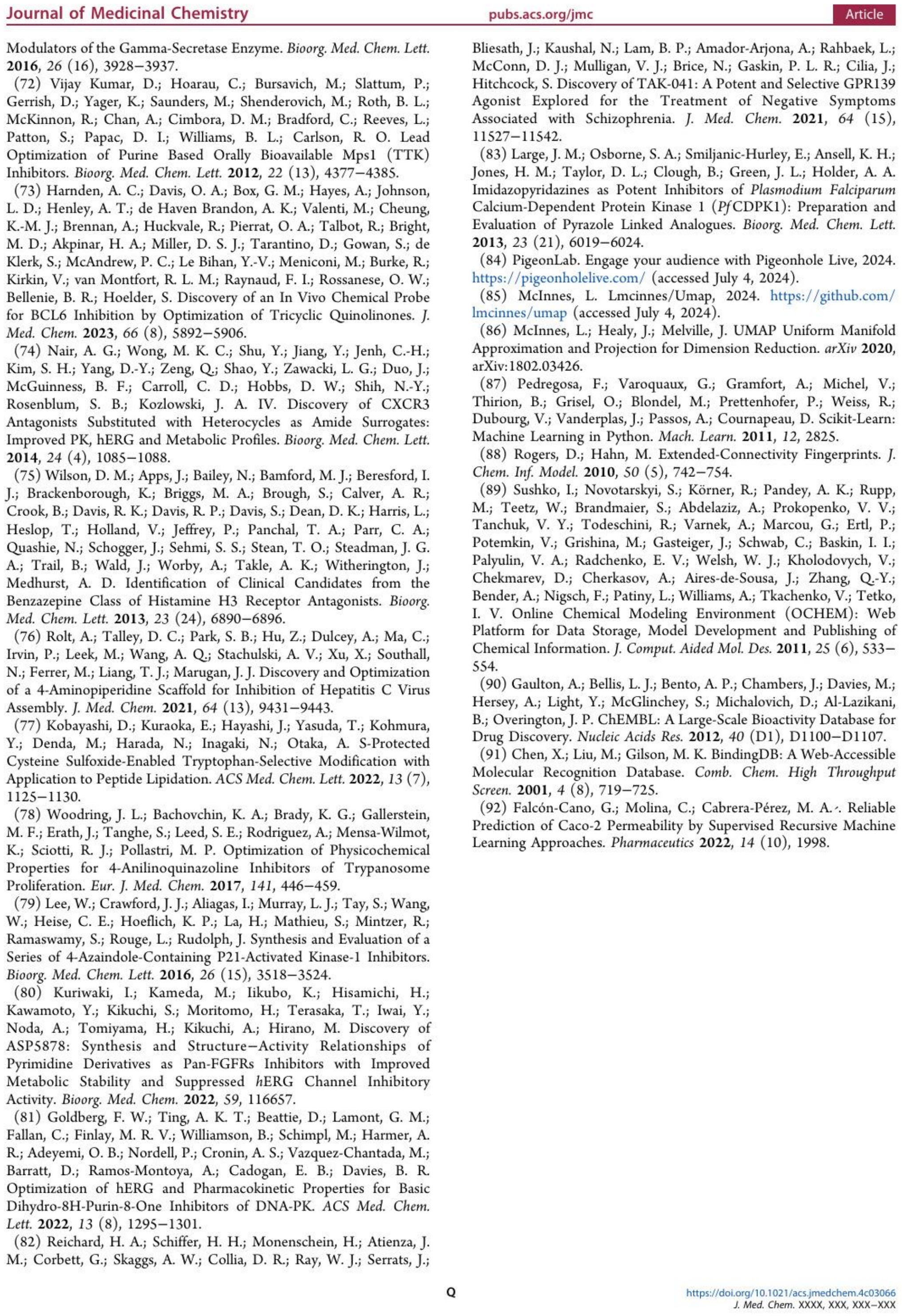
(67) Subbaiah, M. A. M.; Meanwell, N. A. Bioisosteres of the Phenyl Ring: Recent Strategic Applications in Lead Optimization and Drug Design. *J. Med. Chem.* **2021**, *64* (19), 14046−14128.

(68) Huang, S.-C.; Adhikari, S.; Afroze, R.; Brewer, K.; Calderwood, E. F.; Chouitar, J.; England, D. B.; Fisher, C.; Galvin, K. M.; Gaulin, J.; Greenspan, P. D.; Harrison, S. J.; Kim, M.-S.; Langston, S. P.; Ma, L.-T.; Menon, S.; Mizutani, H.; Rezaei, M.; Smith, M. D.; Zhang, D. M.; Gould, A. E. Optimization of Tetrahydronaphthalene Inhibitors of Raf with Selectivity over hERG. *Bioorg. Med. Chem. Lett.* **2016**, *26* (4), 1156−1160.

(69) Kazmierski, W. M.; Anderson, D. L.; Aquino, C.; Chauder, B. A.; Duan, M.; Ferris, R.; Kenakin, T.; Koble, C. S.; Lang, D. G.; Mcintyre, M. S.; Peckham, J.; Watson, C.; Wheelan, P.; Spaltenstein, A.; Wire, M. B.; Svolto, A.; Youngman, M. Novel 4,4-Disubstituted Piperidine-Based C−C Chemokine Receptor-5 Inhibitors with High Potency against Human Immunodeficiency Virus-1 and an Improved Human Ether-a-Go-Go Related Gene (hERG) Profile. *J. Med. Chem.* **2011**, *54* (11), 3756−3767.

(70) Dorado, T. E.; de León, P.; Begum, A.; Liu, H.; Chen, D.; Rajeshkumar, N. V.; Rey-Rodriguez, R.; Hoareau-Aveilla, C.; Alcouffe, C.; Laiho, M.; Barrow, J. C. Discovery and Evaluation of Novel Angular Fused Pyridoquinazolinonecarboxamides as RNA Polymerase I Inhibitors. *ACS Med. Chem. Lett.* **2022**, *13* (4), 608−614.

(71) Rynearson, K. D.; Buckle, R. N.; Barnes, K. D.; Herr, R. J.; Mayhew, N. J.; Paquette, W. D.; Sakwa, S. A.; Nguyen, P. D.; Johnson, G.; Tanzi, R. E.; Wagner, S. L. Design and Synthesis of Aminothiazole



Modulators of the Gamma-Secretase Enzyme. *Bioorg. Med. Chem. Lett.* **2016**, *26* (16), 3928−3937.

(72) Vijay Kumar, D.; Hoarau, C.; Bursavich, M.; Slattum, P.; Gerrish, D.; Yager, K.; Saunders, M.; Shenderovich, M.; Roth, B. L.; McKinnon, R.; Chan, A.; Cimbora, D. M.; Bradford, C.; Reeves, L.; Patton, S.; Papac, D. I.; Williams, B. L.; Carlson, R. O. Lead Optimization of Purine Based Orally Bioavailable Mps1 (TTK) Inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22* (13), 4377−4385.

(73) Harnden, A. C.; Davis, O. A.; Box, G. M.; Hayes, A.; Johnson, L. D.; Henley, A. T.; de Haven Brandon, A. K.; Valenti, M.; Cheung, K.-M. J.; Brennan, A.; Huckvale, R.; Pierrat, O. A.; Talbot, R.; Bright, M. D.; Akpinar, H. A.; Miller, D. S. J.; Tarantino, D.; Gowan, S.; de Klerk, S.; McAndrew, P. C.; Le Bihan, Y.-V.; Meniconi, M.; Burke, R.; Kirkin, V.; van Montfort, R. L. M.; Raynaud, F. I.; Rossanese, O. W.; Bellenie, B. R.; Hoelder, S. Discovery of an In Vivo Chemical Probe for BCL6 Inhibition by Optimization of Tricyclic Quinolinones. *J. Med. Chem.* **2023**, *66* (8), 5892−5906.

(74) Nair, A. G.; Wong, M. K. C.; Shu, Y.; Jiang, Y.; Jenh, C.-H.; Kim, S. H.; Yang, D.-Y.; Zeng, Q.; Shao, Y.; Zawacki, L. G.; Duo, J.; McGuinness, B. F.; Carroll, C. D.; Hobbs, D. W.; Shih, N.-Y.; Rosenblum, S. B.; Kozlowski, J. A. IV. Discovery of CXCR3 Antagonists Substituted with Heterocycles as Amide Surrogates: Improved PK, hERG and Metabolic Profiles. *Bioorg. Med. Chem. Lett.* **2014**, *24* (4), 1085−1088.

(75) Wilson, D. M.; Apps, J.; Bailey, N.; Bamford, M. J.; Beresford, I. J.; Brackenborough, K.; Briggs, M. A.; Brough, S.; Calver, A. R.; Crook, B.; Davis, R. K.; Davis, R. P.; Davis, S.; Dean, D. K.; Harris, L.; Heslop, T.; Holland, V.; Jeffrey, P.; Panchal, T. A.; Parr, C. A.; Quashie, N.; Schogger, J.; Sehmi, S. S.; Stean, T. O.; Steadman, J. G. A.; Trail, B.; Wald, J.; Worby, A.; Takle, A. K.; Witherington, J.; Medhurst, A. D. Identification of Clinical Candidates from the Benzazepine Class of Histamine H3 Receptor Antagonists. *Bioorg. Med. Chem. Lett.* **2013**, *23* (24), 6890−6896.

(76) Rolt, A.; Talley, D. C.; Park, S. B.; Hu, Z.; Dulcey, A.; Ma, C.; Irvin, P.; Leek, M.; Wang, A. Q.; Stachulski, A. V.; Xu, X.; Southall, N.; Ferrer, M.; Liang, T. J.; Marugan, J. J. Discovery and Optimization of a 4-Aminopiperidine Scaffold for Inhibition of Hepatitis C Virus Assembly. *J. Med. Chem.* **2021**, *64* (13), 9431−9443.

(77) Kobayashi, D.; Kuraoka, E.; Hayashi, J.; Yasuda, T.; Kohmura, Y.; Denda, M.; Harada, N.; Inagaki, N.; Otaka, A. S-Protected Cysteine Sulfoxide-Enabled Tryptophan-Selective Modification with Application to Peptide Lipidation. *ACS Med. Chem. Lett.* **2022**, *13* (7), 1125−1130.

(78) Woodring, J. L.; Bachovchin, K. A.; Brady, K. G.; Gallerstein, M. F.; Erath, J.; Tanghe, S.; Leed, S. E.; Rodriguez, A.; Mensa-Wilmot, K.; Sciotti, R. J.; Pollastri, M. P. Optimization of Physicochemical Properties for 4-Anilinoquinazoline Inhibitors of Trypanosome Proliferation. *Eur. J. Med. Chem.* **2017**, *141*, 446−459.

(79) Lee, W.; Crawford, J. J.; Aliagas, I.; Murray, L. J.; Tay, S.; Wang, W.; Heise, C. E.; Hoeflich, K. P.; La, H.; Mathieu, S.; Mintzer, R.; Ramaswamy, S.; Rouge, L.; Rudolph, J. Synthesis and Evaluation of a Series of 4-Azaindole-Containing P21-Activated Kinase-1 Inhibitors. *Bioorg. Med. Chem. Lett.* **2016**, *26* (15), 3518−3524.

(80) Kuriwaki, I.; Kameda, M.; Iikubo, K.; Hisamichi, H.; Kawamoto, Y.; Kikuchi, S.; Moritomo, H.; Terasaka, T.; Iwai, Y.; Noda, A.; Tomiyama, H.; Kikuchi, A.; Hirano, M. Discovery of ASP5878: Synthesis and Structure−Activity Relationships of Pyrimidine Derivatives as Pan-FGFRs Inhibitors with Improved Metabolic Stability and Suppressed *h*ERG Channel Inhibitory Activity. *Bioorg. Med. Chem.* **2022**, *59*, 116657.

(81) Goldberg, F. W.; Ting, A. K. T.; Beattie, D.; Lamont, G. M.; Fallan, C.; Finlay, M. R. V.; Williamson, B.; Schimpl, M.; Harmer, A. R.; Adeyemi, O. B.; Nordell, P.; Cronin, A. S.; Vazquez-Chantada, M.; Barratt, D.; Ramos-Montoya, A.; Cadogan, E. B.; Davies, B. R. Optimization of hERG and Pharmacokinetic Properties for Basic Dihydro-8H-Purin-8-One Inhibitors of DNA-PK. *ACS Med. Chem. Lett.* **2022**, *13* (8), 1295−1301.

(82) Reichard, H. A.; Schiffer, H. H.; Monenschein, H.; Atienza, J. M.; Corbett, G.; Skaggs, A. W.; Collia, D. R.; Ray, W. J.; Serrats, J.; Bliesath, J.; Kaushal, N.; Lam, B. P.; Amador-Arjona, A.; Rahbaek, L.; McConn, D. J.; Mulligan, V. J.; Brice, N.; Gaskin, P. L. R.; Cilia, J.; Hitchcock, S. Discovery of TAK-041: A Potent and Selective GPR139 Agonist Explored for the Treatment of Negative Symptoms Associated with Schizophrenia. *J. Med. Chem.* **2021**, *64* (15), 11527−11542.

(83) Large, J. M.; Osborne, S. A.; Smiljanic-Hurley, E.; Ansell, K. H.; Jones, H. M.; Taylor, D. L.; Clough, B.; Green, J. L.; Holder, A. A. Imidazopyridazines as Potent Inhibitors of *Plasmodium Falciparum* Calcium-Dependent Protein Kinase 1 (*Pf*CDPK1): Preparation and Evaluation of Pyrazole Linked Analogues. *Bioorg. Med. Chem. Lett.* **2013**, *23* (21), 6019−6024.

(84) PigeonLab. Engage your audience with Pigeonhole Live, 2024. https://pigeonholelive.com/ (accessed July 4, 2024).

(85) McInnes, L. Lmcinnes/Umap, 2024. https://github.com/lmcinnes/umap (accessed July 4, 2024).

(86) McInnes, L.; Healy, J.; Melville, J. UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426.

(87) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *Mach. Learn.* **2011**, *12*, 2825.

(88) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(89) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25* (6), 533−554.

(90) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−D1107.

(91) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 719−725.

(92) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. A.·. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14* (10), 1998.

Supporting Information

Harnessing Medicinal Chemical Intuition from Collective Intelligence

Pierre Llompart[1, 2, *], Kwame Amaning[1], Marc Bianciotto[1], Bruno Filoche-Rommé[1], Yann Foricher[3], Pablo Mas[1,4], David Papin[1], Jean-Philippe Rameau[1], Laurent Schio[5], Gilles Marcou[2], Alexandre Varnek[2], Mehdi Moussaid[6,7], Claire Minoletti[1, *], Paraskevi Gkeka[1, *]

[1]Molecular Design Sciences, Integrated Drug Discovery, Sanofi, 1 Imp. Des Ateliers, 94400, Vitry-sur-Seine, France

[2]Laboratory of Chemoinformatics, UMR7140, University of Strasbourg, 67000, Strasbourg, France

[3]Small Molecules Medicinal Chemistry, Integrated Drug Discovery, Sanofi, 1 Imp. Des Ateliers, 94400, Vitry-sur-Seine, France

[4]PASTEUR, Département de chimie, École Normale Supérieure, Université PSL, Sorbonne Université, CNRS, 75005 Paris, France

[5]Integrated Drug Discovery, Sanofi, 1 Imp. Des Ateliers, 94400, Vitry-sur-Seine, France

[6]Center for Adaptive Rationality, Max Planck Institute for Human Development, 14195, Berlin, Germany

[7]School of Collective Intelligence, Université Mohammed VI Polytechnique, 11103, Rabat, Morocco

*Corresponding authors: Paraskevi.Gkeka@sanofi.com, Pierre.Llompart@sanofi.com, Claire.Minoletti@sanofi.com

Contents of SI

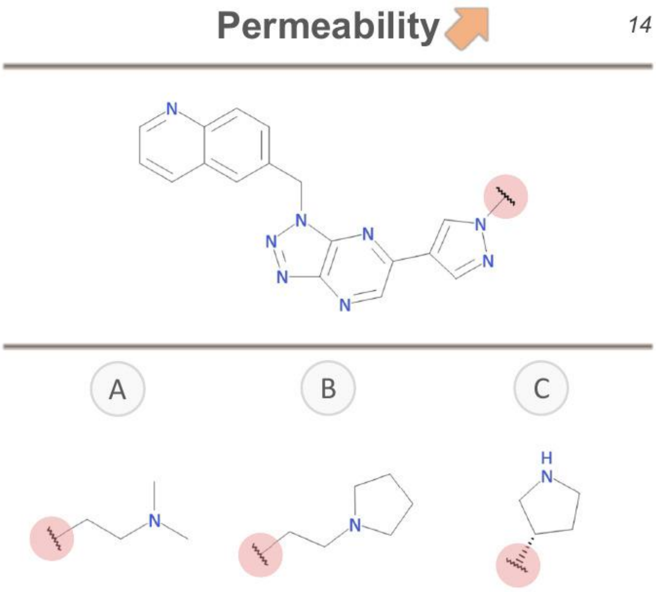


Figure S1: *Example of the way the questions were presented to the participants.* Each question had a title that corresponded to the endpoint of interest (top), one scaffold with an R-group substitution point to be replaced (middle) and three possible substituents (bottom).

