UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE MSII

Le laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie

THÈSE présentée par :

KHOUSSA Khoukha

soutenue le : **01 Octobre 2025**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : **Informatique - Informatique**

Application of Artificial Intelligence to Accelerate the Development of the Active Layer of Organic Photovoltaic Cells

THÈSE dirigée par :

Dr. LEVEQUE Patrick	Maitre de conférences HDR, Université de Strasbourg, France
Pr. BOUBCHIR Larbi	Professeur, Université de Paris 8, France

RAPPORTEURS :

Pr. NAÏT-ALI Amine	Professeur, Université de Paris-Est Créteil, France
Pr. EL YACOUBI Mounîm A	Professeur, Télécom SudParis, France

AUTRES MEMBRES DU JURY :

Pr. BENKHELIFA Elhadj	Professeur, Université de Staffordshire, Royaume Uni
Dr. MARTIN Evelyne	Directrice de recherche, CNRS, France

إهداء

إلى والديَّ العزيزين، رمزَي الحب والعطاء، اللذين كانا دائمًا السند
والدافع الأكبر لي في كل خطوة من حياتي...
إلى زوجي العزيز، رفيق دربي، وسندي في كل خطوة...
إلى عائلتي الكريمة التي منحتني الدعم والتشجيع والصبر...
إلى كل من آمن بي وساعدني على أن أواصل الطريق رغم الصعوبات...
لكم جميعًا أُهدي هذا الإنجاز

Aknowledgement

Writing these acknowledgments is one of the most emotionally complex parts of this thesis. This work represents several years of effort, reflection, and learning, and I am deeply aware that I would not have reached this point without the support, encouragement, and collaboration of many remarkable individuals. As I try to gather my thoughts and express my gratitude, I find myself overwhelmed by the number of people to whom I owe sincere thanks. I extend my deepest apologies to anyone I may have unintentionally forgotten to mention please know that your contribution is not overlooked, and I am truly grateful for everything you have done.

First and foremost, I wish to express my profound gratitude to my thesis supervisors, Dr. Patrick LEVEQUE and Pr. Larbi BOUBCHIR, for their exceptional mentorship, patience, and unwavering support throughout the course of my doctoral research. I feel extremely privileged to have had the opportunity to work under their supervision. Their scientific rigor, insightful feedback, and constant encouragement have been instrumental in shaping both the content of this thesis and my overall approach to research. They always believed in my potential, even during moments of doubt, and provided me with the confidence and motivation to persevere. Beyond their academic guidance, I am also thankful for their kindness, availability, and human qualities, which made my doctoral journey much more meaningful and rewarding.

I would also like to sincerely thank the members of the thesis jury. I am deeply honored that they have agreed to review and evaluate my work, and I truly appreciate the time and effort they will dedicate to reading this manuscript and providing their feedback. Their presence and participation are greatly valued.

My sincere thanks go to all the members of the ICube Laboratory in Cronenbourg, where I had the opportunity to carry out a significant part of my research. I am grateful to the professors, researchers, and fellow PhD students for their warm welcome, technical support, and valuable discussions. I am also deeply appreciative of the support provided by the LIASD Laboratory in Paris, where I found a second academic home. Working within this laboratory allowed me to broaden my research perspectives and benefit from fruitful scientific exchanges.

Beyond academic circles, I wish to thank all the colleagues and friends I met during my doctoral journey both in France and Algeria like Madjed, Ranzo, Achille, Pierre, Cheikh, Thomas, Alaeddine, Mohamed, Sofiane, Chahinez, Leila, Imene, Fatima Zahra and Amina who accompanied me along the way. Your friendship, encouragement, and support made a significant difference during challenging moments and long days of work. I am truly grateful for the shared experiences, laughter, and countless conversations that kept me grounded and motivated. Whether near or far, your presence in my life has been a source of strength.

I would like to express my deepest gratitude to my family, whose love, encouragement, and unwavering belief in me were the foundation of my strength throughout these years. To my parents, I owe everything. Thank you for your endless sacrifices, for always standing by my side, and for instilling in me the values of perseverance and determination. Your trust and love have guided me every step of the way, and this achievement belongs as much to you as it does to me.

To my siblings and sisters in low, thank you for your understanding, your kind words, and your unconditional support, even from afar. You have always encouraged me to push forward, and I am grateful for the strong bond that connects us.

To my husband, thank you for your immense patience, your emotional support, and your unwavering confidence in me during the most demanding phases of this journey. You have been my anchor through the ups and downs, and your presence has been a constant source of comfort and reassurance.

Finally, I wish to thank God, the Almighty, for granting me the strength, courage, and resilience to complete this work. Without divine guidance and support, none of this would have been possible.

To everyone who has contributed in some way to this journey whether through academic collaboration, moral support, friendship, or love thank you from the bottom of my heart.

Summary

List of Abbreviations

Abbreviation	**Meaning**
OPVs	Organic Photovoltaics
V_{oc}	Open Circuit Voltage
J_{sc}	Short Circuit Current Density
PCE	Power Conversion Efficiency
FF	Fill Factor
AL	Active Layer
BHJ	Bulk Heterojunction
D/A	Donor/Acceptor
FMOs	Frontier Molecular Orbitals
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
ETL	Electron Transport Layer
HTL	Hole Transport Layer
SCs	Solar cells
AI	Artificial intelligence
ML	Machine Learning
DL	Deep Learning
GBR	Gradient Boosting Regressor
RF	Random Forest
SVR	Support Vector Regressor
AdaBoost	Adaptive Boosting Regressor
XGBoost	Extreme Gradient Boosting Regressor
CNN	Convolutional Neural Network
LR	Learning Rate
RE	Renewable Energy
GS	Grid Search

Tables list

Figures list

General Introduction

The global demand for energy is rising steadily, driven by industrialization, population growth, and technological advancement. At the same time, humanity faces unprecedented environmental challenges, notably climate change and resource depletion, primarily due to our dependence on fossil fuels [1]. In this context, transitioning toward clean and renewable energy (RE) sources is not merely a strategic priority but a global necessity.

Among the most prominent forms of renewable energy are hydropower, wind energy, and solar energy, each of which plays a crucial role in the global transition toward a cleaner, more sustainable energy future by offering distinctive advantages such as reliability, scalability, and low emissions as well as presenting unique technical, environmental, and economic challenges that must be carefully addressed to realize their full potential. Where:

- **Hydropower**: using flowing or falling water typically from dams to spin turbines and generate electricity. It offers reliable and dispatchable power but can disrupt ecosystems and displace communities when large dams are constructed [2].
- **Wind Energy** captures the kinetic energy of the wind through turbines, converting it into electricity. It is a clean, renewable source with low operational costs and no emissions during use. However, wind power is intermittent and requires suitable geographic conditions and storage or grid integration solutions [3],[4].
- **Solar Energy** harnesses the sun's radiation using photovoltaic panels or solar thermal systems. It is one of the fastest-growing energy sources due to its scalability and decreasing costs. Like wind, solar power is intermittent and depends on weather and daylight availability [5][6].

In the first quarter of 2025, most of the electricity generated from renewable sources in Europe came from wind power, which supplied 42.5% of all renewable electricity production. Hydroelectric power followed with a significant 29.2% share. Notably, solar energy played a crucial role, contributing 18.1%, underscoring its growing importance in Europe’s clean energy transition. Combustible renewable fuels, including biomass and biogas, accounted for 9.8%, while geothermal energy provided a smaller but steady share of 0.5%. Figure 1 illustrate that while wind and hydro remain the largest contributors to renewable electricity in Europe, solar power has established itself as a vital and rapidly expanding source within the region’s energy mix.

From [1] Photovoltaic (PV or Solar cells) technologies offer a direct means to convert sunlight into electricity. While traditional PV cells based on crystalline silicon have achieved impressive efficiencies (above 26% for single-junction silicon cells), their production processes are often energy-intensive and rely on rigid, heavy substrates that limit their use in flexible or lightweight applications [5], [7]. In contrast, organic photovoltaics (OPVs) present a promising alternative. Composed of organic semiconducting materials typically polymers or small molecules, OPVs are lightweight, mechanically flexible, and compatible with low-temperature and solution-based fabrication techniques [5]. These characteristics make them ideal for a wide range of emerging applications, from wearable electronics to building-integrated photovoltaics (BIPV) [8].

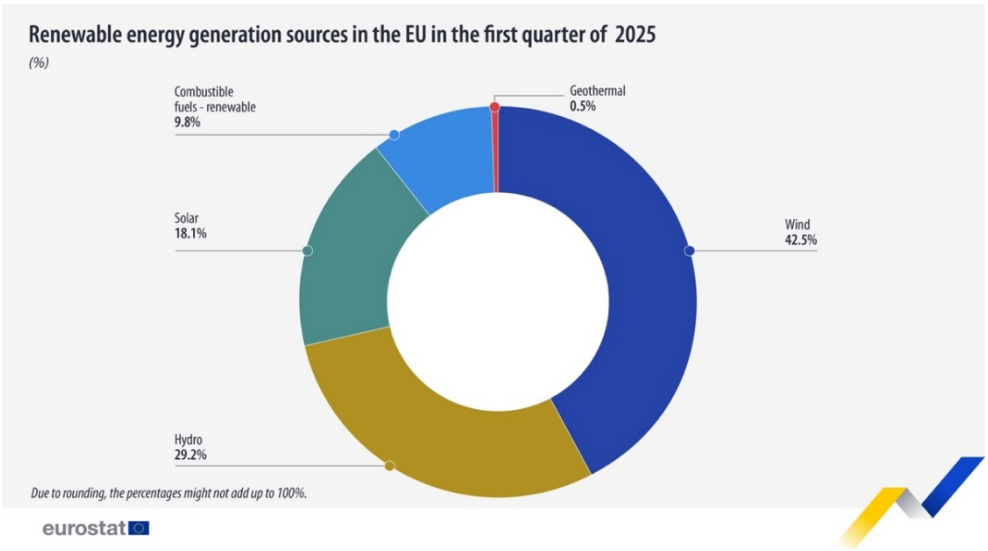


Figure 1. Renewable energy generation sources in the EU in the first quarter of 2025 (From : https://doi.org/10.2908/NRG_CB_PEM)

Despite these advantages, OPVs face key challenges in terms of power conversion efficiency (*PCE*) and operational stability. Although in this year (2025) *PCE*s close to 21% have been achieved in laboratory settings [9], this is still below the performance of inorganic solar cells. Furthermore, the discovery and optimization of new OPV materials remains complex and time consuming. It requires a multidisciplinary effort, encompassing organic synthesis, material characterization, device fabrication, and testing a process that can take months or even years for each new donor/acceptor (D/A) pair.

[1] https://doi.org/10.2908/NRG_CB_PEM

To accelerate this process, researchers have increasingly turned to Artificial Intelligence (AI) and, more specifically, machine learning (ML) [10], [11], [12], [13], [14], [15], [16], [17]. These tools enable the development of predictive models capable of estimating OPV performance metrics such as V_{oc}, J_{sc}, *FF and PCE* based on molecular and device-level descriptors. By capturing complex, non-linear relationships in large datasets, ML can guide material discovery, reduce the need for trial-and-error experimentation, and improve our understanding of structure property relationships in OPVs. Recent advances also include the use of deep learning (DL), particularly convolutional neural networks (CNNs), to interpret molecular structures from 2D images or graphs, bypassing the need for handcrafted features. However, the application of AI in OPV research presents its own set of challenges. These include limited access to high-quality experimental data, variations in device architecture and testing conditions, and the difficulty of extracting meaningful and consistent molecular descriptors. Furthermore, some machine learning models may require large training datasets to avoid overfitting and ensure generalizability. Consequently, careful data curation, validation, and interpretation are essential to developing reliable AI-driven models for OPV performance prediction.

Objectives and contributions of this thesis

The main goal of this thesis is to develop and evaluate AI-based methodologies for the prediction and discovery of high-performance OPV materials. The research is grounded in the analysis of experimentally validated (D/A) pairs and aims to contribute to both theoretical insights and practical tools to support OPV development.

Each part of the thesis contributes to a holistic understanding of how AI techniques can support the design, evaluation, and discovery of OPV materials. The models developed are tested not only on training data but also on unseen experimental (D/A) combinations from recent publications, validating their real-world applicability and generalization.

Structure of the thesis

This thesis is structured into four main chapters:

Chapter 1 – State of the art

Reviews the fundamental concepts of organic photovoltaics, including device architecture, material types, and performance challenges. It also surveys recent applications of AI and ML in the field of materials science, with a focus on OPVs.

Chapter 2 – Chemical descriptors and machine learning models

Details the construction of datasets, extraction of molecular descriptors, selection of ML algorithms, and model evaluation techniques. The predictive performance of different models is compared and interpreted using feature importance method (MrMR).

Chapter 3 – Prediction based on fullerene acceptors and hybrid FA+NFA systems

Explores the modeling of FA-based systems and the effects of combining fullerene and non-fullerene data. This chapter aims to evaluate the model's ability to generalize across different classes of OPV materials.

Chapter 4 – deep learning with 2D structural images

Introduces convolutional neural networks trained on 2D images of donor and acceptor structures. This part explores the potential of image-based deep learning in molecular property prediction and compares it with descriptor-based approaches.

Chapter 1: State of the art: AI and OPVs

1. Introduction

In this chapter, we explore and provide a comprehensive overview of the current state of the Organic Photovoltaics (OPV) and the growing role of Artificial Intelligence (AI) in advancing this field which is a promising technology of next-generation solar cells. We will begin by introducing the basic principles, structure, and laying the groundwork for understanding how these devices work and how they are evaluated using key performance metrics of OPVs. The discussion then shifts to the application of AI in material science, highlighting how techniques such as Machine Learning (ML) and Deep Learning (DL) are being utilized to accelerate the discovery and optimization of materials for OPVs. We also review significant previous work focused on AI-driven prediction of OPV performance. The chapter concludes with a summary of the insights gained and the potential of AI to revolutionize future developments in OPV technology.

2. Organic photovoltaics

2.1. What is organic photovoltaics (OPVs)?

Organic Photovoltaics are a type of solar cell that use carbon-based (organic) small molecules or polymers to capture sunlight or artificial-lighting photons and generate electricity [5]. Unlike conventional silicon-based solar cells, OPVs are made from lightweight, flexible materials and can be manufactured using inexpensive printing methods [18], [19]. These features make them ideal for applications that require flexibility, visual design integration, or quick energy return on investment [5].

2.2. How does organic photovoltaics work?

Organic solar cells are multilayered devices and among these layers, the most important one is the active layer, which plays a central role in light absorption and charge generation. This active layer is typically a blend of at least an electron-donor (D) and an electron-acceptor (A) semiconducting molecule. In traditional efficient organic solar cells, these two distinct organic semiconducting materials are intimately mixed to optimize exciton dissociation and charge transport in a so-called bulk heterojunction morphology. For a long time, the A material (a fullerene derivative) did not absorb efficiently in the solar spectrum wavelength range. To

simplify the description, we will stay in the hypothesis that the D material is primarily responsible for absorbing sunlight photons. Upon photon absorption, the D undergoes photoexcitation, leading to the formation of a tightly bound electron-hole pair known as an exciton[5], [18]. Due to the low dielectric constant of organic semiconductors, these excitons possess high binding energies and must be dissociated to generate free charge carriers. This dissociation occurs at the D/A interface, where the energy level offset between the D and A Lowest Unoccupied Molecular Orbital (LUMO) facilitates free-electron transfer from the donor to the acceptor as mentioned in Figure 2. The energy level offset between D and A Highest Occupied Molecular Orbital (HOMO) helps the free-hole to stay on D as mentioned in Figure 2. The HOMO and the LUMO are often referred to as Frontier Molecular Orbitals (FMOs).

As a result, the electron is transferred to A, while the corresponding hole remains in D [20] forming a charge transfer state (CTS). The spatial separation of the CTS lead to free-charge carriers reducing the recombination rate. The free electron is transported through the acceptor phase toward the cathode layer (the electrode, which collects electrons and completes the circuit), while the hole diffuses through the donor network toward the anode layer (the electrode, which collects holes and allows them to flow through the circuit). This directional movement of electrons and holes, driven by internal electric fields (due to the electrode work function difference) and concentration gradients, generates an electric current, known as the photocurrent. The substrate (the foundational layer on which all other layers are deposited) provides mechanical support for the device and is often made from transparent materials like glass or plastic, allowing light to pass through and reach the active layers [5], [20].

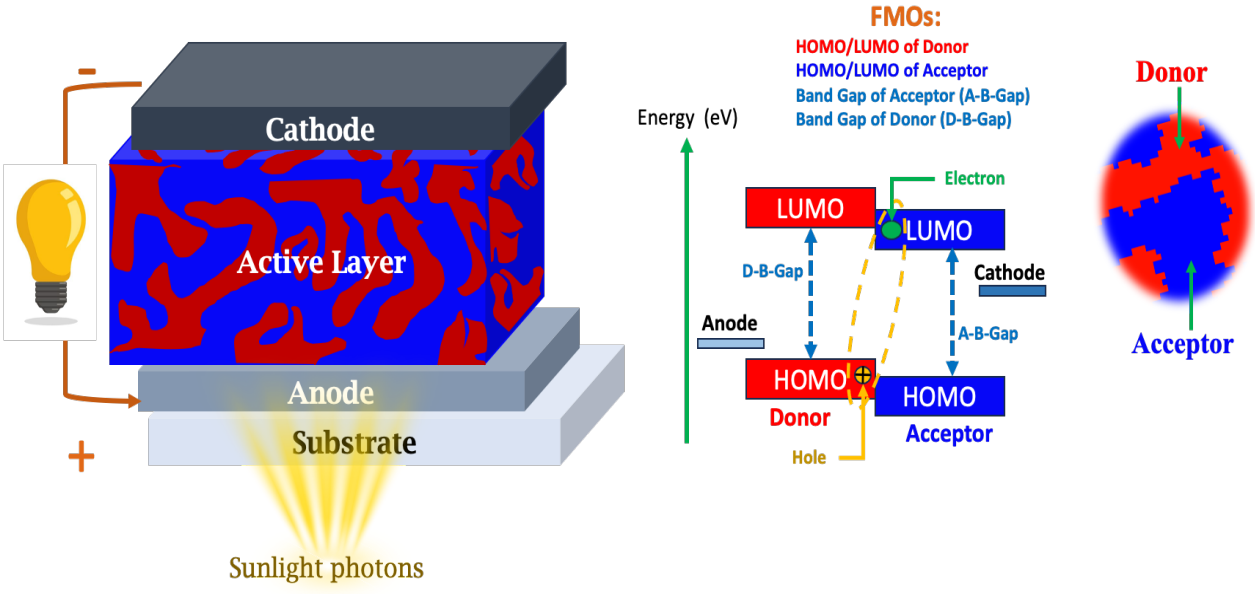


Figure 2. OPV cell Structure

2.3. Keys parameters of OPVs performance

The overall efficiency of organic photovoltaic devices depends on several interconnected processes such as light absorption, exciton diffusion, charge separation and charge transport and collection. To enhance device performance, optimizing the morphology and energy alignment of the FMOs of the donor and acceptor materials is essential. To estimate and calculate the performance of these devices, four key parameters need to be considered: open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (FF), and power conversion efficiency (PCE), as outlined in Table 1 and the curve in Figure 3. V_{oc} (or open-circuit voltage) represents the voltage of a photovoltaic cell under illumination at zero current. J_{sc} (or the short-circuit current-density) reflects the current generated per unit area under illumination when the voltage is zero, showcasing the cell's ability to generate and extract charge carriers. The fill factor quantifies the rectifying behavior of the photovoltaic cell. It is defined as the ratio of the maximum obtainable electrical power per unit area divided by $(V_{oc} \times J_{sc})$. Finaly the overall power conversion efficiency or the *PCE (%)* is the key metric of performance, expressing the percentage of incident solar energy converted into electrical energy, and is calculated as $(V_{oc} \times J_{sc} \times FF) / 100$ where $(V_{oc} \times J_{sc} \times FF)$ is the maximum obtainable electrical power per unit area expressed in mW/cm^2 and 100 mW/cm^2 is the illumination power per unit area under a standard AM1.5G illumination [20].

Table 1. Keys parameters of OPVs performance

Metrics	Full Name	Definition	Units	Significance
V_{oc}	Open Circuit Voltage	Maximum voltage under illumination when no current flows	V	Related mainly to the FMOs of the D and A materials but also on recombination processes.
J_{sc}	Short Circuit Current Density	Current per unit area under illumination when voltage is zero	mA/cm²	Related to the active layer light-absorption, exciton dissociation and free-charge transport and collection at the electrodes
FF	Fill Factor	Ratio of maximum power output to the product of V_{oc} and J_{sc}	Unitless (often %)	Related to the solar-cell rectifying behavior
PCE	Power Conversion Efficiency	Ratio of electrical power output to solar power input:	Percentage (%)	Overall measure of solar cell key performance

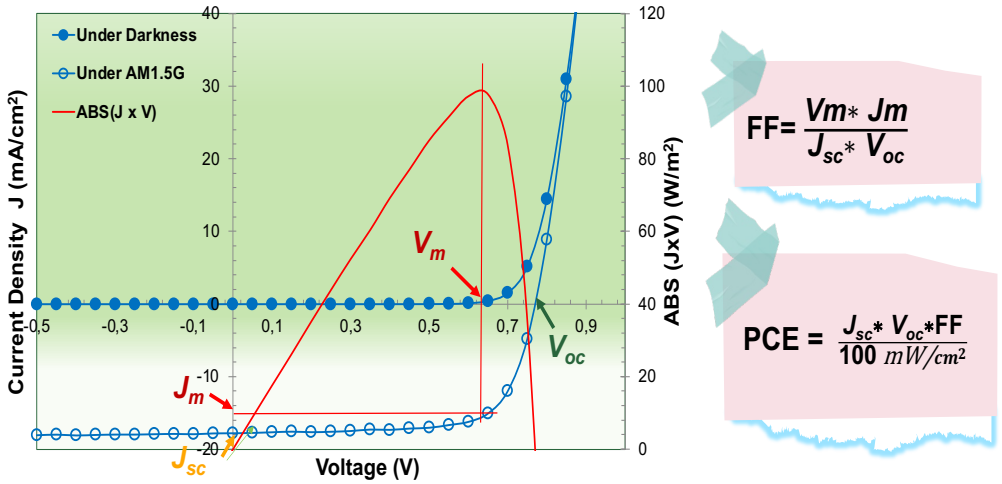


Figure 3. Current Density Vs Voltage Curve under darkness (closed circles) and under standard illumination (open circles) and electrical power under illumination (red curve) [15]

3. Artificial intelligence (AI)

3.1. What is AI?

Artificial Intelligence is the computational discipline focusing on the design and development of adaptive systems capable of perceiving complex data environments, learning implicit patterns, and autonomously optimizing decision-making processes [21], [22], [23]. Unlike conventional algorithmic models that rely on predefined logic, AI systems iteratively improve performance through experience, enabling generalization beyond explicitly programmed instructions [24]. At its core, AI seeks to replicate or extend cognitive functions such as perception, reasoning, and prediction through mathematical modeling, statistical inference, and scalable architectures [22], [24].

3.2. AI evolution path

Within the broad field of artificial intelligence as mentioned in Figure 4, Machine Learning (ML) provides the ability for systems to learn from data and improve over time. Deep Learning (DL), a subfield of ML, utilizes multi-layered neural networks to capture complex data patterns. Building upon DL, Generative AI [25] represents a major advancement, focused on systems that can create new content text, images, audio, or code based on learned patterns from training data. A significant branch of Generative AI is Large Language Models (LLMs) [26], which are trained on massive datasets to understand and generate human like text. LLMs such as GPT [27] (Generative Pre-trained Transformer) can answer questions, write essays, translate languages, and simulate conversations. ChatGPT [28] is a specific application of an LLM, fine-tuned for dialogue and human interaction, making it one of the most impactful demonstrations of AI's capabilities in everyday use. The following diagram illustrates the progression from general AI to ChatGPT:

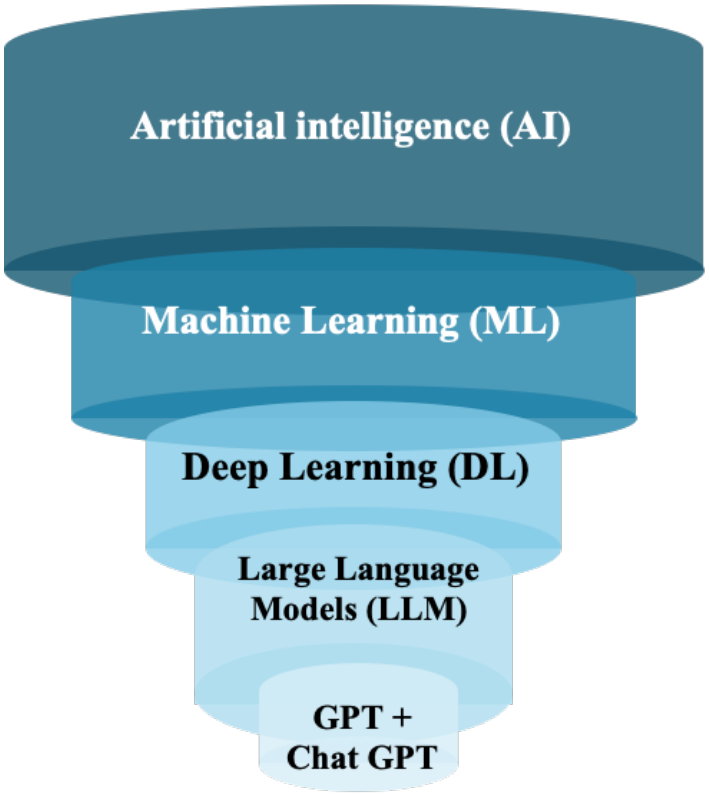


Figure 4. AI evolution path

3.3. AI fields

3.3.1. AI in healthcare and medicine

Artificial Intelligence is revolutionizing the field of medicine by enhancing diagnostic accuracy, accelerating research and optimizing patient care. AI-powered tools assist in medical imaging analysis, such as interpreting X-rays, MRIs, and CT scans, allowing for earlier and more precise detection of conditions like cancer and stroke. In drug development, AI facilitates the discovery of new molecules and significantly reduces the time and cost associated with clinical trials. Personalized medicine also benefits from AI, as algorithms can tailor treatments based on a patient's genetic profile. Additionally, AI supports the real-time monitoring of chronic diseases and improves the management of healthcare resources by streamlining patient flow and hospital operations. These advancements contribute to more efficient, accurate and individualized medical care [29], [30], [31].

3.3.2. AI in agriculture

Artificial Intelligence is increasingly used in agriculture to improve productivity and resource efficiency [32], [33]. It enables farmers to monitor soil conditions, crop health, and moisture levels in real time using data from sensors [32], drones and satellite imagery [33]. This allows for precise adjustments in irrigation, fertilization and treatments, reducing unnecessary waste of water and inputs. AI also helps anticipate weather conditions, allowing farmers to better prepare for climate-related challenges. Additionally, it supports early detection of diseases and pests, enabling timely interventions that protect yields [33]. Through satellite image analysis, AI can assess crop performance and analyze yields, offering valuable insights for planning and decision-making.

3.3.3. AI in transportation

AI is a key driver behind the development of autonomous vehicles and intelligent transportation systems [34]. Self-driving cars rely on real-time vision and decision-making algorithms to analyze their environment and navigate safely without human intervention [35], [36]. These systems process vast amounts of sensor data to detect obstacles, interpret traffic signals and make split-second decisions. AI also contributes to traffic optimization through the use of smart traffic lights and predictive models that forecast congestion and adjust routes accordingly. Beyond road vehicles, AI is used in predictive maintenance for various modes of transportation such as trains and airplanes, continuously monitoring system health to anticipate potential failures and reduce downtime. This integration of AI enhances safety, efficiency and the overall reliability of transport networks [34], [35], [36], [37].

3.3.4. AI in education

AI is reshaping education by enabling personalized learning, improving academic support and assisting educators with customized resources. It adapts content to individual learning styles, offers automatic grading and uses analytics to track student progress. For international learners, real-time translation tools ease language barriers, while AI also helps identify disengagement or academic dishonesty through behavior monitoring [38]. So AI's role in education can be seen through three paradigms: (1) AI-directed, learner-as-recipient, based on behaviorism, where systems like the ACT Programming Tutor deliver structured content, (2) AI-supported, learner-

as-collaborator, grounded in cognitive and social constructivist theories, using tools like QUE that leverage natural language processing and Bayesian networks and (3) AI-empowered, learner-as-leader, rooted in connectivism, promoting adaptive learning through technologies like brain-computer interfaces and real-time MOOC predictive modeling [39].

3.3.5. AI in organic photovoltaics

AI techniques, particularly machine learning and deep learning, are used to accelerate material discovery by predicting the properties of organic compounds and identifying optimal donor-acceptor pairs. AI also aids in optimizing device architecture and fabrication processes, reducing the time and cost typically required for experimental testing [17]. Additionally, predictive modeling enables the simulation of OPV performance under various environmental conditions, guiding the design of more efficient and durable solar cells. By streamlining research and enhancing performance prediction, AI significantly contributes to the development and commercialization of next-generation OPV technologies [15], [40], [14], [18], [41], [42], [43].

Figure 5. AI fields

3.4. AI in material science

Artificial intelligence is playing an increasingly critical role in advancing the field of material science by introducing smarter, data-driven methods for discovering, characterizing and optimizing new materials. Traditionally, developing materials has been a slow, labor-intensive process involving extensive experimentation and simulations. However, with the growing

availability of experimental and computational datasets, AI techniques particularly machine learning (ML) and deep learning (DL) are helping researchers uncover intricate structure property relationships, predict material behavior and guide experimental design with unprecedented efficiency. These advanced tools reduce reliance on trial-and-error approaches by identifying the most promising candidates for synthesis and testing, significantly accelerating innovation in fields like battery technology, catalysis and photovoltaics.[44], [45], [46].

In the field of organic photovoltaics, AI is increasingly transforming device engineering and fabrication workflows, enabling smarter and more efficient development of high-performance solar cells [17], [47] where one major area of impact is the optimization of processing parameters [48]. AI-driven techniques have been employed to systematically refine fabrication variables such as active layer thickness, spin-coating speed, annealing temperature and solvent composition. These techniques can quickly converge on optimal processing conditions that maximize power conversion efficiency, reduce material waste and improve reproducibility across batches [48], [49].

Also, the prediction of organic photovoltaic performance has become increasingly significant due to its potential to accelerate clean energy innovations [12], [13], [42]. Traditional methodologies primarily relied on theoretical calculations and semi-empirical models to estimate frontier molecular orbitals (FMOs) and, by extension, the efficiency of OPV devices. Quantum mechanical approaches like Density Functional Theory (DFT) offer accurate FMO estimations but are computationally demanding and require expert knowledge [50], [51]. In contrast, semi-empirical techniques make use of simplified assumptions and experimental FMO data from donor (D) and acceptor (A) materials to derive photovoltaic parameters more efficiently.

The integration of machine learning (ML) into OPV research has opened new avenues for performance prediction by utilizing experimental datasets [52], [53] and statistical models to uncover complex patterns [54], [10], [12], [16], [48], [55]. However, a key bottleneck is the scarcity of high-quality experimental data, which limits the robustness and generalization capability of ML models [18], [19], [45], [56]. This challenge is compounded by the intrinsic variability and structural diversity of organic compounds. To address this, some researchers have explored the use of transfer learning (TL) [57] and large simulated datasets like the Harvard Clean Energy Project (HCEP) dataset which is a high-throughput virtual screening initiative dedicated to the discovery of novel organic materials for renewable energy

applications, with a strong emphasis on organic photovoltaic devices. The HCEP dataset comprises computational data on over 2.3 million organic molecules, each evaluated for their electronic and photovoltaic properties through quantum chemistry methods. These molecules are systematically generated from a library of molecular building blocks and analyzed using density functional theory (DFT). Key parameters such as the highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), optical excitation energies, and estimated power conversion efficiencies (*PCE*s) are computed for each candidate. The screening process employs a combination of electronic structure calculations and machine learning strategies to identify materials with promising solar energy conversion potential. As one of the largest and most consistent datasets in the field of computational materials science, the HCEP serves as a benchmark resource for data-driven discovery and model development in organic electronics [58], [59]. However, some others used the Harvard Organic Photovoltaic Dataset (HOPV15) which is a comprehensive and meticulously curated dataset designed to support research and development in the field of organic photovoltaics. It integrates both experimental results and computational simulations to provide a rich source of information on the properties and performance of organic materials used in photovoltaic applications. Specifically, the dataset includes data on 350 distinct small molecules and polymers that function as donor materials within organic photovoltaic devices. The HOPV15 dataset covers a wide range of molecular structures, electronic properties, and device performance metrics, making it an invaluable tool for scientists and engineers seeking to understand, predict, and optimize the efficiency of organic solar cells [58]. For example, Olivares-Amaya [60] employed the Random Forest (RF) algorithm on the HCEP dataset and achieved high R^2 values for performance metrics such as *PCE* (0.84), V_{oc} (0.94), and J_{sc} (0.89). Nevertheless, while large computational datasets facilitate initial model training, their applicability to real-world systems can be limited due to the simplified nature of the molecular structures involved. Studies like those by Moore *et al.* [61] demonstrated the advantage of TL using the HCEP and HOPV15 datasets, improving FMO predictions with moderate success (R^2 values of 0.55 for HOMO and 0.63 for LUMO). These findings underscore the superior reliability of models trained on experimental data over those trained exclusively on theoretical or simulated data.

Further developments have focused on expanding experimental datasets to include more diverse OPV configurations. For instance, Random Forest models have been applied to experimental data for binary, tandem, and ternary bulk heterojunction (BHJ) solar cells. Lee and colleagues [62], [63] predicted *PCE*s for tandem OPVs with an R^2 of 0.69 using 70 data points. This dataset

was later extended to 121 samples for ternary cells, yielding a V_{oc} prediction with $R^2 = 0.77$. Notably, these studies focused primarily on FMO descriptors while overlooking morphological and chemical complexity, which can critically influence device performance.

More comprehensive approaches have incorporated descriptors such as TD-DFT-calculated FMO values and van der Waals surface area metrics. TD-DFT (Time-Dependent Density Functional Theory) extends ground-state Density Functional Theory (DFT) to enable the study of excited-state properties in many-electron systems by incorporating their response to time-dependent external perturbations. Built upon the Runge-Gross theorem, TD-DFT establishes that the time-dependent electron density uniquely determines the corresponding time-dependent external potential. This foundation allows electronic excitations to be analyzed through linear-response theory. Due to its favorable balance between computational efficiency and accuracy, TD-DFT is widely applied in quantum chemistry and materials science for simulating absorption spectra, computing excitation energies, and investigating photochemical processes [64]. These were applied to some of the largest curated datasets from literature [14], enabling prediction of key metrics like V_{oc}, J_{sc} and *PCE*. However, even these studies often fail to represent emerging material classes, such as the Y-series non-fullerene acceptors (NFAs) [65], which are prominent in recent high-performance devices [14]. All related works are illustrated in Table 2**.**

A less explored but promising direction involves deep learning (DL) methods based on molecular images rather than numerical descriptors. This shift leverages the pattern recognition strength of convolutional neural networks (CNNs) to analyze 2D chemical structure representations. One such study by Sun *et al.* [43] used a ResNet-based architecture to classify OPV molecules by performance level, achieving over 91% accuracy using HCEP-derived molecular images. However, this work was constrained by the structural simplicity of the HCEP molecules and lacked generalizability to newer, more complex NFAs. These findings suggest that while image-based DL approaches hold great potential, their effectiveness hinges on the quality and representativeness of the training data.

Table 2. related works (Exp : Experimental, Cal : calculated)

Ref	Dataset Size	Data Type	Input Type	Target(s)	ML Model(s)	Validation	Metrics	Results
[62]	70	Exp	FMO	*PCE*	RF, SVR	70/30 Split	R^2, RMSE	$R^2 = 0.69$, RMSE= 2.86%
[63]	121	Exp	FMO	V_{oc}	RF	80/20 Split	R^2	$R^2 = 0.77$
[65]	566	Exp	FMO, ECFP6	*PCE*	RF	K-Fold CV	r	r = 0.85
[60]	2.3M	Cal (HCEP)	CSD	*PCE*, V_{oc}, J_{sc}, *FF*	Linear Regression	Train/Test	R^2	R^2 = 0.84 (*PCE*), 0.94 (V_{oc}), 0.89 (J_{sc}), 0.61 (*FF*)
[49]	51,000	Cal + Exp	Fingerprints	*PCE*	GPR	Train/Test	r	r = 0.43
[66]	2.3M + HOPV15	Cal + Exp	CSD, Fingerprints, FMO	*PCE*	SVR, RF, HDMR	K-Fold + Shuffle	MSE, MAE, R^2, r	MSE = 3.18%, MAE = 1.41%, $R^2 = 0.35$, r = 0.623
[61]	2.3M + HOPV15	Cal + Exp	2D Images	FMO (Donor)	CNN	Train/Test	R^2	R^2 = 0.55 (HOMO), 0.63 (LUMO)
[12]	1242	Exp	CSD, FMO	*PCE*, V_{oc}, J_{sc}	SVR, RF, ANN, GBR	80/20 + LOOCV	RMSE, r	*PCE* : RMSE=2.00% r = 0.79 V_{oc}: RMSE =0.047V r = 0.86
[14]	503	Exp	CSD, FMO (TD-DFT)	*PCE* > 10%	RF	5-Fold CV	R^2, RMSE	$R^2 = 0.28$, RMSE = 1.60%
[43]	2.3M	Cal (HCEP)	2D Image (Donor)	*PCE*	ResNet	70/20/10 Split	Accuracy (Classif.)	91.02%

4. Conclusion

Prior studies have made notable strides in predicting the performance of organic photovoltaic devices using theoretical models, semi-empirical approaches and machine learning techniques. These studies are often based on simulated data or limited experimental inputs and suffer from several critical gaps. In particular, many existing models overlook the complexity of real-world materials and rely heavily on calculated rather than experimental molecular properties.

In this thesis, we aim to bridge this gap by developing predictive models grounded in experimental data, including chemical descriptors of (D/NFA) pairs and experimentally measured frontier molecular orbital (FMO) energies. Distinct from much of the existing literature, our approach leverages these features to build more realistic and generalizable models, that is why we also tested the applicability of our machine learning framework on fullerene acceptor (FA) datasets, evaluating its capacity to predict device performance in traditional OPV systems. Building on these models, we then investigated the potential to discover novel high-performing (D/A) combinations by mixing different donor and acceptor materials and analyzing the predicted outputs. Finally, we explored the use of deep learning with 2D molecular structure images, aiming to capture intricate spatial features and further improve prediction accuracy. Together, these contributions are part of a broader effort to accelerate the discovery and optimization of efficient OPV materials, and each will be developed in the following chapters of this thesis.