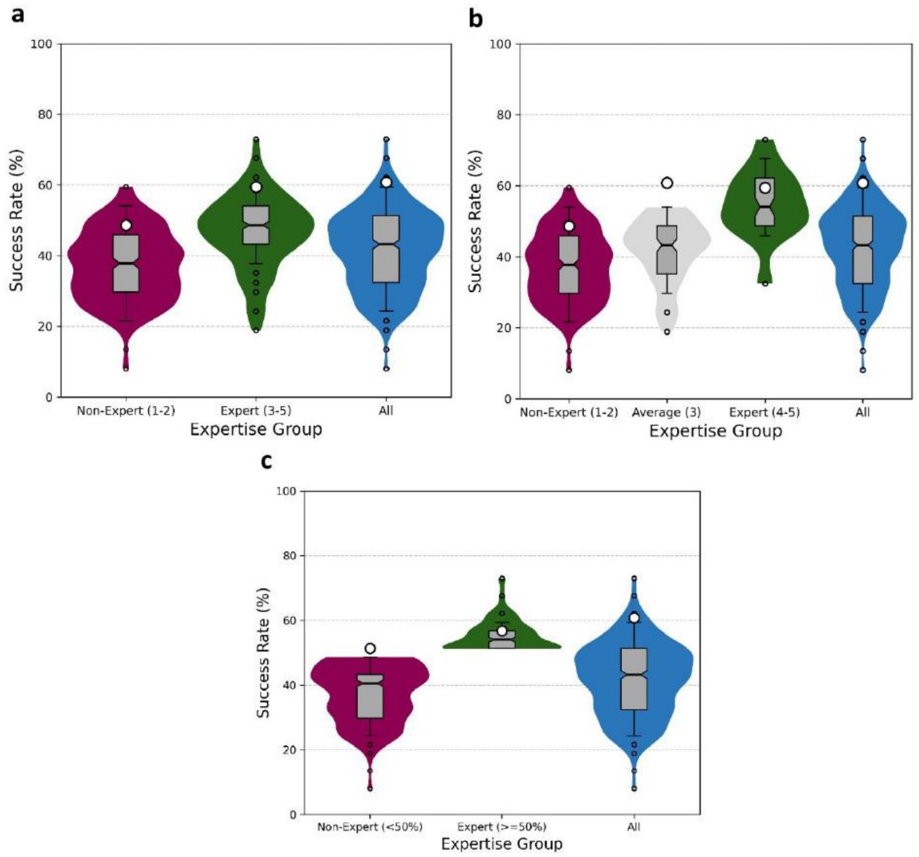


Figure S2: *Violin plots representing the success rate by expertise level in medicinal chemistry*. **a)** SR for groups 1-2 (low or no background), 3-5 (experts) and all the participants. The median is shown as a horizontal line across the thinnest part of the grey boxes. Alongside the boxes, error bars extend from the median line to cover the interquartile range. The collective or democratic SR are shown as white-filled circles. The outliers per group are depicted as small circles. **b)** SR for groups 1-2 (low or no background), 3 (averaged and mixed level), 4-5 (experts) and all the participants. **c)** SR by non-experts, *i.e.*, participants with personal SR less than 50%, and experts, *i.e.,* individuals with SR more than 50%. The violin plot of the SR of all participants is shown again here as a guide for the eye.

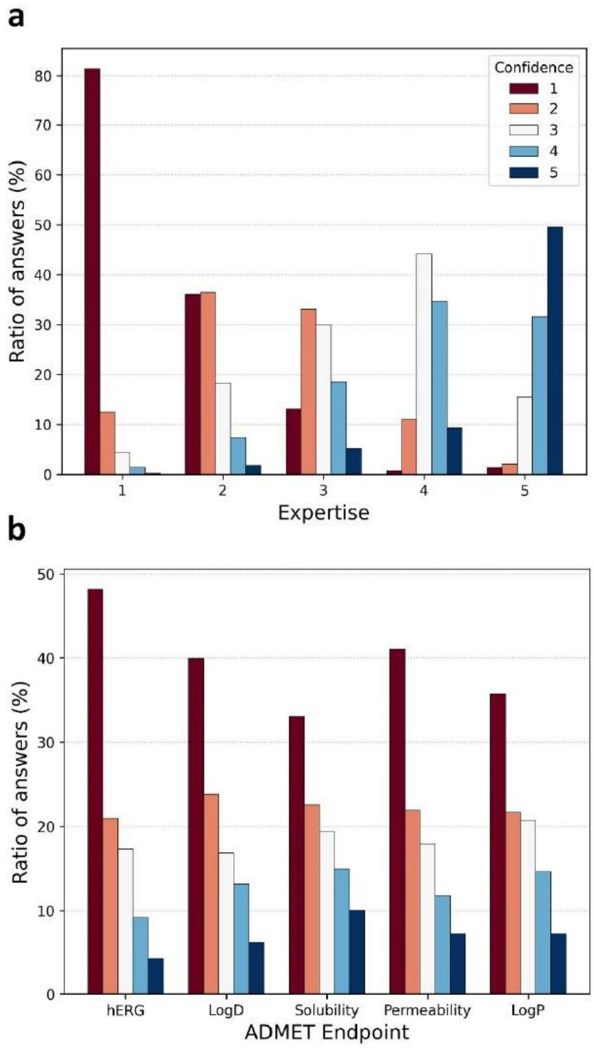


Figure S3: *Distribution of the proportion of answer per confidence level in function of the ADMET endpoint and self-labeled expertise level.* a) Bar plot of the ratio of answers per expertise level grouped by confidence from low (red) to high (blue). b) Bar plot of the ratio of answers per ADMET endpoint.

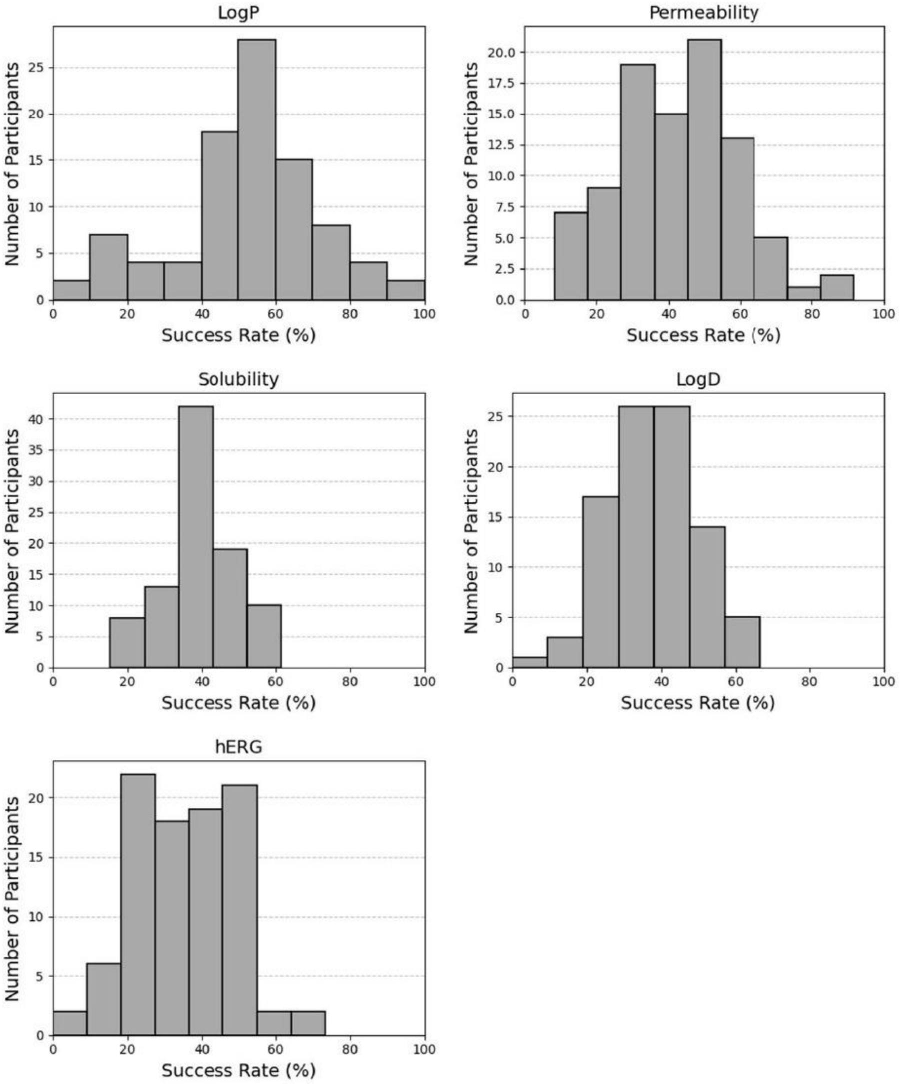


Figure S4: *Success rate distribution for all participants per ADMET endpoint.*

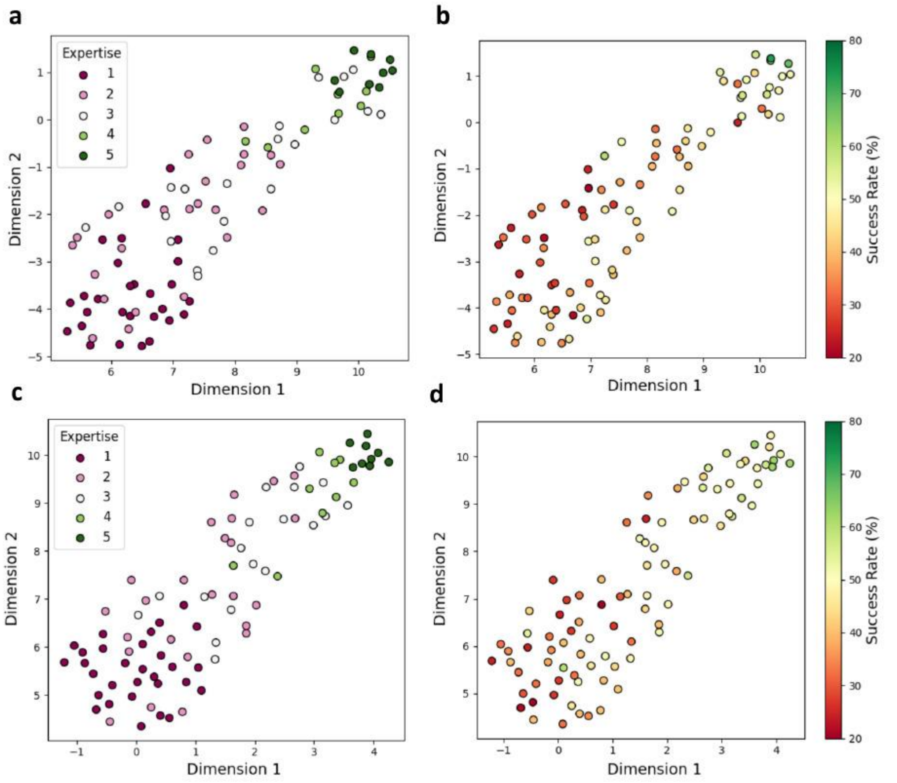


Figure S5: *UMAP of the participant space per session explored using the expertise level and the success rate.* Session one colored by a) expertise level and b) success rate. Session two colored by c) expertise level and d) success rate.

Note: We remind the reader that our experiment was performed in two sessions due to technical limitations of the software we used. More information can be found in the Experimental section of the main manuscript.

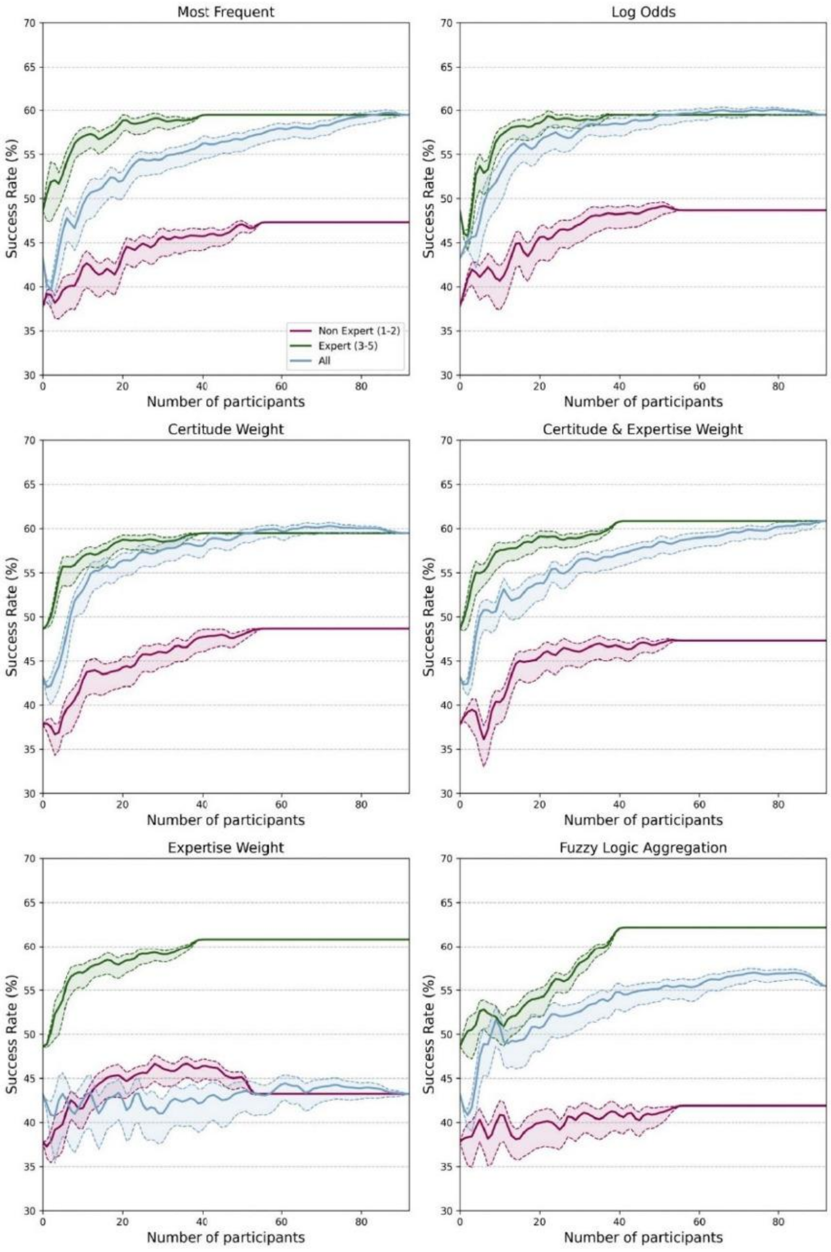


Figure S6: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method and for all endpoints.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue.

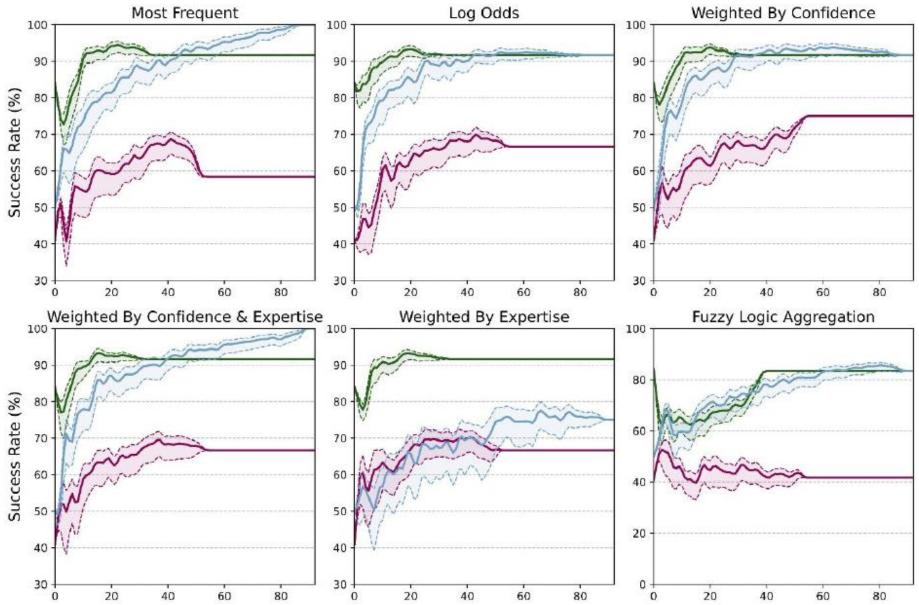


Figure S7: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method for the LogP endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue. Groups are colored as in Figure S6.

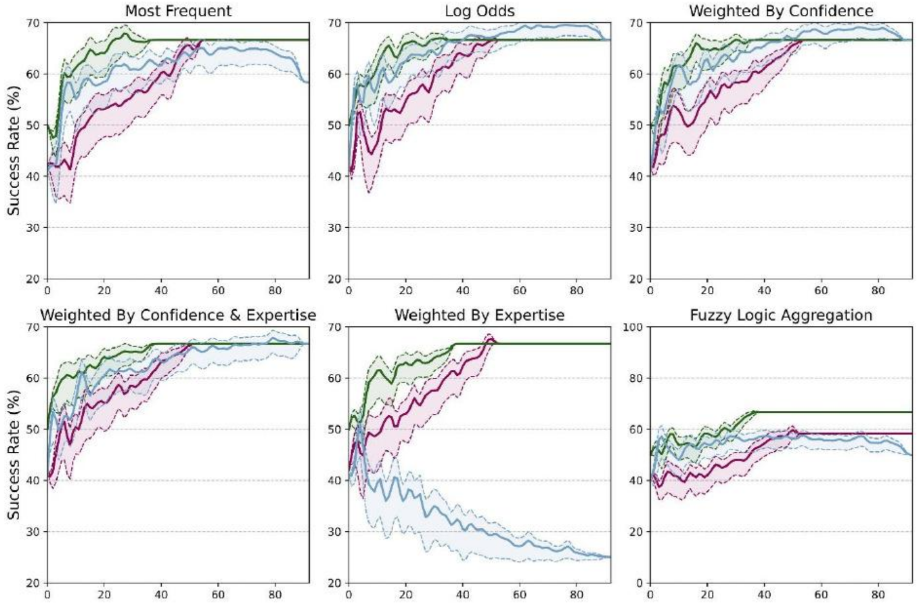


Figure S8: *Evolution of the collective success rate as a function of the number of participants in the population per aggregation method for the permeability endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue. Groups are colored as in Figure S6.

Note: In the "weighted by expertise" method, as the number of participants grows, the SR for all participants declines, whereas SR improves for experts and non-experts individually. This trend likely results from diverging assessments between experts and non-experts. These differences weaken the combined group's SR, as the aggregated answers fail to align with either group's specific preferences. Similar, though less pronounced, effects are seen for solubility and log D.

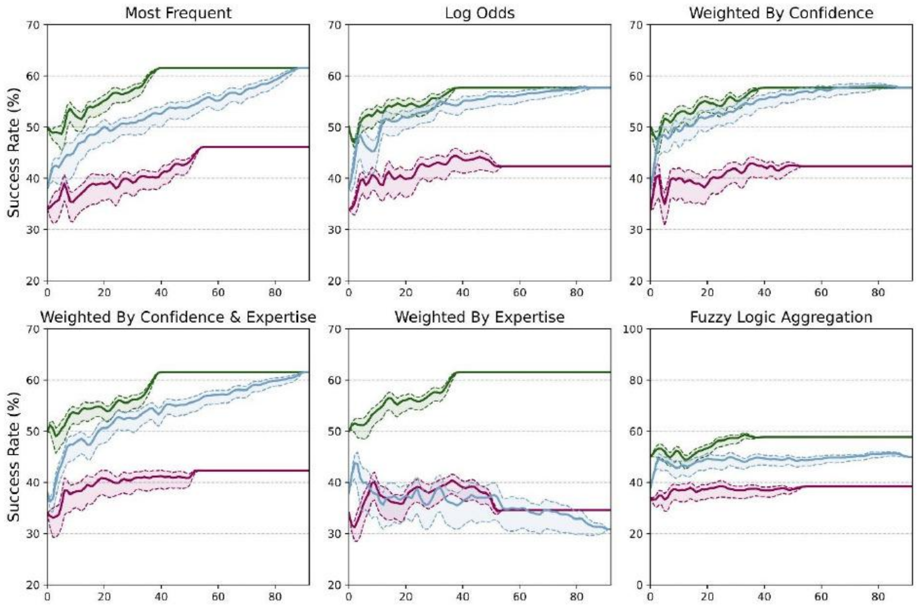


Figure S9: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the solubility endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue. Groups are colored as in Figure S6.

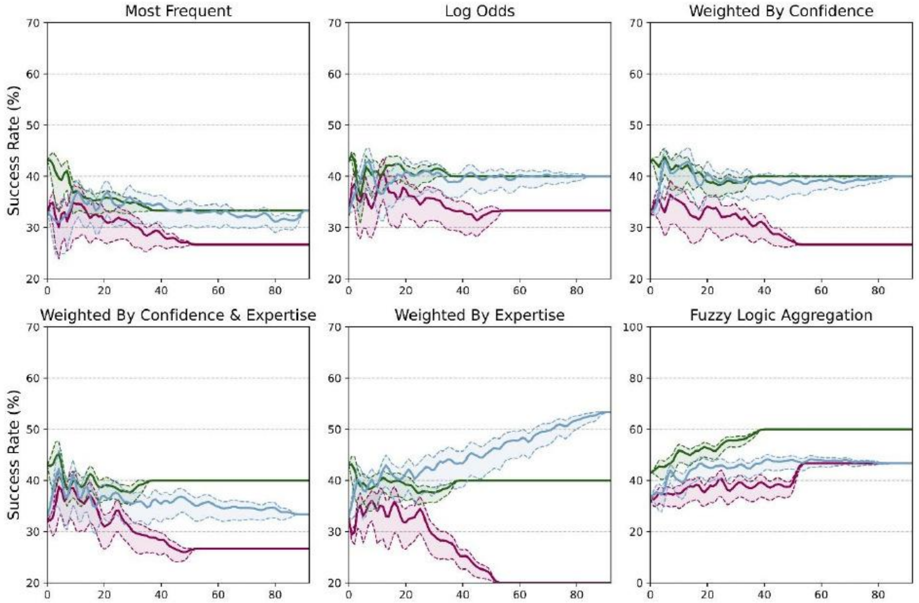


Figure S10: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the hERG endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue. Groups are colored as in Figure S6.

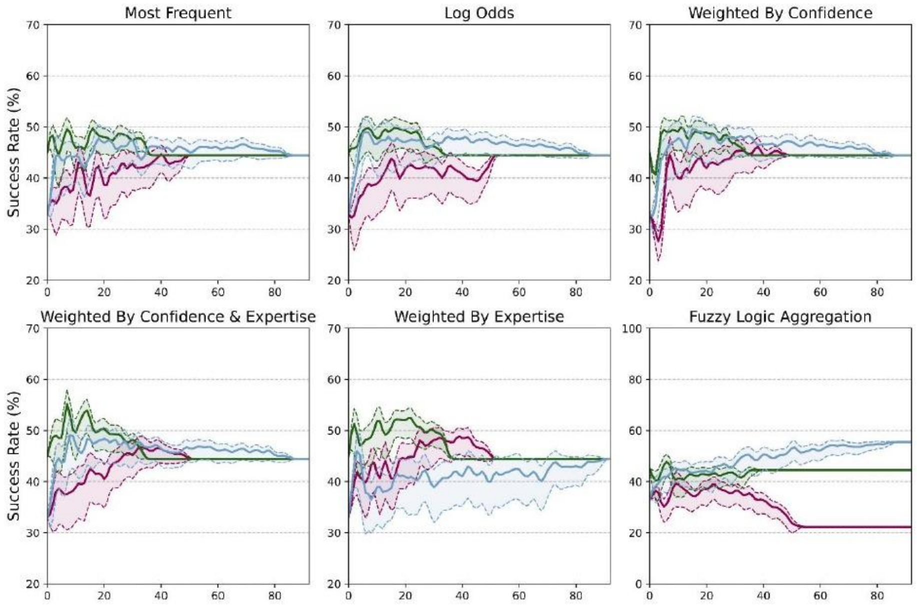


Figure S11: *Evolution of the collective success rate against the number of participants in the population per aggregation method for the LogD endpoint.* The collective SR is denoted per expertise group, from non-expert (1-2) in purple, expert (3-5) in green, and all participants in blue. Groups are colored as in Figure S6.

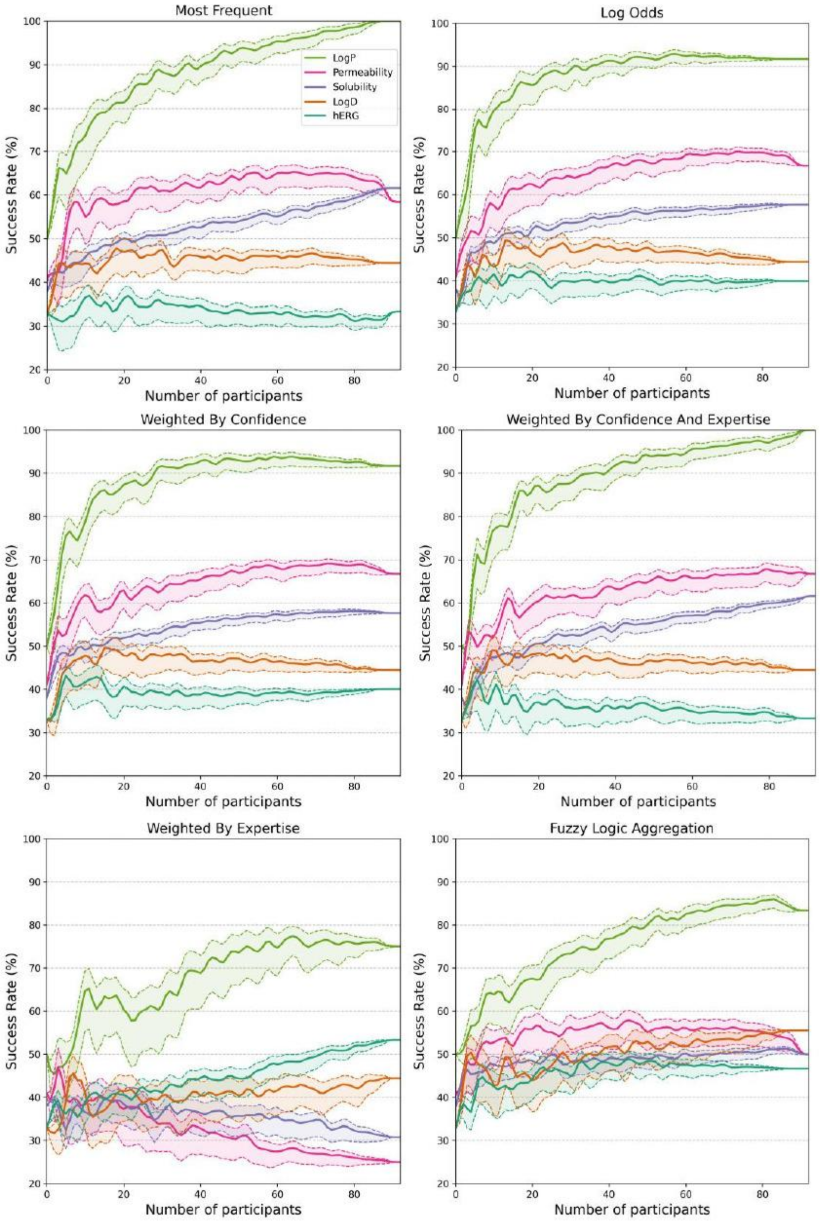


Figure S12: Evolution of the collective success rate as a function of the number of participants in the population per endpoint and per aggregation method.

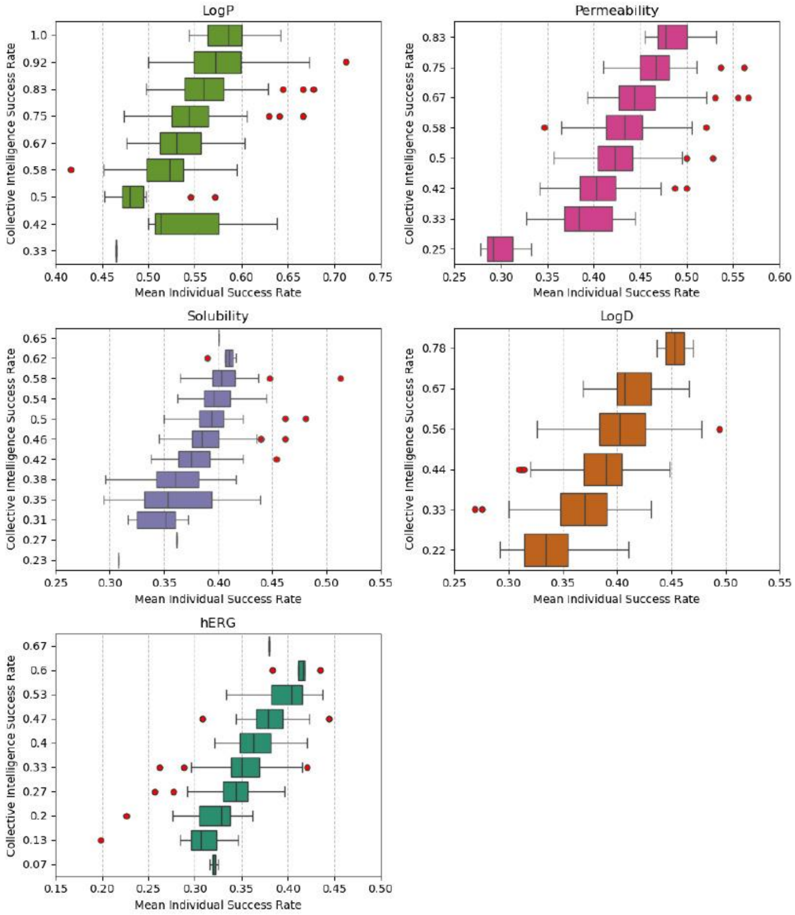


Figure S13: *Boxplots illustrating the relationship between collective intelligence success rate and mean individual success rate across endpoints.* Boxes highlight the median of the distribution of mean individual success rate per CI success rate with red dots indicating outliers.

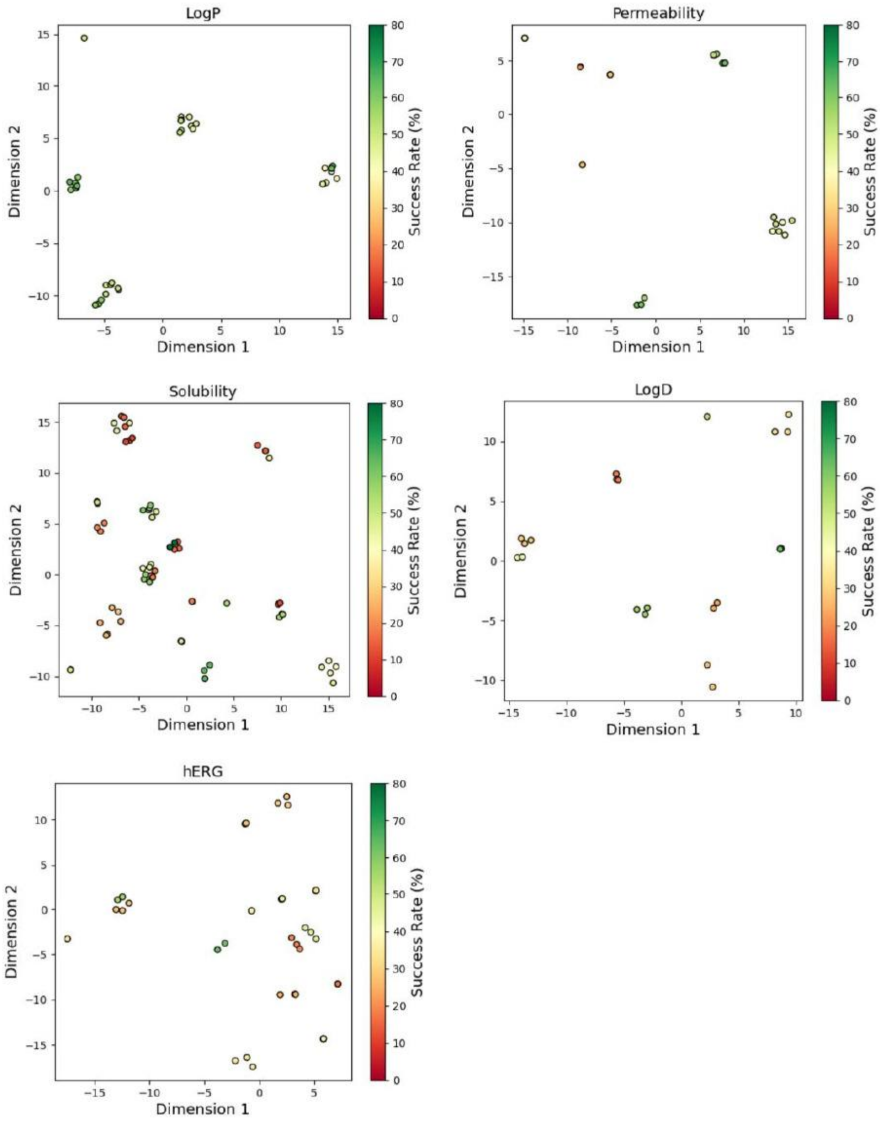


Figure S14: *t-SNE map of the collective intelligence chemical space per endpoint.* Each point represents a unique compound colored by the success rate of the related question using the CI 'most frequent' aggregation method.

Note: the subplots of Figure S14 were prepared using the same data as in Figure 3 of the main manuscript but separated by endpoint.

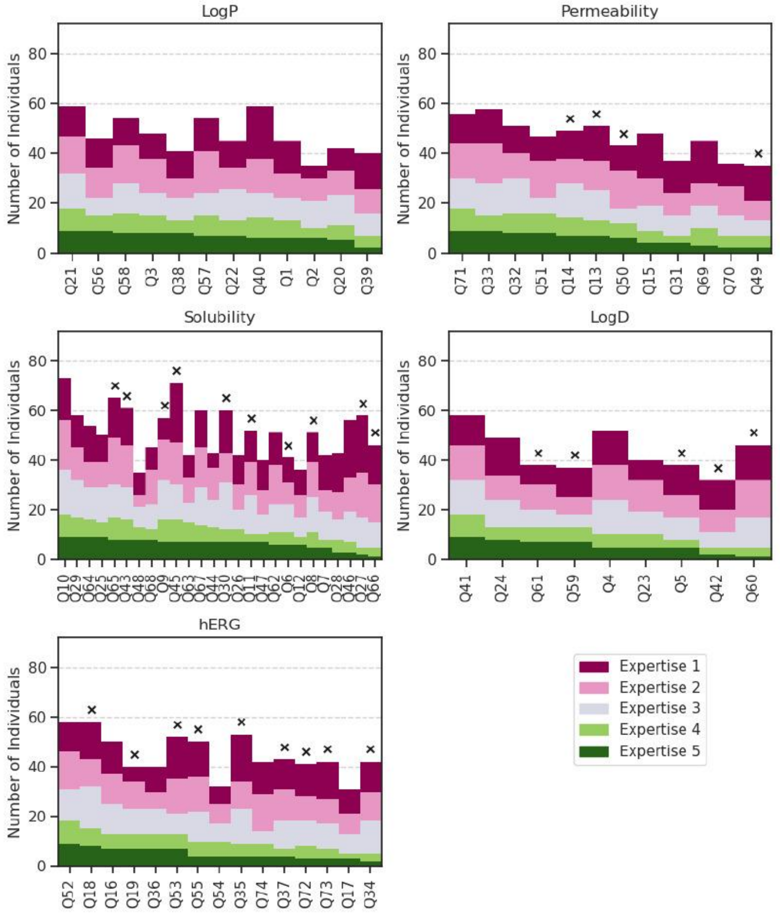


Figure S15: *Distribution of expertise level per question for the most frequently answered responses per endpoint.* Questions are organized by the proportion of participants of expertise level 5 selecting the most frequent answers. The expertise level goes from magenta (level 1) to green (level 5). Crosses have been added for the questions where the most frequent answer was incorrect.