Chapter 2: Predicting OPV Performance using Chemical Descriptors and Machine Learning

1. Introduction

In this chapter, a systematic approach and experimental protocol are provided, which focus on the data-driven estimation of OPV device key performance parameters. The aim of the present study is to construct a predicting tool using the chemical information of (Donor/Acceptor) pairs or (D/A) pairs to effectively predict the efficiency metrics of a device. The pipeline started from gathering and preprocessing the molecular data, where for each (D/A) pair was encoded by means of its SMILES (Simplified Molecular Input Line Entry System) string. These SMILES strings were then used to generate a diverse range of chemical descriptors, capturing molecular features known to influence OPV performance. To streamline the dataset and enhance model interpretability, a feature selection process was implemented using the Minimum Redundancy Maximum Relevance (mRMR) algorithm. This step ensured that the most informative and non-redundant descriptors were retained, thereby reducing noise and focusing the models on the features with the greatest predictive potential. With the refined feature set, multiple machine learning models were trained, and these models were constructed in a carefully preselected pure experimental OPV dataset, in which every (D/A) pair is unique, robust and representative. Lastly, the predictive power of the models developed was compared to modern methods. The increased accuracy achieved with our models confirmed the validity of our methodology and its potential to push the computational device modeling of OPV to new performance standards.

2. State of the art

In recent years, a wide range of machine learning (ML) models has been employed to predict key performance metrics of organic photovoltaic (OPV) devices [1], [11], [18], [45] with power conversion efficiency being a primary target for regression-based modeling efforts. Commonly used algorithms include Random Forest (RF) [43], Support Vector Regression (SVR), Gradient Boosting methods (such as Gradient Boosting Regressor), among others. These models have demonstrated good predictive capability when applied to molecular datasets, particularly those derived from SMILES representations of (D/A) molecular pairs. Molecular descriptors used in these studies are typically computed using cheminformatics tools such as the RDKit python library. To enhance model performance and interpretability, various feature selection techniques are also employed, which are used to identify the most relevant and non-redundant

descriptors, for example Lasso [67]. The combination of carefully selected molecular features with robust regression algorithms has proven its effectiveness in improving the reliability and accuracy of OPV performance predictions. Detailed discussions of the methods and strategies that we used in this study (ML models, feature selection method…etc.) are provided in the following sections.

2.1. Machine learning models

Machine learning has become an increasingly valuable tool in the field of organic photovoltaic research, particularly for modeling and predicting key device performance metrics such as power conversion efficiency (*PCE*) [44], [45], [56]. Given the typically limited size of experimental datasets in this domain, researchers tend to favor ML algorithms that offer strong generalization capabilities and are less prone to overfitting. As a result, simpler yet robust models are often preferred to enhance prediction accuracy and reliability. A range of regression algorithms have been applied in OPV studies, each with unique strengths depending on the dataset characteristics and modeling goals. In this study we used models like Support Vector Regression (SVR), Random Forest (RF), XGBoost, AdaBoost, and Gradient Boosting Regressor (GBR). The following subsections provide an overview of these methods, their underlying principles and their relevance to OPV research.

- **Support Vector Regression (SVR)**

SVR is a regression variant of Support Vector Machines (SVM) and is well-suited for tasks involving small datasets and high-dimensional input spaces. It seeks to find a function that approximates the target values within a specified margin of error (ε) while maintaining a flat and simple structure. Its ability to model non-linear relationships using kernel functions makes it a strong candidate for OPV-related predictions [68].

- **Random Forest (RF)**

Random Forest is an ensemble-based method that constructs multiple decision trees and averages their outputs to produce a final prediction. By aggregating the predictions of several weak learners, RF reduces variance and mitigates overfitting. Its flexibility in handling complex, non-linear data relationships has made it a popular choice in OPV performance modeling [69], [70].

- **Gradient Boosting Regressor (GBR)**

Gradient Boosting Regressor is another powerful ensemble technique that incrementally builds models to minimize residual errors from previous iterations. While it shares similarities with XGBoost, GBR can be more sensitive to overfitting if not properly regularized. Consequently, successful applications in OPV modeling often involve careful hyperparameter tuning to balance model complexity and generalization [71].

- **Extreme Gradient Boosting (XGBoost)**

XGBoost, is an advanced implementation of gradient boosting that emphasizes speed and performance. It includes built-in regularization mechanisms to prevent overfitting and has demonstrated superior accuracy in a variety of regression tasks. XGBoost is particularly effective with structured data and has gained attraction in OPV applications due to its robustness and efficiency [72], [71].

- **Adaptive Boosting (AdaBoost)**

AdaBoost works by sequentially training a series of weak learners typically shallow decision trees where each new model focuses more on correcting the errors of its predecessors. This adaptive approach enhances overall model accuracy and is especially useful for smaller datasets, making it relevant for OPV research where data can be limited [73].

2.2. Machine learning types

In machine learning, tasks are broadly classified based on the type of data available and the objectives of the learning process. Recognizing the nature of each task is essential for selecting suitable models and accurately assessing their performance. This section outlines the principal categories (see Figure 6) of machine learning tasks supervised, unsupervised, semi-supervised, self-supervised and reinforcement learning with an emphasis on those most relevant to this thesis [74].

2.2.1. Supervised learning

Supervised learning is one of the most widely applied ML paradigms. It involves training a model on a dataset that contains both inputs and their corresponding outputs (labels). The objective is to learn a mapping function that can predict the output for new, unseen inputs and

the availability of labeled data allows for direct evaluation of model performance. Supervised learning tasks are typically divided into two categories:

- **Classification**: the target variable is categorical and the model predicts discrete class labels. Common applications include:
  - Spam detection in emails.
  - Image classification (e.g., recognizing handwritten digits).
  - Medical diagnostics (e.g., identifying tumor types).
- **Regression**: the target variable is continuous and the model aims to predict real-valued outputs. Typical use cases involve:
  - Estimating house prices based on features like size and location.
  - Predicting equipment lifespan in industrial setting.
  - Predicting the performance of OPVs based on chemical descriptors [75].

2.2.2. Unsupervised learning

Unsupervised learning focuses on uncovering patterns or structures within data that lacks labeled outcomes. These methods are instrumental in exploratory data analysis, feature extraction and data preprocessing [74]. Key task types include:

- **Clustering**: grouping similar data points based on intrinsic characteristics. This technique is used in:
  - Customer segmentation for marketing.
  - Document categorization by topic.
  - Anomaly detection in cybersecurity.
- **Dimensionality Reduction**: reducing the number of input features while retaining significant information [74]. Techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are commonly applied in:
  - Visualizing high-dimensional data.
  - Noise reduction and performance enhancement.
  - Data compression.

2.2.3. Semi-supervised and self-supervised learning

- **Semi-Supervised Learning** combines a small set of labeled data with a larger pool of unlabeled data to train models more effectively. This approach is particularly useful when labeling data is costly or labor-intensive [74].
- **Self-Supervised Learning** is a rapidly emerging technique where models learn to generate supervisory signals from the input data itself. For instance, language models like BERT are trained to predict missing words in sentences a task that leverages large-scale text data without the need for manual annotation [74].

2.2.4. Reinforcement Learning

Reinforcement Learning (RL) is a goal-driven learning framework inspired by behavioral psychology. In RL, an agent learns to make decisions by interacting with an environment, receiving feedback in the form of rewards or penalties. The aim is to develop a policy that maximizes cumulative rewards over time [76]. Key features of reinforcement learning include a sequential learning process, where each action may influence future outcomes and a trade-off between exploration (trying new actions) and exploitation (leveraging known strategies) [74]. Common applications of RL include:

- Autonomous robotics and self-driving cars [76].
- Dynamic resource allocation and scheduling [77].

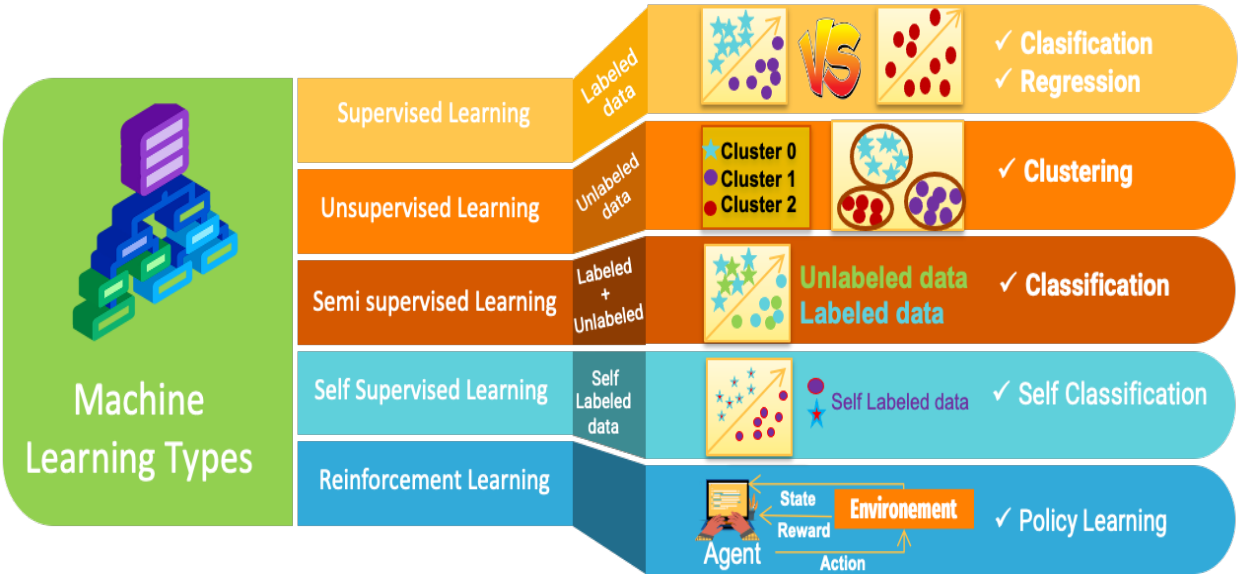


Figure 6. Machine learning types with examples

2.3. Models' performance metrics

To evaluate the performance of the developed regression models, a set of statistical metrics was employed, including the coefficient of correlation (r), root mean square error (RMSE), mean square error (MSE), coefficient of determination (R^2), mean absolute percentage error (MAPE), and mean absolute error (MAE).

The correlation coefficient (r) quantifies the linear relationship between the predicted and the experimental (or actual/real) values as mentioned in the equation (1) where a value close to 1 indicates a strong positive correlation. The RMSE and MSE represented with equations (2) and (3) respectively measure the average squared difference between predicted and actual values, with RMSE providing the result in the same unit as the target variable for example the unit will be *Volts (V)* if the target is the V_{oc}, making it more interpretable. The R^2 (coefficient of determination) in equation (4) indicates the proportion of variance in the dependent variable that is predictable from the independent variables, higher R^2 values reflect better model performance (close to 1). The MAPE with its formula in equation (5) represents the average absolute percentage difference between predicted and observed values, making it useful for interpreting errors in relative terms, especially when the data span different scales. Finally, the equation (6) that represents the MAE which calculates the average magnitude of the errors without considering their direction, offering a simple measure of prediction accuracy. Together, these metrics provide a comprehensive understanding of the model's predictive capabilities and generalization performance [78].

$$r = \frac{\sum_{i=1}^{N}(R_i - \overline{R}_\iota) \times (P_i - \overline{P}_\iota)}{\sqrt{\sum_{i=1}^{N}(R_i - \overline{R}_\iota)^2 \times \sum_{i=1}^{N}(P_i - \overline{P}_\iota)^2}} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(R_i - \overline{P}_\iota)^2}{N}} \quad (2)$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N}(R_i - P_i)^2}{N} \quad (3)$$

$$\mathrm{R}^2 = 1 - \frac{\mathrm{MSE}}{\mathrm{var}(R_i)} \quad (4)$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i - R_i|}{|R_i|} \quad (5)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |R_i - P_i|}{N} \qquad (6)$$

With:

- N: the number of data points in the dataset.
- R_i and P_i: the real (experimental in our case) and predicted values, respectively.
- $\overline{R_i}$ and $\overline{P_i}$: the mean values of the actual and predicted data.
- $\mathrm{var}(R_i)$: the variance of the actual values.

2.4. SMILES

SMILES (Simplified Molecular Input Line Entry System) code is a text-based notation system employed to represent the structure of chemical molecules using compact strings composed of standard ASCII characters [79]. This system encodes essential information about each molecule, including the types and arrangements of atoms, the nature and orientation of chemical bonds, and how these components are interconnected [79]. By expressing molecular structures in a linear, text readable format, this notation provides a way for both humans and computer programs to interpret, analyze, and manipulate complex chemical data efficiently. It plays a crucial role in cheminformatics, enabling tasks such as molecular modeling, database searching, and structure-based prediction to be performed systematically and with high precision [80].

For example, some coding rules are depicted in Figure 7 using some simple molecules. The first example is cyclohexane, a 6-carbon ring with single bounds and with the formulae C_6H_{12}. The SMILES code for cyclohexane is C1CCCCC1. The bounds are assumed to be single bounds by default and the H atoms are omitted. The number (here 1) refers to the opening and closing of the ring. For cyclohexene, there is one double bound and five single ones. The formulae is C_6H_{12} and the SMILES code C=1CCCCC1 where the sign "=" is indicating a double bound. The next example is benzene, a 6-carbon ring with alternatively single and double Carbon/Carbon bounds. Its formulae is C_6H_6 and its SMILES code C1=CC=CC=C1. Nevertheless, as the 6 bounds are considered as equivalent, the SMILES code C1CCCCC1 can also be found sometime for benzene. The last example is nitrobenzene with the following SMILES code: [O-][N+](=O)c1ccccc1. In this code, the parenthesis indicates the branching while the brackets are for charged atoms. This short explanation is only indicative as several different SMILES conventions coexist. Every molecule used in this manuscript can be described by its SMILES code. A toolkit like RDkit is made to extract important chemical parameters by the interpretation of SMILES codes.

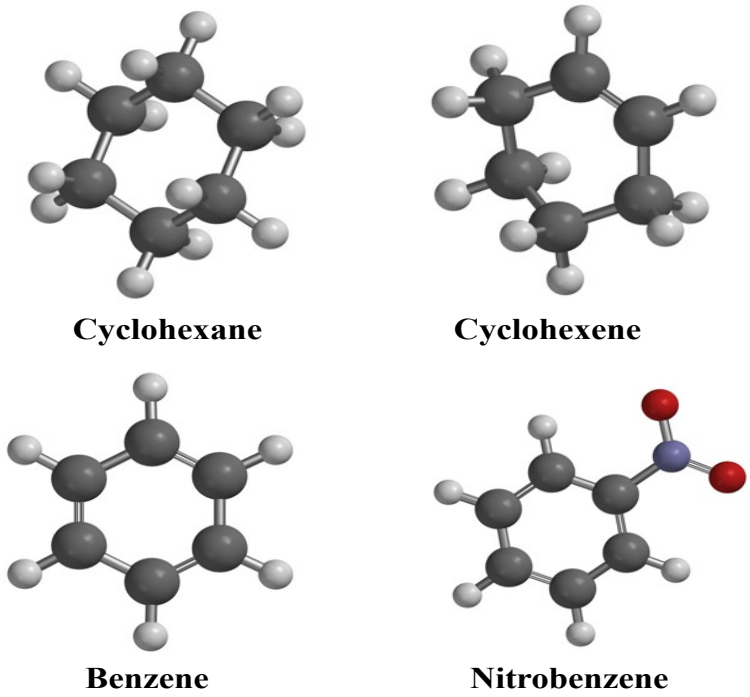


Figure 7. Chemical structure of cyclohexane, cyclohexene, benzene and nitrobenzene

That is why SMILES are particularly suitable for machine learning applications, where data volume and processing speed are often key constraints [18], [45], [56]. Ultimately, the adoption of SMILES in this study aligns with our goal of balancing computational feasibility with molecular complexity, facilitating robust and scalable molecular modeling.

2.5. RDKit cheminformatic toolkit

RDKit is a free and open-source cheminformatics toolkit developed in C++ with Python bindings, designed for the representation, analysis, and manipulation of chemical information[2]. It is widely used in computational chemistry, drug discovery, and machine learning applications involving molecular data [81]. RDKit provides a comprehensive set of tools [3] for:

- **Molecular structure handling**: parsing and writing molecules in various formats (e.g., SMILES, SDF, Mol).
- **Substructure searching**: identifying specific structural motifs within molecules.
- **Molecular fingerprinting**: generating fingerprints (e.g., Morgan fingerprints) for similarity searches and clustering.

[2] https://www.rdkit.org/docs/Overview.html

[3] https://www.rdkit.org/docs/GettingStartedInPython.html

- **Descriptor calculation**: computing molecular descriptors (e.g., molecular weight, logP, topological indices) for quantitative structure-activity relationship (QSAR) modeling.
- **Conformer generation**: building 2D and 3D molecular conformations and performing geometry optimization.
- **Chemical reactions**: encoding, applying, and analyzing chemical transformations using reaction SMARTS.

In this study, we employed the rdkit.Chem module, a core subpackage of the RDKit library that provides essential tools for manipulating chemical structures, including functionalities for handling molecules, atoms, bonds, and chemical reactions. The module is user-friendly and well-documented, as illustrated in Figure 8, with a structured interface that facilitates the search and use of a wide range of chemical descriptors and models. This makes RDKit a versatile and comprehensive toolkit for cheminformatics applications.

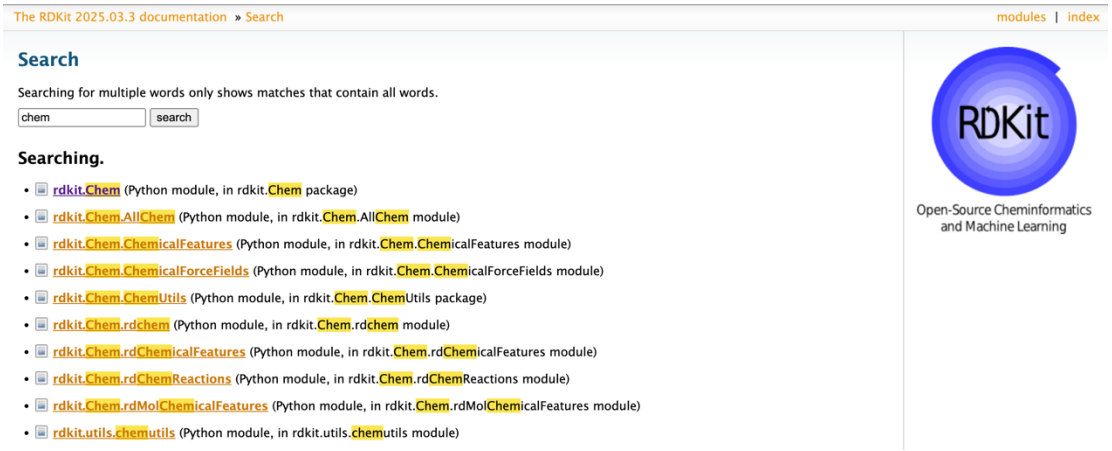


Figure 8. Documentation of RDKit

3. Methodology

3.1. Data analysis and cleaning

In this study, we compiled a comprehensive dataset of 1,225 (donor/non-fullerene acceptor) pairs of organic photovoltaic devices from various peer-reviewed literature sources collected by [14]. The dataset includes crucial photovoltaic parameters such as open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (FF) and power conversion efficiency (PCE). Where available, additional optical and electronic properties such as maximum absorption wavelength (λ_{max}) in solution or film, optical bandgap, ionization energy and electron affinity were also recorded. To standardize and simplify computational analysis, all side chains were

normalized to ethyl groups. This normalization step reduced molecular complexity without significantly impacting the key electronic properties, facilitating descriptor calculation and improving model performance. The dataset encompasses power conversion efficiency values ranging from as low as 0.01% up to a maximum of 18.77% while most of the devices exhibit *PCE*s within the 6% to 14% range. This suggests that most organic photovoltaic systems in the dataset operate at moderate efficiencies, with fewer entries representing either very low or extremely high *PCE* performance.

In addition to efficiency metrics, we also examined the chemical composition of the active layer materials used across the dataset, particularly focusing on donor (D) and acceptor (A) molecules. Figure 9 presents the ten most frequently reported compounds in each category. Among the donor materials, PBDB-T commonly referred in literature as PCE12 emerges as the most prevalent. It is closely followed by PBDB-T-2F, which may also appear under alternative names such as PBDT-T-F or PM6. Other notable donors include PTB7-Th (also known by aliases like PCE10 or PBDTTT-EFT) and the classical polymer P3HT, which continues to be widely studied despite the advent of more efficient materials. We can note that all these prevalent donor molecules are polymers. On the acceptor side, ITIC leads as the most frequently employed compound. It holds historical significance as the first non-fullerene acceptor to surpass fullerene-based materials in performance within the OPV field. Following ITIC, IT-4F (also cited as ITIC-4F or ITIC-2F depending on the source) is also highly utilized. Additional popular non-fullerene acceptors include Y6, a recent breakthrough material recognized for its superior light-harvesting and charge-transport properties, as well as IDIC, which contributes to the evolving landscape of efficient organic acceptor molecules. This overview of commonly used donor and acceptor materials not only highlights the most prominent compounds in the field but also reflects ongoing trends in molecular design and selection aimed at enhancing OPV device performance. To ensure dataset uniqueness and consistency, several refinement steps were applied:

- **Commercial Name Standardization**: molecules reported under different commercial names but identical in structure (e.g., PM6 and PBDB-T-F) were unified. Depending on the research groups and companies, the same molecule can be named differently.
- ***PCE* Validation**: Reported *PCE* values were recalculated from V_{oc}, J_{sc} and *FF* to confirm accuracy and identify inconsistencies due to experimental variation. Indeed, as explained in chapter 1 (see Figure 3 for instance) the four metrics are interdependent following the equation (7):

$$PCE = \frac{V_{oc}J_{sc}FF}{100} \qquad (7)$$

where V_{oc} is in Volts, J_{sc} in mA/cm^2, FF has no units and 100 corresponds to the standard (AM1.5G) illumination of 100 mW/cm^2. Some publications indicate the highest possible metrics, some others the average and some a mix of both. It is therefore important to check that the published PCE, V_{oc}, J_{sc} and FF values follow equation (7). After this validation, only 3 parameters have to be considered, PCE, V_{oc} and J_{sc}.

- **Representative Pair Selection**: for each core structure with multiple entries (e.g., (PM6/Y6)), the device with the highest reported PCE was chosen to represent the potential performance of that (D/A) pair. This selection based on the fact that the experimental device optimization depends on several factors like the amount of molecule available, the nature of the electrodes and of the charge-blocking layers mastered by the research group while we are interested in this study by the real photovoltaic potential of a given (D/A) active layer. Therefore, to consider that the highest reported PCE value for a given (D/A) pair is the real photovoltaic potential of this pair makes sense.

After these refinements, the final dataset consisted of 924 unique (D/A) pairs, incorporating 605 distinct acceptors and 214 donors. Each pair was annotated with corresponding SMILES codes and experimental frontier molecular orbital (*FMO*) information when available.

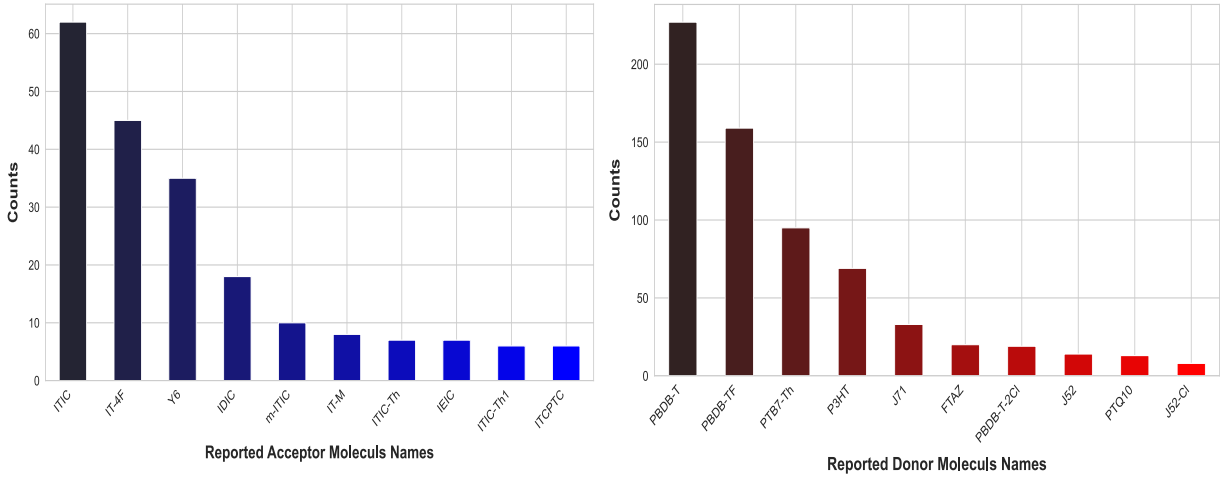


Figure 9. Top 10 common Donor and Acceptor Distribution in our dataset [15]

3.2. Feature engineering and descriptor selection

To transform chemical structures into features suitable for machine learning models, we employed SMILES codes, which provide a compact, standardized textual representation of molecules. These codes were used as input for RDKit, to extract a wide array of molecular descriptors for both donors and acceptors.

In total, over 100 descriptors were calculated per molecule, including:

- **Topological and structural descriptors**: number of single, double, triple bonds, bicyclic rings, Aromatic ring count etc…
- **Electronic descriptors**: molecular weight, EState-VSA (Electrotopological State-Volumetric Surface Area) etc…

These data were compiled into a dataset referred to as **data_1**, which includes molecular descriptors, experimental performance metrics (V_{oc}, J_{sc},*FF and PCE*) and the SMILES codes of each (D/A) pair. To enhance this dataset, the experimentally reported FMOs (HOMO/LUMO energy levels) were incorporated, yielding a new enriched dataset called **data_2**.

To reduce dimensionality and eliminate redundant or non-informative features, we applied the **Minimum Redundancy Maximum Relevance (mRMR)** feature selection method [82] which is designed to select a subset of input features that are both highly relevant to the target variable and minimally redundant with one another. Its objective is to enhance the performance of machine learning models by retaining features that provide the most informative and unique contributions, while filtering out those that are repetitive or overlapping [82], where the maximum relevance component ensures that the chosen features have a strong statistical relationship such as high mutual information or correlation with the target variable, thereby boosting predictive accuracy. At the same time, the minimum redundancy component aims to minimize similarities among selected features, reducing redundancy, avoiding overfitting, and simplifying the model. That is why in our case mRMR ranks features based on their predictive relevance to target variables (V_{oc}, J_{sc} and *PCE*) while minimizing redundancy among all the chemical descriptors. Our results showed that using a carefully selected subset of 20 descriptors (out of over 100) was sufficient to achieve robust prediction performance, results as shown in figures 10, 11, 12 and 13.

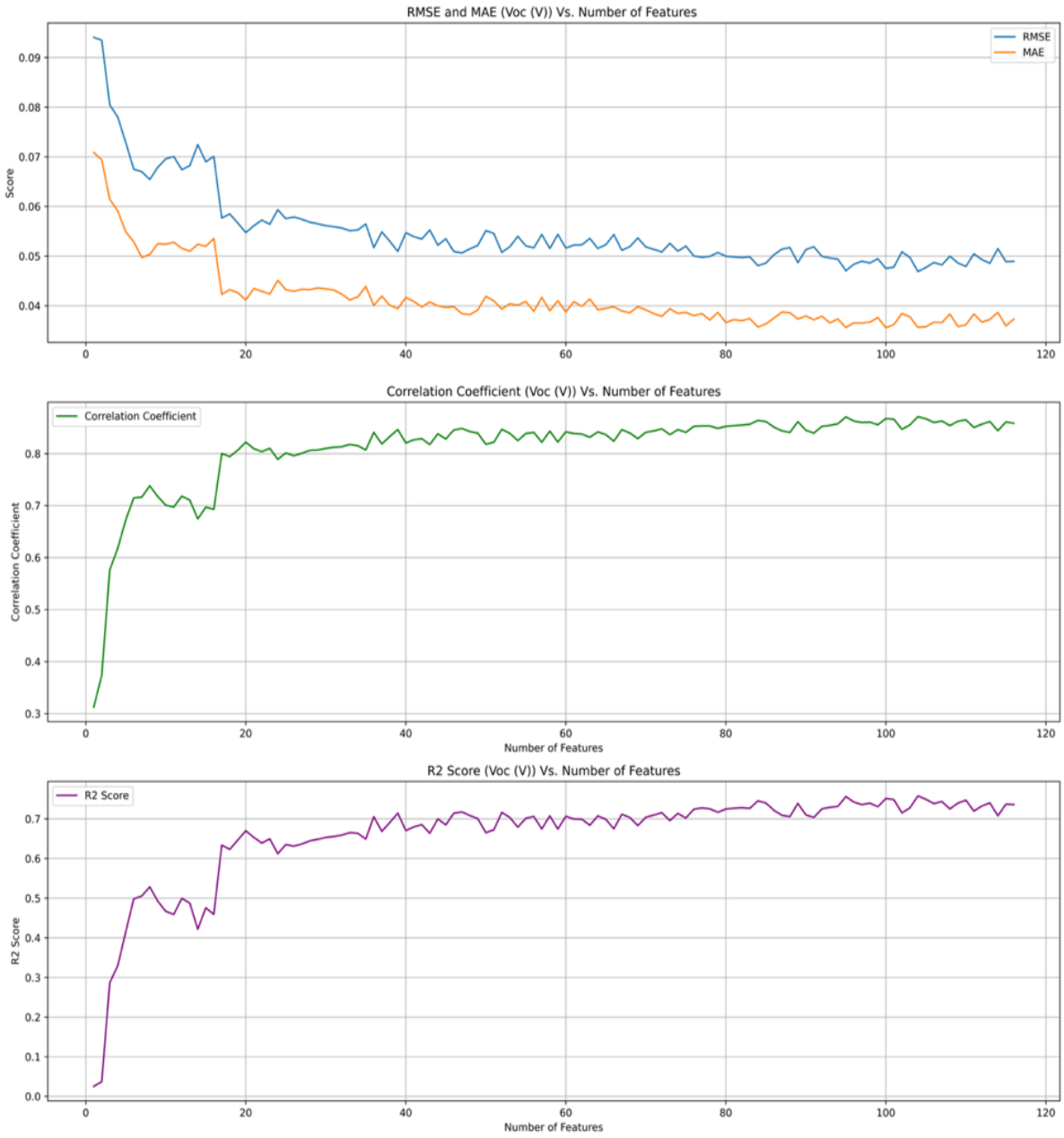


Figure 10. Results of MRMR application to select the important features to predict V_{oc} [15]

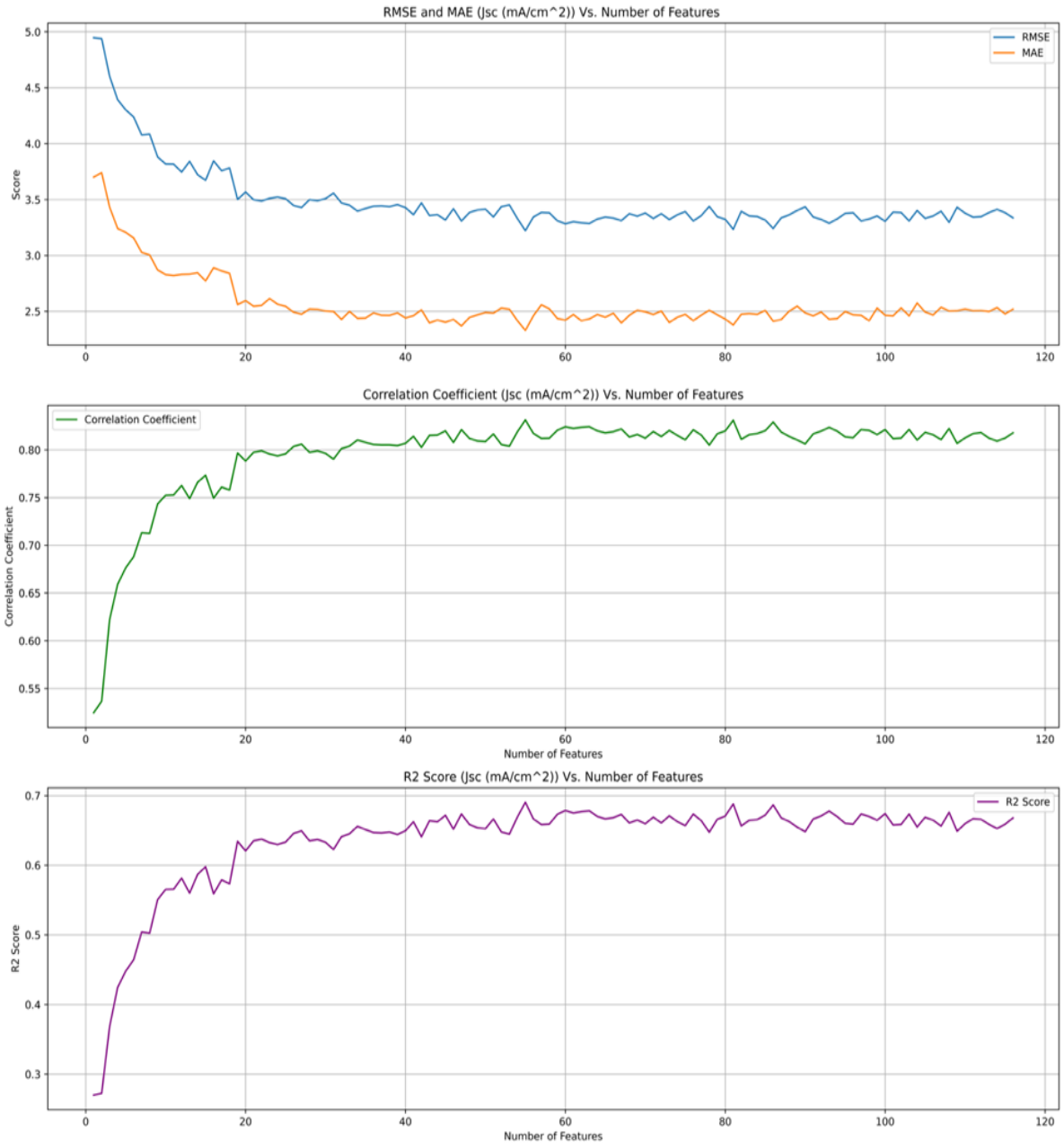


Figure 11. Results of MRMR application to select the important features to predict J_{sc} [15]

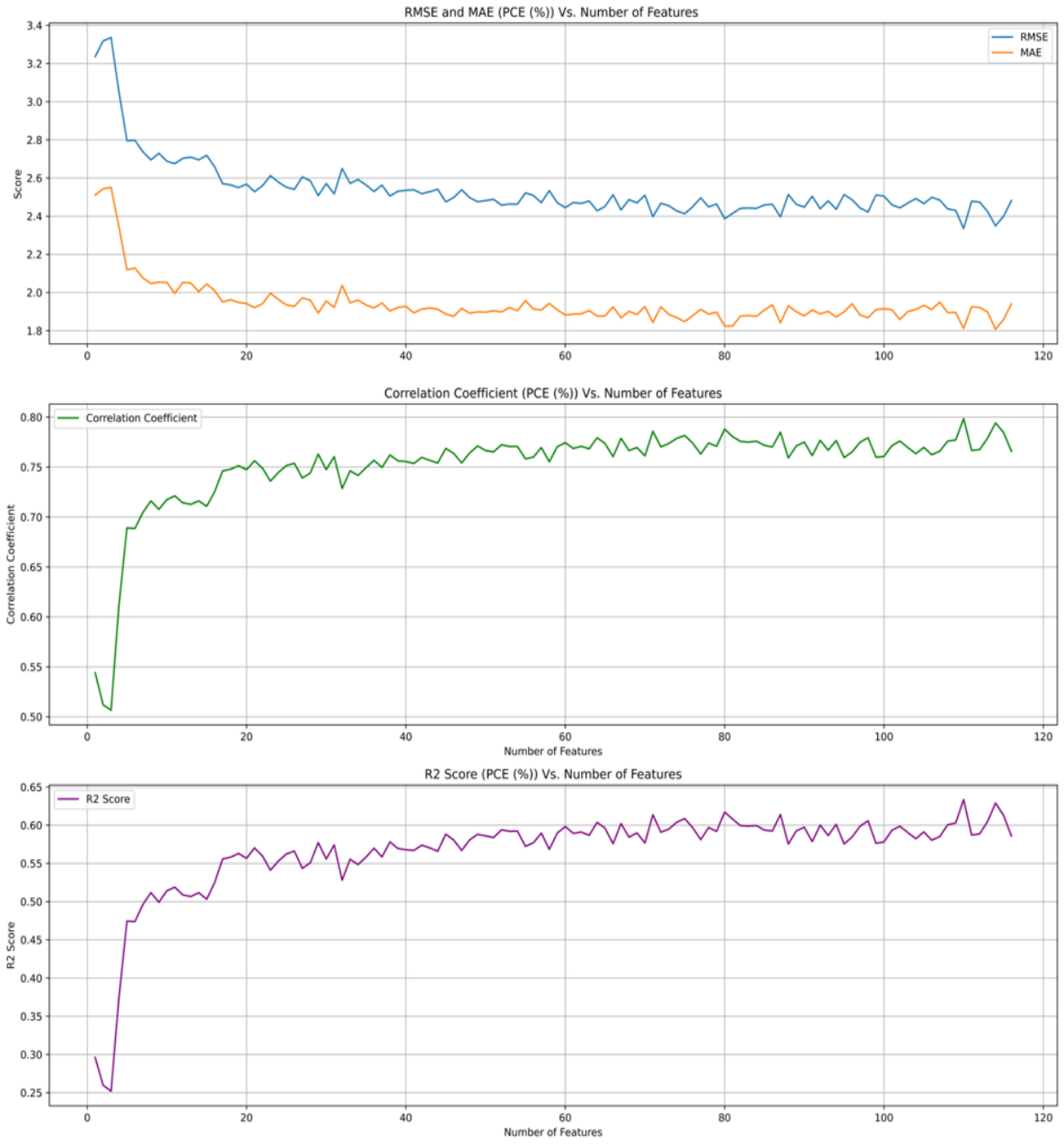


Figure 12. Results of MRMR application to select the important features to predict PCE [15]

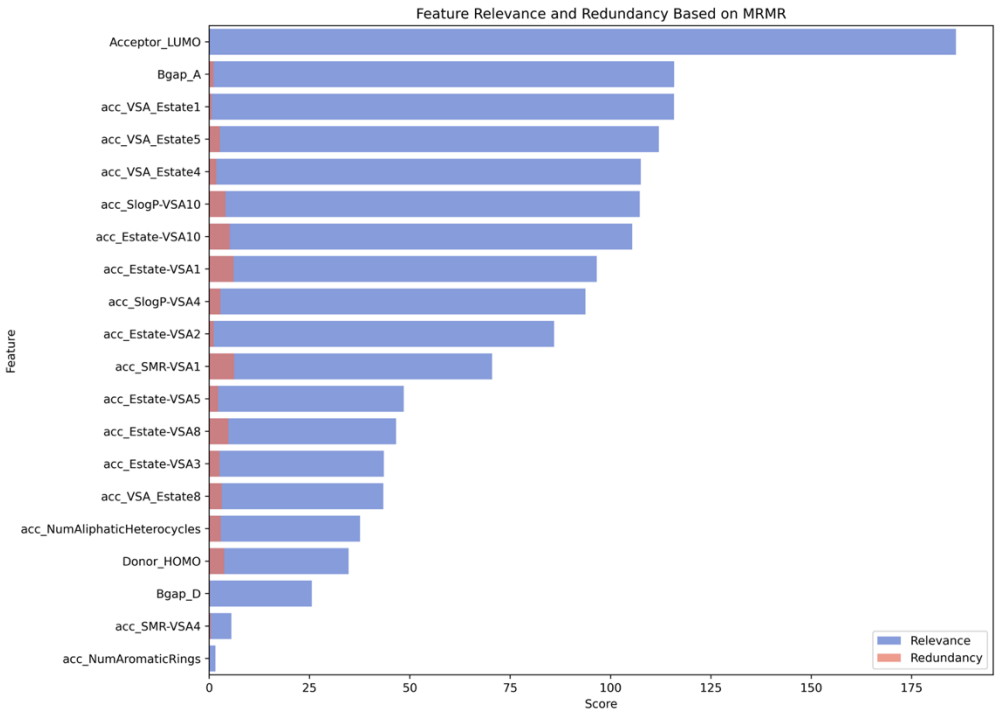


Figure 13. Results of features relevance and redundancy using the MRMR in case of predicting V_{oc} (V) [15]

3.3. Workflow for Predicting the targets

The proposed machine learning workflow for cleaning the data and predicting OPV device performance is summarized in algorithm 1. Where the prediction process comprises three main stages:

3.3.1. Model Development

Five supervised regression algorithms were evaluated on both datasets (data_1 and data_2):

- Support Vector Regression (SVR)
- Random Forest (RF)
- AdaBoost Regressor
- Gradient Boosting Regressor (GBR)
- XGBoost Regressor

Each algorithm was trained to predict V_{oc}, J_{sc} and PCE, using the molecular descriptors (data_1) and the enhanced dataset with the FMOs (data_2). These models were selected for their proven

effectiveness in handling small datasets typically in chemical informatics and also for their ability to prevent overfitting.