Table S1: *Selection of cases from the CI questionnaire where level 3 participants outperformed individual experts.* The table presents the collectively selected compound by average individual experts against the correct selection from level 3 participants, per ADMET endpoint.

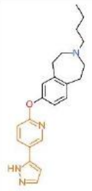
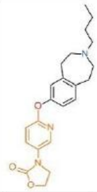
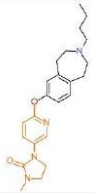
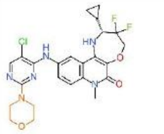
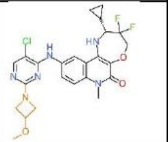
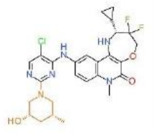
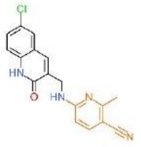
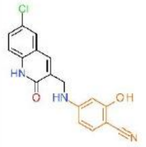
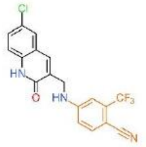
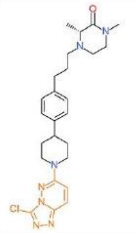
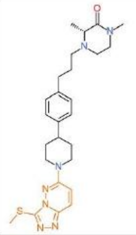
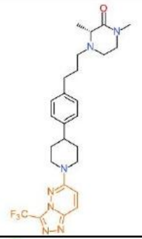
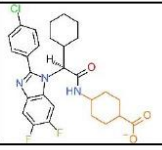
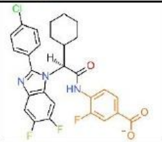
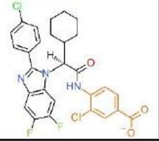
Endpoint	SR 3	SR 5	Expert choice (false)	Group 3 choice (correct)	Third Option	Ref.
LogP	0.50	0.22				1
Permeability	0.44	0.22				2
Solubility	0.67	0.33				3
LogD	0.39	0.22				4
hERG	0.44	0.11				5

Table S2: *Performance of the Graph Neural Networks on the public internal test set on ADMET endpoints.*

Endpoints	Number of unique compounds	R^2	RMSE	MAE
LogP	10,668	0.93	0.47	0.33
Permeability	1,259	0.55	0.56	0.42
Solubility	5,012	0.48	0.68	0.52
LogD	5,347	0.84	0.60	0.44
hERG	8,050	0.56	0.61	0.43

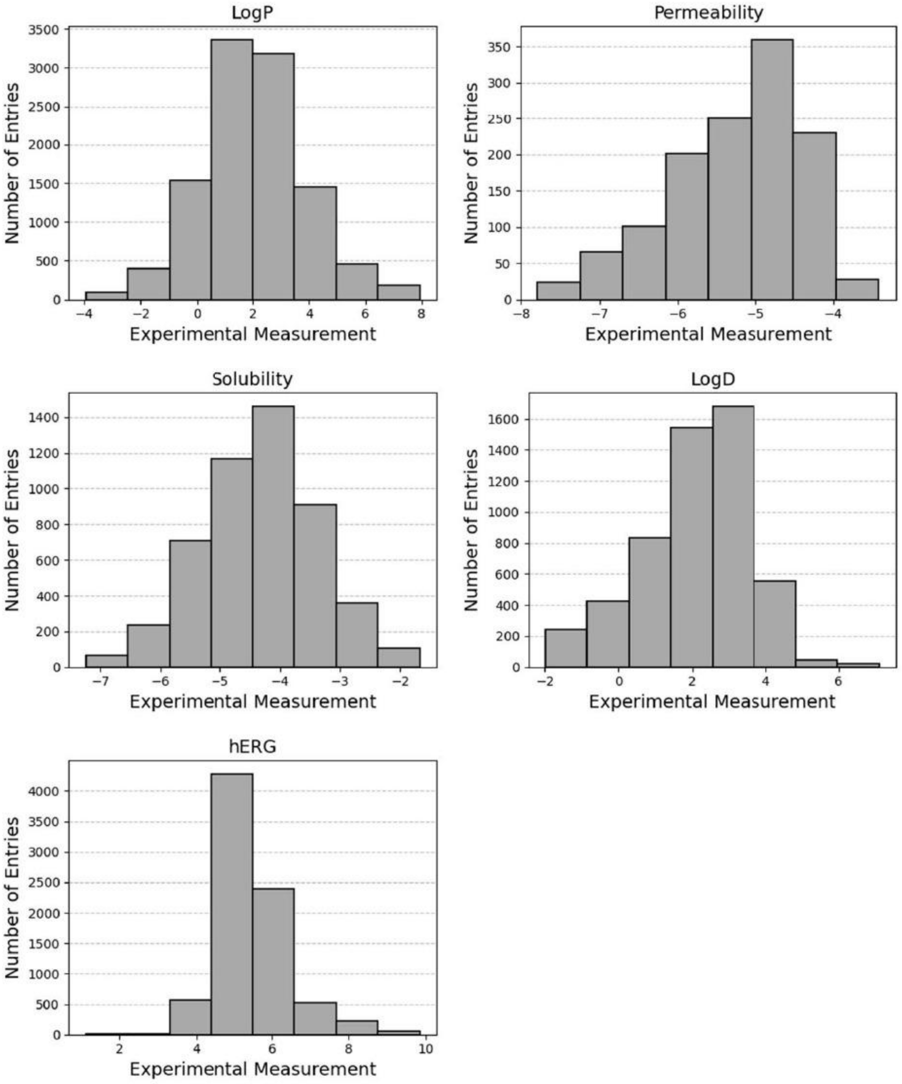


Figure S16: *Distribution of experimental measurements from public data used for modelling purposes.* The present endpoints are expressed in $\log_{10}(C_{octanol}/C_{water})$ for LogP, $\log_{10}(cm/s)$ for permeability, $\log_{10}(C_{buffer})$ for solubility, $\log_{10}(C_{octanol}/C_{buffer})$ for LogD, and pIC50 for hERG inhibition.

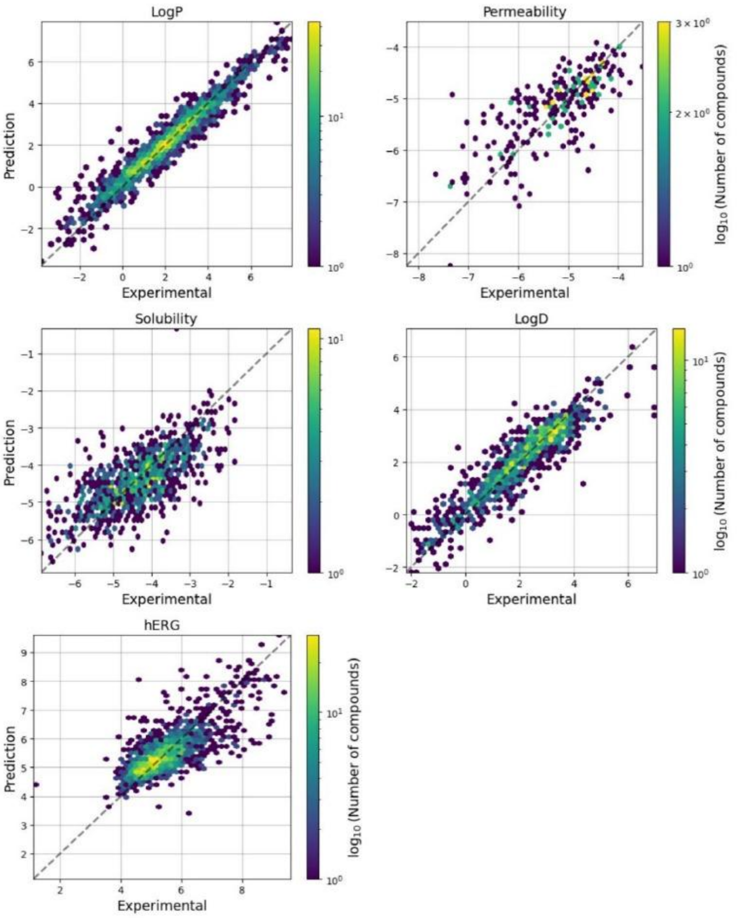


Figure S17: *Correlation between experimental and predicted value per endpoint.* The color scale depicts the density of compounds as the base-10 logarithm of the number of unique compounds.

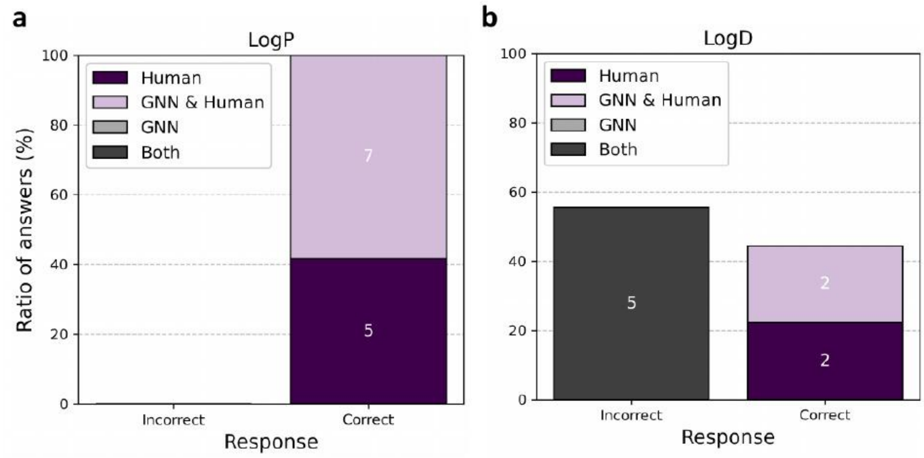
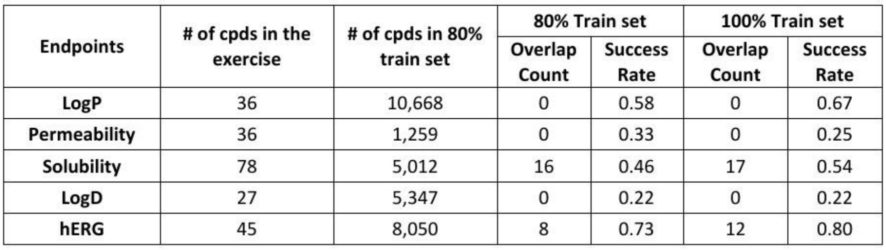


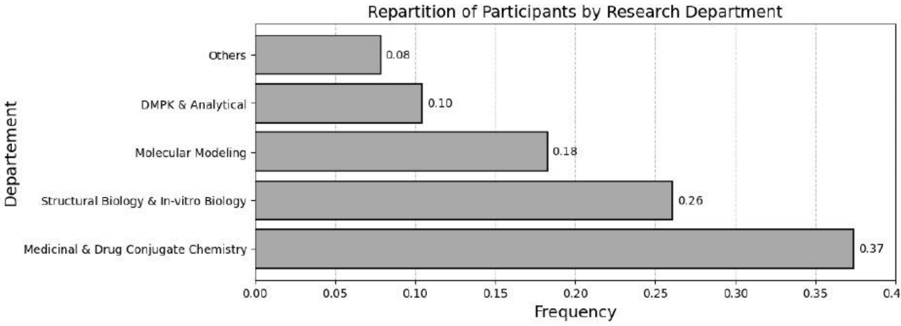
Figure S18: *Answer success and failure ratio (y-axis) and count (number in boxes) for a) logP and for b) logD.* Answers are grouped per source, *i.e.*, human, GNN (predictive model), GNN & Human, and both.

Table S3: *Overlap between the training set of the GNN model and compounds used in the collective intelligence questionnaire.* The table shows the number of unique compounds per question in the exercise and identifies overlaps where compounds from the questionnaire and training sets have matching InChI Keys, standardized using identical protocols. Each set was used to train a GNN model, which was then applied to predict outcomes in the collective intelligence exercise. The GNN success rate for each ADMET endpoint is also reported.



Endpoints	# of cpds in the exercise	# of cpds in 80% train set	80% Train set		100% Train set	
			Overlap Count	Success Rate	Overlap Count	Success Rate
LogP	36	10,668	0	0.58	0	0.67
Permeability	36	1,259	0	0.33	0	0.25
Solubility	78	5,012	16	0.46	17	0.54
LogD	27	5,347	0	0.22	0	0.22
hERG	45	8,050	8	0.73	12	0.80

Figure S19: *Distribution of the participation of each research department in the collective intelligence exercise.* The Medicinal and Drug Conjugate Chemistry departments were merged into the same group such as the Structural Biology with the In-vitro Biology department and the DMPK (Drug Metabolism and Pharmacokinetics) with the Analytical department. The Others group define the Chemistry Process and CMC (Chemistry, Manufacturing and Controls) departments.



References

(1) Wilson, D. M.; Apps, J.; Bailey, N.; Bamford, M. J.; Beresford, I. J.; Brackenborough, K.; Briggs, M. A.; Brough, S.; Calver, A. R.; Crook, B.; Davis, R. K.; Davis, R. P.; Davis, S.; Dean, D. K.; Harris, L.; Heslop, T.; Holland, V.; Jeffrey, P.; Panchal, T. A.; Parr, C. A.; Quashie, N.; Schogger, J.; Sehmi, S. S.; Stean, T. O.; Steadman, J. G. A.; Trail, B.; Wald, J.; Worby, A.; Takle, A. K.; Witherington, J.; Medhurst, A. D. Identification of Clinical Candidates from the Benzazepine Class of Histamine H3 Receptor Antagonists. *Bioorganic & Medicinal Chemistry Letters* **2013**, *23* (24), 6890–6896. https://doi.org/10.1016/j.bmcl.2013.09.090.

(2) Harnden, A. C.; Davis, O. A.; Box, G. M.; Hayes, A.; Johnson, L. D.; Henley, A. T.; de Haven Brandon, A. K.; Valenti, M.; Cheung, K.-M. J.; Brennan, A.; Huckvale, R.; Pierrat, O. A.; Talbot, R.; Bright, M. D.; Akpinar, H. A.; Miller, D. S. J.; Tarantino, D.; Gowan, S.; de Klerk, S.; McAndrew, P. C.; Le Bihan, Y.-V.; Meniconi, M.; Burke, R.; Kirkin, V.; van Montfort, R. L. M.; Raynaud, F. I.; Rossanese, O. W.; Bellenie, B. R.; Hoelder, S. Discovery of an In Vivo Chemical Probe for BCL6 Inhibition by Optimization of Tricyclic Quinolinones. *J. Med. Chem.* **2023**, *66* (8), 5892–5906. https://doi.org/10.1021/acs.jmedchem.3c00155.

(3) Lin, J.; Lu, W.; Caravella, J. A.; Campbell, A. M.; Diebold, R. B.; Ericsson, A.; Fritzen, E.; Gustafson, G. R.; Lancia, D. R. Jr.; Shelekhin, T.; Wang, Z.; Castro, J.; Clarke, A.; Gotur, D.; Josephine, H. R.; Katz, M.; Diep, H.; Kershaw, M.; Yao, L.; Kauffman, G.; Hubbs, S. E.; Luke, G. P.; Toms, A. V.; Wang, L.; Bair, K. W.; Barr, K. J.; Dinsmore, C.; Walker, D.; Ashwell, S. Discovery and Optimization of Quinolinone Derivatives as Potent, Selective, and Orally Bioavailable Mutant Isocitrate Dehydrogenase 1 (mIDH1) Inhibitors. *J. Med. Chem.* **2019**, *62* (14), 6575–6596. https://doi.org/10.1021/acs.jmedchem.9b00362.

(4) Hoveyda, H. R.; Fraser, G. L.; Dutheuil, G.; El Bousmaqui, M.; Korac, J.; Lenoir, F.; Lapin, A.; Noël, S. Optimization of Novel Antagonists to the Neurokinin-3 Receptor for the Treatment of Sex-Hormone Disorders (Part II). *ACS Med. Chem. Lett.* **2015**, *6* (7), 736–740. https://doi.org/10.1021/acsmedchemlett.5b00117.

(5) Richter, H. G. F.; Benson, G. M.; Bleicher, K. H.; Blum, D.; Chaput, E.; Clemann, N.; Feng, S.; Gardes, C.; Grether, U.; Hartman, P.; Kuhn, B.; Martin, R. E.; Plancher, J.-M.; Rudolph, M. G.; Schuler, F.; Taylor, S. Optimization of a Novel Class of Benzimidazole-Based Farnesoid X Receptor (FXR) Agonists to Improve Physicochemical and ADME Properties. *Bioorganic & Medicinal Chemistry Letters* **2011**, *21* (4), 1134–1140. https://doi.org/10.1016/j.bmcl.2010.12.123.

Outline

In this study, we reveal that collective intelligence consistently outperforms individual decision-making in optimizing ADMET endpoints and, in most cases, surpasses AI-driven predictions, except for hERG inhibition, where AI maintains an advantage. Moreover, we identify a complementary relationship between human expertise and machine learning, suggesting that hybrid approaches could enhance predictive accuracy for complex pharmacokinetic and toxicological assessments. This study represents a foundation for future drug discovery research in hybrid intelligence frameworks.

Chapter 8. Conclusions & Perspectives

As of today, modeling in drug discovery has reached a stage where the bottleneck is no longer the prediction method or the descriptors but rather the data itself. In other words, data now drives the choice of the most appropriate ML method based on its size, diversity, and quality. One of the main limitations is quality, which is affected by experimental bias, inherent noise in assays, human error, and a lack of condition homogeneity.

In this thesis, we explored the limitations of modeling experimental data in an industrial drug discovery context. Driving the discovery and development of a drug requires navigating chemical space under multi-objective constraints, most of which are derived from experimental assays. These assays are interdependent, meaning that improving one property may lead to failure in several others.

To evaluate and propose compounds with better potential, projects are supported by modelers and data scientists. As a result, the quality of decisions depends not only on human expertise but also on the accuracy of models, an accuracy that, ultimately, relies on the data bottleneck.

8.1. Data Quality & Modeling in Drug Discovery

Solubility

In this thesis, we explored numerous ways to improve data curation and modeling approaches. These approaches were first applied to one of the most challenging endpoints to model accurately: solubility. Initial work on kinetic solubility revealed that different assays are more similar in measured values than expected. This insight led to the design and development of predictive models suited for early screening campaigns, where companies need to broadly screen large libraries and retain only compounds that may be soluble. This ensures that only testable compounds are used, preventing wasted resources on acquiring insoluble, dry-brick compounds.

However, solubility is not merely a binary decision. During later-stage optimization, solubility, and particularly precise solubility estimation, becomes a key parameter in multi-objective optimization. To guide and validate such a framework, thermodynamic solubility is often estimated, favoring PBS 7.4 or pure water solubility. This assay provides a continuous estimation of a compound's maximum concentration in solution. Despite its importance, it is highly susceptible to various sources of noise. Our work on water solubility exposed the flawed state of existing solubility data, a consequence of years of poor curation, aggregation, and reaggregation of overlapping datasets, along with the failure to adhere to OECD guidelines. Through multiple steps of curation, modeling, and validation, we identified the most error-prone and low-quality data sources and established a guideline for the proper curation of solubility data.

Absorption

Given the multi-objective nature of drug discovery, we explored the next major limitation in a drug's journey: its absorption once solubilized. This process involves numerous interrelated assay properties. To model them accurately, we conducted an in-depth analysis of the factors influencing different permeability assays, including cell lines, inhibitors, and various assay-specific conditions. This research ultimately led us to compare the application of standard single-task ML approaches with GNN-based MTL models.

By leveraging multi-task models, we aimed to exploit task relationships to enhance information extraction from the model. Our study demonstrated a significant improvement in predictive performance with MTL, particularly for small datasets, while predictions on larger datasets remained as accurate as before. Moreover, these models not only delivered better generalization and performance but also provided high-quality representations of compounds. To illustrate this, we featurized thousands of compounds using the model's graph embeddings and projected them onto a chemical space map. This highlighted the model's ability to correctly represent highly related endpoints with fine granularity. Ultimately, this work contributed to the development of novel high-performance models and a highly curated dataset for absorption prediction.

OneADMET

Since decision-making in drug discovery extends beyond a drug's transit from pill to bloodstream, we expanded our approach to account for distribution, metabolism, toxicity, and even activity and selectivity. This involved applying the MTL approach to thousands of endpoints. However, as ML models are highly sensitive to noise and data distribution, an initial round of thorough curation and standardization was necessary. This step led to the creation of OneADMET, the largest and most curated dataset of continuous ADMET and activity data.

This dataset was then used to train a large-scale GNN-MTL model, which was rigorously benchmarked against popular ML approaches. The study demonstrated the broad applicability of MTL models, which not only matched the best optimized SVM on small datasets but also outperformed XGBoost on medium-sized data and remained competitive with single-task GNN models on large datasets. Beyond their power and versatility, these models are also highly efficient. With a single MTL model, we can generate predictions that would otherwise require thousands of individual models and tens of descriptor calculations per compound, not to mention the storage and computational costs of handling such extensive descriptors and models. As of today, GNN-MTL offers the best balance between cost, performance, speed, and applicability. It is not just a predictor but also a featurizer, an interpreter, and an uncertainty estimator for large-scale virtual screening with more reliable decisions.

Collective Intelligence

Hence, these studies raise an important question: which strategies lead to the best decisions? Are they those fully driven by AI, made by a single expert, or decided by a group of individuals?

To gain deeper insights into this problem, we applied the concept of collective intelligence to decision-making in late-stage lead optimization and compared it to decisions solely reliant on state-of-the-art ML models. Our findings revealed that groups composed of both experts and non-experts can make reliable decisions.

When testing different sample sizes, we identified a threshold where cohort size no longer influenced success rates. A group of 10 to 20 individuals proved sufficient to enhance decision-making, often matching or even outperforming both expert-driven and ML-based approaches. However, this collective strategy had its limitations. Its application to complex endpoints, such as hERG inhibition, failed to yield significant advantages. Leading and collaborating within a target project team requires integrating diverse approaches and methods in a collective manner. The potential of collective intelligence is not confined to lead optimization, it can be applied across various stages of the drug discovery process due to its versatility.

Beyond this, other strategies, such as swarm intelligence, ant colony optimization, and crowdsourcing, remain underexplored but have shown early promise. Understanding how to effectively integrate AI, human expertise, and collective strategies in drug discovery remains a critical challenge for the field.

Research Perspectives

Throughout this thesis, it has been made clear that the performance and reliability of predictive models in industrial drug discovery are tightly linked to the quality of input data, the robustness of curation workflows, and the sophistication of computational frameworks. While multi-task learning models have shown promise, further improvements in model performance will rely heavily on resolving persistent challenges in data annotation and fostering more productive, interactive forms of human-AI collaboration. Moving forward, advances in large language models (LLMs) and AI agents offer a compelling path to address these issues.

Enhancing Modeling with AI Agents

A crucial bottleneck in model-driven drug discovery remains the time-consuming and error-prone process of data curation. Recent developments in LLMs provide a scalable solution to automate the extraction and annotation of chemical and biological knowledge from unstructured textual sources such as scientific articles, patents, and lab notebooks. By leveraging transformer-based architecture, LLMs can parse and organize complex domain-specific information with increasing accuracy. For instance, Schilling-Wilhelmi et et al.[178] demonstrated workflows that combine automated annotation with human-in-the-loop corrections to ensure high fidelity, while Ai et al.[179] showcased how fine-tuned models can outperform traditional rule-based systems in extracting synthetic procedures. Similarly, Vangala et al.[180] applied GPT-based models to patent mining, uncovering previously overlooked chemical reactions, and Kosonocky et al.[181] highlighted the capacity of LLMs to infer functional-structural relationships hidden in patent corpora.

Beyond annotation, LLMs hold potential for flagging inconsistencies within datasets through learned recognition of underlying chemical or biological patterns, thus improving overall data integrity. Embedding such tools into curation workflows can dramatically reduce manual overhead, enhance consistency, and accelerate the generation of high-quality datasets.

Human-Agent Collaboration in Drug Discovery

Beyond static annotation tasks, LLMs are starting to be deployed as dynamic, decision-support tools that operate as digital co-scientists. These agents integrate reasoning, planning, and execution modules to autonomously perform and interpret tasks. However, general-purpose LLMs often underperform in generative tasks like multi-step retrosynthesis. To overcome these limitations, purpose-built systems like ChemCrow[182] and ChemLLM[183] have been developed. ChemCrow extends GPT-4 with tool integration through LangChain and chain-of-thought prompting, enabling it to access and use autonomously chemistry-specific tools and reason across multiple steps to complete synthesis or design tasks.

Strategic Outlook & Integration into R&D Pipelines

As drug discovery increasingly relies on large-scale computation, integrating LLM-based annotation tools and autonomous scientific agents into R&D pipelines presents a compelling strategic opportunity. These systems could streamline manual workflows, enhance data integrity, and improve model generalization by reducing inconsistencies in input data. Yet, their current limitations, most notably a reliance on training data and lack of domain reasoning, mean expert oversight remains critical. LLMs should not replace human expertise, but rather augment it, provided their deployment is governed by rigorous standards for validation, transparency, and pharmacological relevance.

Looking ahead, LLM-based agents could extend well beyond data curation to actively support ADMET prediction and decision-making. Acting as digital collaborators, they may assist medicinal chemists and modelers in tasks such as docking compounds into target pockets, summarizing SAR trends, flagging potential liabilities, or preparing compound sets filtered by metabolic or safety criteria. With growing multimodal capacities, these agents could synthesize structural data, bioassay results, and literature to generate context-aware recommendations and refine hypotheses in real time.

To make this vision operational, workflows would need to be modularized into agent-executable steps, domain constraints embedded via fine-tuning or chemical prompts, and predictive backends, such as docking engines or PBPK models, interfaced seamlessly. Crucially, expert feedback mechanisms must be built into interactive tools, allowing users to refine, validate, or redirect outputs on the fly.

This trajectory outlines the rise of an *agentic collective*, a collaborative network of AI systems working in coordination with human scientists, not as isolated utilities but as contextual, task-specialized partners. Rather than replacing decision-making, these agents could serve to sharpen it, enhancing the pace, consistency, and creativity in early drug discovery.

8.2. State of the field

Even though QSAR is now celebrating its 60th anniversary, the field is only now entering its most prolific and active research period. As AI continues to integrate into drug discovery, it raises a critical tension: should these models serve merely as tools to assist experts, or are they gradually shaping a shift toward automation-driven strategies? The field stands at a crossroads where enthusiasm for AI-driven efficiency collides with the reality of its practical constraints. The next section explores how this dynamic is unfolding, tracing the trajectory of AI adoption in drug discovery, from initial breakthroughs to the recalibration of expectations.

From Innovation to Disillusion

The integration of AI into drug discovery follows a well-documented pattern, resembling the rise and fall of the internet boom of the 1990s, the social media explosion of the 2000s, and the blockchain craze of the 2010s.[184,185] Each of these technologies followed a cycle of early innovation, rapid adoption, exaggerated claims, reality check, and eventually stabilization into practical applications.[186] AI in pharmaceutical research is no exception. Initially heralded as game-changing for automation, data analysis, and predictive modeling, AI has been widely promoted as the future of drug discovery. But as with past trends, the enthusiasm has often been accompanied by overstatements, blurring the lines between science and marketing.[187]

The Marketing of Science

The push to accelerate drug development has led pharma, biotech, and contract research organizations (CRO) to adopt AI/ML more widely. Though explored since the 1990s (e.g., AstraZeneca's early infrastructure), recent gains in data, computing, and algorithms have expanded their practical use across the full drug discovery pipeline. In recent years, optimism has surged as AI was credited with revolutionizing drug discovery, from molecular design to toxicity prediction. This wave of enthusiasm has also made these companies prime targets for AI-driven biotech firms.[188] Faced with the choice of building in-house expertise or partnering with external AI companies, pharmaceutical companies must navigate a landscape where true innovation and bold marketing claims, disguised as scientific papers, often mix.

Recently, in 2024, Google DeepMind's AlphaFold 3[189], following version of the successful and performant AlphaFold 2[190], was published as a closed-source model in Nature, only five months after being received by the journal.[191] Many criticized the release as a promotional maneuver rather than a scientific contribution. Mounting pressure eventually forced DeepMind to publish an open-source version months later, revealing tensions between academic transparency and corporate interests.[192] In January 2025, In Silico Medicine reported using quantum computing to characterize KRAS inhibitors, a notoriously hard target under investigation since 1982, with over 100 drug candidates already in development .[193] Their approach combined ultra-large virtual screening with quantum-hybrid generative models, ultimately identifying weak hits (~5 μM IC50). While technically ambitious, the outcome was underwhelming in practical terms. Given the high computational cost and the still-maturing state of quantum hardware, dismissed even by Nvidia CEO Jensen Huang as premature, the effort highlighted a broader issue: the growing gap between technological hype and meaningful pharmacological innovation. Raising the question of whether such methods offer real therapeutic advantage.[194,195]

Earlier, in 2019, In Silico Medicine made bold claims about identifying potent kinase inhibitors within 21 days.[196] The rational of the study was questioned by P. Walters and M. Murcko at the time, pointing out the strong similarity of the In Silico Medecine compounds with the marketed kinase inhibitor Iclusig (ponatinib), questioning the necessity of "fancy" software to substitute an isoxazole for an amide carbonyl, and the relevance of such publication.[197] Meanwhile, they disclosed having reached Phase 1 from target discovery in just 30 months with an AI-discovered drug in 2022. The compound shown some promise in the phase IIa results with a primary endpoint as safety but lacks definitive clinical success.[198]

Similarly, in 2020, Exscientia declared that it had designed a cancer drug candidate (EXS-21546) in only 12 months, designated as the "the first AI-generated drug".[199] While these milestones were widely publicized, the actual clinical outcomes failed to fully meet expectations with two clinical candidates wiped. In 2024, the company was acquired by Recursion.[200] Backed by Nvidia, Recursion, which initially aimed to develop 100 drugs in 10 years reported mixed results for its lead repurposed drug REC-994. While deemed safe, the drug failed Phase II clinical trials.[201]

Additional mentions goes to Atomwise, an AI-driven company founded in 2012 advancing bold claims in 2015, it has yet to send a compound to clinic.[202] Cassava Sciences, once hyped for its experimental Alzheimer's candidate simufilam, failed to show clinical benefit in Phase III trials, leading to the discontinuation of its development in November 2024. Another major AI-driven biotech, BenevolentAI a company valued in 2018 at around $2 billion which stated to have created a "bioscience machine brain" have obtained deceiving results in 2024.[203] Their AI-generated candidate for pan-Trk inhibitor (atopic dermatitis) failed in Phase II trials, performing no better than a placebo.[204–206]

Despite advances in AI-driven drug discovery, its impact remains limited in clinical translation. While AI excels in early-stage tasks like virtual screening and multi-objective optimization, it struggles with biological complexity, data quality issues, and clinical trial unpredictability. Overhyped claims often fail to materialize, as seen in AI-driven biotech firms whose compounds underperform in clinical phases. The reliance on biased, noisy datasets constrains AI's predictive power, making it a decision-support tool rather than a standalone solution. Until AI can reliably address clinical challenges, its role in drug development remains promising but not yet transformative.

Correction to a Reality Check

As the reality check set in, many biotech startups struggled to meet their promises. The outcome has been an industry-wide correction, characterized by failed clinical trials, financial struggles, and large-scale layoffs rounds across the pharmaceutical sector surged by 281% (from 11 rounds in 2023 to 42 rounds in 2024).[207]

In 2024, Exscientia was forced to cut 25% of its workforce, a direct result of financial underperformance and failed partnerships.[208] BenevolentAI laid off 45% of its employees and an immediate unexplained departure of the CEO, only after one year on the job, citing a need to restructure and focus on fewer high-value projects.[209] Other companies, including Atomwise and Recursion, have also reduced their workforce. In total, AI-driven biotech startups have seen over 2,500 job cuts in 2024, as investors and executives reevaluate the practical impact of AI in drug discovery.[210]

Large pharmaceutical companies have also undergone significant layoffs, though for different reasons. Unlike startups that faced direct failures, big pharma layoffs have been driven by two key factors: (**i**) the failure of AI-driven partnerships to yield expected results, and (**ii**) a shift in corporate strategy toward CRO to reduce costs. Bristol Myers Squibb announced a 2,200-employee reduction by the end of 2024. Pfizer eliminated 1,500 positions, including 285 roles at a vaccine R&D site in New York and 120 in Washington. Takeda cut 641 jobs in Massachusetts.

These layoffs in big pharma are indicative of a replacement of departments by experimental or AI-driven CROs. Boards of directors, often distant from the realities of experimental R&D, have embraced AI as a way to reduce expenses. However, this presents a paradox, while AI has not yet delivered on its promise of fully automated drug discovery and revenue are high, companies are already reducing human R&D expertise in favor of AI-driven cost efficiencies.

Communication between R&D and Business Leadership

Many of the layoffs in biotech and big pharma suggest that business executives, often motivated by short-term financial gains satisfying their Key Performance Indicator (KPI), are making decisions without sufficient input from R&D leaders who understand the true capabilities and limitations of AI.[188,211] Strong alignment between Chief Scientific Officers (CSOs), Head of R&D and CEOs is critical to ensure that AI is integrated into drug discovery in a way that enhances, rather than replaces, essential scientific processes and experience-owned knowledges. Historically, successful pharmaceutical companies have maintained direct, science-focused communication between leadership and research divisions. However, as AI hype has taken center stage, some executives have made sweeping AI-driven restructuring decisions, often confusing AI's role in core R&D functions, such as molecular design, predictive modeling, and target validation, with broader AI-driven digitalization efforts in logistics, marketing, and operational efficiency. For AI to contribute meaningfully to drug discovery, companies must adopt a balanced approach.[22]

Toward more Sustainable Approaches

Many biotech firms have historically prioritized aggressive AI-driven narratives to attract and rapidly secure funding, while only a few have focused on incremental advancements supported by rigorous validation. As the industry corrects itself, a more structured approach to AI adoption is emerging. Public-Private Partnerships (PPPs) offer a promising safeguard while providing a mechanism for pooling resources to address shared challenges. PPPs in France, such as the CIFRE program, facilitate industry-academia collaboration by funding PhD students conducting research in private companies. Other key initiatives include PIA (Programme d'Investissements d'Avenir) and France 2030, which invest in AI, biotech, and deep-tech innovation. ANR collaborative programs and Carnot Institutes support joint R&D efforts, while BPI France's DeepTech grants and i-Lab competition fund early-stage biotech startups. The AIChemist program, backed by Horizon Europe, exemplifies international PPPs focused on AI-driven drug discovery.[40]

What comes next for AI in Drug Discovery?

While expectations around AI in drug discovery require adjustment, its potential remains transformative, if applied correctly. The future of AI in pharmaceuticals will likely shift from ambitious claims to practical, results-driven applications.

For investors, the most promising biotech companies prioritize strong datasets first, followed by computing power and, lastly, innovative methodologies. Data is the foundation; without high-quality, validated biological information, even the most advanced AI models will struggle to generate meaningful insights. This reality underscores a key advantage for Big Pharma, which leverages extensive proprietary datasets to enhance the reliability of AI-driven predictions.