3.3.2. Hyperparameter Tuning

To optimize model performance, hyperparameters were fine-tuned using Grid Search (GS), which is a critical step in optimizing machine learning models to achieve the best possible performance [83]. Grid Search is a widely used exhaustive search method that systematically evaluates all possible combinations of predefined hyperparameters. GS identifies the set of hyperparameters that maximize the model's predictive accuracy or other relevant metrics. Despite its simplicity and effectiveness, Grid Search can be computationally expensive, especially when dealing with large parameter spaces. Nonetheless, it remains a fundamental approach for hyperparameter optimization and serves as a baseline for comparing more advanced methods such as Random Search.The Figure 14 clearly explains the concept of grid search with a simple example of 2 hyperparameters and with a range of values (X1,X2, X3 For hyperparameter 1) and (V1, V2,V3,V4 for hyperparameter 2) in each iteration like mentioned in the figure with the 3 colors it trains and then test the model with the combinations (X1,V1), (X1,V2), (X1,V3) and (X1,V4), then in iteration 2 it compares (X2, V1), (X2, V2), (X2, V3), (X2,V4) and in. the final iteration it trains and test with (X3, V1), (X3, V2), (X3, V3), (X3, V4)

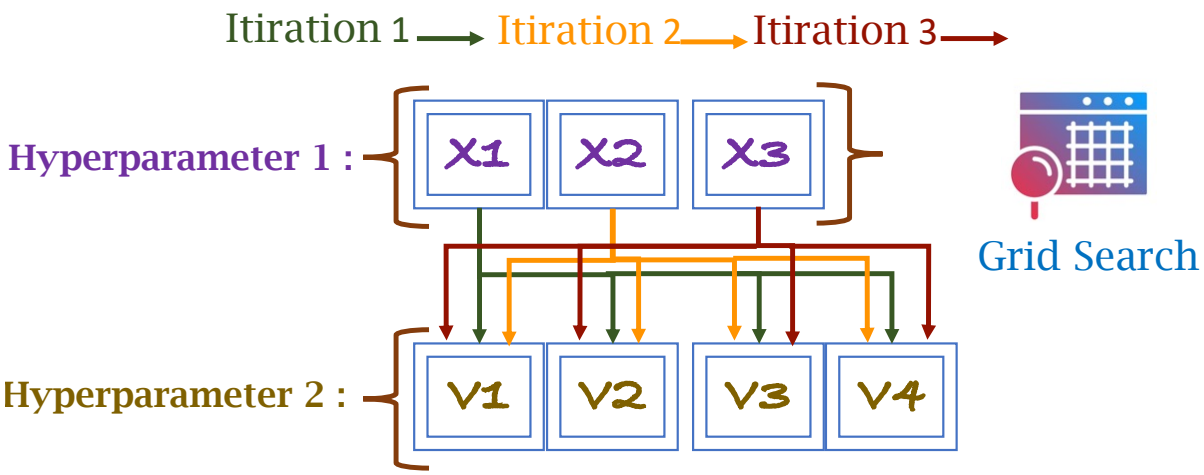


Figure 14. Grid search

3.3.3. Model Evaluation

Models were evaluated using appropriate regression metrics (R^2, MAE, RMSE and r) using train test split and the LOOCV Cross validation method which is a highly effective and robust method for evaluating model performance, especially when working with limited data [84], [85]. By systematically using each data point once as a validation set while training on the rest, LOOCV ensures that every observation contributes to both training and validation. This approach makes full use of the dataset, providing a nearly unbiased estimate of a model's ability to generalize to new data [86]. Its meticulous nature and precision make LOOCV a good choice for assessing model performance, particularly in scenarios where data is scarce and accuracy is critical [87], [89]. This method is explained in Figure 15.

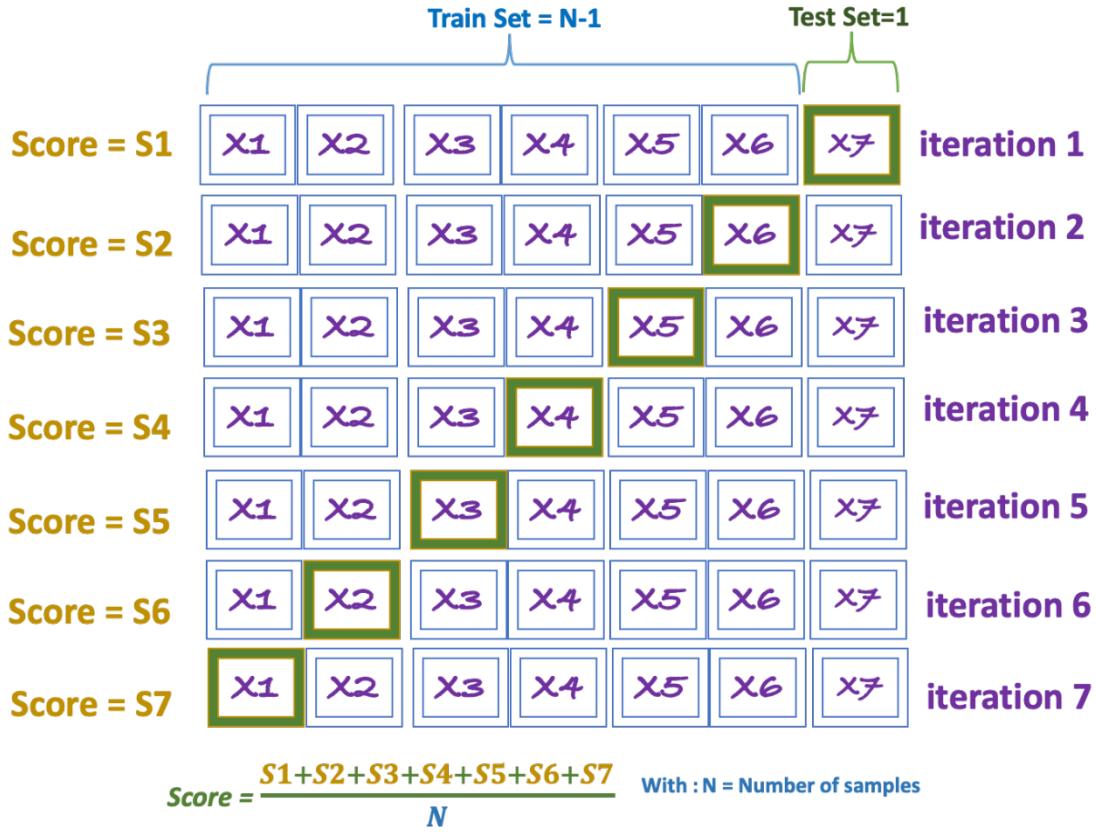


Figure 15. Leave One Out Cross validation (LOOCV) concept

Finally, the used models demonstrate the potential of using molecular structure-based descriptors to reliably predict the photovoltaic performance of organic solar cell devices.

Algorithm 1 The proposed AI-based method for predicting photovoltaic parameters of (D/A) pairs [15].

Input: original_data

Output: Predicted V_{oc}, J_{sc} and *PCE*

```
Begin Algorithm
    # Step 1: Data Pretreatment
        Begin
        # Data Cleaning
            cleaned_data = remove_duplicates(original_data)
            pairs_SMILES = standardize_names(cleaned_data)
        # Feature Extraction using RDKit
            data_1 = extract_descriptors(pairs_SMILES)
            data_2 = add_FMO(data_1, FMO)
        End
        # Step 2: Model training
        Begin
        # Prepare datasets
            datasets = {'data_1': data_1, 'data_2': data_2}
            models={'SVR': SVR(), ' RF': RandomForestRegressor(),
                    'XGBR':XGBRegressor(), 'GBR': GradientBoostingRegressor(),
                    'Adaboost': AdaptiveBoostingRegressor()}
        # Hyperparameter Tuning and Training
            For each data_key in datasets:
                    dataset = datasets[data_key]
                    y= dataset[['Voc', 'Jsc', 'PCE']]
                    X = dataset.drop(columns=['Voc', 'Jsc', 'PCE'])
                    For each model_key in models:
                        model = models [model_key]
                        tuned_model=GridSearchCV(model, param_grid[model_key])
                        tuned_model.fit(X, y)
                        predictions = tuned_model.predict(X)
                        evaluate_performance_train_test_split(predictions, y)
                        evaluate_performance_LOOCV(predictions, y)
```

```
evaluation_results=[]
evaluation_results=evaluation_results.add(
evaluate_performance_train_test_split,
evaluate_performance_LOOCV)
End
# Step 3: Final performance evaluation with the best model
Begin
    For each data_key in datasets:
                best_model = select_best_model(evaluation_results)
# Predict Voc, Jsc, and PCE using the best model
                final_predictions = best_model.predict(X)
                evaluate_performance(final_predictions)
End
```

End Algorithm

---

4. Results and discussion:

4.1. V_{oc} (V) predicting

To evaluate the performance of our predictive models, key statistical metrics were used namely RMSE, MAE, R^2, and the correlation coefficient (r). Lower RMSE and MAE values signify improved model accuracy, while R^2 values approaching 1 and r values near ± 1 are indicative of strong model performance. According to the analysis summarized in Table 3, the Gradient Boosting Regressor (GBR) and XGBoost models produced the most accurate results. These models are known for their robustness against overfitting, which appears to have provided them with an advantage over other methods such as Support Vector Regression (SVR) and Random Forest (RF). Notably, GBR demonstrated superior predictive performance compared to XGBoost. This may be attributed to the way GBR optimizes its loss function in alignment with the specific structure and characteristics of the dataset. When using data_1, which consists solely of RDKit descriptors, the GBR model yielded promising results in predicting the open-circuit voltage. With a training/testing split of 80/20, the model achieved an R^2 score of 0.68, a correlation coefficient of 0.83, and a RMSE of 0.052 V. These outcomes underscore the

effectiveness of RDKit-derived features in capturing relevant molecular characteristics for V_{oc} prediction. To enhance model accuracy, we integrated experimental frontier molecular orbital (FMO) values with RDKit descriptors, resulting in data_2. This hybrid dataset improved the model's performance further, with GBR achieving an R^2 of 0.78 and a reduced RMSE of 0.045 V and approximately a 15% increase in R^2. Figure 16 (a) illustrates the correlation between predicted and actual V_{oc} values. The enhancement aligns with established empirical evidence suggesting a direct relationship between V_{oc} and the energy difference between the donor's and the acceptor's FMOs [90], [91]. Nevertheless, it's important to highlight that RDKit descriptors alone still delivered robust predictions. Variables such as the number of aromatic rings, various VSA descriptors, and aliphatic heterocycle counts proved particularly influential. This is critical since the FMO data used in data_2 are experimentally derived, necessitating molecule synthesis and analytical techniques like cyclic voltammetry. Although theoretical FMO values can be calculated via methods such as DFT or TD-DFT, these are computationally intensive and often diverge from experimental values. Consequently, relying solely on RDKit descriptors (as in data_1) offers a more practical and scalable alternative for screening donor–acceptor pairs. To further validate model generalizability, Leave-One-Out Cross-Validation (LOOCV) was applied. For both data_1 and data_2, the resulting RMSE values (0.044 V and 0.042 V, respectively) confirm consistent predictive accuracy. Moreover, this approach effectively mitigates overfitting risks and ensures that the model maintains reliability when applied to unseen data [85], [84]. In conclusion, while incorporating experimental FMO data enhances model accuracy, the use of RDKit descriptors alone presents a compelling alternative. This approach maintains high predictive performance while simplifying the workflow, making it suitable for early-stage molecular screening in photovoltaic research.

4.2. J_{sc} (mA/cm^2) predicting

The prediction strategy used for V_{oc} was also applied to short-circuit current density, using both descriptor-only (data_1) and descriptors with FMO (data_2) datasets. The models performed comparably well in both cases. With only RDKit descriptors, the model achieved an R^2 of 0.67, a correlation coefficient (r) of 0.82, and a RMSE of 3.33 mA/cm², indicating a strong predictive capability. Incorporating experimental FMO values slightly improved these metrics, yielding an R^2 of 0.70, an r of 0.83 and a RMSE of 3.19 mA/cm². These results suggest that the RDKit descriptors alone are highly effective in capturing the structural features influencing J_{sc}, making them a viable choice when experimental FMO data are not available. However, as illustrated in

Figure 16 (b), the model struggled with accurately predicting J_{sc} for samples with very low current densities (below 2 mA/cm²). This is likely due to their sparse representation in the dataset and the limited optimization of low performing devices in experimental settings. Despite this, such samples were retained in the dataset to ensure that the model remains applicable across the full spectrum of device performances, from suboptimal to highly efficient OPV systems.

4.3. *PCE* (%) predicting

The predictive modeling of power conversion efficiency was conducted using the same established machine learning framework. Initially, the model was trained using only RDKit-derived molecular descriptors. Under these conditions, the model achieved an R^2 of 0.61, a Pearson correlation coefficient of 0.78, and a RMSE of 3.3%. These metrics indicate a solid baseline performance using purely derived molecular features. To assess the impact of incorporating additional electronic information, data_2 was used for training as same as for V_{oc} and J_{sc}. This led to a modest but noticeable improvement in model performance. Specifically, the R^2 value increased to 0.63, the correlation coefficient rose slightly to 0.79 and the RMSE decreased significantly to 2.3%. These improvements are visualized in Figure 16 (c). Overall, these results underscore the effectiveness of machine learning models in predicting *PCE* based solely on molecular descriptors. While incorporating experimental FMO data can enhance performance slightly, the high accuracy achieved with RDKit-derived descriptors alone demonstrates that such models are both robust and reliable. Furthermore, the performance metrics observed here are consistent with, and in some cases comparable to, those reported in the current scientific literature, further validating this modeling approach.

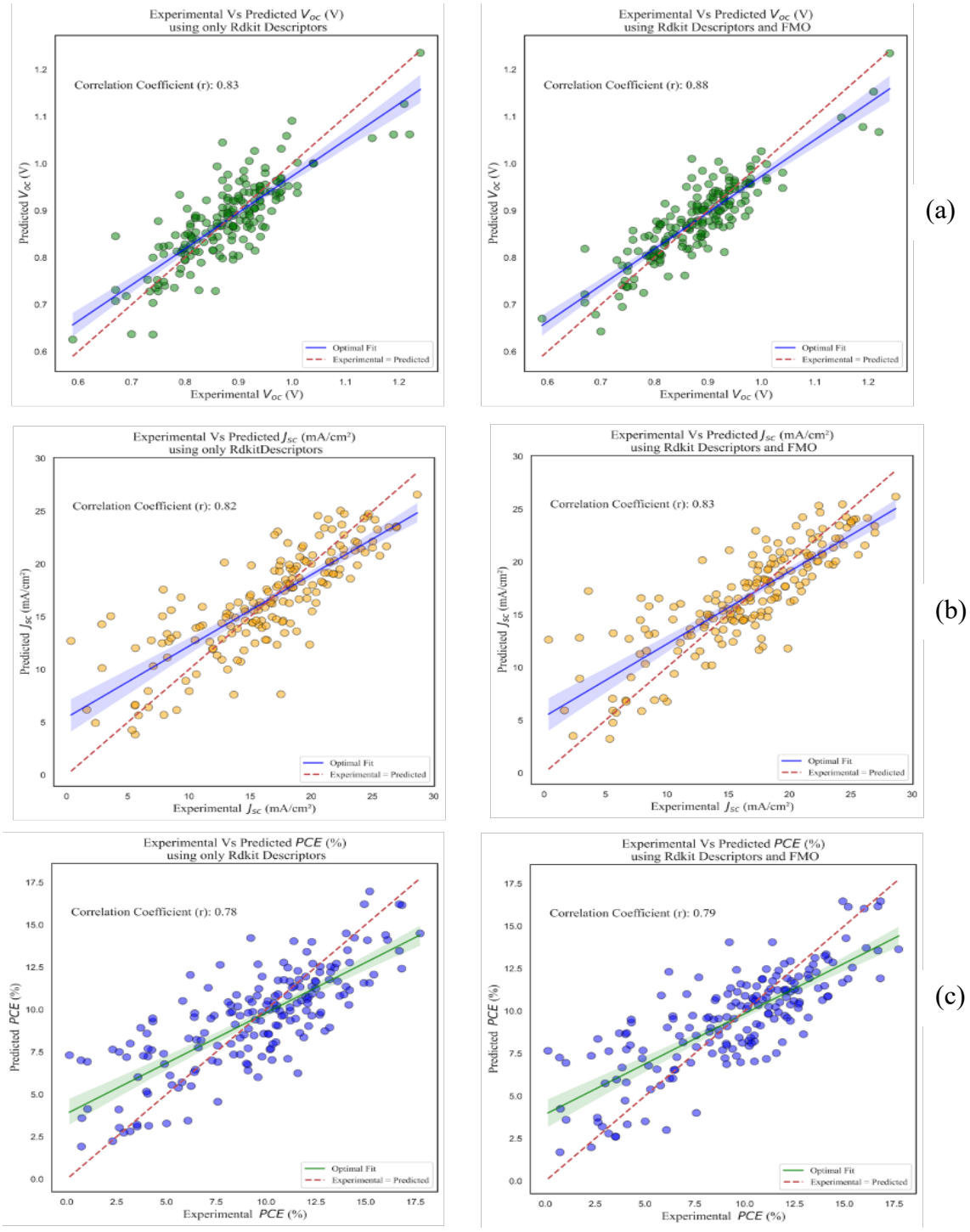


Figure 16. Experimental Vs predicted values of V_{oc} (a), J_{sc} (b) and PCE (c) using GBR model with the chemical descriptors of (D/A) pairs (left) and with the chemical descriptors and the FMOs of (D/A) pairs (right) [15]

Table 3. Prediction of OPV performance results by using different ML models applied to data_2 (With FMO) [13]

Performance Parameters	**V_{oc} (V)**				**J_{sc} (mA/cm^2)**				***PCE* (%)**			
Model evaluation metrics	**R^2**	**MAE (V)**	**RMSE (V)**	**r**	**R^2**	**MAE (mA/cm^2)**	**RMSE (mA/cm^2)**	**r**	**R^2**	**MAE (%)**	**RMSE (%)**	**r**
SVR	0.49	0.051	0.068	0.72	0.55	2.72	3.87	0.76	0.54	1.99	2.60	0.75
RF	0.63	0.048	0.064	0.82	0.65	2.43	3.40	0.81	0.58	1.90	2.49	0.76
XGBoost	0.70	0.040	0.057	0.84	0.67	2.44	3.28	0.82	0.60	1.88	2.43	0.78
GBR	0.78	0.034	0.045	0.88	0.70	2.36	3.19	0.83	0.63	1.84	2.37	0.79

4.4. Targeted evaluation on high-efficiency OPVs (PCE > 10%)

To further validate the robustness and predictive accuracy of our machine learning model, we conducted a targeted evaluation focused exclusively on high-efficiency organic photovoltaic devices specifically, those with experimentally reported power conversion efficiencies exceeding 10%. This performance threshold is particularly relevant as devices in this efficiency range are considered technologically promising and often prioritized in research and development efforts. Our approach mirrors the methodology adopted by Greenstein and Hutchison (2023) [14], who evaluated their Random Forest regression model on a similarly curated subset of high-performance OPV devices. Their model was trained and validated using a five-fold cross-validation scheme, yielding an R^2 value of 0.28 and a root mean square error of 1.6%. These metrics serve as a valuable benchmark for assessing the effectiveness of alternative machine learning frameworks in predicting high-efficiency OPV behavior. Applying the same validation protocol to a Gradient Boosting Regressor (GBR) model with different hyperparameters that were used before, we observed a marked improvement in predictive performance. Specifically, our GBR model achieved an R^2 value of 0.46 and an RMSE of 1.38%, substantially outperforming the RF model in both correlation strength and predictive error. These results demonstrate that our GBR-based approach offers a more reliable and precise tool for forecasting the performance of high-efficiency OPV devices. This targeted analysis underscores the value of advanced ensemble methods like Gradient Boosting in capturing the nuanced structure-property relationships that influence OPV efficiency,

especially in the high-performance devices. Consequently, the GBR model holds a significant promise for accelerating the discovery and optimization of next-generation OPV materials.

4.5. Model validation using previously unseen data

To rigorously test the generalizability of our predictive model, we extended our evaluation to include newly published experimental data on (D/A) material combinations that were entirely absent from the model's training and validation datasets. This approach is crucial for assessing how well the model performs in real-world scenarios where the chemical systems of interest may differ from those previously encountered.

The selected (D/A) pairs represent a diverse range of molecular architectures and performance profiles, encompassing both highly efficient and suboptimal organic photovoltaic (OPV) devices. This diversity allowed us to critically assess the model's capacity to generalize across a broad spectrum of photovoltaic behaviors. For each of these unseen systems, the model was used to predict key device performance parameters. As summarized in Table 4 and visually represented in Figure 17, our model achieved commendable accuracy across all three performance metrics. The predicted values showed strong agreement with experimental results, indicating that the model can effectively infer performance outcomes solely from the molecular structure of the constituent donor and acceptor materials. This capability is particularly notable given that the model had no prior exposure to these specific chemical combinations. These results affirm the model's potential as a powerful tool for forward screening of novel OPV materials. By accurately identifying promising as well as underperforming (D/A) pairs, the model can serve as a valuable decision-making aid in the material design and selection process.

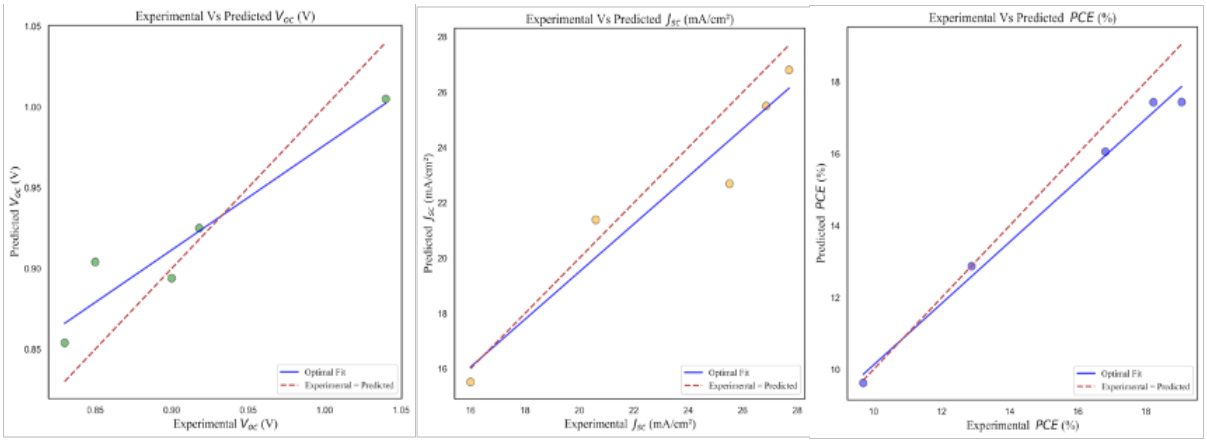


Figure 17. Experimental Vs predicted values of V_{oc} (a), J_{sc} (b) and PCE (c) using GBR model on unseen input data [15]

Table 4. Test of the predicting capabilities of our model on unseen experimental (D/A) pairs comparing the predicted (Pred.) and Experimental (Exp.) values for the mains photovoltaic parameters [15]

Donor	Acceptor	Pred. V_{oc} (V)	Exp. V_{oc} (V)	Pred. J_{sc} (mA/cm^2)	Exp. J_{sc} (mA/cm^2)	Pred. *PCE* (%)	Exp. *PCE* (%)	References
D18	**L8-BO**	0.925	0.92	25.5	26.86	17.437	19.05	[84]
PBDB-TF	**ITIC**	1.005	1.04	15.521	16	9.627	9.7	[85]
D18	**Y6**	0.904	0.85	26.803	27.7	17.432	18.22	[38]
PBDB-TF	**IT-4F**	0.854	0.83	21.386	20.6	12.874	12.88	[86]
PBDB-TF	**Y6-1O**	0.894	0.9	22.692	25.51	16.057	16.81	[87]

5. Discovering novel (donor/acceptor) pairs

In another work [16], we employed a GBR predictive model to evaluate a comprehensive dataset comprising 125867 unique (D/A) material pairs. Among these, 924 pairs have been previously investigated and reported in the scientific literature, providing a valuable benchmark for model validation. The remaining combinations, however, represent unexplored territory, offering a significant opportunity for discovery. Our primary objective was to predict key performance metrics for each (D/A) pair within the context of organic photovoltaic devices. By leveraging our model's predictive capabilities, we aimed to uncover potential high-performing (D/A) combinations that have not yet been synthesized or experimentally tested. The methodology and workflow adopted for this investigation are comprehensively outlined in Figure 18, which illustrates the sequential process from data collection and model training to performance prediction and result analysis.

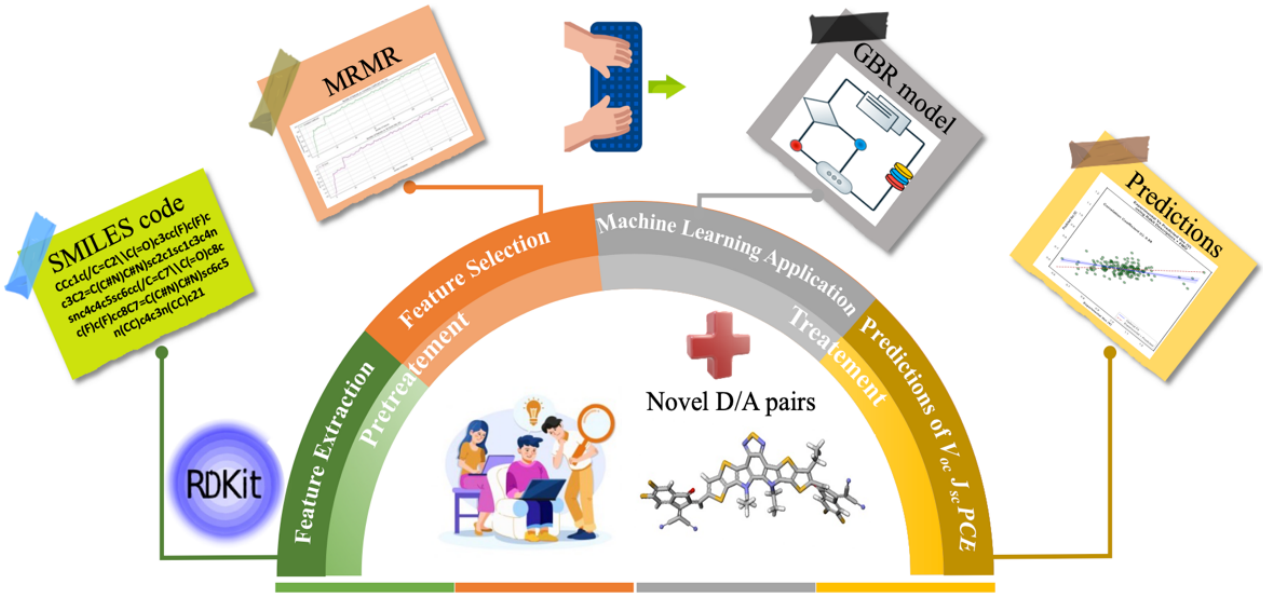


Figure 18. Approach of discovering new D/A pairs [16]

As a result, there was a combination that was not included in the training dataset of our machine learning model namely (D18/Y6-Se) but that was studied experimentally afterward [88]. This (D18/Y6-Se) pair highlights our model's generalization capability. Indeed, the experimental values (second row of Table 3) for this (D/A) pair align well with the ML predicted photovoltaic parameters (first row of Table 3) [15]. Notably, the relative error for the predicted open-circuit voltage remains below 7%, while the predicted short-circuit current density and power conversion efficiency exhibit relative uncertainties around 1%. These precise predictions were made by the Gradient Boosting Regressor. The GBR's strength lies in its ability to identify and emphasize the most influential features within the dataset, enabling reliable forecasting even for previously unseen (D/A) combinations.

Table 5 presents several (D/A) pairs yet to be explored experimentally, but which show highly promising predicted values for V_{oc}, J_{sc} and *PCE*. For example, the donor D18 combined with non-fullerene acceptors such as CH1007 and L8-BO yields predicted *PCE*s exceeding 17%, while PBQx-TF paired with CH1007 achieves a projected *PCE* of 16.5%. These results point to the potential of these combinations as high-performance candidates for organic photovoltaic devices based on NFAs. Collectively, these findings not only demonstrate the reliability and robustness of our ML approach but also underscore its value as a tool for accelerating the identification and development of efficient OPV material systems.

*Table 5. Result of prediction using new (D/A) combination [16]. *from [88]*

Donor	***Acceptor***	***V_{oc} (V)***	***J_{sc} (mA/cm^2)***	***PCE (%)***
D18	Y6-Se	*0.90*	*27.5*	*17.97*
****D18***	**Y6-Se**	***0.84***	***27.98***	***17.7***
D18	CH1007	*0.88*	*28.17*	*17.90*
D18	L8-BO	*0.92*	*25.50*	*17.44*
PBQx-TF	CH1007	*0.82*	*27.13*	*16.50*

6. Conclusion

In this chapter, we developed and evaluated a machine learning framework to predict key photovoltaic performance metrics for organic photovoltaic materials based on donor/non-fullerene acceptor (D/NFA) pairs. Starting with the SMILES representations of donor and acceptor molecules, we employed RDKit to extract chemical descriptors that quantitatively represent molecular structure and properties. These descriptors were used as input features for five different ML models. Among these models, the Gradient Boosting Regressor consistently outperformed the others, demonstrating a good predictive accuracy and robustness across all target parameters. Its ensemble learning architecture effectively captured the non-linear relationships within the descriptor space and emphasized the most relevant molecular features contributing to photovoltaic performance. To further assess the model's generalizability, we applied the trained GBR model to predict the performance of (D/A) new combinations not present in the training set and even not tested in literature and it showed a promising approach to discover some novel (D/A) pairs that can help in elaborating highly performing OPV devices. Therefore, ML has a practical value in materials research for OPV. In this chapter, the focus was on non-fullerene acceptors (NFA), however in the following chapter, we will extend our investigation to fullerene acceptors (FA) to evaluate the generalizability and robustness of our ML approach across different acceptor types.

Chapter 3: Machine Learning-Based Prediction of OPV Efficiency Using (Donor/Fullerene Acceptors) pairs

1. Introduction

In the previous chapter, a machine learning model was developed using datasets composed of (donor/non-fullerene acceptor) or (D/NFA) combinations to predict the power conversion efficiency of organic photovoltaic devices. In this chapter, we explore a complementary and extended approach by focusing on (donor/fullerene acceptors) or (D/FA) pairs. This investigation is carried out in two stages, first we train ML model (GBR) using only (D/FA) data to evaluate whether fullerene-derivative acceptors alone can offer reliable and distinct predictive patterns. This allows us to assess the specific influence of FA on OPV efficiency through a dedicated dataset. Second, we combined the (D/FA) data with the broader (D/NFA) dataset, forming a dataset that includes both types of acceptors. The GBR model is then retrained on this expanded dataset to examine whether the inclusion of fullerene-derivative acceptors enhances model accuracy, generalization, or alters the importance of key molecular descriptors. This dual approach aims to uncover the standalone predictive power of (D/FA) systems and their contribution when merged with more diverse NFA data, offering new insights into material selection and OPV performance modeling.

2. State of the art

In the early phases of organic photovoltaic research, the majority of studies were centered on systems utilizing fullerene-derivative acceptors (FAs). This early focus was driven by the inherent advantages of FAs, including their solubility in common solvents, high electron mobility, favorable energy level alignment with various donor materials, and isotropic charge transport all of which supported efficient exciton dissociation and charge collection [92]. Due to these benefits, FAs-based OPVs became the primary subject of both experimental and computational investigations for many years.

Accordingly, initial applications of machine learning in OPV research were largely confined to FAs-based systems. Early ML models were trained on datasets predominantly composed of (D/A) pairs featuring fullerene derivatives. For example, Padula *et al.* [93] demonstrated a data-driven approach to materials discovery in OPVs by integrating chemical similarity metrics into machine learning models. Using DFT, they computed electronic and structural descriptors for 249 (D/FA) pairs and applied both linear and nonlinear ML algorithms to predict device efficiencies. While electronic and structural parameters individually offered comparable predictive power, combining both improved model performance, achieving correlation

coefficients up to $r \approx 0.7$. This study underscored the value of multi-modal descriptors and provided a foundation for virtual screening strategies. However, while promising, the reliance on DFT-derived features may limit scalability for larger datasets, suggesting a need for faster, more generalizable descriptor generation methods in future work.

Building on this work, Sahu and coworkers [94] later expanded their dataset to include approximately 300 unique small-molecule OPV systems, thereby enhancing the diversity and robustness of data-driven modeling approaches. Simultaneously, attention turned toward polymer donor systems, valued for their ease of processing and mechanical flexibility important attributes for scalable OPV manufacturing. During this time, there was also growing interest in ternary OPV systems, which introduce a third component to improve device performance through mechanisms like energy cascades or charge transfer modulation. In response, *Lee et al.* [95] compiled a dataset of 124 ternary OPV systems, reflecting the increasing complexity and innovation in device architecture.

These foundational datasets focused predominantly on fullerene-based OPVs were instrumental in shaping the early landscape of ML-driven OPV research. They provided critical training data for performance prediction models.

2.1. Fullerene Acceptors

Fullerene acceptors are small molecules based on the spherical or rugby ball-like carbon cage structure called fullerene (C_{60} or C_{70}). The C_{60} and C_{70} are insoluble molecule and they are only deposited by thermal evaporation to form bilayer OPV devices. The fullerene-derivative acceptors, such as [60] PCBM (abbreviation for [6,6]-phenyl-C_{61}-butyric acid methyl ester), are widely used in organic solar cells because they are soluble in common solvents, they efficiently accept electrons from the donor material and transport them to the electrode. Their unique 3D structure enables good electron mobility and compatibility with many donor polymers, making them a standard choice in organic photovoltaics where the bulk heterojunction (BHJ) active layer is deposited in solution. Some examples of the most used FAs can be found in Figure 19. The different FAs have slightly different FMOs and adopt different morphologies when blended with the electron-donor material [96].