Rather than replacing medicinal chemists, AI is evolving into a powerful decision-support system, assisting researchers in SAR modeling, toxicity prediction, and multi-objective optimization. However, for computational predictions to translate into clinically relevant outcomes, AI models must undergo rigorous validation against real-world biological data.

Emerging fields such as RNA-targeted therapeutics and personalized medicine offer promising opportunities for AI integration. In these low-data environments, AI can accelerate target identification and drug design, provided that robust validation methods are in place.

Ultimately, AI's impact on drug discovery will not be defined by hype but by its ability to enhance predictive accuracy, optimize decision-making, and drive clinically meaningful advancements. To realize its full potential, the industry must move beyond speculation and embrace transparent, science-driven innovation.

Chapter 9. List of Abbreviations

3R	Reduction, Refinement, and Replacement
AB-FEP	Absolute Free Energy Perturbation
ACF	Atom-Centered Fragments
AD	Applicability Domain
ADMET	Absorption, Distribution, Metabolism, Elimination, Toxicity
AF3	AlphaFold 3
AGP	α-1-acid glycoproteins
AI	Artificial Intelligence
AL	Active Learning
ANR	Agence Nationale de la Recherche (French National Research Agency)
ATP	Adenosine Tri-Phosphate
BBB	Blood-Brain Barrier
BIO	Biotechnology Innovation Organization
BPI	Banque Publique d'Investissement
Caco-2	Colorectal adenocarcinoma cells
CADD	Computer-Aided Drug Design
CDK	Chemistry Development Kit
CEO	Chief Executive Officer
CERAPP	Collaborative Estrogen Receptor Activity Prediction Project
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
ChemBERTa	Chemical Bidirectional Encoder Representations from Transformers

ChEMBL	Chemical Database at EMBL-EBI
CIFRE	Convention Industrielle de Formation par la Recherche
CNS	Central Nervous System
CoMPARA	Collaborative Modeling Project for Androgen Receptor Activity
Cp	Concentration in plasma
CRO	Contract Research Organization
Cryo-EM	Cryogenic electron microscopy
CSO	Chief Scientific Officer
CV	Cross-Validation
CYP450	Cytochrome P450
DEL	DNA-Encoded Library
DMSO	Dimethylsulfoxide
DMTA	Design, Make, Test, Analyze
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
eADMET	early ADMET
EBI	European Bioinformatics Institute
ECFP	Extended Connectivity Fingerprints
EMA	European Medicines Agency
EMBL	European Molecular Biology Laboratory
EPA	Environmental Protection Agency
FBDD	Fragment-Based Drug Design
FCFP	Functional Class Fingerprints

FDA	Food and Drug Administration
FEP	Free Energy Perturbation
FFN	Feed-Forward Network
FMO	Flavin-Containing Monooxygenase
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GPCR	G Protein-Coupled Receptor
GTM	Generative Topographic Mapping
hERG	Human Ether-A-go-go-Related Gene
HFE	Hydration Free Energy
HMG-CoA	Hydroxymethylglutaryl-Coenzyme A
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
HPLC	High-Performance Liquid Chromatography
HTS	High-Throughput Screening
IA	Intelligence Artificielle (French for AI)
IC_{50}	Half maximal inhibitory concentration
IF	Isolation Forest
IL8	Interleukin 8
InChI	IUPAC International Chemical Identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes
K_i	Inhibition constant
k-NN	k-Nearest Neighbors

KPI	Key Performance Indicator
LBDD	Ligand-Based Drug Design
LBVS	Ligand-Based Virtual Screening
LC-MS	Liquid chromatography–mass spectrometry
LD50	Lethal Dose 50
LE	Ligand Efficiency
LLM	Large Langage Model
LOF	Local Outlier Factor
LogD	Dissociation coefficient
LogP	Partition coefficient
MAA	Marketing Authorization Application
MACCS	Molecular ACCess System
MAE	Mean Absolute Error
MD	Molecular Dynamics
ML	Machine Learning
MOE	Molecular Operating Environment
MOO	Multi-Objective Optimization
MPO	Multi-Parameter Optimization
MSE	Mean Squared Error
MTD	Maximum Tolerated Dose
MTL	Multi-Task Learning
MW	Molecular Weight
NAT	N-Acetyltransferase

Nav1.5	alpha subunit of the voltage-gated sodium channel
NDA	New Drug Application
NME	New Molecular Entity
NMR	Nuclear Magnetic Resonance
NOAEL	No Observed Adverse Effect Level
NPV	Net Present Value
OChem	Online Chemistry Database
OcSVM	One-Class Support Vector Machine
OECD	Organisation for Economic Co-operation and Development
PAMPA	Parallel Artificial Membrane Permeability Assay
PBS	Phosphate-Buffered Saline
PCA	Principal Component Analysis
PD	Pharmacodynamics
P-gP	P-Glycoprotein
PI3K	Phosphoinositide 3-kinase
PIA	Programme d'Investissements d'Avenir
PBPK	Physiologically based pharmacokinetic modeling
PK	Pharmacokinetics
pKa	acid dissociation constant
PPB	Plasma Protein Binding
PPP	Public-Private Partnerships
QED	Quantitative Estimation of Drug-Likeness
QSAR	Quantitative Structure Activity Relationship

QSPR	Quantitative Structure Property Relationship
R&D	Research & Development
R2	Coefficient of Determination
RAND	Research And Development Corporation
RBF	Radial Basis Function
REACH	Registration
RF	Random Forest
RMSE	Root Mean Squared Error
RNA	Ribonucleic Acid
RNN	Recurrent Neural Network
RORγ	Retinoic acid receptor-related Orphan Receptors
SAR	Structure Activity Relationship
SBDD	Structure-Based Drug Design
SGD	Stochastic Gradient Descent
siRNA	Small interfering RNA
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SOM	Self-Organizing Map
SPR	Structure Property Relationship
STL	Single-Task Learning
SVM	Support Vector Machine
TDC	Therapeutic Data Commons
TPSA	Topological Polar Surface Area

t-SNE	t-Distributed Stochastic Neighbor Embedding
UGT	UDP-Glucuronosyltransferase
UMAP	Uniform Manifold Approximation and Projection
UV-Vis	Ultraviolet–visible spectrophotometry
VAE	Variational AutoEncoder
Vd	Volume distribution
VEGF	Vascular Endothelial Growth Factor
VS	Virutal Screening
XGBoost	eXtreme Gradient Boosting

Chapter 10. References

(1) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162* (6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

(2) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2019**, *20* (2), 273–286. https://doi.org/10.1093/biostatistics/kxx069.

(3) *Clinical Development Success Rates and Contributing Factors 2011-2020 | BIO*. https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020 (accessed 2024-11-18).

(4) Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323* (9), 844–853. https://doi.org/10.1001/jama.2020.1166.

(5) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat Rev Drug Discov* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(6) Martin, O. Artificial Intelligence in Drug Discovery and Development. *Advanced Sciences* **2021**, *3* (2), 1–10. https://doi.org/10.69610/j.as.20210822.

(7) Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discovery Today* **2021**, *26* (4), 1040–1052. https://doi.org/10.1016/j.drudis.2020.11.037.

(8) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998**, *21* (1), 203–224. https://doi.org/10.1016/S0925-2312(98)00043-5.

(9) Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*; 2008; pp 413–422. https://doi.org/10.1109/ICDM.2008.17.

(10) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* **2009**, *20* (1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.

(11) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

(12) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076. https://doi.org/10.1021/acs.jcim.7b00146.

(13) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a Curated Reference Set of Aqueous Solubility and 2D Descriptors for a Diverse Set of Compounds. *Sci Data* **2019**, *6* (1), 143. https://doi.org/10.1038/s41597-019-0151-1.

(14) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will We Ever Be Able to Accurately Predict Solubility? *Sci Data* **2024**, *11* (1), 303. https://doi.org/10.1038/s41597-024-03105-6.

(15) Baybekov, S.; Llompart, P.; Marcou, G.; Gizzi, P.; Galzi, J.-L.; Ramos, P.; Saurel, O.; Bourban, C.; Minoletti, C.; Varnek, A. Kinetic Solubility: Experimental and Machine-Learning Modeling Perspectives. *Molecular Informatics* **2024**, *43* (2), e202300216. https://doi.org/10.1002/minf.202300216.

(16) Llompart, P.; Amaning, K.; Bianciotto, M.; Filoche-Rommé, B.; Foricher, Y.; Mas, P.; Papin, D.; Rameau, J.-P.; Schio, L.; Marcou, G.; Varnek, A.; Moussaid, M.; Minoletti, C.; Gkeka, P. Harnessing Medicinal Chemical Intuition from Collective Intelligence. ChemRxiv May 7, 2024. https://doi.org/10.26434/chemrxiv-2024-0hww3.

(17) Hodgson, J. ADMET—Turning Chemicals into Drugs. *Nat Biotechnol* **2001**, *19* (8), 722–726. https://doi.org/10.1038/90761.

(18) *Can open-source R&D reinvigorate drug research? | Nature Reviews Drug Discovery*. https://www.nature.com/articles/nrd2131 (accessed 2025-03-06).

(19) Di, L.; Kerns, E. H. *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*; Academic Press, 2015.

(20) Owens, P. K.; Raddad, E.; Miller, J. W.; Stille, J. R.; Olovich, K. G.; Smith, N. V.; Jones, R. S.; Scherer, J. C. A Decade of Innovation in Pharmaceutical R&D: The Chorus Model. *Nat Rev Drug Discov* **2015**, *14* (1), 17–28. https://doi.org/10.1038/nrd4497.

(21) Morgan, P.; Brown, D. G.; Lennard, S.; Anderton, M. J.; Barrett, J. C.; Eriksson, U.; Fidock, M.; Hamrén, B.; Johnson, A.; March, R. E.; Matcham, J.; Mettetal, J.; Nicholls, D. J.; Platz, S.; Rees, S.; Snowden, M. A.; Pangalos, M. N. Impact of a Five-Dimensional Framework on R&D Productivity at AstraZeneca. *Nat Rev Drug Discov* **2018**, *17* (3), 167–181. https://doi.org/10.1038/nrd.2017.244.

(22) Douglas, F. L.; Narayanan, V. K.; Mitchell, L.; Litan, R. E. The Case for Entrepreneurship in R&D in the Pharmaceutical Industry. *Nat Rev Drug Discov* **2010**, *9* (9), 683–689. https://doi.org/10.1038/nrd3230.

(23) DiMasi, J. A.; Faden, L. B. Competitiveness in Follow-on Drug R&D: A Race or Imitation? *Nat Rev Drug Discov* **2011**, *10* (1), 23–27. https://doi.org/10.1038/nrd3296.

(24) Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Product Drug Discovery. *Nat Rev Drug Discov* **2023**, *22* (11), 895–916. https://doi.org/10.1038/s41573-023-00774-7.

(25) Li, F.-S.; Weng, J.-K. Demystifying Traditional Herbal Medicine with Modern Approach. *Nature Plants* **2017**, *3* (8), 1–7. https://doi.org/10.1038/nplants.2017.109.

(26) Wermuth, C. G. *The Practice of Medicinal Chemistry*; Elsevier, 2003.

(27) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, J. E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat Rev Drug Discov* **2020**, *19* (5), 353–364. https://doi.org/10.1038/s41573-019-0050-3.

(28) Childs-Disney, J. L.; Yang, X.; Gibaut, Q. M. R.; Tong, Y.; Batey, R. T.; Disney, M. D. Targeting RNA Structures with Small Molecules. *Nat Rev Drug Discov* **2022**, *21* (10), 736–762. https://doi.org/10.1038/s41573-022-00521-4.

(29) *RAS-targeted therapies: is the undruggable drugged? | Nature Reviews Drug Discovery*. https://www.nature.com/articles/s41573-020-0068-6 (accessed 2025-03-06).

(30) Sun, D.; Gao, W.; Hu, H.; Zhou, S. Why 90% of Clinical Drug Development Fails and How to Improve It? *Acta Pharmaceutica Sinica B* **2022**, *12* (7), 3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002.

(31) Scott, E. C.; Baines, A. C.; Gong, Y.; Moore, R.; Pamuk, G. E.; Saber, H.; Subedee, A.; Thompson, M. D.; Xiao, W.; Pazdur, R.; Rao, V. A.; Schneider, J.; Beaver, J. A. Trends in the Approval of Cancer Therapies by the FDA in the Twenty-First Century. *Nat Rev Drug Discov* **2023**, *22* (8), 625–640. https://doi.org/10.1038/s41573-023-00723-4.

(32) Hyde, S. C.; Gill, D. R.; Higgins, C. F.; Trezise, A. E. O.; MacVinish, L. J.; Cuthbert, A. W.; Ratcliff, R.; Evans, M. J.; Colledge, W. H. Correction of the Ion Transport Defect in Cystic Fibrosis Transgenic Mice by Gene Therapy. *Nature* **1993**, *362* (6417), 250–255. https://doi.org/10.1038/362250a0.

(33) Payandeh, J.; Volgraf, M. Ligand Binding at the Protein–Lipid Interface: Strategic Considerations for Drug Design. *Nat Rev Drug Discov* **2021**, *20* (9), 710–722. https://doi.org/10.1038/s41573-021-00240-2.

(34) Burke, J. E.; Triscott, J.; Emerling, B. M.; Hammond, G. R. V. Beyond PI3Ks: Targeting Phosphoinositide Kinases in Disease. *Nat Rev Drug Discov* **2023**, *22* (5), 357–386. https://doi.org/10.1038/s41573-022-00582-5.

(35) *Trends in kinase drug discovery: targets, indications and inhibitor design | Nature Reviews Drug Discovery*. https://www.nature.com/articles/s41573-021-00252-y (accessed 2025-03-06).

(36) Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. Trends in GPCR Drug Discovery: New Agents, Targets and Indications. *Nat Rev Drug Discov* **2017**, *16* (12), 829–842. https://doi.org/10.1038/nrd.2017.178.

(37) *Voltage-gated sodium channels in excitable cells as drug targets | Nature Reviews Drug Discovery*. https://www.nature.com/articles/s41573-024-01108-x (accessed 2025-03-06).

(38) Schneider, G. Automating Drug Discovery. *Nat Rev Drug Discov* **2018**, *17* (2), 97–113. https://doi.org/10.1038/nrd.2017.232.

(39) *Drug Discovery and Development in the Era of Big Data: Future Medicinal Chemistry: Vol 8 , No 15 - Get Access*. https://www.tandfonline.com/doi/full/10.4155/fmc-2014-0081 (accessed 2025-03-06).

(40) Scannell, J. W.; Bosley, J.; Hickman, J. A.; Dawson, G. R.; Truebel, H.; Ferreira, G. S.; Richards, D.; Treherne, J. M. Predictive Validity in Drug Discovery: What It Is, Why It Matters and How to Improve It. *Nat Rev Drug Discov* **2022**, *21* (12), 915–931. https://doi.org/10.1038/s41573-022-00552-x.

(41) Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *J. Chem. Inf. Model.* **2023**, *63* (13), 4042–4055. https://doi.org/10.1021/acs.jcim.3c00520.

(42) Franzini, R. M.; Randolph, C. Chemical Space of DNA-Encoded Libraries. *J. Med. Chem.* **2016**, *59* (14), 6629–6644. https://doi.org/10.1021/acs.jmedchem.5b01874.

(43) Lucas, X.; Grüning, B. A.; Bleher, S.; Günther, S. The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, *55* (5), 915–924. https://doi.org/10.1021/acs.jcim.5b00116.

(44) *Full article: A bright future for fragment-based drug discovery: what does it hold?* https://www.tandfonline.com/doi/full/10.1080/17460441.2019.1583643 (accessed 2025-03-06).

(45) Di, L.; Kerns, E. H. Profiling Drug-like Properties in Discovery Research. *Current Opinion in Chemical Biology* **2003**, *7* (3), 402–408. https://doi.org/10.1016/S1367-5931(03)00055-3.

(46) Curatolo, W. Physical Chemical Properties of Oral Drug Candidates in the Discovery and Exploratory Development Settings. *Pharmaceutical Science & Technology Today* **1998**, *1* (9), 387–393. https://doi.org/10.1016/S1461-5347(98)00097-2.

(47) Bhhatarai, B.; Walters, W. P.; Hop, C. E. C. A.; Lanza, G.; Ekins, S. Opportunities and Challenges Using Artificial Intelligence in ADME/Tox. *Nat. Mater.* **2019**, *18* (5), 418–422. https://doi.org/10.1038/s41563-019-0332-5.

(48) Langevin, M.; Bianciotto, M.; Vuilleumier, R. Balancing Exploration and Exploitation in de Novo Drug Design. *Digital Discovery* **2024**, *3* (12), 2572–2588. https://doi.org/10.1039/D4DD00105B.

(49) Swain, S. M.; Shastry, M.; Hamilton, E. Targeting HER2-Positive Breast Cancer: Advances and Future Directions. *Nat Rev Drug Discov* **2023**, *22* (2), 101–126. https://doi.org/10.1038/s41573-022-00579-0.

(50) *A preclinical secondary pharmacology resource illuminates target-adverse drug reaction associations of marketed drugs | Nature Communications*. https://www.nature.com/articles/s41467-023-40064-9 (accessed 2025-03-06).

(51) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat Rev Drug Discov* **2012**, *11* (12), 909–922. https://doi.org/10.1038/nrd3845.

(52) Schneider, P.; Schneider, G. De Novo Design at the Edge of Chaos. *J. Med. Chem.* **2016**, *59* (9), 4077–4086. https://doi.org/10.1021/acs.jmedchem.5b01849.

(53) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat Rev Drug Discov* **2013**, *12* (12), 948–962. https://doi.org/10.1038/nrd4128.

(54) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat Rev Drug Discov* **2004**, *3* (8), 711–716. https://doi.org/10.1038/nrd1470.