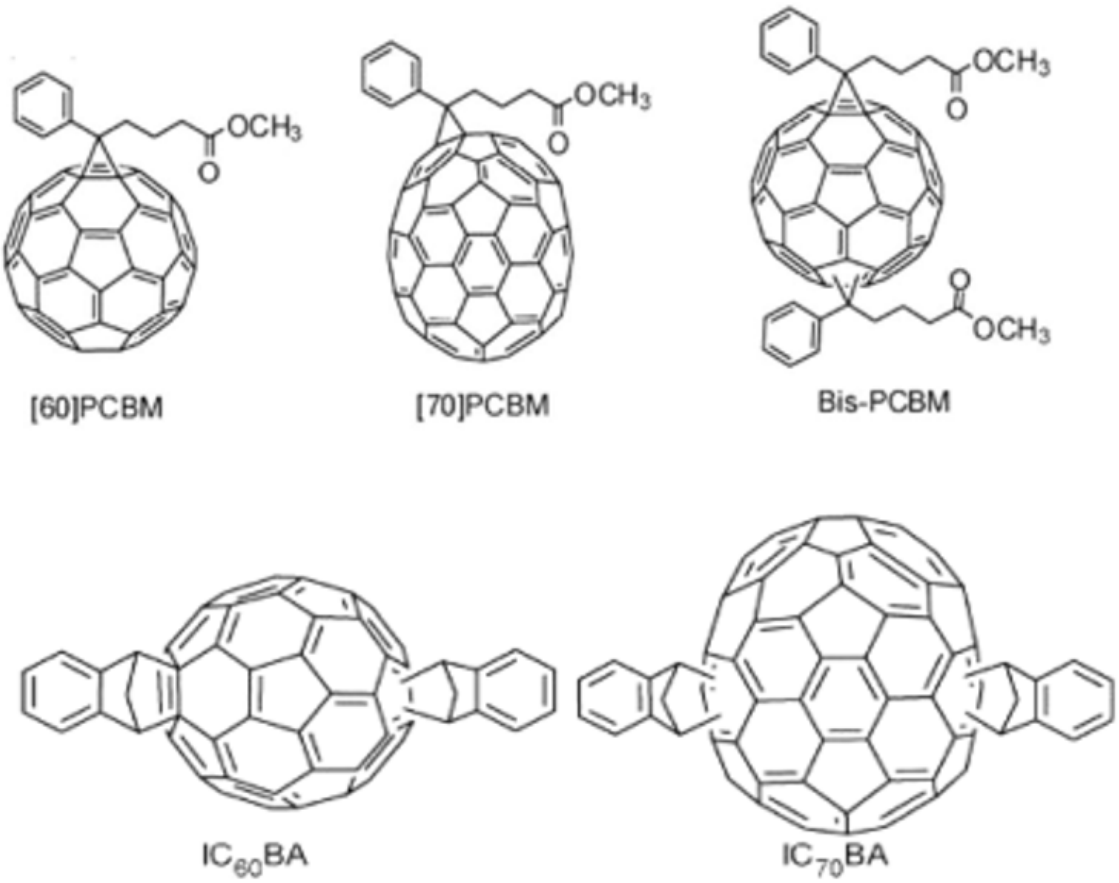


Figure 19. Different PCBM acceptors (top line) and non-PCBM acceptors (bottom line) used in OPV (adapted from [96])

3. Methodology

3.1. FA Dataset

In this study, we utilized a curated dataset comprising 249 experimentally characterized (donor/acceptor) pairs for organic photovoltaic systems, collected from peer reviewed literature published between 2013 and 2017 [93]. The dataset primarily features bulk heterojunction (BHJ) solar cells, with a minor inclusion of 8 bilayer devices, offering a representative snapshot of experimentally investigated OPV materials during that time period, where each entry in the dataset includes a comprehensive set of experimental photovoltaic performance parameters (V_{oc}, J_{sc}, *PCE* and *FF*). To complement the experimental data, density functional theory (DFT) calculations were performed to extract key electronic and structural descriptors, including HOMO and LUMO of each donor and acceptor.

The dataset is deliberately designed with a low variability in acceptor molecules, focusing exclusively on fullerene-based acceptors: C_{60}, $PC_{61}BM$, (or $PC_{60}BM$ or [60]PCBM) and $PC_{71}BM$ (or $PC_{70}BM$ or [70]PCBM). This constrained variation mirrors the experimental approach commonly used in OPV research, where donor materials are scanned while keeping the acceptor fixed, allowing for more controlled performance comparisons.

To begin the dataset analysis, we examined the distribution of (D/A) pairs with respect to the type of fullerene acceptor used. Among the 249 entries, $PC_{71}BM$ is by far the most common, appearing in 119 pairs, followed by $PC_{61}BM$ with 42 entries, and PCBM with 22 which is the same as the $PC_{61}BM$. Less frequently used acceptors include $PC_{70}BM$ (8 pairs) which is the same as the $PC_{71}BM$ and C_{60} (only 2 pairs). This uneven distribution reflects the experimental trends during the 2013–2017 period, where researchers often fixed the acceptor most commonly $PC_{71}BM$ and varied the donor material to explore performance (Figure 20). The predominance of $PC_{71}BM$ is due to its favorable optical properties, which made it a standard choice in many studies. Indeed, $PC_{61}BM$ absorbs light only in the UV while $PC_{71}BM$ absorbs decently photons in the low vavelength range of the solar-spectrum wavelength range.

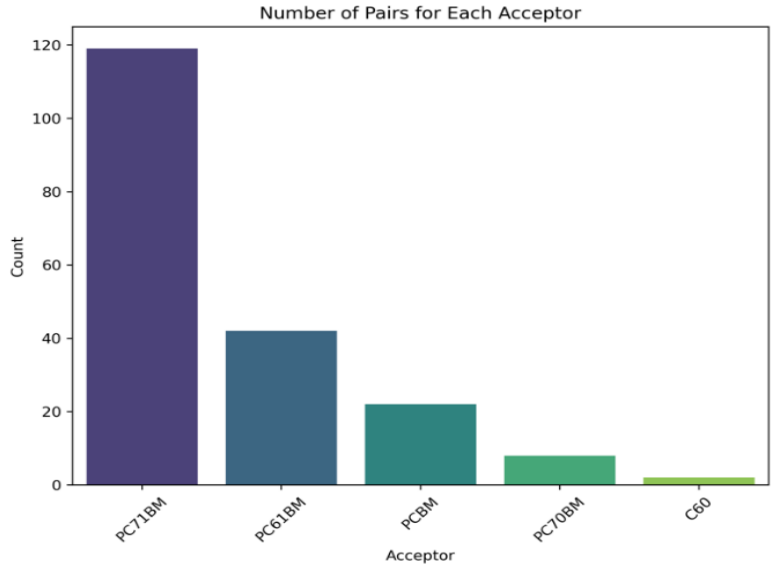


Figure 20. Distribution of FAs in the pairs in the dataset

The dataset exhibits a broad distribution of key photovoltaic parameters as mentioned in Figure 21. The open-circuit voltage ranges from 0.1 to 1.1 (V), with the majority of values concentrated between 0.6 and 1.0 (V). The short-circuit current density spans from 0 to 15 mA/cm², displaying a fluctuating pattern across this range, with values between 2 and 4 mA/cm² appearing most frequently. Similarly, the fill factor (*FF*) varies between 20% and 70% and its distribution pattern resembles that of J_{sc}, with a scattered spread and recurring peaks in the mid-

range. The power conversion efficiency values (*PCE*) lie between 0 and 7%, with a noticeable concentration of data points in the 0 to 4% range, suggesting that a significant portion of the samples demonstrate low to moderate performance.

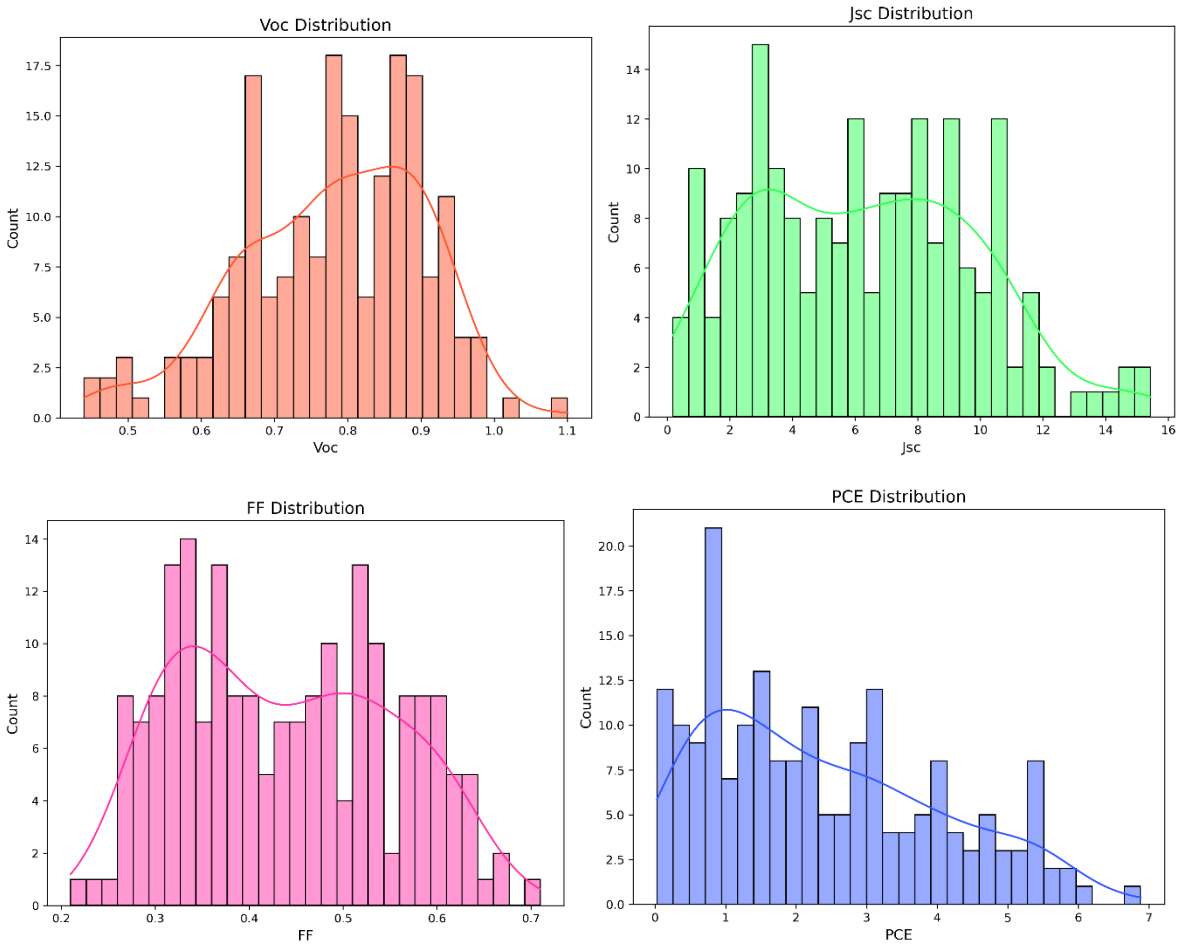


Figure 21. Distribution of the OPV performance key parameters in the dataset

3.2. NFA + FA dataset

To enhance the diversity of the dataset, we combined both the NFA and FA datasets, resulting in a total of 1111 data pairs. As illustrated in Figure 22, the distribution of V_{oc}, J_{sc}, and *PCE* values reflects the combined characteristics of both subsets. Notably, the FA-based devices generally exhibit lower performance in terms of power conversion efficiency compared to their NFA counterparts, highlighting the performance disparity between the two material systems. This is expected as NFAs were introduced to absorb efficiently photons in the solar spectral range.

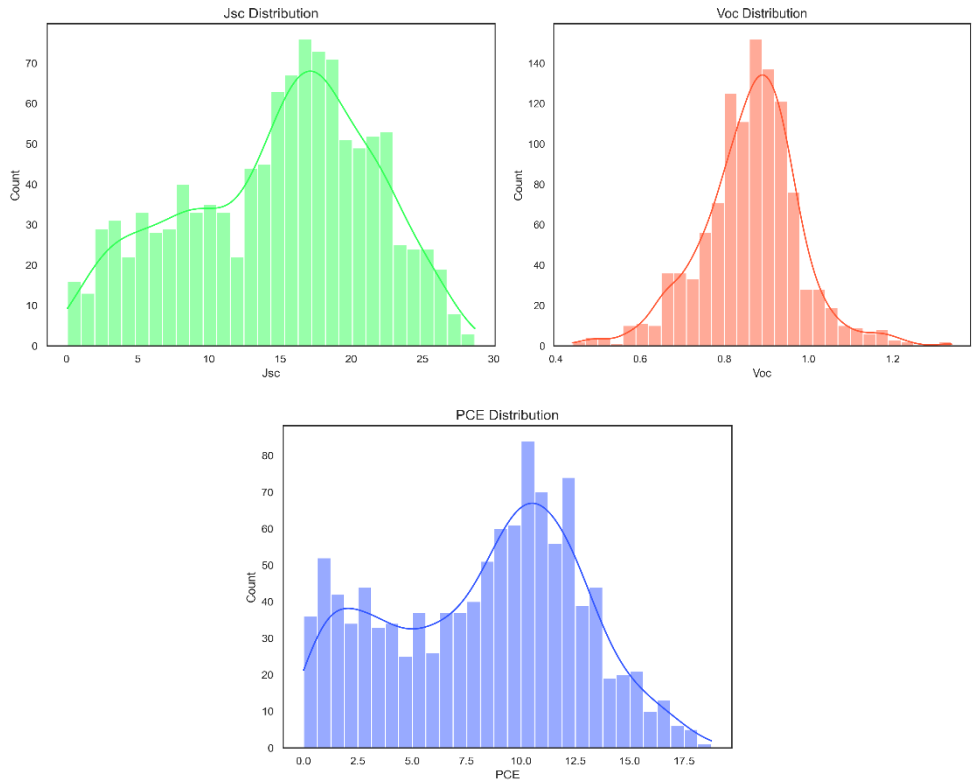


Figure 22. Distribution of the OPV performance key parameters in the NFA & FA dataset

3.3. Results and discussion

For the FA dataset, a data cleaning process was conducted to ensure the quality and reliability of the results. This involved removing duplicate donor/acceptor pairs as well as eliminating entries containing missing values (NaN). After this preprocessing step, we retained a total of 193 unique (D/A) combinations, each accompanied by its corresponding SMILES representation. To describe the chemical structures of the materials, we extracted over 100 relevant molecular descriptors using RDKit, following the same feature extraction methodology employed in our previous studies. These descriptors capture a wide range of physicochemical, structural, and electronic properties essential for predicting the performance of organic photovoltaic (OPV) devices.

For the predictive modeling task, we employed the Gradient Boosting Regressor which is a robust ensemble learning method known for its high accuracy and resistance to overfitting as explained in the previous chapters. The model was carefully tuned through hyperparameter optimization to achieve the best possible predictive performance across all target parameters, using grid search method.

As presented in Figure 23, we evaluated the predictive performance of the GBR model using a train test split and with two different training strategies: one using only the FA dataset with the chemical descriptors and the experimental FMOs as input and the other using the combined dataset, which includes both FA and NFA entries for enhanced diversity also with their corresponding FMOs and descriptors. The results show that the model trained on the FA dataset alone achieved competitive predictive accuracy. Specifically, for power conversion efficiency, the model obtained a root mean square error of 1.51% and a coefficient of correlation (r) of 0.69. These results are comparable to those reported by Padula *et al.* [93], who achieved a r value of 0.68 using a kernel ridge regression (KRR) model and DFT-based molecular descriptors.

Furthermore, the model demonstrated good predictive performance for the open-circuit voltage, with an RMSE of 0.07 V and an r value of 0.62. For the short-circuit current density, the GBR model achieved a r value of 0.72 and an RMSE of 2.36 mA/cm². While these results are encouraging and validate the predictive capacity of our molecular descriptor-based approach, the limited size and diversity of the FA dataset impose constraints on broader generalizability. We overcome these limitations and introduce more chemical and performance diversity into the training data, by training with the NFA & FA combined dataset. Retraining the GBR model on this expanded dataset led to significant improvements across all predictive targets. We paid particular attention to a careful mixing of the dataset in a way to get 20% of the data size as a test set where 10% represent FA pairs and the other 10% NFA pairs. For *PCE*, the model achieved an RMSE of 2.27% and a substantially improved r value of 0.85, reflecting a stronger correlation between predicted and actual values. Similarly, for V_{oc}, the RMSE remained consistent at 0.07 V, while the r value improved to 0.83. For the J_{sc}, the model reached an RMSE of 3.10 mA/cm² with a r value of 0.88, marking a clear enhancement in prediction accuracy. These results underscore the importance of dataset diversity in machine learning models for OPV performance prediction. By incorporating a broader range of (donor/acceptor) combinations, the model gains greater exposure to the variability inherent in molecular structures and device behaviors, leading to more robust and reliable predictions, and these findings highlight the trade-off between error magnitude and model generalization. While the error slightly increased due to broader value ranges, the correlation improvements clearly indicate a stronger and more reliable prediction trend. All results are mentioned in Table 6.

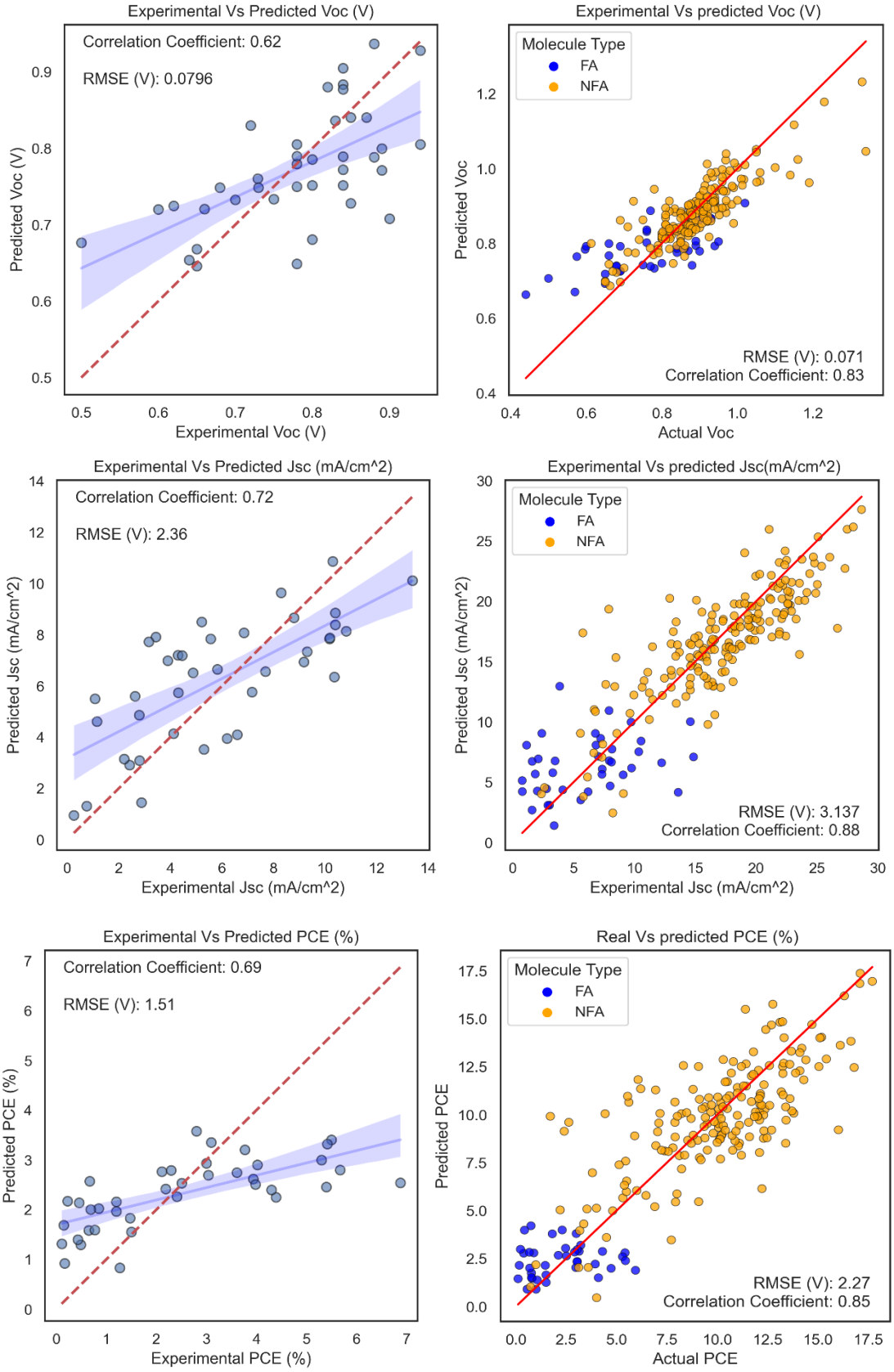


Figure 23. Results of prediction using only FAs (left) and using FAs and NFAs (right)

Table 6. Result of prediction using GBR model with FA and the combined FA & NFA dataset

Target Parameter	Dataset	RMSE	r (Correlation Coefficient)
PCE	**FA-only**	1.51%	0.69
	Combined (FA+NFA)	2.27%	0.85
V_{oc}	**FA-only**	0.07 V	0.62
	Combined (FA+NFA)	0.07 V	0.83
J_{sc}	**FA-only**	2.36 mA/cm²	0.72
	Combined (FA+NFA)	3.10 mA/cm²	0.88

4. Conclusion

In this chapter, we have explored the process of constructing and analyzing datasets for the prediction of organic photovoltaic device performance using machine learning techniques. Starting with the FA dataset, we demonstrated the potential of molecular descriptors derived from SMILES representations in capturing essential structure/property relationships relevant to OPV efficiency. The Gradient Boosting Regressor model trained solely on the FA dataset showed promising predictive capabilities, achieving respectable accuracy for key performance metrics. These results not only validated the methodological approach but also aligned well with comparable studies in literature, highlighting the effectiveness of descriptor-based machine learning in this field.

However, the limitations inherent to the relatively small and chemically narrow FA dataset became apparent, constraining the model's ability to generalize across a broader range of (donor/acceptor) pairs. Recognizing the critical role of dataset diversity and size in machine learning performance, we expanded our study by combining the FA and NFA datasets (1111 pairs). This larger, more diverse dataset introduced greater variability in molecular structures and device characteristics, which significantly enhanced the predictive power of the GBR model. Notably, we observed marked improvements in correlation coefficients and consistent root mean square errors across all key parameters when compared to the FA-only results. This confirmed that incorporating a wider range of molecular systems facilitates the development of more robust and reliable predictive models.

The findings from this chapter emphasize the importance of comprehensive and diverse datasets in advancing the application of machine learning for materials discovery and device

optimization in organic photovoltaics. Furthermore, they demonstrate that descriptor-based approaches, when combined with powerful regression algorithms like GBR, can effectively predict complex device properties with good accuracy, potentially accelerating the design and screening of new OPV materials.

Overall, this work lays a strong foundation for future studies aimed at integrating additional data sources, exploring advanced descriptors and applying more sophisticated machine learning architectures. Such efforts will be crucial in pushing the boundaries of OPV performance prediction and ultimately contributing to the development of highly efficient, cost-effective, and sustainable solar energy technologies.

Building on this work, the next chapter explores an alternative strategy: the use of deep learning models based on two-dimensional (2D) molecular structure images. By leveraging convolutional neural networks (CNNs), we investigate whether meaningful features can be automatically learned from visual representations of molecular structures, potentially removing the need for handcrafted descriptors and offering a scalable, automated path toward efficient OPV material screening.

Chapter 4: Deep Learning-Based Prediction of OPV Efficiency Using 2D Images of (Donor/Acceptor) Pairs

1. Introduction

In this fourth chapter, we investigate the potential of image-based deep learning models as an alternative to traditional descriptor-driven approaches for predicting the performance of Organic Photovoltaic devices. Building on our earlier work which relied extensively on experimental Frontier Molecular Orbital (FMO) values and handcrafted chemical descriptors we explore in the present chapter whether a convolutional neural network (CNN) trained solely on 2D molecular images of (D/A) pairs can match or surpass the predictive accuracy of these conventional models, all without the need for computationally intensive inputs. Such models often depend on the availability and relevance of predefined molecular features, which may not fully capture the nuances of molecular interactions or structural morphology. In contrast, CNNs offer the ability to autonomously learn complex spatial and structural patterns from visual data, potentially uncovering features that are overlooked or difficult to encode through traditional descriptors. We begin by outlining our data preparation workflow, including the conversion of SMILES strings into standardized 2D molecular images using RDkit cheminformatic tool. We then detail the architecture of our CNN model, with particular emphasis on its feature extraction mechanisms, training protocols, and evaluation metrics and compare its performance both with and without the inclusion of FMO values. Ultimately, we aim to determine whether image-based models can provide a scalable, descriptor-free pathway for accurate OPV performance prediction, thereby supporting more efficient and automated materials discovery in organic electronics.

2. State of the art

Despite the rapid advancements in machine learning and deep learning technologies, their application in the domain of organic photovoltaic performance prediction using two dimensional images of (donor/acceptor) pairs remain relatively underexplored [75], [18]. Most existing research efforts have predominantly focused on numerical descriptors or textual representations such as SMILES strings or molecular fingerprints to estimate key performance metrics. In contrast, the utilization of image-based inputs, particularly visual representations of molecular structures, offers a unique opportunity to exploit spatial and structural information that conventional descriptors may fail to capture. This image-driven approach has gained limited attention in literature, with only a few pioneering studies addressing it directly [1], [18], [61], [75]. One notable example is the work by Sun *et al.* [43] who demonstrated the potential

of deep learning for the rapid evaluation of OPV materials through chemical structure images. Leveraging the extensive Harvard Clean Energy Project (HCEP) dataset, which contains over 2.3 million candidate molecules, they developed a convolutional neural network based on the ResNet architecture to classify molecules by predicted PCE performance. The model was initially trained and validated on a smaller subset of 5,000 molecules and later scaled to a larger subset of 50,000 compounds, divided into training, validation, and test sets. It achieved an impressive classification accuracy of 91.02%, distinguishing between low-performance (PCE $< 5\%$) and high-performance (PCE $\geq 5\%$) categories. Sun *et al.* study underscored the potential of DL models to accelerate the materials discovery process by bypassing traditional computationally expensive quantum chemical simulations or experimental screenings. Their results validated the feasibility of using image-based learning to extract meaningful patterns directly from molecular structures. However, their model was trained on relatively simple molecular configurations and its ability to generalize across more chemically diverse and structurally complex OPV systems remains to be further investigated. Although the application of 2D images of donor and acceptor materials within deep learning frameworks for OPV performance prediction remains relatively rare, our findings and prior research suggest that this approach holds significant potential.

In this work, we explored various tools and methodologies to predict the performance of organic photovoltaic (OPV) devices, with a particular focus on the use of convolutional neural networks (CNNs). We investigated how different representations of donor and acceptor materials can impact predictive performance, including the novel use of 2D molecular images.

2.1. Convolutional Neural Networks

Convolutional Neural Networks are a powerful class of deep learning models widely used for analyzing visual data. They are particularly well-suited for tasks involving image input, including both classification and prediction problems. These networks are structured with multiple layers, such as convolutional layers, pooling layers and fully connected layers, that allow them to automatically extract hierarchical features from raw images and learn complex patterns [97] [98]. For CNNs to achieve optimal performance on a specific task, whether it's recognizing objects in photographs or predicting continuous values from chemical images, the careful selection and tuning of hyperparameters is crucial. These hyperparameters include learning rate, number of layers, kernel size, stride, batch size, activation functions and regularization methods, among others. Each of these factors plays a significant role in shaping

the model's learning process and its final performance [97]. Table 7 provides a comprehensive overview of these hyperparameters and their impact on the model's effectiveness.

Table 7. Hyperparameters of CNN

Term	Meaning
Learning Rate	Controls how fast the model updates its weight during training
Epochs	One full pass through the entire training dataset and more epochs allows the model to better learn from the input data
Batch Size	Number of samples processed at once before weights are updated, it impacts memory usage and model stability during training
Loss Function	Measures the error between predicted and actual values (e.g: MSE = Mean Squared Error)
Convolutional Layer	Extracts features (Feature Map) from the input data
Pooling Layer	Reduces spatial size of data to focus on key features that influence predictions so it will speed up computation and avoids overfitting
Activation Function (e.g.: ReLU, sigmoid, Linear...)	It introduces non-linearity, allowing the network to learn complex relationships in the data and helps tailor the model for either classification or regression tasks
Dropout	Randomly disables neurons during training to prevent overfitting

Furthermore, Table 8 also illustrates the key distinction between classification and regression tasks. In classification, the goal is to assign input data into predefined discrete categories or labels such as identifying whether an image contains a cat, dog or a car. In contrast, regression tasks involve predicting a continuous output based on input data for instance, estimating the age of a person from a facial image or predicting the performance of OPVs using chemical structure of (D/A) pairs. Table 8 highlights how different model configurations and evaluation metrics are applied depending on whether the task is categorical (classification) or continuous (regression). As a result, selecting the right hyperparameters and understanding the nature of the task (classification Vs. regression) are both essential steps in developing a successful CNN-based system.

Table 8. Key difference between classification and regression tasks

Feature	**Classification Task**	**Regression Task**
Problem Type	Predicting discrete class labels	Predicting continuous numerical values
Examples	Cat vs Dog, Handwritten digit recognition, Tumor detection…	Age prediction from face, House price estimation, Steering angle prediction…
Output Type	One or more categories (labels)	Real-valued numbers
Final Layer Output	Vector of class scores or probabilities	One or more numeric values (e.g., [25.7], [5.1, 3.8])
Final Activation	Softmax (multi-class), Sigmoid (binary or multi-label)	Linear activation (identity function)
Loss Function	Cross-Entropy Loss (e.g., Categorical or Binary)	Mean Squared Error (MSE), Mean Absolute Error (MAE)
Evaluation Metrics	Accuracy, Precision, Recall, F1 Score, AUC	RMSE, MAE, R^2 Score, MSE
Typical Output Layer	Fully connected layer with size equal to number of classes	Fully connected layer with size equal to number of outputs (e.g., 1)
Use Case Domains	Medical diagnosis, Object detection, Sentiment analysis	Price prediction, Risk estimation, Sensor calibration
Activation in Hidden Layers	Usually ReLU, Tanh, or LeakyReLU	Same (ReLU, Tanh, etc.)

2.2. The Calculated Atomic Radius

The atomic radius of a chemical element is defined as the distance from the nucleus to the outermost boundary of its electron cloud. Due to the probabilistic nature of electron distribution as described by quantum mechanics, this boundary is not sharply defined, resulting in multiple, context-dependent definitions of atomic radius—such as covalent, van der Waals, metallic, and calculated radii. In this work, the calculated atomic radius was selected to visually differentiate atoms without relying on additional cues like color. This choice ensures a consistent and theoretically grounded comparison across elements, as calculated radii are derived from

quantum mechanical models of isolated, neutral atoms. In contrast, covalent and van der Waals radii can vary significantly with chemical environment and bonding. The calculated values used here were sourced from reputable references, including the table from [4] entries compiled from peer-reviewed literature and validated computational data. By using this definition, the visual representation of atomic sizes remains standardized, environment-independent, and scientifically robust. For example (see Figure 24), the calculated atomic radius of carbon (C) is 0.67 Å [99] , whereas its covalent radius is about 0.75 Å, and its van der Waals radius is around 1.70 Å [100] a substantial variation depending on context. Similarly, hydrogen has a calculated radius of about 0.53 Å, while its covalent and van der Waals radii are approximately 0.32 Å and 1.20 [100] or 1.10 Å [101], respectively. These discrepancies highlight the influence of chemical interactions on measured atomic sizes. By using calculated radii, which are consistent across all elements and independent of bonding conditions, the visual representation of atomic sizes remains standardized and scientifically robust.

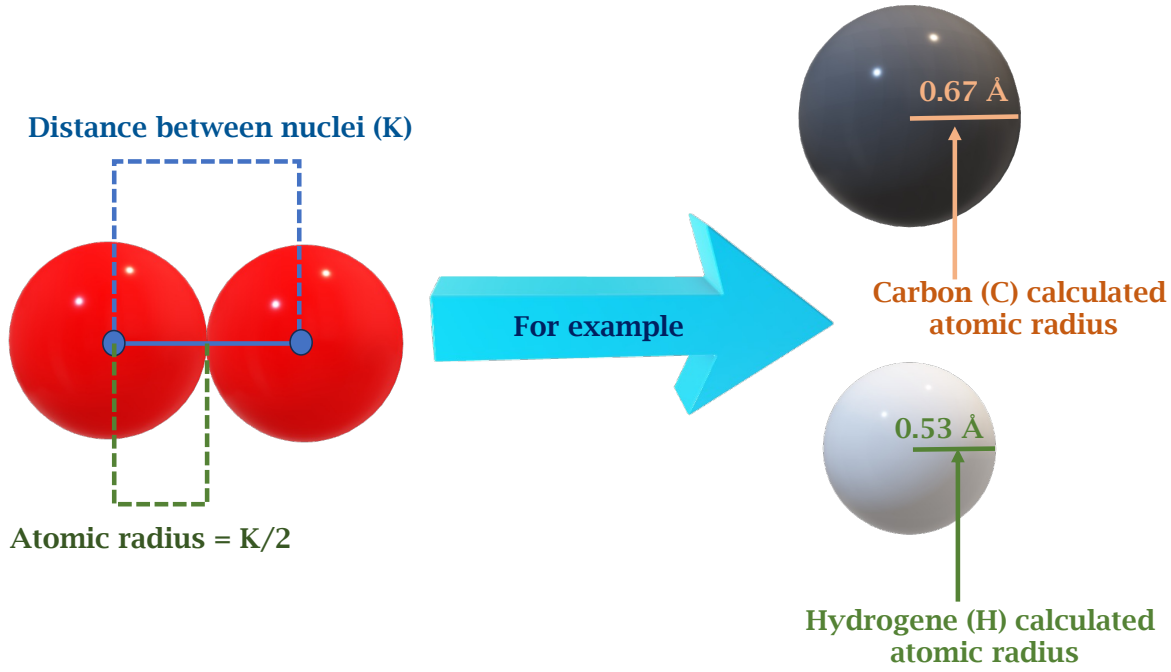


Figure 24. Atomic Radius with example

[4] https://en.wikipedia.org/wiki/Atomic_radii_of_the_elements_(data_page)

3. Methodology

3.1. Data preparation

3.1.1. Molecular representation

- **Conversion to molecular objects**: when RDKit receives a cleaned SMILES string, it transforms it into a Mol object through a structured multi-step process. Initially, it parses the SMILES notation to reconstruct the molecular graph, identifying atoms, bond types, rings, branches, charges, and stereochemistry. Each atom is assigned attributes such as atomic number, hybridization, and formal charge, while bonds are categorized by type (e.g., single, double, aromatic) and, when applicable, annotated with stereochemical information. Following parsing, RDKit performs a sanitization step. This involves checking atom valencies, inferring implicit hydrogens, validating the molecular structure against chemical rules, and identifying aromatic systems. Although SMILES strings do not include coordinate data, RDKit can optionally generate 2D or 3D coordinates using built-in layout and embedded algorithms. The result is a fully initialized Mol object: a graph-based internal representation that serves as a foundation for visualization, querying, and a wide range of cheminformatics analyses. Based on the analysis illustrated in Figure 25, which provides a Python code that demonstrates a function capable of generating an RDKit Mol object from a SMILES string. And also extract and print relevant molecular information contained in the Mol object, like 2D coordinates.

```
>>> m2 = Chem.MolFromSmiles('C1CCC1')
>>> print(Chem.MolToMolBlock(m2))

     RDKit          2D

  4  4  0  0  0  0  0  0  0  0999 V2000
    1.0607    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0000   -1.0607    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0607    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    1.0607    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  1  0
  3  4  1  0
  4  1  1  0
M  END
```

Figure 25. Code for creating 2D coordinates via RDKit [5]

[5] https://www.rdkit.org/docs/GettingStartedInPython.html

- **2D coordinate generation**: we generated 2D coordinates for each molecule extracted from the Mol Object to create planar visual representations. RDKit's tools facilitated this step, enabling us to visualize the molecules in two dimensions. As mentioned in Figure 25, the code used to create 2D coordiantes using RDKit.

3.1.2. Custom visualization

To go beyond RDKit's default symbolic molecular visualizations, we developed a custom rendering approach that provides a more chemically meaningful and visually informative representation of molecules.

- **Customized atomic representation**

Instead of representing atoms using simple textual symbols, each atom was visualized as a circle with a radius proportional to its calculated atomic radius. This approach allows users to intuitively perceive the relative sizes of different atomic species and offers a scalable method of emphasizing chemical diversity at a glance.

This design choice is motivated by the fact that atomic radius is a chemically relevant property and tends to differ significantly across elements. For example, carbon, oxygen, and sulfur atoms exhibit distinguishable radii, making this visual cue effective for differentiating atom types without relying solely on labels.

- **Bond representation and use of color**

While in theory the difference in bond types (single, double, triple) can be visually inferred from their lengths, since bond length generally decreases with increasing bond order, this difference is often too subtle in 2D projections to be reliably distinguished, especially when visualized at small or fixed scales.

To address this, we introduced a color-coded bond representation:

  - **Black** for single bonds
  - **Red** for double bonds
  - **Cyan** for triple bonds

This explicit color-coding compensates for the imperceptible differences in bond lengths on-screen and enables faster recognition of bond order during visual inspection. It enhances clarity, particularly in dense molecular graphs where spatial resolution is limited and overlapping bonds are common.

- **Use of side chains and computational cost optimization**

To evaluate the impact of solubilizing chains on both visualization and computational processing, we generated two sets of molecular images:

  - With side chains (full structure)
  - Without side chains (core/backbone only)

Taking the example of a PM6-like molecule, the conjugated backbone might contain approximately 70 atoms, while the solubilizing alkyl chains can add over 200 additional atoms, depending on their length and number. Including all atoms in a computation increases the atom count by a factor of ~4, which significantly impacts the computational cost of tasks such as geometry optimization, quantum property prediction, and deep learning-based modeling.

In terms of algorithmic complexity, many molecular modeling algorithms scale roughly as $O(N^2)$ to $O(N^3)$ where N is the number of atoms. Thus, going from 70 to 270 atoms could increase computation time by a factor of 16–50×, depending on the method [102].

Although these chains do not, most of the time, directly contribute to electronic delocalization (which is typically confined to the π-conjugated backbone), they can affect molecular behavior through steric effects, solubility, and morphology, potentially influencing packing and charge transport. Nonetheless, for electronic structure studies where computational cost must be managed, focusing on the backbone offers a justifiable trade-off between accuracy and efficiency.

3.1.3. Image preparation

To prepare the images for input into our Convolutional Neural Network (CNN), we determined the maximum X and Y coordinates from the 2D molecular representations. All images were then resized to a uniform resolution of 80 pixels per dimension (80 PDI), with final dimensions of 259 * 480 pixels. This standardization ensures that the input images have consistent size and aspect ratio, which is critical for effective CNN training.

3.1.4. Final dataset

The final dataset consists of images of donor-acceptor pairs, each visualized with our custom style to enhance clarity and differentiation of atom and bond types. This preprocessed data is now prepared for use in CNN-based analysis and modeling.

Figure 26 represents an example that illustrates the difference between traditional and our customized 2D images of materials (Ex. PM6).

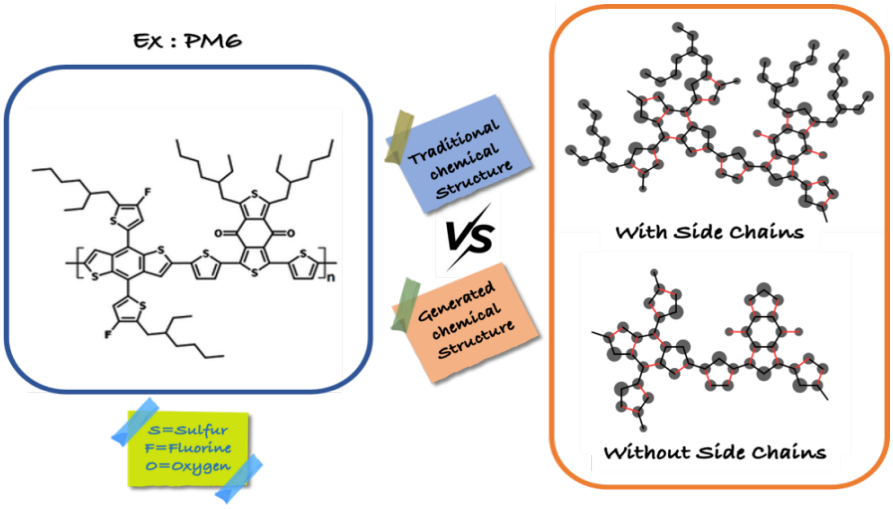


Figure 26. Traditional (Left) Vs our customized representation (Right) of the chemical structure of materials

3.2. Deep learning model architecture

In this study, a convolutional neural network served as the core of a deep learning model designed to predict the three critical photovoltaic performance metrics. The second phase of our approach was designed around leveraging image data representations of donor and acceptor materials specifically, which were processed through parallel CNN pipelines. Each of these image inputs was passed independently through a shared CNN architecture, allowing the model to extract high-dimensional and informative features from each material. The architecture implemented for each input image consisted of five convolutional layers interleaved with four max-pooling layers where each convolutional layer utilized standard 2D convolution operations with ReLU activation function, progressively capturing low to high level hierarchical patterns relevant to the electronic and structural properties encoded within the image data. Max-pooling was introduced after each convolutional block except the last to reduce the spatial dimensionality while retaining the most significant features. This structure allowed for a more compact representation of each image, mitigating overfitting and reducing computational burden without sacrificing the learning capacity of the model. Following the CNN processing, the final feature maps were flattened into one-dimensional vectors, each representing the learned features of the donor and acceptor images. These two 1D vectors were then concatenated to form a single combined vector representing the interaction between donor and

acceptor features. This composite vector served as the input to the dense or what called the fully connected layers, which acted as a regressor to predict the *PCE,* V_{oc}, and J_{sc} values. Experiments were conducted at multiple epoch settings notably 20, 50, 100 and 300 to evaluate convergence behavior and training stability. A distinguishing feature of our approach is that the entire model training and evaluation pipeline was conducted using only CPU resources, with no reliance on GPU acceleration and the hardware utilized was an HPC server where we used only 16 GB of RAM, which underscores the accessibility and scalability of our method. The ability to train and evaluate a deep learning model using only these resources while still achieving competitive results demonstrates the cost efficiency and practical feasibility of this method for broader research and deployment contexts.

To further explore and potentially improve the model's performance, a second experimental configuration was introduced. In this extension, we incorporated four numerical values corresponding to the Frontier Molecular Orbitals properties such as HOMO and LUMO energies into the predictive pipeline. These FMO descriptors were concatenated with the combined feature vector derived from the donor and acceptor CNN outputs. This enriched vector was then fed into the dense layers to perform the same regression task. The rationale behind this secondary approach was to assess the contribution of traditional molecular descriptors when used in conjunction with deep-learned image features. Comparing both configurations the image only CNN approach and the hybrid CNN + FMO model allowed for a robust analysis of whether the CNN alone could learn sufficient feature representations from the visual data, or whether performance was meaningfully improved by including conventional numerical descriptors. Figure 27 illustrates our methodology and the architecture of the used CNN model.

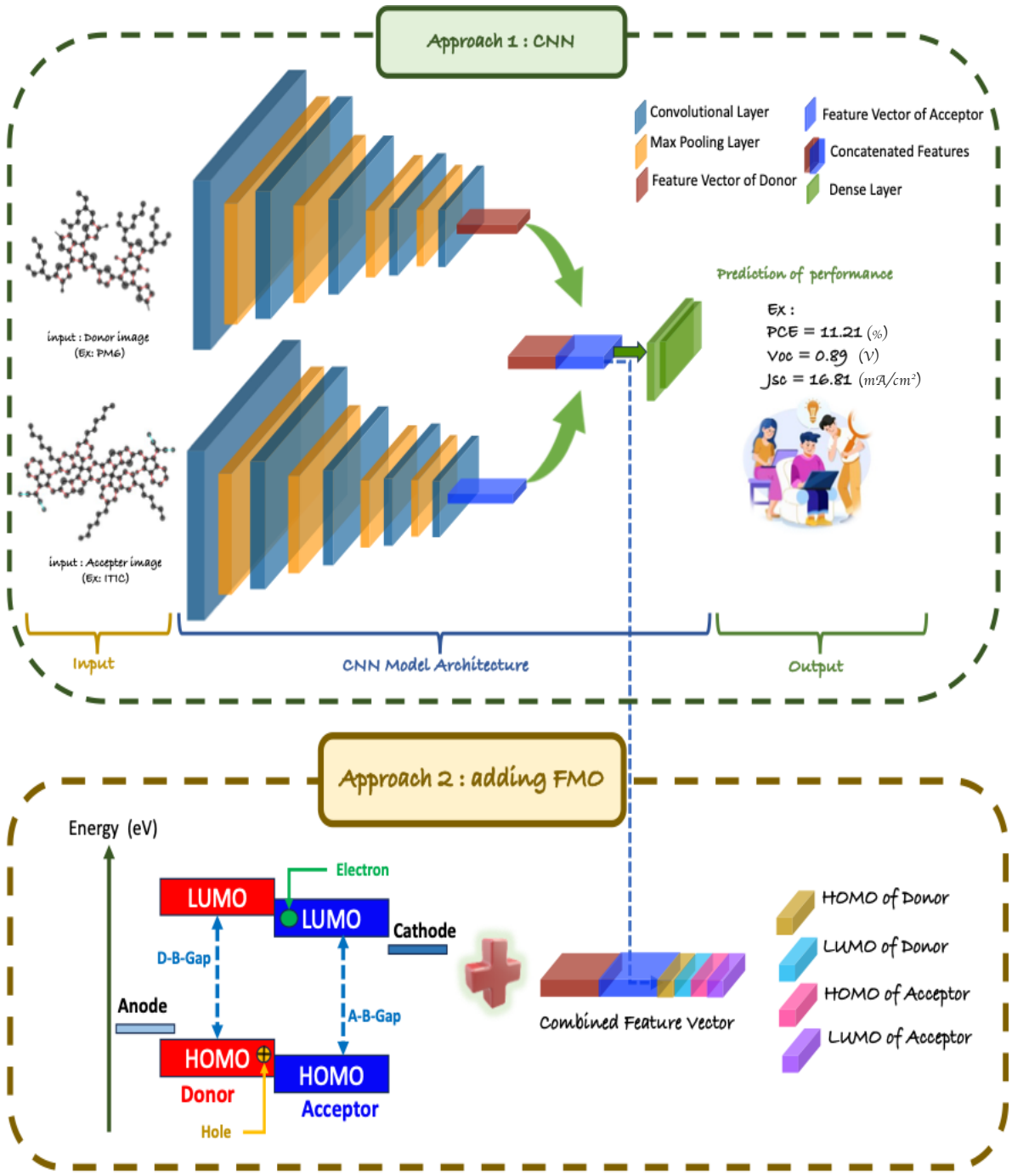


Figure 27. Our approach

In conclusion, the methodology described herein highlights a low-cost, computationally efficient, and highly scalable framework for deep learning-based prediction in organic photovoltaic research. By capitalizing on convolutional architectures and minimal hardware requirements, this work lays the foundation for accessible machine learning tools in the field of materials science and renewable energy prediction.

4. Results and Discussion

In our initial strategy for optimizing model performance, we concentrated on fine-tuning key hyperparameters, with a particular focus on the learning rate and kernel size. By systematically exploring various configurations, we identified an optimal set of parameters that delivered the highest predictive accuracy across the three primary performance indicators of organic photovoltaic devices. Notably, all training procedures were conducted on CPUs, with each epoch requiring approximately 2 to 5 seconds to complete. This approach provided a highly cost-effective and computationally efficient alternative to GPU-accelerated training. To evaluate the impact of training duration on model performance, we experimented with a range of epoch counts. Interestingly, even a relatively low number of epochs around 20 was sufficient to yield encouraging predictive results, suggesting that the model could converge rapidly and effectively without extensive training. Remarkably, this image-based convolutional neural network model achieved a good predictive outcome, even without the inclusion of frontier molecular orbital values. Compared to our earlier work [15], which utilized machine learning algorithms trained on chemical descriptors derived from SMILES strings, the CNN-based approach delivered comparable if not superior results. For instance, in predicting the *PCE*, our model reached a root mean square error (RMSE) of 2.53%, a Pearson correlation coefficient (r) of 0.73, a mean absolute error (MAE) of 1.87%, and a coefficient of determination (R^2) of 0.54. These metrics closely align with those achieved when FMO descriptors were included in the input, underscoring the CNN's inherent ability to learn critical structural and electronic features from image data alone.

As anticipated, the addition of FMO values contributed more significantly to improving the accuracy of V_{oc} predictions. However, its impact on *PCE* prediction remained relatively marginal. This observation reinforces the conclusion that, for certain performance metrics such as *PCE*, image-based models can independently capture the majority of influential molecular features without requiring supplementary quantum chemical descriptors. Importantly, our findings also highlight that deep learning models like CNNs are capable of internal feature selection, thereby eliminating the need for external dimensionality reduction techniques such as Minimum Redundancy Maximum Relevance. The network autonomously identifies and prioritizes the most informative visual patterns associated with device performance, streamlining the prediction pipeline and reducing reliance on expert driven feature engineering. Collectively, these results emphasize the potential of convolutional neural networks as robust

and scalable tools for predicting the key parameters of OPV performance using only graphical molecular data. The success of this approach not only broadens the scope of machine learning applications in materials science but also paves the way for faster, more accessible screening of candidate materials in the absence of complex chemical descriptor calculations. Detailed outcomes of these experiments are presented in Table 9 and visualized in Figure 28.

Table 9. Results of prediction of our two approaches

Performance Parameters	***PCE* (%)**				**J_{sc} (mA/cm^2)**				**V_{oc} (V)**			
Model evaluation metrics	**R^2**	**MAE (%)**	**RMSE (%)**	**r**	**R^2**	**MAE (mA/cm^2)**	**RMSE (mA/cm^2)**	**r**	**R^2**	**MAE (V)**	**RMSE (V)**	**r**
Approach 1 (Only CNN)	0.54	1.87	2.53	0.73	0.57	2.63	3.37	0.78	0.66	0.05	0.06	0.84
Approach 2 (CNN With FMO)	0.55	1.89	2.52	0.73	0.58	2.57	3.36	0.79	0.69	0.04	0.05	0.86

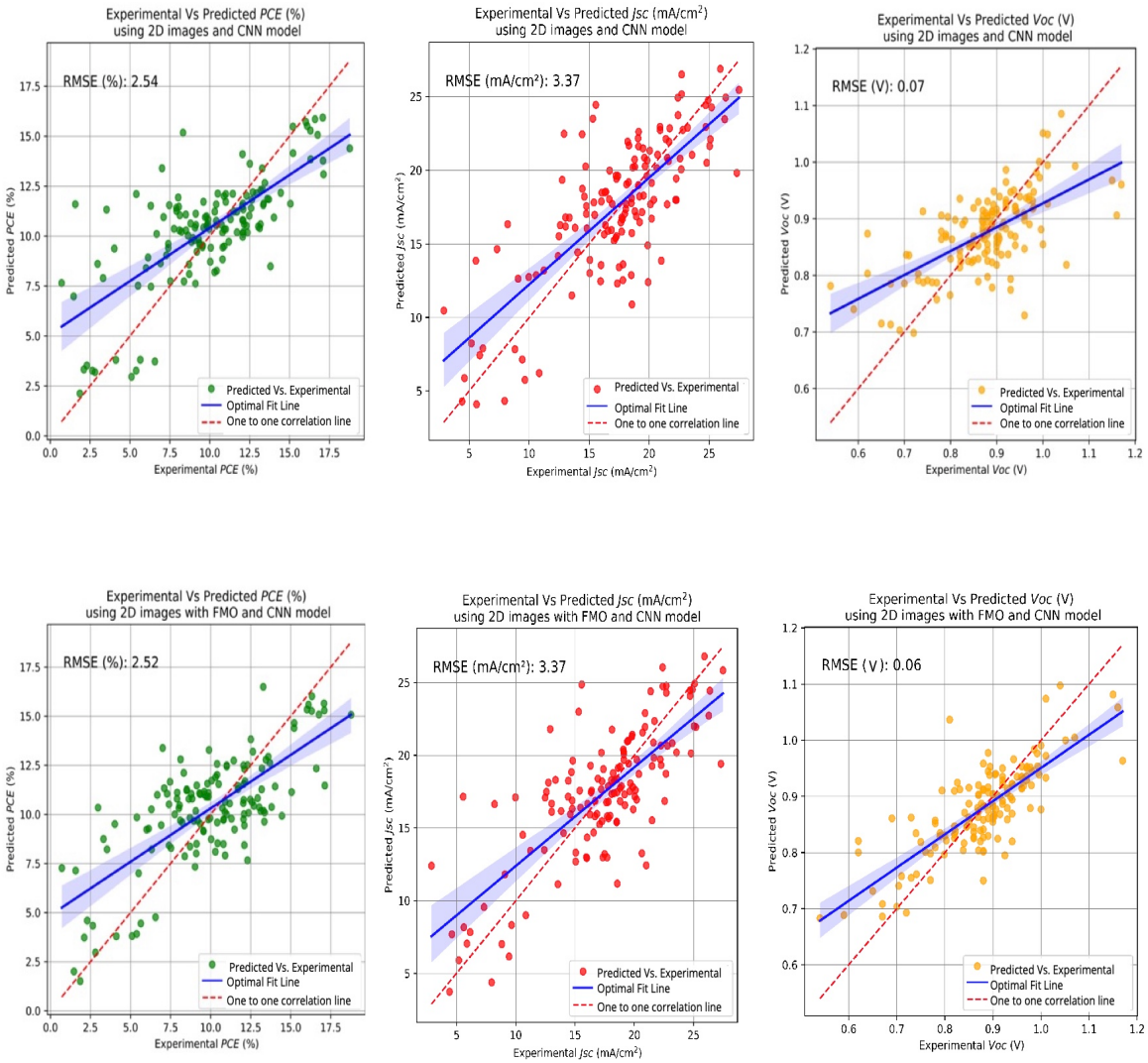


Figure 28. Results of prediction: (a) using 2D images and CNN, (b) using 2D images with FMOs and CNN

5. Conclusion

As a conclusion, this chapter demonstrates the promising potential of image-based deep learning models, specifically convolutional neural networks, as a viable alternative to traditional descriptor-driven methods for predicting Organic Photovoltaic device performance. By leveraging 2D molecular images of (donor/acceptor) pairs, our approach effectively bypasses the need for computationally expensive molecular descriptors like FMOs while still capturing complex structural features relevant to device efficiency. The results highlight that CNNs can achieve comparable, and in some cases superior, predictive accuracy relative to models relying on handcrafted descriptors and FMO values. This advancement opens the door

to more scalable and automated strategies for materials screening and design in organic electronics, emphasizing the value of visual data representations in enhancing predictive modeling capabilities.

General Conclusion

The global shift toward sustainable energy sources has become imperative in light of the growing energy demand and the severe environmental consequences associated with fossil fuel dependence. Renewable energy technologies, such as solar, wind, and hydropower, offer a pathway to mitigate climate change and promote energy security. Among them, solar energy has gained substantial momentum, not only due to its abundance and scalability but also due to the rapid technological advancements in photovoltaic systems. Organic photovoltaics (OPVs), in particular, have emerged as a promising class of solar technology owing to their unique advantages, including mechanical flexibility, lightweight construction, and compatibility with low-cost, solution-based manufacturing processes.

Despite these advantages, the commercialization and widespread deployment of OPVs are hindered by limitations in power conversion efficiency (*PCE*) and operational stability. Traditional approaches to improving OPV performance such as trial-and-error synthesis and empirical material screening are often time-consuming, resource-intensive, and unable to keep pace with the rapidly evolving demands of the energy sector. The increasing complexity of OPV systems, especially with the rise of non-fullerene acceptors and hybrid active layers, further complicates material discovery and performance optimization.

This thesis addresses these challenges by leveraging the capabilities of Artificial Intelligence (AI), specifically machine learning (ML) and deep learning (DL), to accelerate the discovery and evaluation of high-performance OPV materials. The research focuses on developing and assessing AI-driven models that can predict OPV performance metrics based on both molecular descriptors and structural information. Through a multidisciplinary approach that combines materials science, computational modeling, and data analytics, this work contributes to novel methodologies and insights to the field of renewable energy research.

The contributions of this thesis are structured across four core chapters, each building upon the previous to form a comprehensive framework for AI-assisted OPV design. In Chapter 1, the foundational concepts of OPV technology were reviewed, with an emphasis on device architecture, active layer materials, and the limitations inherent in traditional development processes. The chapter also surveyed the current landscape of AI applications in materials science, highlighting the growing interest in data-driven research and the potential of ML to transform the field.

Chapter 2 focused on the development of machine learning models using chemically meaningful descriptors. This included the construction of curated experimental dataset, feature extraction using mRMR method, and rigorous model evaluation using metrics such as R^2, r,

MAE, and RMSE. Various regression algorithms like SVR, RF, GBR, AdaBoost and XgBoost were tested, and feature importance analysis was employed to interpret the influence of molecular properties on device performance. The findings from this chapter revealed key molecular factors that govern OPV efficiency and demonstrated the predictive potential of ML models trained on well-characterized data.

In Chapter 3, the scope was expanded to include fullerene and non-fullerene acceptor systems, as well as hybrid combinations of both. This allowed for the evaluation of model generalizability across different classes of OPV materials a critical step in ensuring that predictive models are not overfitted to narrow data subsets. The models trained in this chapter showed promising performance not only on training data but also on unseen (D/A) (donor/acceptor) pairs extracted from recent experimental publications, further validating their practical applicability.

Chapter 4 introduced deep learning techniques, particularly convolutional neural networks (CNNs), to predict OPV performance based on 2D molecular structure images. This image-based approach offers a powerful alternative to traditional descriptor-based methods by allowing models to automatically learn relevant features from molecular representations. The CNN models developed in this chapter achieved competitive performance and demonstrated the feasibility of using visual molecular inputs in predictive modeling, reducing reliance on handcrafted descriptors and expert-driven feature engineering.

Together, these chapters present a holistic framework for integrating AI into the OPV material development pipeline. The methodologies explored in this thesis not only improve predictive accuracy but also reduce the time and resources required for material screening. By combining descriptor-based modeling with image-based deep learning, this research bridges the gap between traditional computational chemistry approaches and modern AI tools.

However, several challenges and limitations remain. One of the most significant obstacles is the availability and consistency of high-quality experimental data. Variability in device architectures, fabrication conditions, and testing environments can introduce noise and hinder model transferability. Moreover, certain machine learning models, especially deep neural networks, require large datasets to achieve robust generalization something that is not always feasible in the OPV research community. Addressing these limitations will require continued efforts in data standardization, collaborative data sharing, and the development of hybrid models that can integrate physical principles with statistical learning.

Future work should aim to incorporate new AI techniques to improve model transparency and foster greater trust among experimental researchers. Additionally, transfer learning and few-shot learning methods may provide promising avenues for leveraging small datasets while maintaining model accuracy.

Looking ahead, the integration of AI in OPV research is expected to deepen, with future developments likely to include automated experimentation (e.g., self-driving labs), multi-modal data fusion (e.g., combining images, spectra, and text), and closed-loop optimization strategies. These innovations will further accelerate the discovery cycle and support the deployment of next-generation photovoltaic technologies.

In conclusion, this thesis demonstrates the substantial potential of AI-driven approaches to transform the way organic photovoltaic materials are discovered, analyzed, and optimized. By combining robust data-driven methodologies with domain-specific knowledge, it contributes to the ongoing pursuit of high-efficiency, low-cost, and scalable solar energy solutions. The insights and models developed herein not only advance the field of OPVs but also lay the groundwork for broader applications of AI in renewable energy and materials science.

References

[1] Ebru Kondolot, S.; Erdal, I. Advances in Organic Photovoltaic Cells: A Comprehensive Review of Materials, Technologies, and Performance. *RSC Advances.* **13,** pp 12244–12269 (2023). https://doi.org/10.1039/d3ra01454a.

[2] Yüksel, I. Hydropower for Sustainable Water and Energy Development. *Renewable and Sustainable Energy Reviews.* **14**, pp 462–469 (2010). https://doi.org/10.1016/j.rser.2009.07.025.

[3] Joselin Herbert, G. M.; Iniyan, S.; Sreevalsan, E.; Rajapandian, S. A Review of Wind Energy Technologies. *Renewable and Sustainable Energy Reviews.* **11**, pp 1117–1145 (2007). https://doi.org/10.1016/j.rser.2005.08.004.

[4] Campos, I.; Wittmayer, J. M.; Hielscher, S.; Oliveira, F.; Moreira Alves, F.; Progscha, S.; Wientjes, A. Just Participation in Wind Energy: The Role of Social Innovations. *Renewable and Sustainable Energy Reviews.* **209**, p. 115146 (2025). https://doi.org/10.1016/j.rser.2024.115146.

[5] Su, Y.-W.; Lan, S.-C.; Wei, K.-H. Organic Photovoltaics. *Materials Today.* **15** , pp 554–562 (2012). https://doi.org/10.1016/S1369-7021(13)70013-0.

[6] Ibrahim, T.; Durillon, B.; Faraj, J.; Ali, S.; Saudemont, C.; Harion, J.; Khaled, M. A Comprehensive Review of Solar Energy Systems: Technical, Economic, and Environmental Perspectives for Sustainable Development. *International Communications in Heat and Mass Transfer.* **165**, p. 109095 (2025). https://doi.org/10.1016/j.icheatmasstransfer.2025.109095.

[7] Kamel, M. S. A.; Al-jumaili, A.; Oelgemöller, M.; Jacob, M. V. Inorganic Nanoparticles to Overcome Efficiency Inhibitors of Organic Photovoltaics: An in-Depth Review. *Renewable and Sustainable Energy Reviews.* **166**, p. 112661 (2022). https://doi.org/10.1016/j.rser.2022.112661.

[8] Faes, A.; Virtuani, A.; Quest, H.; Maturi, L.; Scognamiglio, A.; Frontini, F.; Schlueter, A.; Martin-Chivelet, N.; Reinders, A.; Ballif, C. Building-Integrated Photovoltaics. *Nature Reviews Clean Technology.* **1** ,pp 333–350 (2025).

https://doi.org/10.1038/s44359-025-00059-9.

[9] Jiang, Y.; Liu, K.; Liu, F.; Ran, G.; Wang, M.; Zhang, T.; Xu, R.; Liu, H.; Zhang, W.; Wei, Z.; Cui, Y.; Lu, X.; Hou, J.; Zhu, X. 20.6% Efficiency Organic Solar Cells Enabled by Incorporating a Lower Bandgap Guest Non fullerene Acceptor Without Open-Circuit Voltage Loss. *Advanced Materials.* **37**, p. 2500282 (2025).
https://doi.org/10.1002/adma.202500282.

[10] Cai, J.; Chu, X.; Xu, K.; Li, H.; Wei, J. Machine Learning-Driven New Material Discovery. *Nanoscale Advances.* **2**, pp 3115–3130 (2020).
https://doi.org/10.1039/d0na00388c.

[11] Liu, X.; Zhang, X.; Sheng, Y.; Zhang, Z.; Xiong, P.; Ju, X.; Zhu, J.; Ye, C. Advancing Organic Photovoltaic Materials by Machine Learning-Driven Design with Polymer-Unit Fingerprints. *NPJ Computational Materials.* **11**, p. 107 (2025).
https://doi.org/10.1038/s41524-025-01608-3.

[12] Suthar, R.; Abhijith, T.; Karak, S. Machine-Learning-Guided Prediction of Photovoltaic Performance of Non-Fullerene Organic Solar Cells Using Novel Molecular and Structural Descriptors. *Journal of Materials Chemistry A*. **11**, pp 22248–22258 (2023).
https://doi.org/10.1039/d3ta04603f.

[13] Huang, G.; Guo, Y.; Chen, Y.; Nie, Z. Application of Machine Learning in Material Synthesis and Property Prediction. *Materials.* **16**, p. 5977 (2023).
https://doi.org/10.3390/ma16175977.

[14] Greenstein, B. L.; Hutchison, G. R. Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms. *Journal of Physical Chemistry C*. **127**, pp 6179–6191 (2023).
https://doi.org/10.1021/acs.jpcc.3c00267.

[15] Khoussa, K.; Boubchir, L.; Leveque, P. Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs. (2024).
https://doi.org/10.2139/ssrn.4997197.

[16] Khoussa, K.; Leveque, P.; Boubchir, L. On the Use of Machine Learning to Discover Novel Donor-Acceptor Pairs For Organic Photovoltaic Devices. *2024 IEEE International Conference on Big Data.* pp 4659–4662 (2024).
https://doi.org/10.1109/BigData62323.2024.10825948.

[17] Khoussa, K.; Chapuis, Y.-A.; Lachiche, N. Optimisation de Matériaux et Dispositifs Pour l'énergie à Partir de Concepts d'intelligence Artificielle Pour Small Data. (2023)

https://hal.science/hal-04944011v1.

[18] Kumar, G.; Chen, F. C. A Review on Recent Progress in Organic Photovoltaic Devices for Indoor Applications. *Journal of Physics D: Applied Physics.* **56**, p. 353001 (2023). https://doi.org/10.1088/1361-6463/acd2e5.

[19] Ghorab, M.; Fattah, A.; Joodaki, M. Fundamentals of Organic Solar Cells: A Review on Mobility Issues and Measurement Methods. *Optik (Stuttg),* **267**, p. 169730 (2022). https://doi.org/10.1016/j.ijleo.2022.169730.

[20] Leclerc, N; Lévêque, P. Photovoltaique Organique-Ingénierie des matériaux, de La Nano-Morphologie et Des Dispositifs. *Technique de l'ingenieur.* (2023), NM5205 v2. https://doi.org/ 10.51257/a-v2-nm5205

[21] Brynjolfsson, E.; Mcafee, A. ARTIFICIAL INTELLIGENCE, FOR REAL. *Harvard business review.* **1**, (2017).

[22] Nakkeeran, J. Artificial Intelligence-The Next Revolution; pp. 37 (1984). https://www.researchgate.net/publication/389004322.

[23] James H Fetzer. AI. Artificial intelligence: its scope and Limits. *Springer nature links.* (1990).

[24] Chiu, T. K. F.; Ahmad, Z.; Ismailov, M.; Sanusi, I. T. What Are Artificial Intelligence Literacy and Competency? A Comprehensive Framework to Support Them. *Computers and Education Open.* **6**, p. 100171 (2024). https://doi.org/10.1016/j.caeo.2024.100171.

[25] Feuerriegel, S.; Hartmann, J.; Janiesch, C.; Zschech, P. Generative AI. *Business and Information Systems Engineering.* **66**, pp 111–126 (2024). https://doi.org/10.1007/s12599-023-00834-7.

[26] Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; Xie, X. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology.* **15**, p 1-45 (2024). https://doi.org/10.1145/3641289.

[27] Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT Understands, Too. *AI Open.* **5**, 208–215 (2024). https://doi.org/10.1016/j.aiopen.2023.08.012.

[28] Singh, S. K.; Kumar, S.; Mehra, P. S. Chat GPT & Google Bard AI: A Review. *2023 International Conference on IoT, Communication and Automation Technology (ICICAT);* IEEE. pp 1–6 (2023).

https://doi.org/10.1109/ICICAT57735.2023.10263706.

[29] Hamet, P.; Tremblay, J. Artificial Intelligence in Medicine. *Metabolism.* **69**, pp S36–S40 (2017).
https://doi.org/10.1016/j.metabol.2017.01.011.

[30] Rashidi, H. H.; Pantanowitz, J.; Hanna, M. G.; Tafti, A. P.; Sanghani, P.; Buchinsky, A.; Fennell, B.; Deebajah, M.; Wheeler, S.; Pearce, T.; Abukhiran, I.; Robertson, S.; Palmer, O.; Gur, M.; Tran, N. K.; Pantanowitz, L. Introduction to Artificial Intelligence and Machine Learning in Pathology and Medicine: Generative and Nongenerative Artificial Intelligence Basics. *Modern Pathology.* **38** , p. 100688 (2025).
https://doi.org/10.1016/j.modpat.2024.100688.

[31] Mohsin Khan, M.; Shah, N.; Shaikh, N.; Thabet, A.; alrabayah, T.; Belkhair, S. Towards Secure and Trusted AI in Healthcare: A Systematic Review of Emerging Innovations and Ethical Challenges. *International Journal of Medical Informatics.* **195**, p. 105780 (2025).
https://doi.org/10.1016/j.ijmedinf.2024.105780.

[32] Shiravale, S.; Bhagat, S. M. Wireless Sensor Networks in Agriculture Sector-Implementation and Security Measures. *International Journal of Computer Applications.* **92**, (2014).
https://doi.org/10.5120/16069-5217.

[33] Jha, K.; Doshi, A.; Patel, P.; Shah, M. A Comprehensive Review on Automation in Agriculture Using Artificial Intelligence. *Artificial Intelligence in Agriculture.* **2**, pp 1–12 (2019).
https://doi.org/10.1016/j.aiia.2019.05.004.

[34] Abduljabbar, R.; Dia, H.; Liyanage, S.; Bagloee, S. A. Applications of Artificial Intelligence in Transport: An Overview. *Sustainability (Switzerland).* **11**, p. 189 (2019).
https://doi.org/10.3390/su11010189.

[35] Okrepilov, V. V.; Kovalenko, B. B.; Getmanova, G. V.; Turovskaj, M. S. Modern Trends in Artificial Intelligence in the Transport System. *Transportation Research Procedia.* **61**, pp 229–233 (2022).
https://doi.org/10.1016/j.trpro.2022.01.038.

[36] Machin, M.; Sanguesa, J. A.; Garrido, P.; Martinez, F. J. On the Use of Artificial Intelligence Techniques in Intelligent Transportation Systems. *2018 IEEE Wireless Communications and Networking Conference Workshops, WCNCW 2018.* pp 332–337 (2018).
https://doi.org/10.1109/WCNCW.2018.8369029.

[37] Parveen, S.; Kumar, P.; Chadha, R. S.; Singh, J. Artificial Intelligence in Transportation Industry. *International Journal of Innovative Science and Research Technology.* **7**, pp 1274-1283 (2022). https://www.researchgate.net/publication/380000645.

[38] Chen, L.; Chen, P.; Lin, Z. Artificial Intelligence in Education: A Review. *IEEE Access 2020*, **8**, pp 75264–75278 (2020). https://doi.org/10.1109/ACCESS.2020.2988510.

[39] Ouyang, F.; Jiao, P. Artificial Intelligence in Education: The Three Paradigms. *Computers and Education: Artificial Intelligence.* **2**, p. 100020 (2021). https://doi.org/10.1016/j.caeai.2021.100020.

[40] Maleki, R.; Asadnia, M.; Razmjou, A. Artificial Intelligence-Based Material Discovery for Clean Energy Future. *Advanced Intelligent Systems.* **4**, p. 2200073 (2022). https://doi.org/10.1002/aisy.202200073.

[41] Liu, Q.; Jiang, Y.; Jin, K.; Qin, J.; Xu, J.; Li, W.; Xiong, J.; Liu, J.; Xiao, Z.; Sun, K.; Yang, S.; Zhang, X.; Ding, L. 18% Efficiency Organic Solar Cells. *Science Bulletin.* **65**, pp 272–275 (2020). https://doi.org/10.1016/j.scib.2020.01.001.

[42] Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine Learning Property Prediction for Organic Photovoltaic Devices. *NPJ Computational Materials.* **6**, p. 166 (2020). https://doi.org/10.1038/s41524-020-00429-w.

[43] Sun, W.; Li, M.; Li, Y.; Wu, Z.; Sun, Y.; Lu, S.; Xiao, Z.; Zhao, B.; Sun, K. The Use of Deep Learning to Fast Evaluate Organic Photovoltaic Materials. *Advanced Theory and Simulation.* **2**, p. 1800116 (2019). https://doi.org/10.1002/adts.201800116.

[44] Zhao, Y.; Mulder, R. J.; Houshyar, S.; Le, T. C. A Review on the Application of Molecular Descriptors and Machine Learning in Polymer Design. *Polymer Chemistry.* **14**, pp 3325–3346 (2023). https://doi.org/10.1039/d3py00395g.

[45] Bhatti, S.; Manzoor, H. U.; Michel, B.; Bonilla, R. S.; Abrams, R.; Zoha, A.; Hussain, S.; Ghannam, R. Machine Learning for Accelerating the Discovery of High Performance Low-Cost Solar Cells: A Systematic Review. (2022). *arXiv preprint* arXiv:2212.13893.

[46] Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science.* **12**, p. e1608 (2022). https://doi.org/10.1002/wcms.1608.

[47] Du, X.; Lüer, L.; Heumueller, T.; Wagner, J.; Berger, C.; Osterrieder, T.; Wortmann, J.; Langner, S.; Vongsaysy, U.; Bertrand, M.; Li, N.; Stubhan, T.; Hauch, J.; Brabec, C. J. Elucidating the Full Potential of OPV Materials Utilizing a High-Throughput Robot-Based Platform and Machine Learning. *Joule.* **5** , pp 495–506 (2021). https://doi.org/10.1016/j.joule.2020.12.013.

[48] Kirkey, A.; Luber, E. J.; Cao, B.; Olsen, B. C.; Buriak, J. M. Optimization of the Bulk Heterojunction of All-Small-Molecule Organic Photovoltaics Using Design of Experiment and Machine Learning Approaches. *ACS Applied Materials & Interfaces.* **12**, pp 54596-54607 (2020). https://doi.org/10.1021/acsami.0c14922

[49] Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule.* **1** , pp 857–870 (2017). https://doi.org/10.1016/j.joule.2017.10.006.

[50] Usman Khan, M.; Shafiq, F.; Ramzan Saeed Ashraf Janjua, M.; Khalid, M.; Yaqoob, J.; Arshad, M.; Alshehri, S. M.; Ahmad Khan, R. Predicting Benzodithiophene Based Donor Materials with Enhanced 19.09% PCE, Open-Circuit Voltage and Optoelectronic Attributes for Solar Cell Applications: Photochemical Insights from DFT. *Journal of Photochemistry and Photobiology A: Chemistry.* **446**, p.115115 (2024). https://doi.org/10.1016/j.jphotochem.2023.115115.

[51] Zanlorenzi, C.; Akcelrud, L. Theoretical Studies for Forecasting the Power Conversion Efficiencies of Polymer-Based Organic Photovoltaic Cells. *Journal of Polymer Science Part B: Polymer Physics.* **55** , pp. 919–927 (2017). https://doi.org/10.1002/polb.24338.

[52] Güleryüz, C.; Sumrra, S. H.; Hassan, A. U.; Mohyuddin, A.; Waheeb, A. S.; Awad, M. A.; Jalfan, A. R.; Noreen, S.; Kyhoiesh, H. A. K.; El Azab, I. H. A Machine Learning and DFT Assisted Analysis of Benzodithiophene Based Organic Dyes for Possible Photovoltaic Applications. Journal of Photochemistry and Photobiology A: Chemistry. Elsevier. **460**, p. 116157 (2025). https://doi.org/10.1016/j.jphotochem.2024.116157.

[53] Yu, H.-N.; Chen, H.-Y.; Sharma, G. D.; Cheng, Y.-J.; Hsu, C.-S.; Chu, T.-Y.; Lu, J.; Chen, F.-C. Computational Modeling of Indoor Organic Photovoltaics: Dataset Curation, Predictive

Analysis, and Machine Learning Approaches. *Archives of Computational Methods in Engineering.* pp.1-14 (2025).
https://doi.org/10.1007/s11831-025-10310-y.

[54] Togun, H.; Basem, A.; Jweeg, M. J.; Biswas, N.; Abed, A. M.; Paul, D.; Mohammed, H. I.; Chattopadhyay, A.; Sharma, B. K.; Abdulrazzaq, T. Advancing Organic Photovoltaic Cells for a Sustainable Future: The Role of Artificial Intelligence (AI) and Deep Learning (DL) in Enhancing Performance and Innovation. *Solar Energy.* **291**, p.113378 (2025).
https://doi.org/10.1016/j.solener.2025.113378.

[55] Wu, Y.; Guo, J.; Sun, R.; Min, J. Machine Learning for Accelerating the Discovery of High-Performance Donor/Acceptor Pairs in Non-Fullerene Organic Solar Cells. *NPJ Computational Materials.* **6**, p.120 (2020).
https://doi.org/10.1038/s41524-020-00388-2.

[56] Tarique, W. B.; Uddin, A. A Review of Progress and Challenges in the Research Developments on Organic Solar Cells. *Materials Science in Semiconductor Processing.* **15**, p.107541 (2023).
https://doi.org/10.1016/j.mssp.2023.107541.

[57] Nie, C.; Wang, K.; Zhou, H.; Deng, J.; Chen, Z.; Zhang, K.; Chen, L.; Huang, D.; Liang, J.; Zhao, L. Combination of Transfer Learning and Chemprop Interpreter with Support of Deep Learning for the Energy Levels of Organic Photovoltaic Materials Prediction and Regulation. *ACS Applied Materials & Interfaces.* **16**, pp 66316–66326 (2024).
https://doi.org/10.1021/acsami.4c15835.

[58] Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A. The Harvard Organic Photovoltaic Dataset. *Scientific data.* **3**, pp.1-7 (2016).
https://doi.org/10.1038/sdata.2016.86.

[59] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *Journal of Physical Chemistry Letters.* **2**, pp 2241–2251 (2011).
https://doi.org/10.1021/jz200866s.