(55) Nicolaou, C. A.; Brown, N. Multi-Objective Optimization Methods in Drug Design. *Drug Discovery Today: Technologies* **2013**, *10* (3), e427–e435. https://doi.org/10.1016/j.ddtec.2013.02.001.

(56) Pognan, F.; Beilmann, M.; Boonen, H. C. M.; Czich, A.; Dear, G.; Hewitt, P.; Mow, T.; Oinonen, T.; Roth, A.; Steger-Hartmann, T.; Valentin, J.-P.; Van Goethem, F.; Weaver, R. J.; Newham, P. The Evolving Role of Investigative Toxicology in the Pharmaceutical Industry. *Nat Rev Drug Discov* **2023**, *22* (4), 317–335. https://doi.org/10.1038/s41573-022-00633-x.

(57) Eichler, H.-G.; Bloechl-Daum, B.; Abadie, E.; Barnett, D.; König, F.; Pearson, S. Relative Efficacy of Drugs: An Emerging Issue between Regulatory Agencies and Third-Party Payers. *Nat Rev Drug Discov* **2010**, *9* (4), 277–291. https://doi.org/10.1038/nrd3079.

(58) Urvas, L.; Chiesa, L.; Bret, G.; Jacquemard, C.; Kellenberger, E. Benchmarking AlphaFold-Generated Structures of Chemokine–Chemokine Receptor Complexes. *J. Chem. Inf. Model.* **2024**, *64* (11), 4587–4600. https://doi.org/10.1021/acs.jcim.3c01835.

(59) Sellami, A.; Réau, M.; Langenfeld, F.; Lagarde, N.; Montes, M. Chapter 5 - Virtual Libraries for Docking Methods: Guidelines for the Selection and the Preparation. In *Molecular Docking for Computer-Aided Drug Design*; Coumar, M. S., Ed.; Academic Press, 2021; pp 99–117. https://doi.org/10.1016/B978-0-12-822312-3.00017-5.

(60) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57* (2), 225–242. https://doi.org/10.1002/prot.20149.

(61) Wei, H.; McCammon, J. A. Structure and Dynamics in Drug Discovery. *npj Drug Discov.* **2024**, *1* (1), 1–8. https://doi.org/10.1038/s44386-024-00001-2.

(62) Mollica, L.; Theret, I.; Antoine, M.; Perron-Sierra, F.; Charton, Y.; Fourquez, J.-M.; Wierzbicki, M.; Boutin, J. A.; Ferry, G.; Decherchi, S.; Bottegoni, G.; Ducrot, P.; Cavalli, A. Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein–Ligand Residence Times. *J. Med. Chem.* **2016**, *59* (15), 7167–7176. https://doi.org/10.1021/acs.jmedchem.6b00632.

(63) *Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits | Scientific Reports*. https://www.nature.com/articles/s41598-022-06306-4 (accessed 2025-03-06).

(64) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(65) Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A. Integrating QSAR Modelling and Deep Learning in Drug Discovery: The Emergence of Deep QSAR. *Nat Rev Drug Discov* **2024**, *23* (2), 141–155. https://doi.org/10.1038/s41573-023-00832-0.

(66) *Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking | Nature Protocols*. https://www.nature.com/articles/s41596-021-00659-2 (accessed 2025-03-06).

(67) Sturm, N.; Desaphy, J.; Quinn, R. J.; Rognan, D.; Kellenberger, E. Structural Insights into the Molecular Basis of the Ligand Promiscuity. *J. Chem. Inf. Model.* **2012**, *52* (9), 2410–2421. https://doi.org/10.1021/ci300196g.

(68) Catacutan, D. B.; Alexander, J.; Arnold, A.; Stokes, J. M. Machine Learning in Preclinical Drug Discovery. *Nat Chem Biol* **2024**, *20* (8), 960–973. https://doi.org/10.1038/s41589-024-01679-1.

(69) Reynès, C.; Host, H.; Camproux, A.-C.; Laconde, G.; Leroux, F.; Mazars, A.; Deprez, B.; Fahraeus, R.; Villoutreix, B. O.; Sperandio, O. Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors Using Machine-Learning Methods. *PLOS Computational Biology* **2010**, *6* (3), e1000695. https://doi.org/10.1371/journal.pcbi.1000695.

(70) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the Performance of 3D Virtual Screening Protocols: RMSD Comparisons, Enrichment Assessments, and Decoy Selection—What Can We Learn from Earlier Mistakes? *J Comput Aided Mol Des* **2008**, *22* (3), 213–228. https://doi.org/10.1007/s10822-007-9163-6.

(71) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49* (3), 678–692. https://doi.org/10.1021/ci8004226.

(72) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat Rev Drug Discov* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.

(73) Schneider, G. Generative Models for Artificially-Intelligent Molecular Design. *Molecular Informatics* **2018**, *37* (1–2), 1880131. https://doi.org/10.1002/minf.201880131.

(74) *Deep generative models for ligand-based de novo design applied to multi-parametric optimization - Perron - 2022 - Journal of Computational Chemistry - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.26826 (accessed 2025-03-06).

(75) Team, D. D. W. *Chemical Space, High Throughput Screening and The World Of Blockbuster Drugs*. Drug Discovery World (DDW). https://www.ddw-online.com/chemical-space-high-throughput-screening-and-the-world-of-blockbuster-drugs-1528-201304/ (accessed 2025-03-09).

(76) Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the Decline in Pharmaceutical R&D Efficiency. *Nat Rev Drug Discov* **2012**, *11* (3), 191–200. https://doi.org/10.1038/nrd3681.

(77) Gloriam, D. E. Bigger Is Better in Virtual Drug Screens. *Nature* **2019**, *566* (7743), 193–194. https://doi.org/10.1038/d41586-019-00145-6.

(78) Luttens, A.; Gullberg, H.; Abdurakhmanov, E.; Vo, D. D.; Akaberi, D.; Talibov, V. O.; Nekhotiaeva, N.; Vangeel, L.; De Jonghe, S.; Jochmans, D.; Krambrich, J.; Tas, A.; Lundgren, B.; Gravenfors, Y.; Craig, A. J.; Atilaw, Y.; Sandström, A.; Moodie, L. W. K.; Lundkvist, Å.; van Hemert, M. J.; Neyts, J.; Lennerstrand, J.; Kihlberg, J.; Sandberg, K.; Danielson, U. H.; Carlsson, J. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main

Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J. Am. Chem. Soc.* **2022**, *144* (7), 2905–2920. https://doi.org/10.1021/jacs.1c08402.

(79) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

(80) Lagarde, N.; Zagury, J.-F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, *55* (7), 1297–1307. https://doi.org/10.1021/acs.jcim.5b00090.

(81) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Research* **2012**, *40* (D1), D400–D412. https://doi.org/10.1093/nar/gkr1132.

(82) *ChEMBL: a large-scale bioactivity database for drug discovery | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/40/D1/D1100/2903401 (accessed 2025-03-06).

(83) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Combinatorial Chemistry & High Throughput Screening* **2001**, *4* (8), 719–725. https://doi.org/10.2174/1386207013330670.

(84) Oprisiu, I.; Novotarskyi, S.; Tetko, I. V. Modeling of Non-Additive Mixture Properties Using the Online CHEmical Database and Modeling Environment (OCHEM). *J Cheminform* **2013**, *5* (1), 4. https://doi.org/10.1186/1758-2946-5-4.

(85) Attene-Ramos, M. S.; Miller, N.; Huang, R.; Michael, S.; Itkin, M.; Kavlock, R. J.; Austin, C. P.; Shinn, P.; Simeonov, A.; Tice, R. R.; Xia, M. The Tox21 Robotic Platform for the Assessment of Environmental Chemicals – from Vision to Reality. *Drug Discovery Today* **2013**, *18* (15), 716–723. https://doi.org/10.1016/j.drudis.2013.05.015.

(86) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. arXiv August 28, 2021. https://doi.org/10.48550/arXiv.2102.09548.

(87) Wu, Z.; Ramsundar, B.; N. Feinberg, E.; Gomes, J.; Geniesse, C.; S. Pappu, A.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.

(88) Wognum, C.; Ash, J. R.; Aldeghi, M.; Rodríguez-Pérez, R.; Fang, C.; Cheng, A. C.; Price, D. J.; Clevert, D.-A.; Engkvist, O.; Walters, W. P. A Call for an Industry-Led Initiative to Critically Assess Machine Learning for Real-World Drug Discovery. *Nat Mach Intell* **2024**, *6* (10), 1120–1121. https://doi.org/10.1038/s42256-024-00911-w.

(89) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010. https://doi.org/10.1021/jm4004285.

(90) Guner, O. F.; Hughes, D. W.; Dumont, L. M. *An integrated approach to three-dimensional information management with MACCS-3D*. ACS Publications. https://doi.org/10.1021/ci00003a007.

(91) *RDKit*. https://www.rdkit.org/ (accessed 2025-01-02).

(92) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design* **2006**, *12* (17), 2111–2120. https://doi.org/10.2174/138161206777585274.

(93) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J Cheminform* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(94) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Molecular Informatics* **2010**, *29* (12), 855–868. https://doi.org/10.1002/minf.201000099.

(95) *Molecular Operating Environment (MOE) | MOEsaic | PSILO*. https://www.chemcomp.com/en/Products.htm (accessed 2025-03-07).

(96) *rdkit.Avalon.pyAvalonTools module — The RDKit 2024.09.6 documentation*. https://www.rdkit.org/docs/source/rdkit.Avalon.pyAvalonTools.html (accessed 2025-03-07).

(97) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(98) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *Atom pairs as molecular features in structure-activity studies: definition and applications*. ACS Publications. https://doi.org/10.1021/ci00046a002.

(99) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(100) Gedeck, P.; Rohde, B.; Bartels, C. QSAR − How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46* (5), 1924–1936. https://doi.org/10.1021/ci050413p.

(101) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58* (1), 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

(102) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv October 23, 2020. https://doi.org/10.48550/arXiv.2010.09885.

(103) Abdi, H.; Williams, L. J. Principal Component Analysis. *WIREs Computational Statistics* **2010**, *2* (4), 433–459. https://doi.org/10.1002/wics.101.

(104) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579–2605.

(105) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 18, 2020. https://doi.org/10.48550/arXiv.1802.03426.

(106) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Computation* **1998**, *10* (1), 215–234. https://doi.org/10.1162/089976698300017953.

(107) Kaski, S. Self-Organizing Maps. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2010; pp 886–888. https://doi.org/10.1007/978-0-387-30164-8_746.

(108) Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping in Drug Design. *Drug Discovery Today: Technologies* **2019**, *32–33*, 99–107. https://doi.org/10.1016/j.ddtec.2020.06.003.

(109) Brown, A. C.; Fraser, T. R. 4. On the Changes Produced by Direct Chemical Addition on the Physiological Action of Certain Poisons. *Proceedings of the Royal Society of Edinburgh* **1869**, *6*, 228–232. https://doi.org/10.1017/S0370164600045843.

(110) *Revue des travaux scientifiques*; Imprimerie nationale, 1894.

(111) Meyer, H. Zur Theorie der Alkoholnarkose. *Archiv f. experiment. Pathol. u. Pharmakol* **1899**, *42* (2), 109–118. https://doi.org/10.1007/BF01834479.

(112) *Membrane Permeability: 100 Years Since Ernest Overton*; Academic Press, 1999.

(113) Dearden, J. C. The History and Development of Quantitative Structure-Activity Relationships (QSARs): Addendum. *IJQSPR* **2017**, *2* (2), 36–46. https://doi.org/10.4018/IJQSPR.2017070104.

(114) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman and Hall/CRC: New York, 2017. https://doi.org/10.1201/9781315139470.

(115) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(116) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **1997**, *55* (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

(117) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001**, *29* (5), 1189–1232.

(118) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.

(119) Goldberger, J.; Hinton, G. E.; Roweis, S.; Salakhutdinov, R. R. Neighbourhood Components Analysis. In *Advances in Neural Information Processing Systems*; MIT Press, 2004; Vol. 17.

(120) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.

(121) Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*; Lechevallier, Y., Saporta, G., Eds.; Physica-Verlag HD: Heidelberg, 2010; pp 177–186. https://doi.org/10.1007/978-3-7908-2604-3_16.

(122) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. arXiv January 30, 2017. https://doi.org/10.48550/arXiv.1412.6980.

(123) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*. https://doi.org/10.3389/fenvs.2015.00080.

(124) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(125) *Generalized biomolecular modeling and design with RoseTTAFold All-Atom | Science*. https://www.science.org/doi/full/10.1126/science.adl2528 (accessed 2025-03-06).

(126) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J Cheminform* **2021**, *13* (1), 12. https://doi.org/10.1186/s13321-020-00479-8.

(127) Tran, H. N. T.; Joshua Thomas, J.; Malim, N. H. A. H.; Ali, A. M.; Huynh, S. B. Graph Neural Networks in Cheminformatics. In *Intelligent Computing and Optimization*; Vasant,

P., Zelinka, I., Weber, G.-W., Eds.; Springer International Publishing: Cham, 2021; pp 823–837. https://doi.org/10.1007/978-3-030-68154-8_71.

(128) Wang, Y.; Gu, Y.; Lou, C.; Gong, Y.; Wu, Z.; Li, W.; Tang, Y.; Liu, G. A Multitask GNN-Based Interpretable Model for Discovery of Selective JAK Inhibitors. *J Cheminform* **2022**, *14* (1), 16. https://doi.org/10.1186/s13321-022-00593-9.

(129) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discovery Today: Technologies* **2020**, *37*, 1–12. https://doi.org/10.1016/j.ddtec.2020.11.009.

(130) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63* (16), 8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959.

(131) Breiman, L. Bagging Predictors. *Mach Learn* **1996**, *24* (2), 123–140. https://doi.org/10.1007/BF00058655.

(132) *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S002200009791504X (accessed 2025-03-08).

(133) Caruana, R. Multitask Learning. *Machine Learning* **1997**, *28* (1), 41–75. https://doi.org/10.1023/A:1007379606734.

(134) Gong, T.; Lee, T.; Stephenson, C.; Renduchintala, V.; Padhy, S.; Ndirango, A.; Keskin, G.; Elibol, O. H. A Comparison of Loss Weighting Strategies for Multi Task Learning in Deep Neural Networks. *IEEE Access* **2019**, *7*, 141627–141632. https://doi.org/10.1109/ACCESS.2019.2943604.

(135) Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; Rabinovich, A. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR, 2018; pp 794–803.

(136) Lin, X.; Zhen, H.-L.; Li, Z.; Zhang, Q.-F.; Kwong, S. Pareto Multi-Task Learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2019; Vol. 32.

(137) Mao, Y.; Wang, Z.; Liu, W.; Lin, X.; Xie, P. MetaWeighting: Learning to Weight Tasks in Multi-Task Learning. In *Findings of the Association for Computational Linguistics: ACL 2022*; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2022; pp 3436–3448. https://doi.org/10.18653/v1/2022.findings-acl.271.

(138) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. arXiv February 6, 2015. https://doi.org/10.48550/arXiv.1502.02072.

(139) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Current Pharmaceutical Design* **2007**, *13* (34), 3494–3504. https://doi.org/10.2174/138161207782794257.

(140) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29* (6–7), 476–488. https://doi.org/10.1002/minf.201000061.

(141) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition$. *SAR and QSAR in Environmental Research* **2016**, *27* (11), 865–881. https://doi.org/10.1080/1062936X.2016.1250229.

(142) Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R. Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J. Chem. Inf. Model.* **2016**, *56* (11), 2243–2252. https://doi.org/10.1021/acs.jcim.6b00291.

(143) *Chemoinformatic Classification Methods and their Applicability Domain - Mathea - 2016 - Molecular Informatics - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1002/minf.201501019 (accessed 2025-03-06).

(144) Manevitz, L. M.; Yousef, M. One-Class Svms for Document Classification. *J. Mach. Learn. Res.* **2002**, *2*, 139–154.

(145) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29* (2), 93–104. https://doi.org/10.1145/335191.335388.

(146) *Test No. 105: Water Solubility | OECD*. https://www.oecd.org/en/publications/test-no-105-water-solubility_9789264069589-en.html (accessed 2025-03-06).

(147) *Pegylated Phosphine Ligands in Iridium(I) Catalyzed Hydrogen Isotope Exchange Reactions in Aqueous Buffers - Martinelli - 2024 - Chemistry – A European Journal - Wiley Online Library*. https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/chem.202402038 (accessed 2025-04-03).

(148) Kerns, E. H.; Di, L.; Petusky, S.; Farris, M.; Ley, R.; Jupp, P. Combined Application of Parallel Artificial Membrane Permeability Assay and Caco-2 Permeability Assays in Drug Discovery. *Journal of Pharmaceutical Sciences* **2004**, *93* (6), 1440–1453. https://doi.org/10.1002/jps.20075.

(149) Matijašić, M.; Munić Kos, V.; Nujić, K.; Čužić, S.; Padovan, J.; Kragol, G.; Alihodžić, S.; Mildner, B.; Verbanac, D.; Eraković Haber, V. Fluorescently Labeled Macrolides as a Tool for Monitoring Cellular and Tissue Distribution of Azithromycin. *Pharmacological Research* **2012**, *66* (4), 332–342. https://doi.org/10.1016/j.phrs.2012.06.001.

(150) Trainor, G. L. The Importance of Plasma Protein Binding in Drug Discovery. *Expert Opinion on Drug Discovery* **2007**, *2* (1), 51–64. https://doi.org/10.1517/17460441.2.1.51.

(151) Abbott, N. J.; Patabendige, A. A. K.; Dolman, D. E. M.; Yusof, S. R.; Begley, D. J. Structure and Function of the Blood–Brain Barrier. *Neurobiology of Disease* **2010**, *37* (1), 13–25. https://doi.org/10.1016/j.nbd.2009.07.030.