[60] Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated Computational Discovery of High-

Performance Materials for Organic Photovoltaics by Means of Cheminformatics. *Energy and Environmental Science.* **4**, pp 4849–4861 (2011). https://doi.org/10.1039/c1ee02056k.

[61] Moore, G. J.; Bardagot, O.; Banerji, N. Deep Transfer Learning: A Fast and Accurate Tool to Predict the Energy Levels of Donor Molecules for Organic Photovoltaics. *Advanced Theory Simulation.* **5**, p.2100511 (2022). https://doi.org/10.1002/adts.202100511.

[62] Lee, M. H. Performance and Matching Band Structure Analysis of Tandem Organic Solar Cells Using Machine Learning Approaches. *Energy Technology.* **8**, p.1900974 (2020). https://doi.org/10.1002/ente.201900974.

[63] Lee, M.-H. A Machine Learning–Based Design Rule for Improved Open-Circuit Voltage in Ternary Organic Solar Cells. *Advanced Intelligent Systems.* **2**, p. 1900108 (2020). https://doi.org/10.1002/aisy.201900108.

[64] Gross, E. K. U.; Kohn, W. Time-Dependent Density-Functional Theory. *Advances in quantum chemistry.* **21**, pp 255–291 (1990). https://doi.org/10.1016/S0065-3276(08)60600-0.

[65] Kranthiraja, K.; Saeki, A. Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells. *Advanced Functional Materials.* **31**, p. 2011168 (2021). https://doi.org/10.1002/adfm.202011168.

[66] Eibeck, A.; Nurkowski, D.; Menon, A.; Bai, J.; Wu, J.; Zhou, L.; Mosbach, S.; Akroyd, J.; Kraft, M. Predicting Power Conversion Efficiency of Organic Photovoltaics: Models and Data Analysis. *ACS Omega.* **6**, pp 23764–23775 (2021). https://doi.org/10.1021/acsomega.1c02156.

[67] Fonti, V.; Belitser, E. Feature Selection Using LASSO. *VU Amsterdam research paper in business analytics.* **30**, pp.1-25. (2017).

[68] Basak, D.; Pal, S.; Patranabis, D. C. Support Vector Regression. *Neural Information Processing-Letters and Reviews.* **11**, pp 203-224 (2007).

[69] Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of chemical information and computer sciences.* **43** , pp 1947–1958 (2003). https://doi.org/10.1021/ci034160g.

[70] Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *Journal of Physical Chemistry Letters.* **9** , 2639–2646 (2018). https://doi.org/10.1021/acs.jpclett.8b00635.

[71] Natekin, A.; Knoll, A. Gradient Boosting Machines, a Tutorial. Front Neurorobot. *Frontiers* **7**, p. 21 (2013). https://doi.org/10.3389/fnbot.2013.00021.

[72] Wiens, M.; Verone-Boyle, A.; Henscheid, N.; Podichetty, J. T.; Burton, J. A Tutorial and Use Case Example of the EXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clinical and Translational Science.* **18**, p e70172 (2025). https://doi.org/10.1111/cts.70172.

[73] Wang, R. AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. *Physics Procedia.* **25**, pp 800–807 (2012). https://doi.org/10.1016/j.phpro.2012.03.160.

[74] Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN computer science.* **2** , p. 160 (2021). https://doi.org/10.1007/s42979-021-00592-x.

[75] Mahmood, A.; Wang, J. L. Machine Learning for High Performance Organic Solar Cells: Current Scenario and Future Prospects. *Energy and Environmental Science.* **14**, pp 90–105 (2021). https://doi.org/10.1039/d0ee02838j.

[76] Elallid, B. Ben; Benamar, N.; Hafid, A. S.; Rachidi, T.; Mrani, N. A Comprehensive Survey on the Application of Deep and Reinforcement Learning Approaches in Autonomous Driving. *Journal of King Saud University - Computer and Information Sciences.* **34**, pp 7366–7390 (2022). https://doi.org/10.1016/j.jksuci.2022.03.013.

[77] Shyalika, C.; Silva, T.; Karunananda, A. Reinforcement Learning in Dynamic Task Scheduling: A Review. *SN Computer Science.* **1**, p. 306 (2020). https://doi.org/10.1007/s42979-020-00326-5.

[78] Liu, X.; Zhang, X.; Sheng, Y.; Zhang, Z.; Xiong, P.; Ju, X.; Zhu, J.; Ye, C. Advancing Organic Photovoltaic Materials by Machine Learning-Driven Design with Polymer-Unit Fingerprints. *NPJ Computational Materials.* **11**, p. 107 (2025). https://doi.org/10.1038/s41524-025-01608-3.

[79] Lunnon, W. F.; Brunvoll, J.; Cyvin, S. J.; Cyvin, B. N.; Balaban, A. T. Topological Properties of Benzenoid Systems-The Boundary Code. Springer. **28**. pp 18-24., (1988).

[80] Quirós, M.; Gražulis, S.; Girdzijauskaitė, S.; Merkys, A.; Vaitkus, A. Using SMILES Strings for the Description of Chemical Connectivity in the Crystallography Open Database. *Journal of cheminformatics.* **10**, p. 23 (2018). https://doi.org/10.1186/s13321-018-0279-6.

[81] Lovrić, M.; Molero, J. M.; Kern, R. PySpark and RDKit: Moving towards Big Data in Cheminformatics. *Molecular informatics.* **38**, p.1800082 (2019). https://doi.org/10.1002/minf.201800082.

[82] Hanchuan Peng; Fuhui Long; Ding, C. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on pattern analysis and machine intelligence.* **27**, pp 1226–1238 (2005). https://doi.org/10.1109/TPAMI.2005.159.

[83] Shekar, B. H.; Dagnew, G. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. pp 1–8 (2019). https://doi.org/10.1109/ICACCP.2019.8882943.

[84] Kanagawa, M. Fast Computation of Leave-One-Out Cross-Validation for k-NN Regression. (2024). *arXiv preprint arXiv:2405.04919*.

[85] Cheng, J.; Dekkers, J. C. M.; Fernando, R. L. Cross-Validation of Best Linear Unbiased Predictions of Breeding Values Using an Efficient Leave-One-out Strategy. *Journal of Animal Breeding and Genetics.* **138** , pp 519–527 (2021). https://doi.org/10.1111/jbg.12545.

[86] Wong, T.-T. Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation. *Pattern Recognittion.* **48**, pp 2839–2846 (2015). https://doi.org/10.1016/j.patcog.2015.03.009.

[87] Lumumba, V. W.; Kiprotich, D.; Lemasulani Mpaine, M.; Grace Makena, N.; Daniel Kavita, M. Comparative Analysis of Cross-Validation Techniques: LOOCV, K-Folds Cross-Validation, and Repeated K-Folds Cross-Validation in Machine Learning Models. *SSRN Electronic Journal* (2025). https://doi.org/10.2139/ssrn.5266507.

[88] Zhang, Z.; Li, Y.; Cai, G.; Zhang, Y.; Lu, X.; Lin, Y. Selenium heterocyclic electron acceptor with small urbach energy for as-cast high-performance organic solar cells *J. Am. Chem. Soc.*, **142**, pp. 18741-18745 (2020). https://doi.org/10.1021/jacs.0c08557.

[89] Iglesias, D.; Sorrel, M. A.; Olmos, R. Cross-Validation and Predictive Metrics in Psychological Research: Do Not Leave out the Leave-One-Out. *Behavior Research Methods.* **57** ,p. 85 (2025). https://doi.org/10.3758/s13428-024-02588-w.

[90] Scharber, M. C.; Sariciftci, N. S. Efficiency of Bulk-Heterojunction Organic Solar Cells. *Progress in Polymer Science.* **38**, pp 1929–1940 (2013). https://doi.org/10.1016/j.progpolymsci.2013.05.001.

[91] Scharber, M. C.; Mühlbacher, D.; Koppe, M.; Denk, P.; Waldauf, C.; Heeger, A. J.; Brabec, C. J. Design Rules for Donors in Bulk-Heterojunction Solar Cells - Towards 10 % Energy-Conversion Efficiency. *Advanced Materials.* **18** ,pp 789–794 (2006). https://doi.org/10.1002/adma.200501717.

[92] Zhao, Z.; Geng, Y.; Troisi, A.; Ma, H. Performance Prediction and Experimental Optimization Assisted by Machine Learning for Organic Photovoltaics. *Advanced Intelligent Systems.* **4**, pp p.2100261 (2022). https://doi.org/10.1002/aisy.202100261.

[93] Padula, D.; Simpson, J. D.; Troisi, A. Combining Electronic and Structural Features in Machine Learning Models to Predict Organic Solar Cells Properties. *Materials Horizons.* **6** , pp 343–349 (2019). https://doi.org/10.1039/c8mh01135d.

[94] Sahu, H.; Ma, H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *The journal of physical chemistry letters.* **10**, pp. 7277–7284 (2019). https://doi.org/10.1021/acs.jpclett.9b02772.

[95] Lee, M. Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. *Advanced Energy Materials* . **9,** p. 1900891 (2019). https://doi.org/10.1002/aenm.201900891.

[96] Yuan, S.; Luo, W.; Xie, M.; Peng, H. Progress in Research on Organic Photovoltaic Acceptor Materials. *RSC Advances.* **15** , pp. 2470–2489 (2025). https://doi.org/10.1039/D4RA08370A.

[97] Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. *ISPRS Journal of Photogrammetry and Remote Sensing.* **173**, pp 24–49 (2021). https://doi.org/10.1016/j.isprsjprs.2020.12.010.

[98] Sahu, M.; Dash, R. A Survey on Deep Learning: Convolution Neural Network (Cnn). In Smart Innovation, Systems and Technologies; *Springer Science and Business Media Deutschland GmbH.* **153**, pp 317–325 (2021). https://doi.org/10.1007/978-981-15-6202-0_32.

[99] Clementi, E.; Raimondi, D. L.; Reinhardt, W. P. Atomic Screening Constants from SCF Functions. II. Atoms with 37 to 86 Electrons. *Journal of Chemical Physics.* **47**, pp. 1300–1307 (1967). https://doi.org/10.1063/1.1712084.

[100] Bondi, A. Van Der Waals Volumes and Radii. *Journal of Chemical Physics.* **68**, pp 441–451 (1964). https://doi.org/10.1021/j100785a001.

[101] Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. Consistent van Der Waals Radii for the Whole Main Group. *The Journal of Physical Chemistry A*. **113**, pp 5806–5812 (2009). https://doi.org/10.1021/jp8111556.

[102] Papadimitriou, C. Algorithms, Complexity, and the Sciences. *Proceedings of the National Academy of Sciences,* **111** , pp 15881–15887 (2014). https://doi.org/10.1073/pnas.1416954111.

Appendix

Appendix A : SHapley Additive exPlanations (SHAP)

To gain deeper insight into the relationship between molecular descriptors and photovoltaic parameters, we carried out a SHAP (SHapley Additive exPlanations) analysis (Figure **29**). As previously discussed, the inclusion of the frontier molecular orbitals of both donor and acceptor molecules among the top 20 most influential descriptors is expected. These energy levels are critical for photon absorption processes (through the bandgap of both A and D) and play a key role in the exciton dissociation driving force. The open-circuit voltage (Voc) is highly dependent on the FMOs, but it is also influenced by charge-carrier recombination a factor that is less easily captured through standard descriptors. Similarly, the short-circuit current density (Jsc) is affected not only by the absorption characteristics of the donor and acceptor (related to their bandgap), but also by the efficiency of exciton dissociation (tied to HOMO and LUMO offsets), and by the effectiveness of charge-carrier extraction, which is shaped by factors such as carrier mobility, morphology of the active layer, and domain purity. Although our study does not directly predict the fill factor (FF), this parameter reflecting the rectification behavior of the solar cell is also influenced by charge transport and extraction mechanisms. However, since our analysis is limited to the active layer, effects from the cell contacts are not considered.

The power conversion efficiency (PCE), calculated from Voc, Jsc, and FF, reflects a combination of multiple factors, including HOMO and LUMO energies of D and A, as well as processes such as charge generation, transport, and recombination. These latter mechanisms are closely linked to the active layer morphology encompassing aspects such as domain size and purity, polymer chain orientation, and the formation of percolation pathways. Connecting such morphological traits to simple molecular descriptors remains challenging. Nevertheless, some tentative correlations can be proposed. For instance, the number of aromatic rings in the donor molecule (don_NumAromaticRings, Figure **29**) may indicate its conjugation length. In contrast, the number of single bonds in the acceptor molecule (acc_NumSingleBondss, Figure **29**) could reflect a trade-off between molecular planarity (associated with fused ring systems) and structural flexibility, which helps mitigate excessive crystallization within acceptor domains. Additionally, the hydrophobic character of the donor and acceptor molecules (as captured by descriptors like acc_MolLogP and don_MolLogP in Figure **29**) appears to be a relevant factor. Other significant descriptors relate to van der Waals surface areas (VSA), including metrics linked to hydrophobicity (SlogP), polarizability (SMR), and electrotopological properties (Estate). These latter parameters are more complex to interpret, as they describe characteristics

of both the donor and acceptor molecules and their mutual interactions. They may influence the solubility and miscibility of the two materials, ultimately impacting the nanostructured morphology of the active layer.

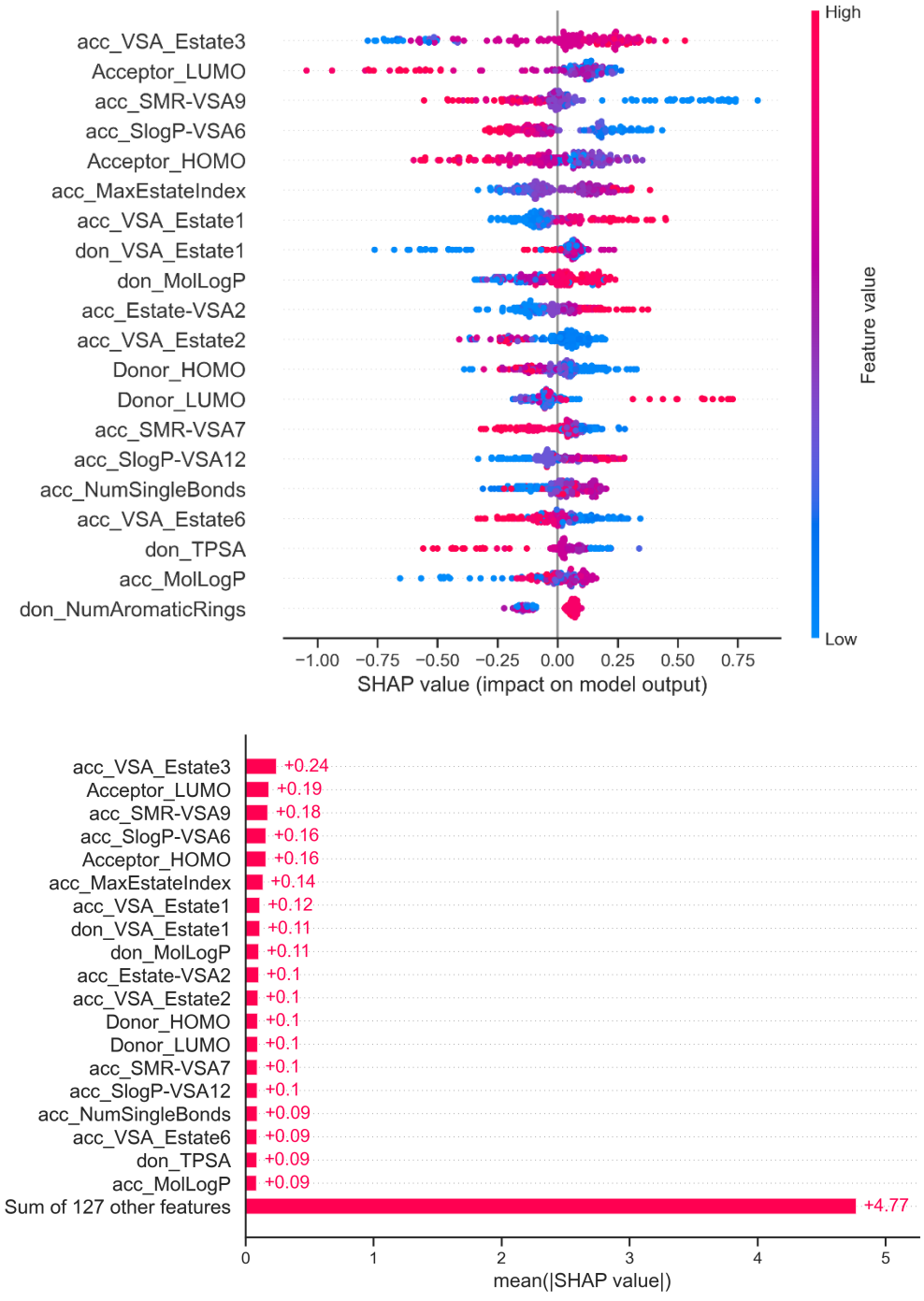


Figure 29. Example of SHAP summary and bar plots for PCE predictions.

*Table 10. Table of comparaison between our works and other works.(**15 In bold, K. Khoussa's work**;*: Not Published Yet)*

Ref	Dataset Size	Data Type	Input Type	Target(s)	ML Model(s)	Validation	Metrics	Results
[62]	70	Exp	FMO	*PCE*	RF, SVR	70/30 Split	R^2, RMSE	R^2 = 0.69, RMSE= 2.86%
[63]	121	Exp	FMO	V_{oc}	RF	80/20 Split	R^2	R^2 = 0.77
[65]	566	Exp	FMO, ECFP6	*PCE*	RF	K-Fold CV	r	r = 0.85
[60]	2.3M	Cal (HCEP)	CSD	*PCE*, V_{oc}, J_{sc}, *FF*	Linear Regression	Train/Test	R^2	R^2 = 0.84 (*PCE*), 0.94 (V_{oc}), 0.89 (J_{sc}), 0.61 (*FF*)
[49]	51,000	Cal + Exp	Fingerprints	*PCE*	GPR	Train/Test	r	r = 0.43
[66]	2.3M + HOPV15	Cal + Exp	CSD, Fingerprints, FMO	*PCE*	SVR, RF, HDMR	K-Fold + Shuffle	MSE, MAE, R^2, r	MSE = 3.18%, MAE = 1.41%, R^2 = 0.35, r = 0.623
[61]	2.3M + HOPV15	Cal + Exp	2D Images	FMO (Donor)	CNN	Train/Test	R^2	R^2 = 0.55 (HOMO), 0.63 (LUMO)
[12]	1242	Exp	CSD, FMO	*PCE*, V_{oc}, J_{sc}	SVR, RF, ANN, GBR	80/20 + LOOCV	RMSE, r	*PCE :* RMSE=2.00% r = 0.79 V_{oc}: RMSE =0.047V r = 0.86
[14]	503	Exp	CSD, FMO (TD-DFT)	*PCE* > 10%	RF	5-Fold CV	R^2, RMSE	R^2 = 0.28, RMSE = 1.60%
[43]	2.3M	Cal (HCEP)	2D Image (Donor)	*PCE*	ResNet	70/20/10 Split	Accuracy (Classif.)	91.02%
[15]	924	Exp	CSD, FMO (Exp)	*PCE*, V_{oc}, J_{sc}	SVR, RF, Adaboost, XGBoost, GBR	80/20 + LOOCV + New Data	RMSE, MAE, R^2, r	R^2 = 0.63 (*PCE*), 0.78 (V_{oc}), 0.70 (J_{sc})
*	707	Exp	2D Images (D/A Pairs)	*PCE*, V_{oc}, J_{sc}	CNN	80/20 Split	RMSE, MAE, R^2, r	R^2 = 0.54, RMSE = 2.53%, r = 0.73 (*PCE*)

Appendix B: List of publications

- **Journal article**
    - Khoukha Khoussa, Larbi Boubchir, Patrick Lévêque, "Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs", Engineering Reports, 2025. DOI: 10.1002/eng2.70334
- **Journal article under revision**
    - Khoukha Khoussa, Patrick Lévêque, Larbi Boubchir, "Deep Learning Approach for Predicting Efficiency in Organic Photovoltaics from 2D Molecular Images of D/A pairs", Advanced Theory and Simulations (under revision)
- **International conference article**
    - Khoukha Khoussa, Patrick Lévêque, Larbi Boubchir, "On the use of Machine Learning to Discover Novel Donor-Acceptor Pairs For Organic Photovoltaic Devices," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 4659-4662, doi: 10.1109/BigData62323.2024.10825948.

Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs

Khoukha Khoussa[1], Larbi Boubchir[2], Patrick Lévêque[1,]*

K. Khoussa, P. Lévêque
ICube Laboratory, University of Strasbourg-CNRS, 23 rue de Loess, 67037 Strasbourg, France
E-mail: Patrick.leveque@unistra.fr

L. Boubchir
LIASD Laboratory, University of Paris 8, 2 rue de la Liberté, 93526 Saint-Denis, France

Keywords: artificial intelligence, organic photovoltaics, artificial learning, feature engineering, molecular chemical structure descriptors, prediction

Organic solar cells (OSCs) can achieve power conversion efficiencies around 20%. Yet, further improvements in efficiency and long-term stability are necessary to rival the dominant silicon technology. Key factors influencing OSC performance include device architecture and the active-layer semiconducting organic materials. In this study, we utilize Artificial Intelligence (AI) techniques to analyze an experimental dataset of organic semiconductors used in the active-layer of OSCs. We propose an AI-based methodology to predict the performance of OSCs using the chemical structure of Donor-Acceptor (D/A) pairs. The method employs Simplified Molecular Input Line Entry System (SMILES) representations to extract molecular features. These features, selected according to maximum relevance and minimum redundancy criteria, are used by supervised machine learning regression algorithms to predict the main photovoltaic parameters. Our AI model demonstrate significant predictive power. Further, we use our model to predict the photovoltaic parameters of (D/A) pairs that were not included in our initial dataset. These findings highlight the potential of AI-driven analysis to accurately estimate the photovoltaic potential of new (D/A) pairs before synthesizing them and therefore to accelerate the development of commercially viable OPV devices and to lower the materials research cost.

1. Introduction

Renewable energy (RE) sources are on a steady rise, offering a sustainable and environmentally friendly alternative to conventional energy (CE) sources. Embracing these alternatives brings numerous advantages for both the environment and the society. Firstly, RE aids in combating climate change by slashing greenhouse gas emissions, thus mitigating the adverse effects of global warming like extreme weather events and rising sea levels.[1] Moreover, sources such as solar and wind power are abundant and never-ending, unlike finite CE sources, ensuring long-term energy security and stability. Additionally, investing in renewables drives economic growth and generates employment opportunities across various sectors like manufacturing, installation and maintenance of renewable energy infrastructure. Furthermore, utilizing RE reduces reliance on imported CE, bolstering energy independence and resilience to supply disruptions. Overall, transitioning to renewable energy is imperative for constructing a sustainable future, fostering environmental stewardship, and ensuring fair access to clean and affordable energy for all.[2,3]

Solar cells, also known as photovoltaic cells, are integral to this renewable energy process. They convert sunlight into electricity through the photovoltaic effect.[4] Their types vary, including traditional silicon-based inorganic cells and newer organic cells made of polymers or small molecules.[1] Traditional single-junction Photovoltaics cells (PVs) predominantly utilize crystalline silicon, either in a single crystal or polycrystalline structure. Renowned for their efficiency and durability, they can however incur high production costs, especially during the silicon purification process that requires high temperatures. On the other hand, organic photovoltaics diverge in their production method, employing a thin layer of organic semiconductors deposited at low temperature onto a rigid or flexible substrate, which offers the significant advantage of adaptability to various shapes and sizes.[1,4,5] Organic photovoltaics currently exhibit lower power conversion efficiency and shorter lifespans, around 30 years estimated, compared to their traditional inorganic counterpart.[6] Nevertheless, they represent a promising leap towards sustainable energy solutions, offering a range of advantages over inorganic PVs. OPVs with an active layer consisting of semiconducting organic molecules (small molecules or polymers), boast unparalleled flexibility, being lightweight and potentially transparent. This flexibility opens numerous possibilities, enabling integration into various applications such as wearable electronics, portable solar chargers, and Building Integrated Photovoltaics (BIPV).[2,7] Moreover, their manufacturing process, utilizing mainly low-temperature solution processing techniques and printing, is not only cost-effective but also environmentally friendly, with significantly lower energy requirements compared to traditional

PVs. Furthermore, OPVs exhibit remarkable advancements in efficiency, with recent developments pushing *PCE* slightly below 20% efficiency.[1,7,8] While still slightly lower than traditional PVs, (26.1% certified *PCE* for non-concentrated single crystal silicon solar cells)[9] the *PCE* gap is rapidly narrowing, propelled by ongoing research and innovation in the field of organic electronics.

Recent trends in materials science involve integrating material properties with photovoltaic parameters through the utilization of artificial intelligence to develop robust quantitative structure-activity relationships (QSARs).[10] AI is now present in many fields of science, such as physics, chemistry, biology, healthcare, material science and many others. It is also increasingly used in all sectors of industry such as automotive, aerospace, semiconductors, energy etc...[11]. In materials science, particularly in OPV technology, AI serves as a powerful tool for generating new molecular structures, allowing researchers to efficiently explore extensive chemical spaces, and this capability not only accelerates the discovery and development of novel compounds with desirable properties but also saves significant time and reduces costs.[12] It can also develop predictive models to understand the relationship between the molecular structure of materials and their performance and many other uses AI to optimize the OPV development.[13-16] Therefore, collaborations between groups working in AI and materials science are essential to unlock the full potential of these technologies. However, many problems can be faced when applying AI to chemical data which is complex to explore.[17,18] For example, some AI techniques, in particular deep learning, may require large amounts of experimental data.[15,19,20]

The present paper introduces a novel methodology for predicting the photovoltaic parameters (V_{oc}, J_{sc}, *PCE*) of organic photovoltaic devices. We consider only bulk heterojunction (BHJ) organic solar cells with binary blends i.e. an electron-donor (D) and an electron-acceptor (A) molecules (see definition below). We also consider only non-fullerene acceptors (NFAs) in the active layer and not electron-acceptors that are fullerene derivatives. We utilize the D and A chemical structure descriptors in combination with experimental data. Unlike previous studies that often depend on calculated or simulated data, our approach is characterized by the exclusive use of authentic experimental data, including Frontier Molecular Orbitals, specifically the HOMO (Highest Occupied Molecular Orbital) and the LUMO (Lowest Unoccupied Molecular Orbital) to ensure the authenticity and applicability of our findings to practical scenarios. Moreover, unlike other works that frequently include duplicate (D/A) pairs due to their reliance on datasets collected from the literature, where slight variations in performance for the same (D/A) across different laboratories are often ignored, we strictly focused on using unique ones

to eliminate redundancy and to mitigate the risk of overfitting, ensuring that our predictive models remain robust and reliable. Additionally, our dataset encompasses a diverse range of high and low performing OPVs, in contrast to many existing studies that predominantly focus on high-performing devices. This diversity enhances our model ability to predict performance across a broader spectrum of OPVs, making it applicable to both low-performance and high-performance systems. We also incorporate in our work both small molecules and polymers to achieve a level of variety that is rarely addressed in previous research and which allows our methodology to accommodate a wider range of OPV material systems, improving its generalizability and predictive accuracy.

Deep pretreatment process is then used with several machine learning models and a careful tuning of their hyperparameters in order to predict the photovoltaic parameters of OPVs. The proposed methodology yields comparable results to other recent approaches from literature.[12,21] Furthermore, by employing a minimal number of features to train our models, we achieved accurate prediction results, demonstrating the efficiency of our methodology. Finally, five different experimental recent data, not included in the initial dataset of the present work, were used to highlight in a very conclusive way the predictive capability of our model.

2. Experimental Section

2.1. Organic Photovoltaic Parameters

Organic solar cells are composed of multiple layers with an organic material layer, also known as Active Layer (AL), positioned between two electrodes as shown in **Figure 1.** The device structure is most of the time more complex with charge-carrier selective layers between the AL and the electrodes.

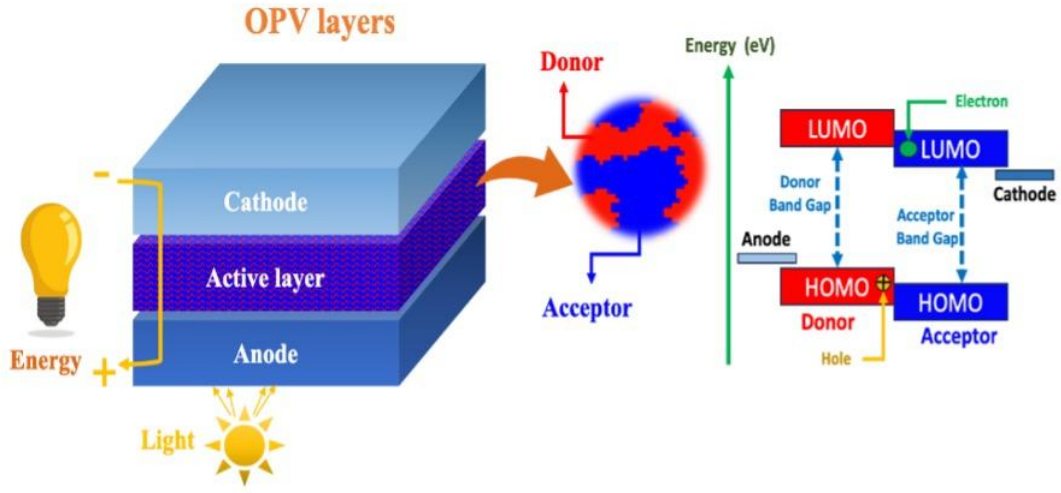


Figure 1. Organic photovoltaics device general structure.

The AL is the part of the OPVs where the process of converting sunlight into charge-carriers occurs. It typically consists of a blend of semiconducting organic materials (an electron-donor (D) one and an electron-acceptor (A) one: Figure 1) that can absorb photons. If a photon is absorbed by the electron-donor in the AL (i.e. if the photon energy is higher that the D band-gap), an exciton (a bond electron-hole pair) is generated. This exciton can then diffuse to an interface between D and A where the electron can be transferred from the LUMO of D to the LUMO of the A while the hole will stay in the HOMO of D, leading to the generation of a Charge-carrier Transfer State (CTS). This CTS will potentially lead to the creation of free charge carriers that can move in the AL under the electric field due to the electrode work function difference.[22]

Several morphologies exist for the AL. The first efficient one was the bilayer structure with A and D layers on top of each other,[23] then Bulk Heterojunction was introduced later to maximize the D/A interface and favor exciton dissociation.[24] Each design aims to enhance device performance and recently, modified bilayer elaboration was shown to be quite efficient (*PCE*s over 19%).[25] In the present work, we focus on the BHJ structure, by far the most employed in the last years. We further analyze data measured on binary blends.

Evaluating the efficiency of OPVs include measuring different photovoltaic parameters under standard (*AM1.5G* (*100 mW/cm*^{2})) illumination conditions. The first important photovoltaic parameter is the short-circuit current (I_{sc}) which is the current under standard illumination at zero bias voltage. For sake of comparison between cells of different surface, the short-circuit current density (J_{sc}) which is I_{sc} divided by the surface of the cell is often used. The second parameter is the open-circuit voltage (V_{oc}) or the voltage for a zero current under illumination. The third one is the Fill Factor (*FF*) which is the ratio between the maximum electrical power by surface unit and the product of V_{oc} by J_{sc} and quantifies the rectifying behavior of the OPV solar-cell as mentioned in **Figure 2.**

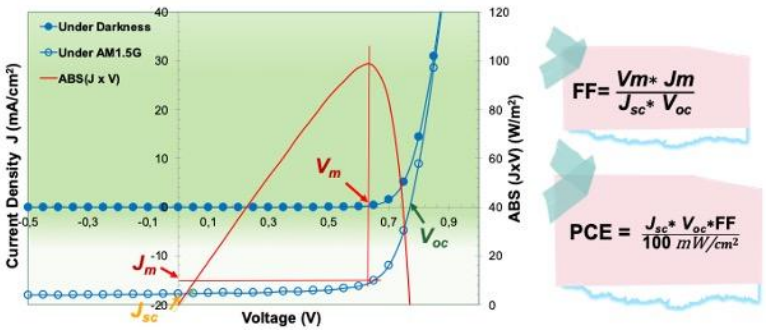


Figure 2. Current Density Vs Voltage Curve under darkness (closed circles) and under standard illumination (open circles) and electrical power under illumination (red curve).

Finally, the power conversion efficiency (*PCE*) is the electrical power (by surface unit) divided by the light power (by surface unit). Under standard illumination conditions (AM1.5G (100 mW/cm^2)), *PCE* (in %) is simply equal to the product of J_{sc} (in mA/cm^2) by V_{oc} (in V) by *FF* (no unit).[24] Together, these metrics provide valuable insights into the performance of OPVs. In the present paper, we focused on the prediction of V_{oc}, J_{sc}, *PCE* and the fourth photovoltaic parameter (*FF*) was calculated.

2.2. Related Work

In recent years, the prediction of organic photovoltaics performances has garnered significant attention due to its potential to revolutionize renewable energy technologies.[16] Semi-empirical and theoretical models have traditionally served as the cornerstone for estimating FMOs and ultimately OPV performances. Theoretical models, such as those based on density functional theory (DFT) or other quantum mechanical approaches, allow calculating the best molecular conformation and the corresponding FMOs in vacuum or in solution in different solvents. Theoretical models are computationally intensive and require domain-specific expertise.[26,27]. Semi-empirical models rely on FMOs of the donor material (D) and the acceptor one (A) and on some experimental hypothesis to quickly estimate, using predictive equations, the photovoltaic parameters of BHJ solar cells.[28] Recently, machine learning (ML) has emerged as a powerful tool to complement traditional approaches.[11,12,21,29,30] Leveraging experimental datasets and advanced algorithms, machine learning can uncover complex patterns and relationships, enhancing predictive accuracy and efficiency in estimating OPV performance.

However, OPVs research poses a significant challenge for ML applications due to the limited experimental data availability. This strong limitation restricts the robustness and generalizability of ML models, hindering their effectiveness in accurately predicting material properties, device performance, and optimizing manufacturing processes.[31] Further, the inherent complexity and variability of organic materials exacerbate this issue, as ML requires large and diverse datasets for comprehensive model training. Consequently, researchers face obstacles in achieving reliable predictions and identifying subtle trends within the data. Addressing this challenge requires innovative approaches such as using transfer learning (TL) and simulated or calculated datasets as the Harvard Clean Energy Project (HCEP) dataset,[32] which is the largest computational dataset, gathered by scientists at the Harvard University to help finding new materials for clean energy technologies, like solar panels or batteries. It contains data on 2.3 million of organic molecules and their properties. This huge dataset was

used by Olivares-Amaya who applied the Random Forest algorithm (RF) to predict solar cell parameters with high correlation scores: *PCE* (0.84), V_{oc} (0.94), J_{sc} (0.89), and *FF* (0.61) with the numbers between parenthesis being the R^2 score.[33] Gaussian Process Regressor (GPR)[34] and RF[35] algorithms were also used to predict *PCE* using the HCEP dataset. However, the prediction results on the validation set were in both case lower than the one using the HCEP, achieving a coefficient of correlation r = 0.43 (49 experimental data) and r = 0.62 (350 experimental data).

AI technics have also been used to predict the electronic properties of donor molecules (FMOs). G.J. Moore, O. Bardagot and N. Banerji[15] applied a Deep Transfer Learning on the HCEP and on the Harvard Organic photovoltaics dataset (HOPV15) which is a real experimental dataset collected from the literature.[36] They trained first their model on the HCPE dataset, and then, by using the TL, they retrained the model and obtained a HOMO prediction with a R^2 of 0.55 and a LUMO prediction with a R^2 of 0.63. Their study highlights the dataset significance for advanced predictive models for OPVs. Consequently, emphasizing the importance of experimental data for model training over calculated datasets like the HCEP one, we advocate for the utilization of experimental datasets. Grounded in empirical evidence, these datasets provide more accurate representations of the complexities inherent in real systems, thereby enhancing the reliability and applicability of our models in practical settings.[6,21] Several studies using experimental data to predict the performance of binary BHJ solar cells but also tandem solar cells (2 solar cells connected in series) and ternary solar cells (3 different organic semiconductors in the active layer) can be found in the literature. For instance, M. Lee applied random Forest using experimental FMOs of (D/A) pairs to predict *PCE*s using 70 experimental data for tandem OSCs.[37] The achieved prediction results showed a R^2 score of 0.69. Afterwards, in a more recent work,[38] the dataset was upgraded to 121 ternary dataset where 26 are polymer donors and 90 are a third component collected from literature and applied the same model to predict only V_{oc}. The generated model showed an R^2 value of 0.77, a good result considering that small size of the input dataset. Both studies exclusively focused on utilizing FMO of organic semiconductors as descriptors, while overlooking other molecular descriptors and disregarding the potential influence of thin film morphology. The random forest model was also trained on 566 pairs and demonstrated impressive predictive capabilities for power conversion efficiency.[30] However, it is important to highlight that this dataset did not include the Y-series acceptors,[12] which is a recent class of materials increasingly incorporated into many high-performing OSCs. Other searchers used the FMO properties calculated with Time-dependent density-functional theory (TD-DFT) and the molecular chemical descriptors like van

der Waals surface area descriptors of D and A to predict V_{oc}, J_{sc} and the *PCE* using two of the largest experimental datasets to date collected from the literature.[12,21] All the aforementioned results are presented in **Table 1**.

Table 1. Related works from literature using experimental (Exp) and calculated (Cal) data and our work.