(152) Rendic, S.; Carlo, F. J. D. Human Cytochrome P450 Enzymes: A Status Report Summarizing Their Reactions, Substrates, Inducers, and Inhibitors. *Drug Metabolism Reviews* **1997**. https://doi.org/10.3109/03602539709037591.

(153) Bhutani, P.; Joshi, G.; Raja, N.; Bachhav, N.; Rajanna, P. K.; Bhutani, H.; Paul, A. T.; Kumar, R. U.S. FDA Approved Drugs from 2015–June 2020: A Perspective. *J. Med. Chem.* **2021**, *64* (5), 2339–2381. https://doi.org/10.1021/acs.jmedchem.0c01786.

(154) *Bile Acid Metabolism and Signaling - Chiang - Major Reference Works - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1002/cphy.c120023 (accessed 2025-03-06).

(155) Fermini, B.; Fossa, A. A. The Impact of Drug-Induced QT Interval Prolongation on Drug Discovery and Development. *Nat Rev Drug Discov* **2003**, *2* (6), 439–447. https://doi.org/10.1038/nrd1108.

(156) Schmidtke, P.; Ciantar, M.; Theret, I.; Ducrot, P. Dynamics of hERG Closure Allow Novel Insights into hERG Blocking by Small Molecules. *J. Chem. Inf. Model.* **2014**, *54* (8), 2320–2333. https://doi.org/10.1021/ci5001373.

(157) Istvan, E. S.; Deisenhofer, J. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. *Science* **2001**, *292* (5519), 1160–1164. https://doi.org/10.1126/science.1059344.

(158) Testa, B.; Krämer, S. D. *The Biochemistry of Drug Metabolism: Volume 2: Conjugations, Consequences of Metabolism, Influencing Factors*; Wiley, 2010.

(159) Dorato, M. A.; Engelhardt, J. A. The No-Observed-Adverse-Effect-Level in Drug Safety Evaluations: Use, Issues, and Definition(s). *Regulatory Toxicology and Pharmacology* **2005**, *42* (3), 265–274. https://doi.org/10.1016/j.yrtph.2005.05.004.

(160) Villoutreix, B. O.; Taboureau, O. Computational Investigations of hERG Channel Blockers: New Insights and Current Predictive Models. *Advanced Drug Delivery Reviews* **2015**, *86*, 72–82. https://doi.org/10.1016/j.addr.2015.03.003.

(161) Li, Q.; Jørgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG Classification Model Based on a Combination of Support Vector Machine Method and GRIND Descriptors. *Mol. Pharmaceutics* **2008**, *5* (1), 117–127. https://doi.org/10.1021/mp700124e.

(162) Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Steen Jørgensen, F.; Vermeulen, N. P. E.; Oostenbrink, C. Virtual Screening and Prediction of Site of Metabolism for Cytochrome P450 1A2 Ligands. *J. Chem. Inf. Model.* **2009**, *49* (1), 43–52. https://doi.org/10.1021/ci800371f.

(163) Tyzack, J. D.; Kirchmair, J. Computational Methods and Tools to Predict Cytochrome P450 Metabolism for Drug Discovery. *Chemical Biology & Drug Design* **2019**, *93* (4), 377–386. https://doi.org/10.1111/cbdd.13445.

(164) Plonka, W.; Stork, C.; Šícho, M.; Kirchmair, J. CYPlebrity: Machine Learning Models for the Prediction of Inhibitors of Cytochrome P450 Enzymes. *Bioorganic & Medicinal Chemistry* **2021**, *46*, 116388. https://doi.org/10.1016/j.bmc.2021.116388.

(165) Goldwaser, E.; Laurent, C.; Lagarde, N.; Fabrega, S.; Nay, L.; Villoutreix, B. O.; Jelsch, C.; Nicot, A. B.; Loriot, M.-A.; Miteva, M. A. Machine Learning-Driven Identification of Drugs Inhibiting Cytochrome P450 2C9. *PLOS Computational Biology* **2022**, *18* (1), e1009820. https://doi.org/10.1371/journal.pcbi.1009820.

(166) *Towards Non-Animal Testing in European Regulatory Toxicology: An Introduction to the REACH Framework and Challenges in Implementing the 3Rs | European Journal of Risk Regulation | Cambridge Core*. https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/towards-nonanimal-testing-in-european-regulatory-toxicology-an-introduction-to-the-reach-framework-and-challenges-in-implementing-the-3rs/A6E22DA7D85EB03886BFDC121EA1E88A (accessed 2025-03-06).

(167) Gleeson, M. P.; Modi, S.; Bender, A.; Robinson, R. L. M.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The Challenges Involved in Modeling Toxicity Data In Silico: A Review. *Current Pharmaceutical Design* **2012**, *18* (9), 1266–1291. https://doi.org/10.2174/138920012799362819.

(168) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting Drug Metabolism: Experiment and/or Computation? *Nat Rev Drug Discov* **2015**, *14* (6), 387–404. https://doi.org/10.1038/nrd4581.

(169) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012**, *52* (3), 617–648. https://doi.org/10.1021/ci200542m.

(170) Kirchmair, J.; Howlett, A.; Peironcely, J. E.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van Buren, L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted? *J. Chem. Inf. Model.* **2013**, *53* (2), 354–367. https://doi.org/10.1021/ci300487z.

(171) Galton, F. Vox Populi. *Nature* **1907**, *75* (1949), 450–451. https://doi.org/10.1038/075450a0.

(172) *The Delphi Method: An Experimental Study of Group Opinion*. https://apps.dtic.mil/sti/citations/trecms/AD0690498 (accessed 2025-03-06).

(173) Levy, P.; Bononno, R. *Collective Intelligence: Mankind's Emerging World in Cyberspace*; Perseus Books: USA, 1997.

(174) Surowiecki, J. *The Wisdom of Crowds*; Anchor, 2005.

(175) Ungar, L.; Mellers, B.; Satopaa, V. A.; Tetlock, P.; Baron, J. The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions; 2012.

(176) *De novo protein design by citizen scientists | Nature*. https://www.nature.com/articles/s41586-019-1274-4 (accessed 2025-03-06).

(177) Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A. M.; Alberga, D.; Alves, V. M.; Andersson, P. L.; Andrade, C. H.; Bai, F.; Balabin, I.; Ballabio, D.; Benfenati, E.; Bhhatarai, B.; Boyer, S.; Chen, J.; Consonni, V.; Farag, S.; Fourches, D.; García-Sosa, A. T.; Gramatica, P.; Grisoni, F.; Grulke, C. M.; Hong, H.; Horvath, D.; Hu, X.; Huang, R.; Jeliazkova, N.; Li, J.; Li, X.; Liu, H.; Manganelli, S.; Mangiatordi, G. F.; Maran, U.; Marcou, G.; Martin, T.; Muratov, E.; Nguyen, D.-T.; Nicolotti, O.; Nikolov, N. G.; Norinder, U.; Papa, E.; Petitjean, M.; Piir, G.; Pogodin, P.; Poroikov, V.; Qiao, X.; Richard, A. M.; Roncaglioni, A.; Ruiz, P.; Rupakheti, C.; Sakkiah, S.; Sangion, A.; Schramm, K.-W.; Selvaraj, C.; Shah, I.; Sild, S.; Sun, L.; Taboureau, O.; Tang, Y.; Tetko, I. V.; Todeschini, R.; Tong, W.; Trisciuzzi, D.; Tropsha, A.; Van Den Driessche, G.; Varnek, A.; Wang, Z.; Wedebye, E. B.; Williams, A. J.; Xie, H.; Zakharov, A. V.; Zheng, Z.; Judson, R. S. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives* **2020**, *128* (2), 027002. https://doi.org/10.1289/EHP5580.

(178) Schilling-Wilhelmi, M.; Ríos-García, M.; Shabih, S.; Victoria Gil, M.; Miret, S.; T. Koch, C.; A. Márquez, J.; Maik Jablonka, K. From Text to Insight: Large Language Models for Chemical Data Extraction. *Chemical Society Reviews* **2025**, *54* (3), 1125–1150. https://doi.org/10.1039/D4CS00913D.

(179) Ai, Q.; Meng, F.; Shi, J.; Pelkie, B.; W. Coley, C. Extracting Structured Data from Organic Synthesis Procedures Using a Fine-Tuned Large Language Model. *Digital Discovery* **2024**, *3* (9), 1822–1831. https://doi.org/10.1039/D4DD00091A.

(180) Vangala, S. R.; Krishnan, S. R.; Bung, N.; Nandagopal, D.; Ramasamy, G.; Kumar, S.; Sankaran, S.; Srinivasan, R.; Roy, A. Suitability of Large Language Models for Extraction of High-Quality Chemical Reaction Dataset from Patent Literature. *J Cheminform* **2024**, *16* (1), 131. https://doi.org/10.1186/s13321-024-00928-8.

(181) *Mining Patents with Large Language Models Demonstrates Congruence of Functional Labels and Chemical Structures | OpenReview*. https://openreview.net/forum?id=IhD1rBHhDy (accessed 2025-04-04).

(182) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting Large Language Models with Chemistry Tools. *Nat Mach Intell* **2024**, *6* (5), 525–535. https://doi.org/10.1038/s42256-024-00832-8.

(183) Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.; Zhou, D.; Zhang, S.; Su, M.; Zhong, H.-S.; Li, Y. ChemLLM: A Chemical Large Language Model. arXiv April 25, 2024. https://doi.org/10.48550/arXiv.2402.06852.

(184) Jordan, A. M. Artificial Intelligence in Drug Design—The Storm Before the Calm? *ACS Med. Chem. Lett.* **2018**, *9* (12), 1150–1152. https://doi.org/10.1021/acsmedchemlett.8b00500.

(185) *AI: a great crash of hype into reality*. Drug Target Review. https://www.drugtargetreview.com/article/108086/artificial-intelligence-ai-a-great-crash-of-hype-into-reality/ (accessed 2025-04-03).

(186) Vijayan, R. S. K.; Kihlberg, J.; Cross, J. B.; Poongavanam, V. Enhancing Preclinical Drug Discovery with Artificial Intelligence. *Drug Discovery Today* **2022**, *27* (4), 967–984. https://doi.org/10.1016/j.drudis.2021.11.023.

(187) AI's Potential to Accelerate Drug Discovery Needs a Reality Check. *Nature* **2023**, *622* (7982), 217–217. https://doi.org/10.1038/d41586-023-03172-6.

(188) Pitt, W. R.; Bentley, J.; Boldron, C.; Colliandre, L.; Esposito, C.; Frush, E. H.; Kopec, J.; Labouille, S.; Meneyrol, J.; Pardoe, D. A.; Palazzesi, F.; Pozzan, A.; Remington, J. M.; Rex, R.; Southey, M.; Vishwakarma, S.; Walker, P. Real-World Applications and Experiences of AI/ML Deployment for Drug Discovery. *J. Med. Chem.* **2025**, *68* (2), 851–859. https://doi.org/10.1021/acs.jmedchem.4c03044.

(189) Masters, M. R.; Mahmoud, A. H.; Lill, M. A. Do Deep Learning Models for Co-Folding Learn the Physics of Protein-Ligand Interactions? bioRxiv June 4, 2024, p 2024.06.03.597219. https://doi.org/10.1101/2024.06.03.597219.

(190) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(191) *Accurate structure prediction of biomolecular interactions with AlphaFold 3 | Nature*. https://www.nature.com/articles/s41586-024-07487-w (accessed 2025-03-06).

(192) *AI protein-prediction tool AlphaFold3 is now more open*. https://www.nature.com/articles/d41586-024-03708-4 (accessed 2025-03-06).

(193) Joshi, S. *The Future of KRAS Targeting Cancer Therapies Beyond G12C*. DelveInsight Business Research. https://www.delveinsight.com/blog/kras-inhibitors-beyond-g12c (accessed 2025-03-06).

(194) *Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors | Nature Biotechnology*. https://www.nature.com/articles/s41587-024-02526-3 (accessed 2025-03-06).

(195) Betting on Qubits. *Nat Electron* **2025**, *8* (1), 1–2. https://doi.org/10.1038/s41928-025-01346-w.

(196) *Deep learning enables rapid identification of potent DDR1 kinase inhibitors | Nature Biotechnology*. https://www.nature.com/articles/s41587-019-0224-x (accessed 2025-03-06).

(197) *Assessing the impact of generative AI on medicinal chemistry | Nature Biotechnology*. https://www.nature.com/articles/s41587-020-0418-2 (accessed 2025-03-06).

(198) *Insilico Medicine, an early force in AI-driven drug R&D, lets IPO plans peter out for second year in a row*. https://endpts.com/insilico-medicine-an-early-force-in-ai-driven-drug-rd-lets-ipo-plans-peter-out-for-second-year-in-a-row/ (accessed 2025-03-06).

(199) Artificial Intelligence-Created Medicine to Be Used on Humans for First Time. January 30, 2020. https://www.bbc.com/news/technology-51315462 (accessed 2025-03-06).

(200) Armstrong, A. *Exscientia fires CEO Andrew Hopkins effective immediately*. https://www.fiercebiotech.com/biotech/exscientia-fires-ceo-andrew-hopkins-over-inappropriate-relationships-2-employees (accessed 2025-03-06).

(201) Sunny, M. E. Nvidia-Backed Recursion's Shares Fall on Mixed Data for Rare Disorder Drug. *Reuters*. September 3, 2024. https://www.reuters.com/business/healthcare-pharmaceuticals/recursions-blood-vessel-disorder-drug-meets-main-goal-mid-stage-study-2024-09-03/ (accessed 2025-03-06).

(202) Buhr, S. *YC-Backed Atomwise Discovers Drugs For Diseases That Don't Even Exist Yet*. TechCrunch. https://techcrunch.com/2015/03/06/y-combinator-backed-atomwise-discovers-drugs-for-diseases-that-dont-even-exist-yet/ (accessed 2025-03-06).

(203) BenevolentAI. *BenevolentAI Raises $115 Million to Extend Its Leading Global Position in the Field of AI Enabled Drug Development*. https://www.prnewswire.com/news-releases/benevolentai-raises-115-million-to-extend-its-leading-global-position-in-the-field-of-ai-enabled-drug-development-680180573.html (accessed 2025-03-06).

(204) Taylor, N. P. *BenevolentAI flunks midphase eczema trial to dent deal plans*. https://www.fiercebiotech.com/biotech/benevolentai-cruel-rd-ai-enabled-drug-flunks-midphase-eczema-trial-dent-deal-plans (accessed 2025-03-06).

(205) Post; Share; Post; Print; Email; License. *Believe the hype? Mixed signals from AI's impact on drug development*. PharmaVoice. https://www.pharmavoice.com/news/artificial-intelligence-hype-ai-drug-development/704153/ (accessed 2025-03-06).

(206) Correspondent, A. R., Chief Business. *How tech entrepreneur is taking BenevolentAI back to its roots*. https://www.thetimes.com/article/how-tech-entrepreneur-is-taking-benevolentai-back-to-its-roots-zth6ffjwv (accessed 2025-03-06).

(207) Masson, G. *Big Pharma layoff rounds jump 281% in 2024*. https://www.fiercebiotech.com/biotech/big-pharma-layoff-rounds-jump-281-24-while-total-biopharma-staff-cuts-similar-23 (accessed 2025-03-06).

(208) Armstrong, A. *Exscientia cuts a quarter of staff while preserving AI pipeline*. https://www.fiercebiotech.com/biotech/exscientia-cuts-quarter-staff-while-preserving-ai-generated-pipeline (accessed 2025-03-06).

(209) *BenevolentAI lays off around 180 staffers, cuts pipeline programs in reorg*. Endpoints News. https://endpts.com/benevolentai-lays-off-around-180-staffers-cuts-pipeline-programs-in-reorg/ (accessed 2025-03-06).

(210) *Layoffs Continued Across Biopharma in 2024*. BioSpace. https://www.biospace.com/job-trends/layoffs-continued-across-biopharma-in-2024 (accessed 2025-03-06).

(211) Buntz, B. *Analyzing the biotech and pharma layoffs in 2024*. Drug Discovery and Development. https://www.drugdiscoverytrends.com/mapping-2024-biotech-and-pharma-layoffs/ (accessed 2025-03-06).

Pierre Llompart

Conception moléculaire par IA multitâche et exploration chémographique

Résumé

Cette thèse vise à faire progresser le rôle de la modélisation in silico dans la recherche pharmaceutique, en abordant les défis persistants liés aux échecs tardifs et aux inefficacités du développement de médicaments. L'évaluation ADMET (Absorption, Distribution, Métabolisme, Élimination et Toxicité) intervient souvent trop tard dans le processus, augmentant ainsi les coûts et ralentissant la progression. Pour remédier à ces problèmes, la modélisation in silico, en particulier la prédiction précoce des propriétés ADMET (eADMET), est devenue essentielle pour rationaliser la prise de décision dès les premières étapes de la découverte de médicaments. Cependant, la complexité de la biologie humaine, l'évolution des modèles expérimentaux et les incohérences des données exigent des modèles prédictifs à la fois précis, adaptables et interprétables. Cette thèse propose une approche systématique pour l'amélioration de la modélisation eADMET, en s'appuyant sur le nettoyage des données, l'apprentissage multi-tâches, l'application à grande échelle et la collaboration humain-machine.

Résumé en anglais

This thesis is dedicated to advancing the role of in silico modeling in pharmaceutical research, addressing the persistent challenges of late-stage failures and inefficiencies in drug development. ADMET (Absorption, Distribution, Metabolism, Elimination, and Toxicity) testing often occurs too late in the pipeline, driving up costs and delaying progress. To mitigate these issues, in silico modeling, particularly early ADMET (eADMET) prediction, has become essential for streamlining decision-making in early drug discovery. However, the complexity of human biology, evolving assays, and data inconsistencies necessitate predictive models that are not only accurate but also adaptable and interpretable. This thesis presents a systematic approach to refining eADMET modeling through data curation, multi-task learning, large-scale applicability, and human–machine collaboration.