References	Dataset	Exp or Cal (HCEP)	Input Type	Target	ML models	Validation	Validation Metrics	Results
[37]	70	Exp	FMO	*PCE*	RF and SVR	Train: 70% Test: 30%	R^2 RMSE	$R^2 = 0.69$ RMSE = 2.86 %
[38]	121	Exp	FMO	V_{oc}	RF	Train: 80% Test: 20%	R^2	$R^2 = 0.77$
[30]	566	Exp	FMO, Weights, fingerprint (ECFP6)	*PCE*	RF	K-Fold CV	r	r = 0.85
[33]	2.3 million	HCEP	CSD	*PCE* V_{oc} J_{sc} *FF*	Linear Regression	Train test split	R^2	$R^2 = 0.84$ (*PCE*) $R^2 = 0.94$ (V_{oc}) $R^2 = 0.89$ (J_{sc}) $R^2 = 0.61$ (*FF*)
[35]	51,000	HCEP +Exp	Fingerprints	*PCE*	GPR	Train test split	r	r = 0.43
[34]	2.3 million + HOPV15	HCEP, HOPV15	CSD and fingerprints +FMO	*PCE*	SVR, RF, HDMR	Train test Split shuffle +K-Fold CV	MSE MAE R^2 r	MSE = 3.18% MAE = 1.41% $R^2 = 0.35$ r = 0.623
[15]	2.3 million+ HOPV15	HCEP + Exp	2D chemical structure images	FMO of Donor	CNN	Train test split	R^2	$R^2 = 0.55$ (HOMO) $R^2 = 0.63$ (LUMO)
[21]	1242	Exp	CSD+FMO	*PCE* V_{oc} J_{sc}	SVR, RF, ANN, GBR	Train: 80% Test: 20% +LOOCV	RMSE /r	2.004 %/0.79 (*PCE*) 0.047 V/0.86 (V_{oc})
[12]	503	Exp	CSD +FMO (TDs-DFT)	*PCE* > 10	RF	5 Fold CV	R^2 RMSE	R^2 = 0.28 RMSE = 1.60%
Our Work	924	Exp	CSD+FMO (Exp)	V_{oc} J_{sc} *PCE*	SVR, RF, Adaboost, Xgboost, GBR	Train: 80% Test: 20% +LOOCV +New data	RMSE MAE R^2 r	GBR best model $R^2 = 0.78$ (V_{oc}) $R^2 = 0.70$ (J_{sc}) $R^2 = 0.63$ (*PCE*)

In order to improve the efficiency and the deep understanding of organic photovoltaics, there has been a convergence of experimental and computational approaches. In many studies, data from the HCEP is utilized to construct machine learning models to predict OPVs performance.

However, the molecules described in the OPV literature are often significantly more complex than those found in the HCEP dataset. Importantly, our work focuses on unique pairs of D and A molecules (we have excluded duplicates to ensure data integrity) using non-fullerene acceptor (NFA) based organic photovoltaic materials which have shown excellent performances in recent studies. Despite notable advancements in this domain, a comprehensive understanding of the intricate interplay between molecular structure, molecular electronic properties, molecular chemical descriptors and device performance remains a challenge. Our study endeavors to address this gap by focusing on the predictive potential of experimental data and molecular descriptors using different types of materials for both weak and high performance OPVs.

2.3. Dataset

2.3.1. Molecular representation

In the realm of molecular research, the representation of molecules is pivotal for leveraging the capabilities of AI. From traditional SMILES codes, which encode molecular structures as a text string (more details are provided in the Supporting Information and **Figure S2**), to more complex 2D images and graph-based representations, each approach offers unique insights into molecular structure and properties.[39] Additionally, advancements in three-dimensional conformations representation of materials have expanded the repertoire of tools available for molecular analysis and design. In this study, we focus on utilizing SMILES codes presented in **Figure 3** as our primary molecular representation as these codes can be easily used with cheminformatics software to calculate a wide range of molecular descriptors (Ex: molecular weights, number of bonds with their types, van der Waals surface area descriptors, etc.), and it can represent complex molecular structures in a compact form. This simplification helps in handling large datasets and reduces computational complexity.[39]

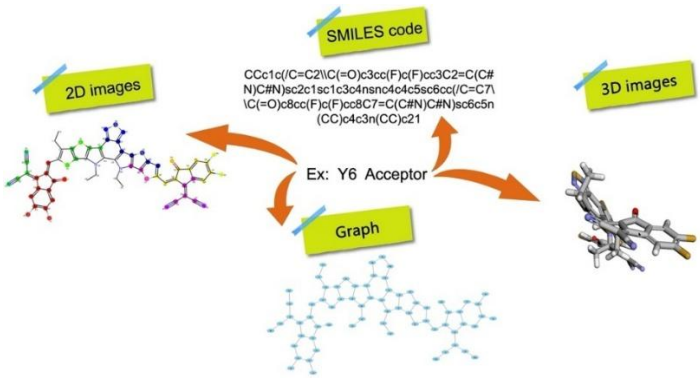


Figure 3. Different representations of the same molecule (Y6) to train AI models on the active-layer of organic solar cell.

2.3.2. Dataset details

1225 experimental donor/NFA devices have been collected from various literature sources.[12] This dataset includes crucial parameters such as V_{oc}, J_{sc}, *FF* and *PCE*, also when available, λ_{max} in solution or film, the optical bandgap, the ionization energy and the electron affinity. To simplify computational analysis, they standardized all side chains to an ethyl group, thus reducing computational complexity.

To ensure data uniqueness and enhance accuracy, the dataset underwent refinement to include only distinct (D/A) pairs. If several data were available for the same (D/A) pair like for instance for PM6/Y6 devices, the highest *PCE* was chosen. Indeed, most of the results published for each pair are already an average over a few cells, but the quantity of data for each (D/A) pair remains too small to make a selection based on a statistical study, as for example in the study by Dang *et al.* covering 579 articles published between 2002 and 2010 on P3HT/PCBM.[40] Dang *et al.* study shows the influence of the P3HT characteristics (M_n, M_w, polydispersity…) but also of the solvents, of the deposition method, of the post-elaboration annealing and of the solar-cell structure on the *PCE* for P3HT/PCBM bulk heterojunction solar-cells. The *PCE* was found to be, in average, between 3.5 and 4% with some publications reporting *PCE* values up to 6%. In the present work, we would have chosen a *PCE* of 6% for P3HT/PCBM The aim of this approach is to consider the maximum photovoltaic potential of each pair (D/A) based on the molecular structure and not on molecule batch-to-batch variations or non-optimized OPV devices elaboration or structure.

The refined dataset consists of 999 unique (D/A) pairs, comprising 605 distinct acceptors and 214 donors, each accompanied by their respective SMILES codes. This dataset encompasses both small molecules and polymers, adding diversity compared to other datasets used for similar purposes. Exclusion of all pairs which have no reported experimental FMO, resulted in a final dataset of 924 containing unique pairs of donor and acceptor with their corresponding SMILES code and electronic properties (FMO).

The dataset contains *PCE* (%) values between 0.01% and 18.77%, with the most frequent values between 6% and 14% as shown in the histogram represented in **Figure 4**.

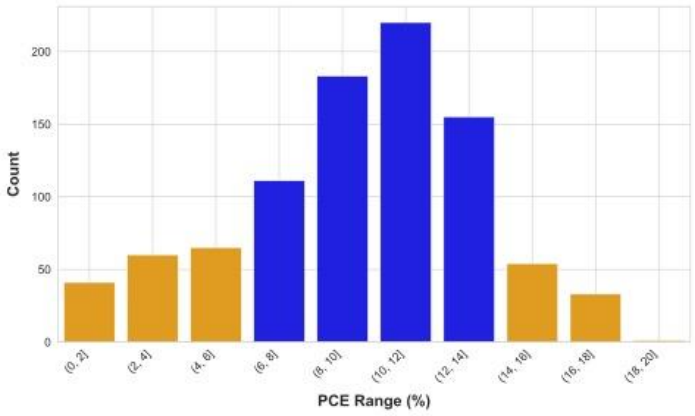


Figure 4. Distribution of *PCE* (%) values in the dataset.

We have also plotted the top 10 molecules used as donor and as acceptors. As shown in **Figure 5**, the most common donors are PBDB-T (also known as PCE12), PBDB-T-2F (or PBDT-T-F or PM6 in different works), PTB7-Th (or PCE10 or PBDTTT-EFT) and P3HT and the most common acceptors are ITIC which is the first non-fullerene acceptor to out-perform fullerene acceptors in organic photovoltaics, IT-4F (or ITIC-4F or ITIC-2F), Y6 and IDIC.

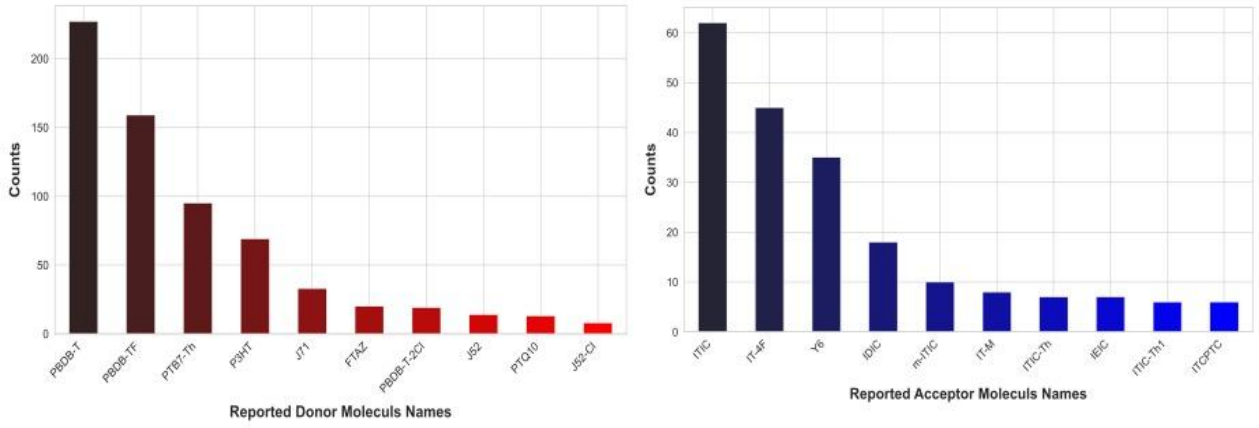


Figure 5. Top 10 common Donor and Acceptor distribution in our dataset.

2.4. Method

We present here a novel method for predicting the performance of OPV devices using the chemical structures of D/A pairs. The proposed method includes three primary phases: (2.4.1) data pre-processing, (2.4.2) feature extraction and selection, and (2.4.3) prediction process using ML algorithms. **Figure 6** depicts an illustrative diagram of the proposed method. The following sections describe all the steps constituting this method.

2.4.1. Data pre-processing

In the initial stages of any machine learning process, cleaning and preprocessing the dataset are crucial to ensure compatibility with our ML models. The first task involves removing duplicate entries and considering the commercial names of D and A molecules. Occasionally, we encounter identical molecules with distinct commercial names like PM6 which is the same molecule as the PBDB-T-F due to various laboratories independently publishing their works simultaneously and naming the same molecule differently. Consequently, it is essential to meticulously filter the dataset to retain only unique (D/A) pairs of molecules. It is also important to check the reported *PCE* values by recalculating them using the provided V_{oc}, J_{sc} and *FF*. This step is necessary as sometimes, the reported *PCE* values might be inflated or optimized under different experimental conditions, leading to inconsistencies. By recalculating the *PCE* we ensure that the reported values are accurate and consistent. This verification helps maintaining the integrity and reliability of the dataset for subsequent machine learning analysis.

2.4.2. Feature extraction

We have extracted more than 100 molecular descriptors using RDKit which is an open-source toolkit for cheminformatics that provides functionalities to handle chemical informatics tasks such as reading and writing molecules, calculating molecular descriptors and performing substructure searches and the SMILES code strings of donor and acceptor as input. For example, we have extracted the molecular weight of D and A, the number of single, double and triple bonds of D and A, the Estate-VSA (Electrotopological State Volumetric Surface Area) and the number of bicyclic rings etc. After the extraction step, we saved these extracted features in a dataset referred as "data_1". This dataset contains various RDKit descriptors with the reported performance *(PCE*, V_{oc}, J_{sc} and *FF*) and the SMILES code of each donor and acceptor pair. Subsequently, we added the reported experimental Frontier Molecular Orbitals along with the data_1 and created a new dataset, data_2.

2.4.3. Feature selection

In our study, we utilized the MRMR (Minimum Redundancy Maximum Relevance) feature selection method to identify the most pertinent chemical descriptors from the extensive dataset generated by RDKit. This approach was instrumental in enhancing the predictive accuracy of our machine learning models. MRMR operates by prioritizing features that exhibit strong correlations with the target variables while concurrently minimizing redundancy among selected descriptors.[41] By focusing on these informative yet non-redundant descriptors, we

ensured that our models were not only more accurate but also more interpretable, facilitating deeper insights into the underlying chemical factors influencing photovoltaic performance. This methodological choice not only optimized model efficiency but also underscored our commitment to rigorous feature selection, essential for robust predictive modeling in complex chemical systems. The results of the MRMR on the extracted features to predict V_{oc}, J_{sc} and *PCE* can be found in the Supporting Information. We can conclude that the MRMR helps us to use only nonredundant features and with only 20 features out of more than 100 we can achieve good predictive results as mentioned in **Figure S3**, **S4** and **S5** in the **Supporting Information**. A graph that represents the 20 descriptors showing their relevance and redundancy scores in predicting V_{oc} *(V)* is shown in **Figure S6**.

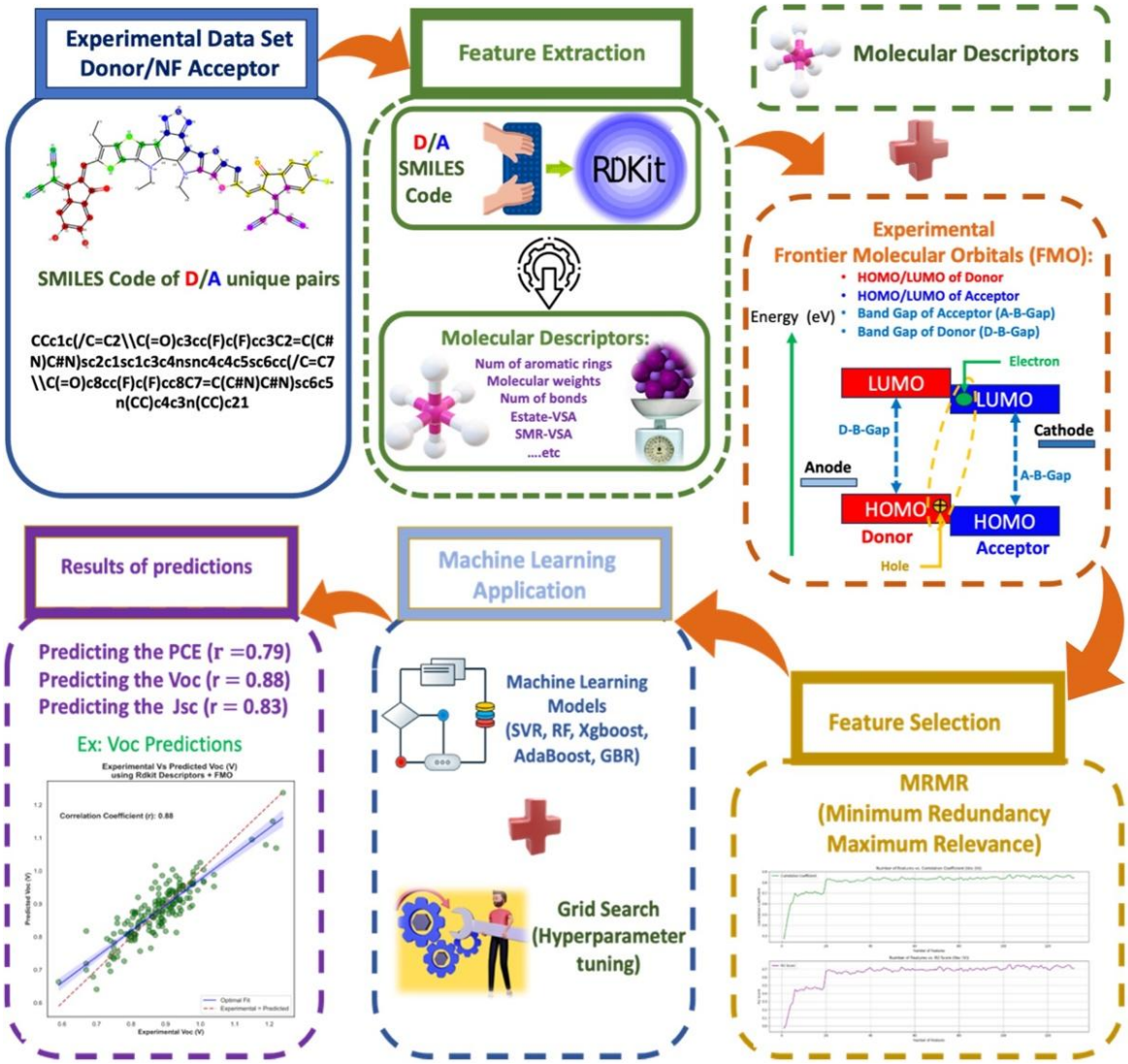


Figure 6. Schematic description of the proposed approach.

2.4.4. Prediction Process

The prediction process is realized by employing ML algorithms on data_1 and data_2**.** Indeed, we have applied five supervised ML regression models: SVR (Support Vector Regressor), RF (Random Forest), Adaboost Regressor (Adaptive Boosting), XGBoost Regressor (Extreme gradient boosting regressor) and GBR (Gradient Boosting Regressor) on the two distinct datasets data_1 and data_2, while carefully tuning the hyperparameter of each model, which are a set of adjustable parameters that significantly influence model behavior and predictive outcomes. Effective tuning of these hyperparameters through methods like Grid Search enables us to identify the optimal configuration that maximizes predictive accuracy and generalization capability. Also, it can enhance our model's robustness, mitigate overfitting to training data and improve the model's ability to capture complex relationships within the data.

The proposed methodology is highlighted in **Figure 6** as well as the **Algorithm S1 (Supporting Information)** that summarizes all the steps constituting it.

3. Results and discussion

3.1. Results obtained on the whole dataset

To assess the performance of our model, we have employed various evaluation metrics such as the coefficient of determination (R2 score or R^2), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and the correlation coefficient (r). The equations for these metrics can be found in the Supporting Information section.

Table 2. Prediction of OPV performance results by using different ML models applied to data_2.

Performance Parameters	***V_{oc}* (V)**				***J_{sc}* (mA/cm^2)**				***PCE* (%)**			
Model evaluation metrics	R^2	MAE (V)	RMSE (V)	r	R^2	MAE (mA/cm^2)	RMSE (mA/cm)	r	R^2	MAE (%)	RMSE (%)	r
SVR	0.49	0.051	0.068	0.72	0.55	2.72	3.87	0.76	0.54	1.99	2.60	0.75
RF	0.63	0.048	0.064	0.82	0.65	2.43	3.40	0.81	0.58	1.90	2.49	0.76
XGBoost	0.70	0.040	0.057	0.84	0.67	2.44	3.28	0.82	0.60	1.88	2.43	0.78
GBR	**0.99 (Train) and 0.78 (Test)**	**0.007 (Train) and 0.034 (Test)**	**0.01 (Train) and 0.045 (Test)**	**0.99 (Train) and 0.88 (Test)**	**0.93 (Train) and 0.70 (Test)**	**1.12 (Train) and 2.36 (Test)**	**1.59 (Train) and 3.19 (Test)**	**0.97 Train) and 0.83 (Test)**	**0.85 (Train) and 0.63 (Test)**	**1.12 (Train) and 1.84 (Test)**	**1.50 (Train) and 2.37 (Test)**	**0.92 (Train) and 0.79 (Test)**

As shown in **Table 2**, the best predictions are obtained using GBR and XGBoost. These two models are thought to manage overfitting efficiently compared to SVR and RF. Further, GBR outperforms XGBoost and this is due to the optimization of the loss function employed in GBR which is well adapted to the nature and the structure of the data used. For V_{oc} prediction with data_1, using 20% as a test set and 80% as a train test, we obtained notably good results. Specifically, the GBR model achieved a R^2 value of 0.68, a r of 0.83 and a RMSE of 0.052 V. These metrics indicate a good predictive capability, demonstrating the efficiency of RDKit descriptors in capturing the essential molecular features required for accurate V_{oc} predictions. However, when we expanded the dataset to include experimental FMO data alongside RDKit descriptors (data_2), the prediction performances improved. As mentioned in **Table 2**, the R^2 relative increase was almost 15% using experimental FMOs with a score of 0.78 for test and 0.99 for train, the RMSE became as low as 0.045 V for test and 0.01 V for train. The experimental versus predicted values of the V_{oc} are mentioned in **Figure 7 (a)**. The significant increase in the accuracy of V_{oc} prediction was expected as it is well known from empirical studies that V_{oc} is directly proportional to the FMOs of the donor and acceptor. Indeed, in the case of weakly absorbing fullerene acceptors, V_{oc} is proportional to the anergy difference between the HOMO of the donor and the LUMO of the acceptor.[28] Nevertheless, without FMOs (data_1), the RDKit descriptors alone are also robust returning reliable predictions for V_{oc}. This is crucial as the FMO incorporated in data_2 are experimental and therefore need the molecules to be synthesized and experimentally analyzed (using cyclic voltammetry for instance). We could also calculate FMOs in order to avoid molecules synthesis but calculating FMOs (for instance using DFT and TD-DFT) is also time consuming and leads often to values substantially different from the measured ones. Therefore, using our model to predict the photovoltaic performances of a given (D/A) pair without synthesizing it and consequently without the experimental FMOs may be more convincing when trained on data_1 instead of trained with data_2. Overall, the findings highlight that while incorporating FMO data can enhance model performance marginally, the use of RDKit descriptors alone is justified by their efficiency and adequacy in producing high-quality predictive models for V_{oc} for a given (D/A) pair. This approach not only simplifies the computational process but also ensures that the predictions remain robust and reliable. Additionally, to rigorously evaluate the model's performance, we also used the Leave-One-Out Cross-Validation (LOOCV). LOOCV is a technique where each data point is used as a validation set while the rest of the data acts as the training set in turn. This process iterates for each data point, ensuring that every sample serves once as a test set (**Figure S1**).

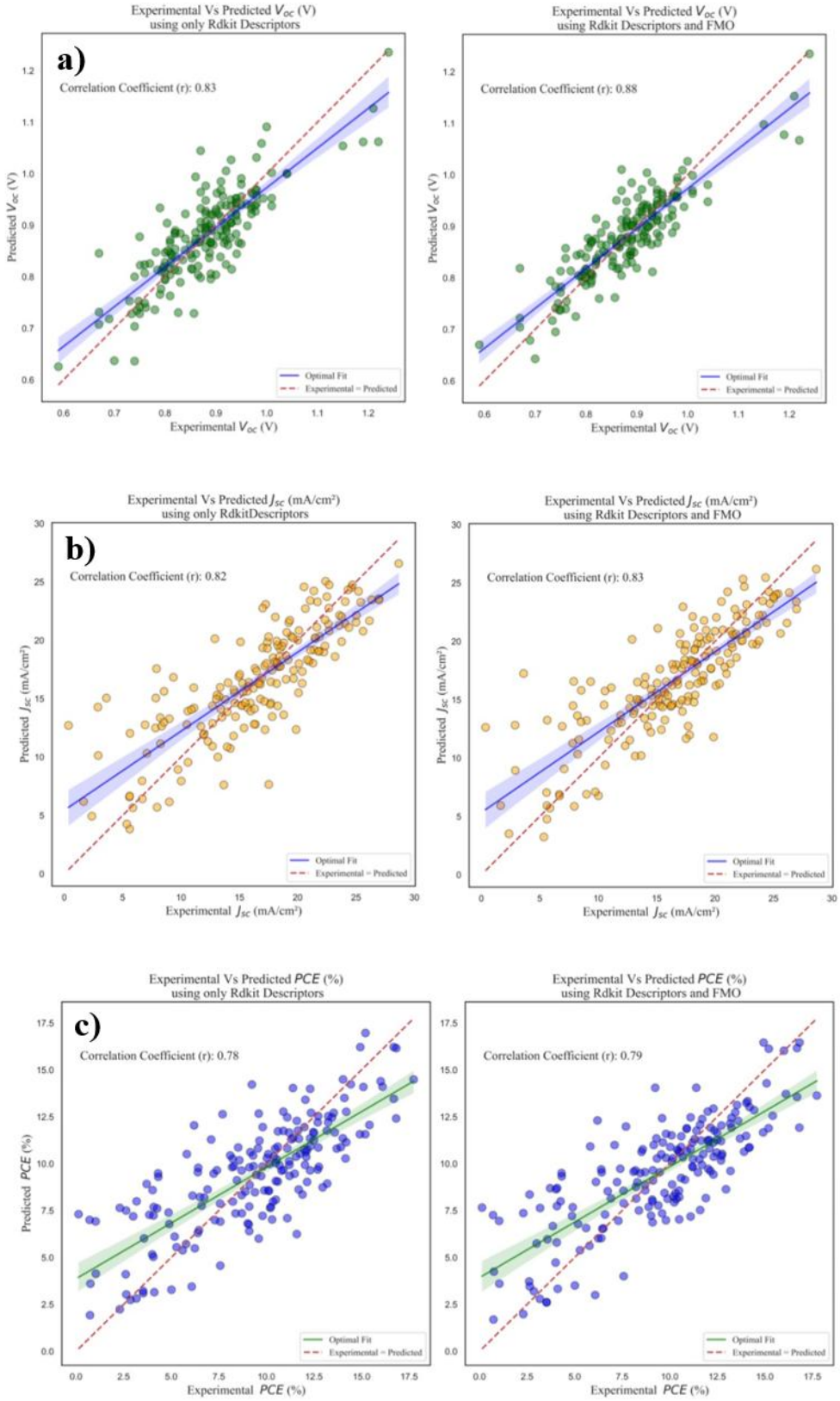


Figure 7. Experimental Vs predicted values of V_{oc} (a), J_{sc} (b) and *PCE* (c) using GBR model with the chemical descriptors of (D/A) pairs (left) and with the chemical descriptors and the FMOs of (D/A) pairs (right)

This method provides a robust assessment of the model's generalization capability, especially beneficial for small datasets. Applying LOOCV to our models trained on datasets featuring RDKit descriptors alone and RDKit descriptors combined with FMO, we observed consistent trends. The models exhibited strong predictive accuracy across most samples, as indicated by relatively low RMSE values (0.042 V and 0.044 V, respectively) for the V_{oc}. Also, by leveraging LOOCV, we can effectively mitigate the risk of overfitting, as the model is tested on each individual data point, ensuring that it generalizes well to unseen data.[42,43]. This method provides a reliable estimate of the model predictive capability and ensures that our results are not overly tailored to a specific subset of the data, thus enhancing the overall validity and reliability of our findings.

We employed the same method applied to V_{oc} with J_{sc}, while comparing both dataset predictions (with or without experimental FMO) achieving nearly comparable results, showing a R^2 value of 0.67, a r of 0.82 and a RMSE of 3.33 mA/cm² which indicate a good performance compared to the state of the art using only RDKit descriptors and a R^2 value of 0.70, a r of 0.83 and a RMSE of 3.19 mA/cm² when adding the FMO to the input data. This suggests that RDKit descriptors alone also allows an accurate prediction for the J_{sc} parameter. However, some samples, particularly those with J_{sc} values below 2 mA/cm², as shown in **Figure 7 (b)**, were poorly predicted. This is likely due to their limited distribution in the dataset or to the fact that low-performance solar cells are most of the time less optimized experimentally than the more promising ones. Nonetheless, these samples were included in our analysis to comprehensively explore diverse types of OPV devices, encompassing both high and low-performing variants. Furthermore, we implemented the same method to predict the *PCE* which showed comparable result to the state of the art with a R^2 value of 0.61, a r of 0.78 and a RMSE of 3.3% using only RDKit descriptors (data_1) and a R^2 value of 0.63, a r of 0.79 and a RMSE of 2.3% when adding the FMO features (data_2) as shown in **Figure 7 (c)**.

In order to better understand the links between the molecular descriptors and the photovoltaic parameters, we performed a SHAP (SHapley Additive exPlanations) analysis (**Figure S7**). As already explained, it is not surprising to find the FMOs of both A and D molecules in the 20 more important parameters for prediction as these frontier energy levels plays a major role in the photon absorption (energy band-gap of A and D) as well as in the exciton dissociation driving force. V_{oc} is strongly depending on the FMOs but also on the charge-carrier recombination that is much harder to anticipate from the main descriptors. The short-circuit current-density (J_{sc}) is also strongly depending on the A and D absorption properties (linked to

the energy band-gap) but also on the exciton-dissociation efficiency (depending on HOMO and LUMO offsets), on the free charge-carrier extraction (charge-carrier mobility, active layer morphology, domain purity…). The *FF*, which is not directly predicted in our work measures the rectifying behavior of the organic solar cell and depends also on the charge-carrier extraction (the cell contacts are not taken into account as the active layer only is considered in our work). Finally, the *PCE* is experimentally calculated from V_{oc}, J_{sc} and *FF* and, in our study, it is linked to the HOMO and LUMO of D and A as well as to free charge-carrier generation, transport and recombination. Free charge-carrier generation, transport and recombination are depending on the active layer morphology (domain size, domain purity, polymer orientation relative to the substrate, percolation paths…) and to relate the molecular descriptors to such material bulk properties is quite hazardous. Nevertheless, the D number of aromatic rings (don_NumAromaticRings in **Figure S7**) may be tentatively linked to the conjugation length of the electron-donor molecule while the A number of single bonds (acc_NumSingleBondss in **Figure S7**) could be a measure of the balance to be found between the A planarity (linked to the proportion of fused rings) and flexibility in order to avoid macro crystallization in the A domains. The hydrophobicity of both A and D molecules seems also to be an important property (acc_MolLogP and don_MolLogP in **Figure S7**). The other parameters are related to the sum of the van der Walls surface area for each atom in a molecule (VSA) and focus on the hydrophobicity (SlogP), the polarizability (SMR) and the electropological character (Estate). [44-46] These parameters are difficult to interpret as they are related to the D and A molecules and to their interaction. They may be connected to the solubility of the D and A molecule as well as to their mutual interactions and therefore on the morphology of the active layer at a nanoscopic scale.

3.2. Model Validation on only experimental PCE > 10%

To compare our best predicting model with the literature, we also used the GBR model on a smaller set of values including only the experimental *PCE* values higher than 10%. This corresponds to the work of Greenstein[12] (see **Table 1**) where they used a RF model with a 5-fold Cross Validation on the (D/A) pairs leading to a *PCE* > 10%. We used the very same validation on our GBR model and obtained a *PCE* prediction with a R^2 value of 0.46 (compared to 0.28) and a RMSE of 1.38% (compared to 1.6%). This direct comparison highlights the predictive capability of our GBR model.

3.3. Model Validation on unseen (D/A) pairs

To further evaluate our model, we selected several new experimental results on (D/A) pairs that were not included in our initial dataset. Our model yielded good results in terms of predicting V_{OC}, J_{SC} and *PCE*, as shown in **Table 3** and **Figure 8**. These findings demonstrate that our model performs well with both high and poor performance organic photovoltaics (OPVs) and can accurately identify them based on the chemical structure of the (D/A) pairs that constitute the active layer.

Table 3. Predicting capabilities of our model on unseen experimental (D/A) pairs comparing the predicted (Pred.) and Experimental (Exp.) values for the mains photovoltaic parameters.

Donor	Acceptor	Pred. V_{oc} (V)	Exp. V_{oc} (V)	Pred. J_{sc} (mA/cm^2)	Exp. J_{sc} (mA/cm^2)	Pred. *PCE* (%)	Exp. *PCE* (%)	References
D18	L8-BO	0.92	0.92	25.50	26.86	17.44	19.05	[47]
PBDB-TF	ITIC	1.00	1.04	15.52	16.00	9.63	9.70	[48]
D18	Y6	0.90	0.85	26.80	27.70	17.43	18.22	[49]
PBDB-TF	IT-4F	0.85	0.83	21.39	20.60	12.87	12.88	[50]
PBDB-TF	Y6-1O	0.89	0.90	22.69	25.51	16.06	16.81	[51]

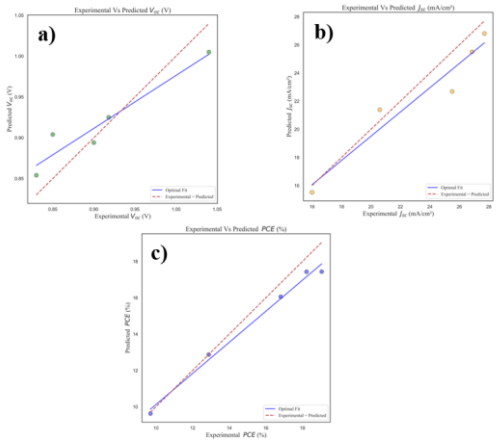


Figure 8. Experimental Vs predicted values of V_{oc} (a), J_{sc} (b) and *PCE* (c) using GBR model on unseen input data.

Five examples alone are not sufficient to fully establish generalizability. Nevertheless, their inclusion demonstrates the model's applicability to real-world systems beyond those it was trained on. This external validation complements our primary evaluation methods, which

include an 80/20 train/test split and Leave-One-Out Cross-Validation (LOOCV), both designed to rigorously assess predictive performance and minimize overfitting.

4. Conclusion

This study demonstrates the significant potential of artificial intelligence in accelerating the development of organic photovoltaics, particularly in the identification and optimization of donor/acceptor materials pairs of the active layer. By utilizing a Gradient Boosting Regression model in conjunction with the chemical structure descriptors and the Frontier Molecular Orbitals of these materials, we have predicted the key performance metrics (i.e., V_{oc}, J_{sc} and *PCE*). The RDKit library was employed to extract chemical structure descriptors from SMILES representations. These descriptors alone proved to be highly effective in predicting OPV performances when used as inputs to the GBR model. The addition of experimental FMO descriptors offered only a modest improvement. For instance, on the *PCE*, the R^2 value increases from 0.61 to 0.63 (relative increase of only 3%) when FMO descriptors are included. Therefore, while FMO provides additional useful information, the chemical structure descriptors are robust predictors on their own. To further validate our model, it was tested against real literature results that were not included in the initial dataset. This test confirms our model accuracy and reliability in predicting OPV performance metrics. These findings have significant implications. Firstly, they demonstrate the feasibility of using AI models to expedite the discovery and optimization of materials in the field of photovoltaics. Secondly, the methodology established here provides a solid framework for future research aimed at identifying and developing new D/A pairs, potentially leading to more efficient and cost-effective OPV materials. Overall, the integration of AI into the research and development process of OPVs holds great promise for achieving rapid advancements and innovations in renewable energy technologies. Future work should focus on expanding the dataset, refining the models, exploring additional descriptors and molecular features to further enhance predictive accuracy and material discovery efforts, finding potentially efficient (D/A) pairs and exploring the use of a third component in the active-layer of efficient OPV cells.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author. It includes a comparison between organic solar cells and silicon solar cells, the definition of the machine learning algorithm used in the study, an algorithmic representation of our method,

the definition of the evaluation metrics used and of the SMILES as well as the results obtained by minimum redundancy maximum relevance procedure.

Acknowledgements

K.K. would like to express her sincere gratitude to Dr. Olivier Bardagot for the insightful discussions and valuable input that greatly contributed to the development of this work.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author.

ORCID

Pr. Larbi Boubchir: 0000-0002-5668-6801

Dr. Patrick Lévêque: 0000-0002-6927-9025

Received: ((will be filled in by the editorial staff))

Revised: ((will be filled in by the editorial staff))

Published online: ((will be filled in by the editorial staff))

References

[1] E.K. Solak, E. Irmak. *RSC Adv.* **2023**, *13*, 12244-12269.
[2] G. Kumar, F.C. Chen. *J. Phys. D: Appl. Phys.* **2023**, *56*, 353001.
[3] M.R.S.A. Janjua, *J. Phys. Chem. Solids*, **2022**, *160*, 110352.
[4] M.S.A. Kamel, A. Al-jumaili, M. Oelgemöller, M.V. Jacob, *Renew. Sust. Energ. Rev.* **2022**, *166*, 112661.
[5] Y.-W. Su, S.-C. Lan, K.-H. Wei, *Mater. Today.* **2012** *15*, 554-562.
[6] a) Y. Li, X. Huang, K. Ding, H.K.M. Sheriff, L. Ye, H. Liu, C.Z. Li, H. Ade, S.R. Forrest, *Nat. Commun.* **2021**, *12*, 5419; b) Y. Li, T. Li, Y. Lin, *Mater. Chem. Front.* **2021**, *5*, 2907-2930.
[7] W.B. Tarique, A. Uddin. *Mater. Sci. Semicond. Process.* **2023**, *163*, 107541.
[8] C. Guo, Y. Sun, L. Wang, C. Liu, C. Chen, J. Cheng, W. Xia, Z. Gan, J. Zhou, Z. Chen, J. Zhou, D. Liu, J. Guo, W. Li, T. Wang, *Energy Environ. Sci.* **2024**, *17*, 2492-2499.
[9] Photovoltaic Research » Best Research-Cell Efficiency Chart https://www.nrel.gov/pv/cell-efficiency.html, accessed February, **2025**.
[10] E.N. Muratov, J. Bajorath, R.P. Sheridan, I.V. Tetko, D. Filimonov, V. Poroikov, T.I. Oprea, I.I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D.A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rev.* **2020**, *49*, 3525-3564.
[11] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, *Nanoscale Adv.* **2020**, *2*, 3115-3130.
[12] B.L. Greenstein, G.R. Hutchison, *J. Phys. Chem. C*, **2023**, *127*, 6179-6191.
[13] E. Abbasi Jannat Abadi, H. Sahu, S.M. Javadpour, M. Goharimanesh, *Mater. Today Energy*, **2022**, *25*, 100969.
[14] Y. Zhou, G. Long, A. Li, A. Gray-Weale, Y. Chen, T. Yan, *J. Mater. Chem C*, **2018**, *6*, 3276-3287.
[15] G.J. Moore, O. Bardagot, N. Banerji, *Adv. Theory Simul*. **2022**, *5*, 2100511.
[16] Z. Zhao, Y. Geng, A. Troisi, H. Ma, *Adv. Intell. Syst.* **2022**, *4*, 2100261.
[17] Y. Zhang, C. Ling. *Npj Comput. Mater.* **2018**, *4*, 25.
[18] A. Kirkey, E.J. Luber, B. Cao, B.C. Olsen, J.M. Buriak, *ACS Appl. Mater. Interfaces*, **2020**, *12*, 54596-54607.
[19] M. Jeong, J.F. Joung, J. Hwang, M. Han, C.W. Koh, D.H. Choi, S. Park, S. *Npj Comput. Mater.* **2022**, *8*, 147.
[20] H. Wang, J. Feng, Z. Dong, L. Jin, M. Li, J. Yuan, Y. Li, *Npj Comput. Mater.* **2023**, *9*, 200.
[21] R. Suthar, T. Abhijith, S. Karak, *J. Mater. Chem. A*, **2023**, *11*, 22248-22258.
[22] Y.-W. Su, S.-C., Lan, K.-H. Wei, *Mater. Today* **2012**, *15*, 554-562.
[23] C.W. Tang. *Appl. Phys. Lett.* **1986**, *13*, 183.
[24] N. Leclerc, P. Lévêque, *Techniques de l'ingénieur* **2023**, NM5205 v2.
[25] C. Xie, X. Zeng, C. Li, X. Sun, S. Liang, H. Huang, B. Deng, X. Wen, G. Zhang, P. You, C. Yang, Y. Han, S. Li, G. Lu, H. Hu, N. Li, Y. Chen, *Energy Environ. Sci.* **2024** *17*, 2441-2452.
[26] M. Usman Khan, F. Shafiq, M. Ramzan Saeed Ashraf Janjua, M. Khalid, J. Yaqoob, M. Arshad, S.M. Alshehri, R. Ahmad Khan, *J. Photochem. Photobiol.* **2024**, *446*, 115115; M.R.S.A. Janjua, W. Guan, L. Yan, Z.-M. Su, M. Ali, I. H. Bukhari, *J. Mol. Graph.* **2010**, *28*, 735-745; M.R.S.A. Janjua, *J. Iran. Chem. Soc.* **2017**, *14*, 2041-2054.
[27] C. Zanlorenzi, L. Akcelrud, *J. Polym. Sci. B: Polym. Phys.* **2017**, *55*, 919–927.
[28] a) M.C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A.J. Heeger, C.J. Brabec, *Adv. Mater.* **2006**, *18*, 789–794; b) M.C. Scharber, N.S. Sariciftci, *Prog. Polym. Sci.* **2013**, *38*, 1929–1940.

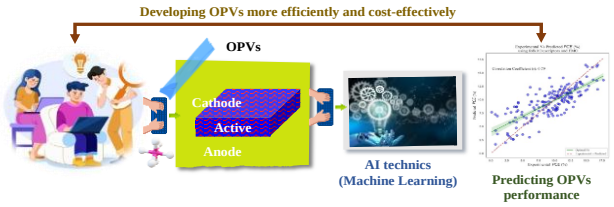
[29] E. Abbasi Jannat Abadi, H. Sahu, S.M. Javadpour, M. Goharimanesh, *Mater. Today Energy*, **2022**, *25*, 100969.
[30] K. Kranthiraja, A. Saeki. *Adv. Funct. Mater.* **2021**, *31*, 2011168.
[31] S. Bhatti, H.U. Manzoor, B. Michel, R.S. Bonilla, R. Abrams, A. Zoha, S. Hussain, R.S. Ghannam, *(Preprint) arxiv*, https://doi.org/10.48550/arXiv.2212.13893, submitted: December **2026**.
[32] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.
[33] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R.S. Sánchez-Carrera, L. Vogt, A. Aspuru-Guzik, *Energy Environ. Sci.* **2011**, *4*, 4849-4861.
[34] S.A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, A. Aspuru-Guzik, *Joule*, **2017**, *1*, 857-870.
[35] A. Eibeck, D. Nurkowski, A. Menon, J. Bai, J. Wu, L. Zhou, S. Mosbach, J. Akroyd, M. Kraft, *ACS Omega*, **2021**, *6*, 23764-23775.
[36] S.A. Lopez, E.O. Pyzer-Knapp, G.N. Simm, T. Lutzow, K. Li, L.R. Seress, J. Hachmann, A. Aspuru-Guzik, *Sci. Data*, **2016**, *3*, 160086.
[37] M.-H. Lee, *Energy Technol.* **2020**, *8*, 1900974.
[38] M.-H. Lee, *Adv. Intell. Syst.* **2020**, *2*, 1900108.
[39] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science*, **2018**, *361*, 360-365.
[40] M.T. Dang, L. Hirsch and G. Wantz, *Adv. Mater.* **2011**, *23*, 3597-3602.
[41] H. Peng, F. Long, C. Ding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2005**, *27*, 1226-1238.
[42] J. Cheng, J.C.M. Dekkers, R.L. Fernando, *J. Anim. Breed. Genet.* **2021**, *138*, 519–527.
[43] V.W. Lumumba, D. Kiprotich, M.L. Mpaine, N.G. Makena, M.D. Kavita, *AJTAS*, **2024**, *13*, 127-137.
[44] P. Labute, *J. Mol. Graph.* **2000**, *18*, 464-477.
[45] S.A. Wildman, G.M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.
[46] L.H. Hall, B. Mohney, L.B. Kier, *J Chem Inf Comput Sci.* **1991**, *31*, 76-82.
[47] Y. Wei, Z. Chen, G. Lu, N. Yu, C. Li, J. Gao, X. Gu, X. Hao, G. Lu, Z. Tang, J. Zhang, Z. Wei, X. Zhang, H. Huang, *Adv. Mater.* **2022**, *34*, 2204718.
[48] Y. Wang, Q. Fan, X. Guo, W. Li, B. Guo, W. Su, X. Ou, M. Zhang, *J. Mater. Chem. A*, **2017**, *5*, 22180-22185.
[49] Q. Liu, Y. Jiang, K. Jin, J. Qin, J. Xu, W. Li, J. Xiong, J. Liu, Z. Xiao, K. Sun, S. Yang, X. Zhang, L. Ding, *Sci. Bull.* **2020**, *65*, 272.
[50] T.R. Andersen, F. Zhao, Y. Li, M. Dickinson, H. Chen, *Solar RRL*, **2020**, *4*, 2000246.
[51] Y. Chen, H. Meng, L. Ding, J. Tang, J. Yi, J. Zhang, Z. Wang, R. Ma, Z. Li, L. Lyu, X. Xu, R. Li, Q. Peng, H. Yan, H.Y. Hu, *Chem. Mater.* **2022**, *34*, 10144.

TOC

This study shows the interest of using Artificial Intelligence to predict the photovoltaic performance of organic solar cells. The molecular structure of the semiconducting organic components of the active layer is used to estimate accurately the main photovoltaic parameters. The Artificial Intelligence model enables rapid evaluation of novel materials, reducing research costs while accelerating the development of organic photovoltaic technologies.

Khoukha Khoussa, Larbi Boubchir, Patrick Lévêque*

Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs



Supporting Information

Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs

Khoukha Khoussa, Larbi Boubchir, Patrick Lévêque

Contents

1. Organic Solar Cells *vs* Silicon Solar Cells

Table S1. Main differences between organic photovoltaic and traditional photovoltaics.

Photovoltaics Type / Features	Organic Photovoltaics (OPVs)	Traditional Photovoltaics (PVs)
Material	Organic molecules or polymers	Inorganic semiconductors (e.g., silicon)
Manufacturing Process	Solution processing, printing, roll-to-roll fabrication at low temperature (below 200°C)	High-temperature processing (more than 1500°C)
Flexibility	Lightweight, flexible, and potentially transparent	Rigid and heavy panels
Efficiency	Generally lower (up to ~ 20% in recent literature for small experimental cells)	Higher (commercially ~ 19-23%, lab > 24% for large surface modules)
Applications	Wearable electronics, portable solar chargers, BIPV	Large-scale solar farms, residential and commercial rooftops, photovoltaic power station
Cost of Production	Potentially lower due to simpler, low-temperature processes	Higher due to complex, energy-intensive processes
Stability and Longevity	Generally, less stable, ongoing improvements (estimated 30 years)	Highly stable, long operational life (25 + guaranteed years)
Environmental Impact	Lower energy input for production	Higher energy input, but long-term benefits
Market Maturity	Emerging technology, niche applications	Established technology, wide market acceptance

2. Machine Learning Algorithms

- **Gradient Boosting Regressor**: GBR is a machine learning ensemble technique where multiple weak predictive models (typically decision trees) are iteratively trained and combined, with each new model correcting errors made by its predecessor thereby minimizing the overall prediction error.
- **Support Vector Regressor**: SVR is a machine learning algorithm used for regression tasks that finds a function which best fits the data by maximizing the margin between predicted values and actual target values, often using non-linear transformations of input data.
- **Random Forest**: RF is a machine learning algorithm that builds multiple decision trees during training and merges them together to get more accurate and stable predictions for regression tasks.
- **Leave-one-out cross-validation**: LOOCV is a technique used to evaluate the performance of a machine learning model by training it on (N−1) samples and testing it on the one left-out sample, repeating this process N times; where N is the number of samples [S1,S2]. It provides a robust estimate of model performance, especially useful for small datasets where every data point is valuable. **Figure S1** describes schematically LOOCV.

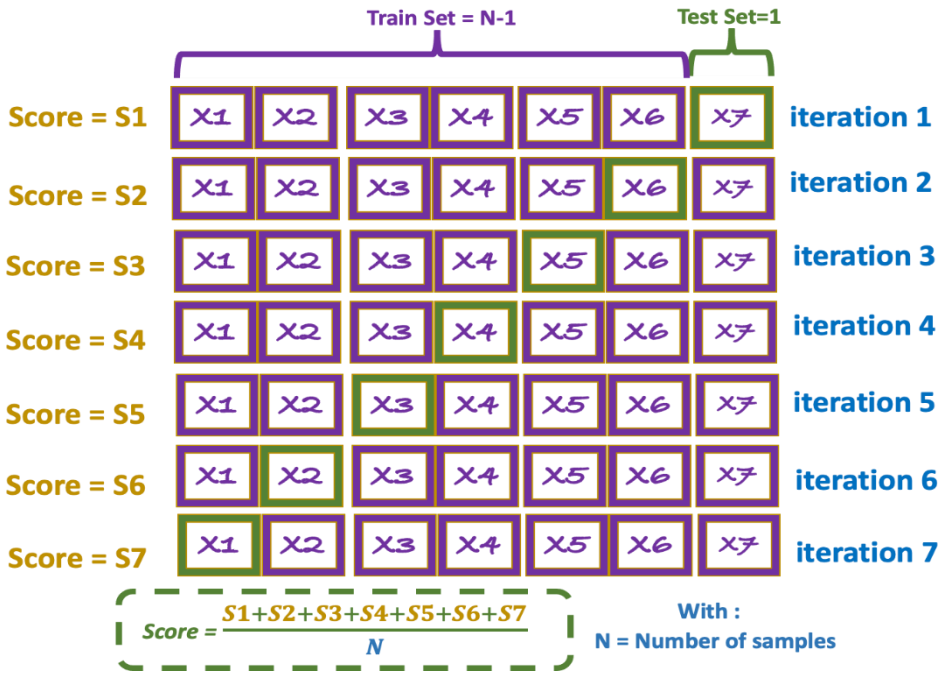


Figure S1. Leave-one-out cross-validation.

3. Algorithm representation of the proposed AI method

Algorithm S1 The proposed AI-based method for predicting photovoltaic parameters of D/A pairs.

Input: original_data

Output: Predicted V_{oc}, J_{sc} and *PCE*

Begin Algorithm

```
    # Step 1: Data Pretreatment
        Begin
        # Data Cleaning
            cleaned_data = remove_duplicates(original_data)
            pairs_SMILES = standardize_names(cleaned_data)
        # Feature Extraction using RDKit
            data_1 = extract_descriptors(pairs_SMILES)
            data_2 = add_FMO(data_1, FMO)
        End
        # Step 2: Model training
        Begin
        # Prepare datasets
            datasets = {'data_1': data_1, 'data_2': data_2}
            models={'SVR': SVR(), ' RF': RandomForestRegressor(),
                    'XGBR':XGBRegressor(), 'GBR': GradientBoostingRegressor(),
                    'Adaboost': AdaptiveBoostingRegressor()}
        # Hyperparameter Tuning and Training
            For each data_key in datasets:
                    dataset = datasets[data_key]
                    y= dataset[['Voc', 'Jsc', 'PCE']]
                    X = dataset.drop(columns=['Voc', 'Jsc', 'PCE'])
                    For each model_key in models:
                        model = models [model_key]
                        tuned_model=GridSearchCV(model, param_grid[model_key])
                        tuned_model.fit(X, y)
                        predictions = tuned_model.predict(X)
                        evaluate_performance_train_test_split(predictions, y)
```

evaluate_performance_LOOCV(predictions, y)

evaluation_results=[]

evaluation_results=evaluation_results.add(

evaluate_performance_train_test_split,

evaluate_performance_LOOCV)

End

 # Step 3: Final performance evaluation with the best model

Begin

For each data_key in datasets:

best_model = select_best_model(evaluation_results)

Predict Voc, Jsc, and PCE using the best model

final_predictions = best_model.predict(X)

evaluate_performance(final_predictions)

End

End Algorithm

---

4. Evaluation Metrics

To assess the performance of our model, we have employed various evaluation metrics such as the coefficient of determination (R2 or R^2 score), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and the correlation coefficient (r). The equations for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad \text{(S1)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad \text{(S2)}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \quad \text{(S3)}$$

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y}) \times (\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2} \times \sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \quad \text{(S4)}$$

With

- N Sample size
- y_i Experimental value of sample i
- $\hat{y}_i$ Predicted value of sample i
- $\bar{y}$ Mean of the Experimental values
- $\bar{\hat{y}}$ Mean of the Predicted values

When evaluating our model performances, lower RMSE and MAE values indicate better results, while a R^2 score value close to 1 reflects better model performance. The correlation coefficient (r) values for good model performances must be close to -1 or to +1.

5. Simplified Molecular Input Line Entry System (SMILES)

SMILES is a textual representation of chemical structures that encodes molecular information in a concise and human-readable format. It consists of a series of characters that describe the atoms, bonds, and connectivity within a molecule (**Figure S2**).

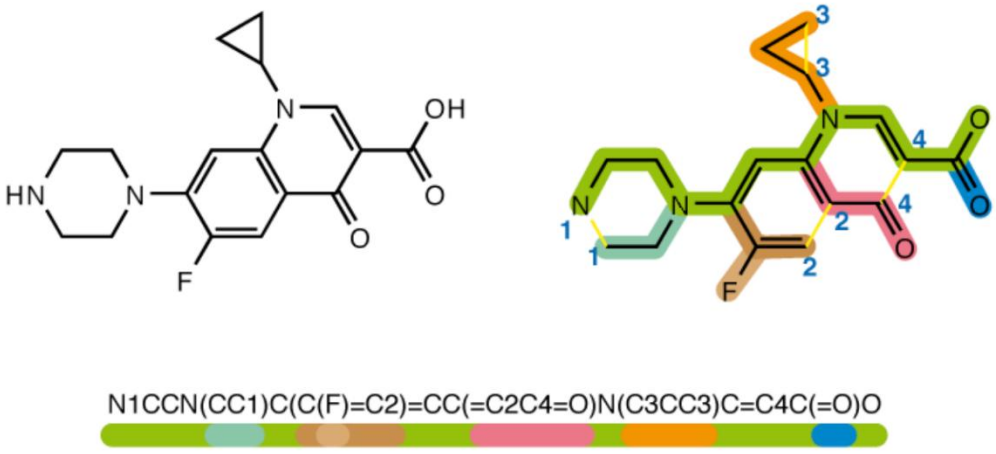


Figure S2. Chemical structure and SMILES specification of the chemical *ciprofloxacin* (from https://commons.wikimedia.org/wiki/File:SMILES.png).

6. Minimum Redundancy Maximum Relevance (MRMR) results

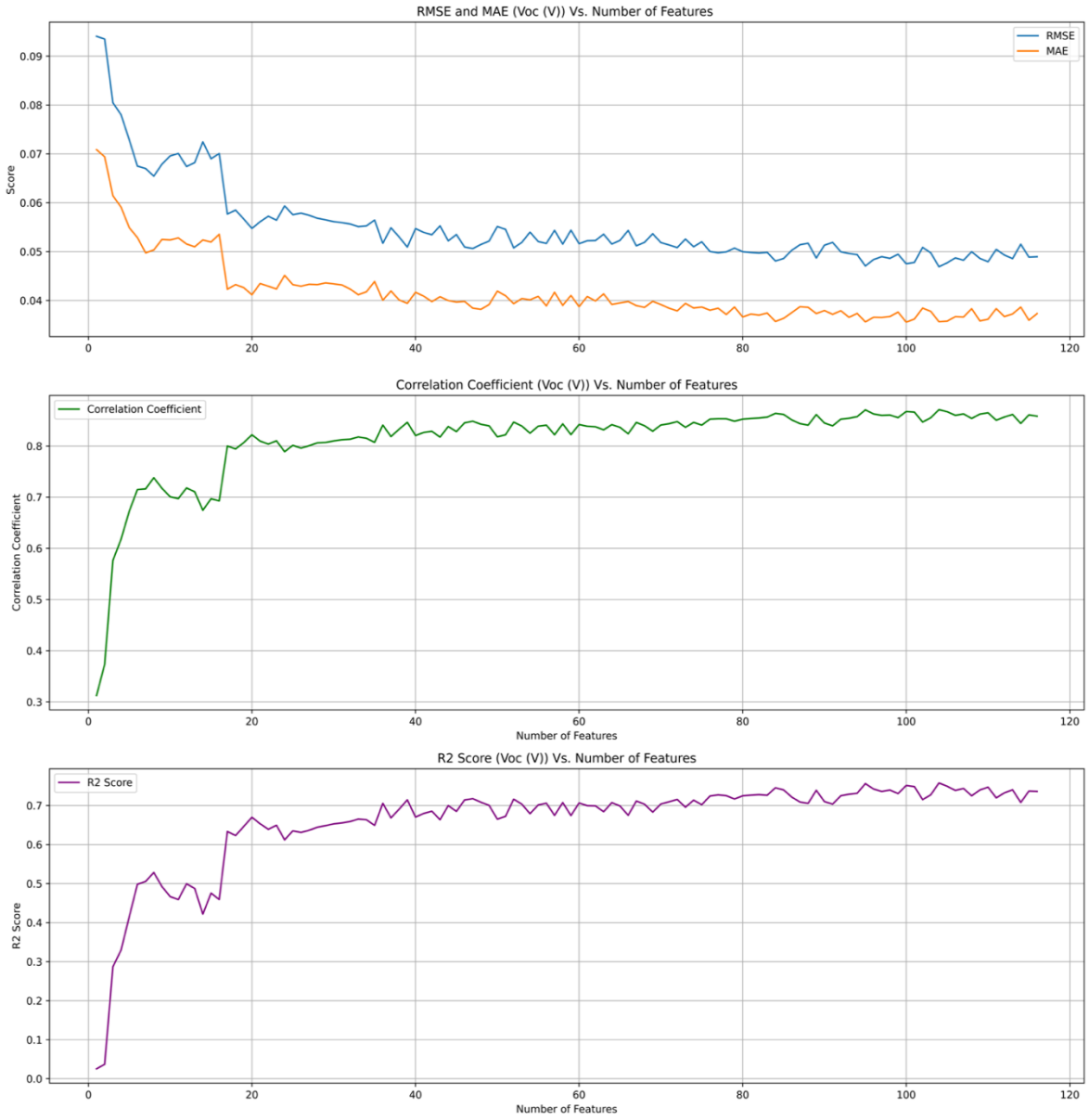


Figure S3. Results of MRMR application to select the important features to predict V_{oc}.

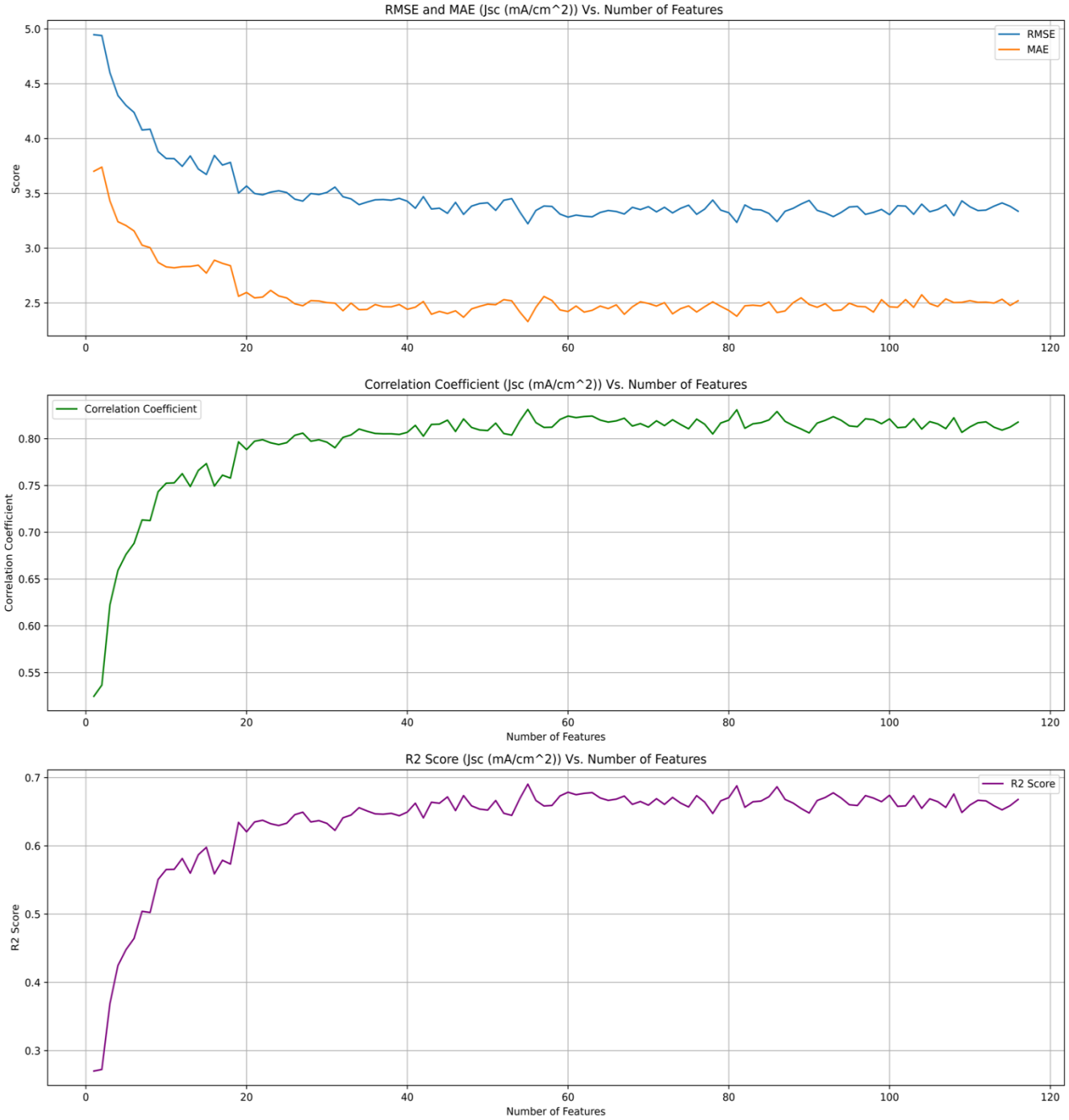


Figure S4. Results of MRMR application to select the important features to predict J_{sc}.

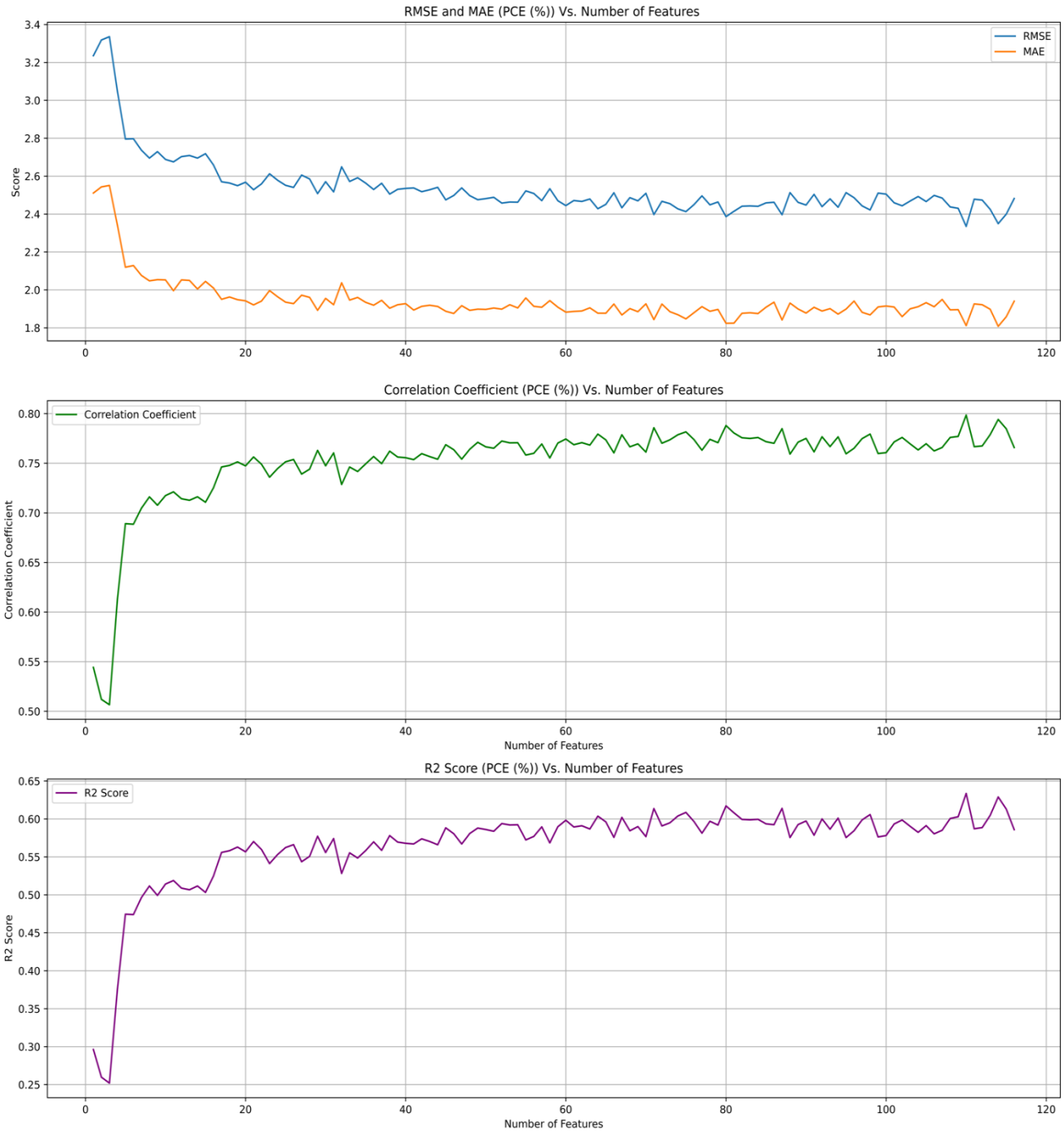


Figure S5. Results of MRMR application to select the important features to predict *PCE*.

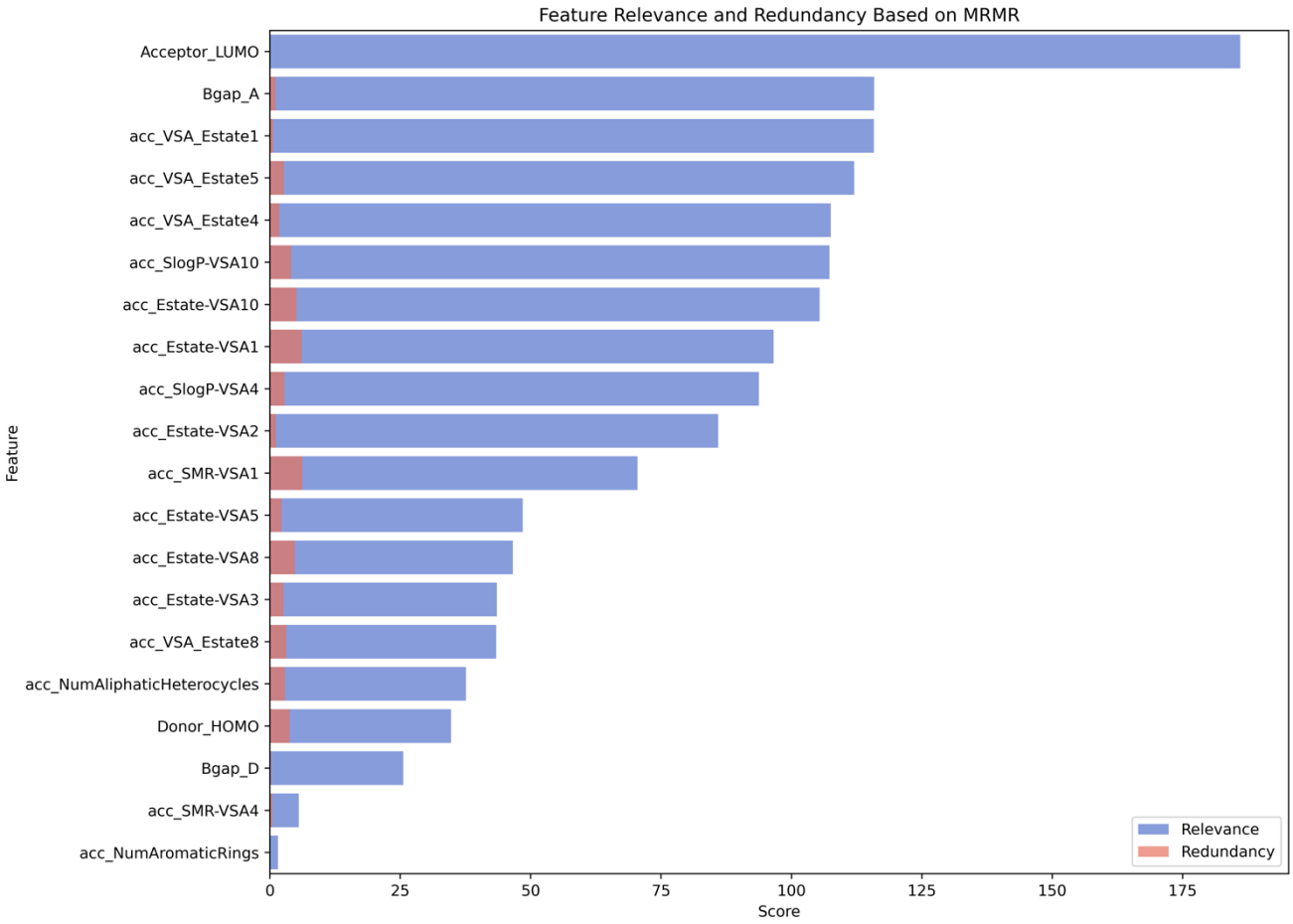


Figure S6. Results of features relevance and redundancy using the MRMR in case of predicting V_{oc} *(V)*.

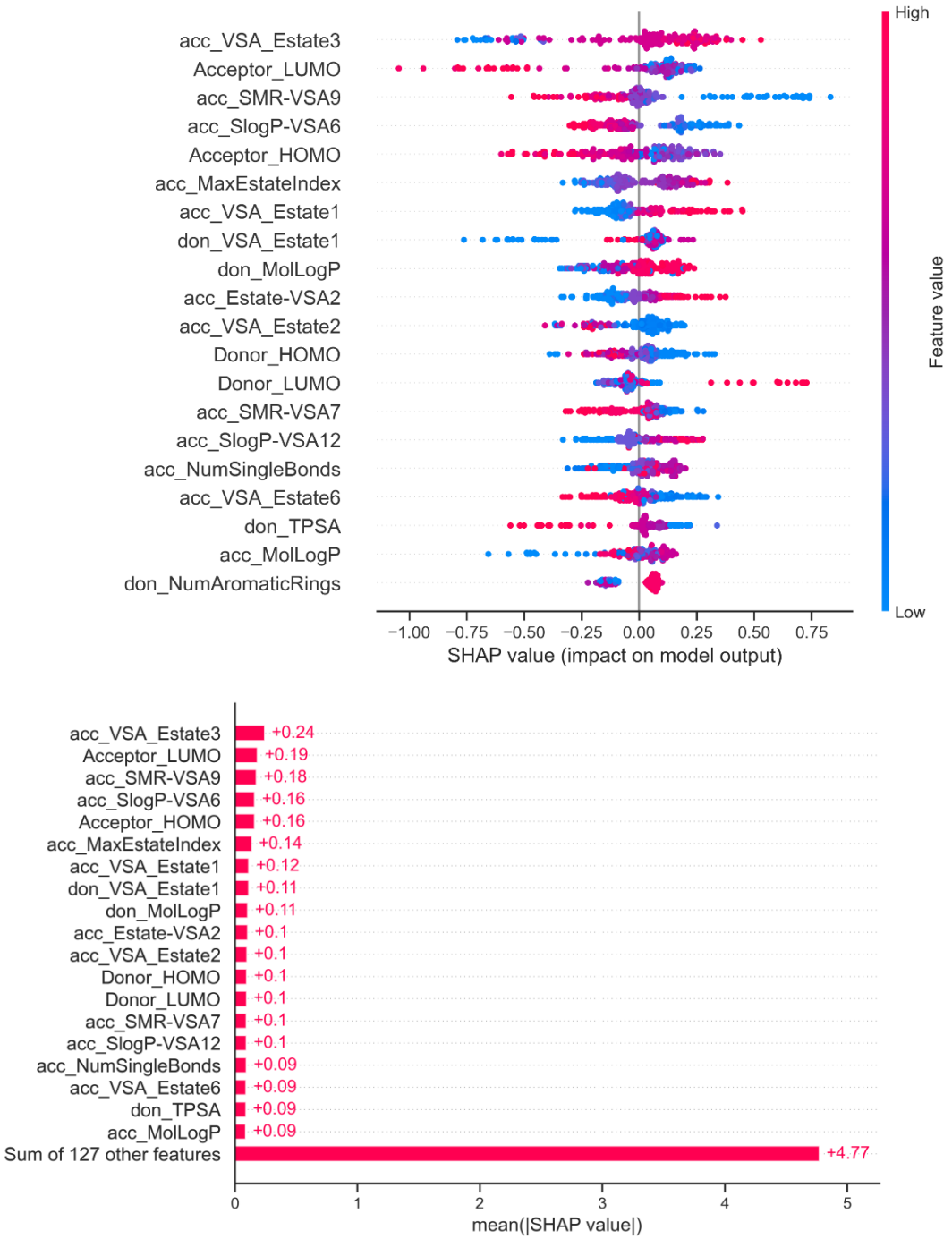


Figure S7. Example of SHAP summary and bar plots for PCE predictions.

7. References

[S1] J. Cheng, J.C.M. Dekkers, R.L. Fernando, *J. Anim. Breed. Genet.* **2021**, *138*, 519–527.
[S2] M. Kanagawa. (Preprint). 2024 May 8, Available from:http://arxiv.org/abs/2405.04919

On the use of Machine Learning to Discover Novel Donor-Acceptor Pairs For Organic Photovoltaic Devices

Khoukha KHOUSSA
ICube Laboratory
University of Strasbourg-CNRS
23 rue de Loess, 67037
Strasbourg, France
khoukha.khoussa@etu.unistra.fr

Patrick LEVEQUE
ICube Laboratory
University of Strasbourg-CNRS
23 rue de Loess, 67037
Strasbourg, France
patrick.leveque@unistra.fr

Larbi BOUBCHIR
LIASD Laboratory
University of Paris 8
2 rue de la Liberté, 93526
Saint-Denis, France
larbi.boubchir@univ-paris8.fr

***Abstract*—The search for efficient organic photovoltaics materials is crucial for advancing solar energy technologies due to their potential for low-cost, lightweight, and flexible solar cells compared to traditional inorganic photovoltaics. In this study, we employed a machine learning (ML) approach to predict the key photovoltaic parameters, namely the open-circuit voltage (Voc), the short-circuit current density (Jsc) and the power conversion efficiency (PCE) of organic semiconductors used in the active- layer of organic solar cells. We trained our ML model on a comprehensive dataset of known donor-acceptor (D/A) pairs and their respective photovoltaic properties. Using this trained model, we generated and evaluated numerous novel (D/A) combinations and predicted their Voc, Jsc and PCE values. This high-throughput screening enabled us to identify promising (D/A) pairs that have not yet been explored in the literature. As a result, our findings demonstrate the power of machine learning in accelerating the discovery and optimization of new materials combinations for organic photovoltaics, potentially leading to more efficient and cost-effective solar cells, thus advancing the viability of sustainable energy solutions.**

***Index Terms*—Artificial Intelligence, Organic Photovoltaics, Donor/Acceptor Pairs, Molecular Chemical Structure, New Materials, Prediction.**

I. Introduction

Renewable energy sources like solar and wind are essential to the European Green Deal due to their clean and sustainable characteristics [1]. Traditional inorganic photovoltaics, such as silicon-based solar cells, dominate the market due to their high efficiencies and extended lifespans. However, these technologies are associated with significant manufacturing costs and a lack of flexibility, which restrict their use in diverse applications [1]. In contrast, organic photovoltaics (OPVs) have recently gained attention as a viable alternative.

OPVs boast several advantages, including lower production costs, mechanical flexibility, and lightweight characteristics. These properties make them ideal for a wide array of uses, ranging from portable electronic devices to building-integrated photovoltaics. Nonetheless, OPVs face considerable challenges in terms of efficiency and stability when compared to their inorganic counterparts, which hinders their broader adoption [2], [3], [4].

A significant challenge in advancing organic photovoltaic technology is the identification and optimization of efficient materials. Part of this challenge is to identify the organic semiconductors used in the active layer, at least an electron-donating semiconductor (D) paired with an electron-accepting one (A). Indeed, the photovoltaic performances of OPVs is strongly influenced by the combination (D/A) pair that affect the critical parameters like the open-circuit voltage, the short-circuit current density and the power conversion efficiency. Traditional experimental methods for discovering new donor-acceptor pairs are both time-consuming and resource-intensive, underscoring the need for more efficient approaches [4], [5].

To tackle this challenge, our study employs a machine learning approach to expedite the discovery and optimization of novel organic photovoltaic materials. Initially, we developed a ML model trained on an extensive dataset of known donor-acceptor pairs and their corresponding photovoltaic properties, as described in our previous work[6] .This model was tested on (D/A) pairs that were not included in the dataset but published afterward [6]. In the present work, we use this ML model to predict the Voc, the Jsc, and the PCE of various new donor-acceptor combinations through high-throughput screening. Our results highlight several promising donor-acceptor pairs that have not yet been explored experimentally in the literature demonstrating the power of ML in not only speeding up the discovery process but also in potentially uncovering materials with superior photovoltaic properties. This approach could also significantly enhance the efficiency and reduce the cost of OPVs materials discovery, thereby advancing the viability of sustainable energy solutions. This paper is organized as follows:

- Section 2 provides a comprehensive review of the existing literature related to organic photovoltaics and the application of machine learning in materials discovery.
- Section 3 outlines the methodology employed in our study, detailing the data collection process, the training

of our machine learning model, and the high- throughput screening approach used to evaluate novel donor-acceptor combinations.
- In Section 4, we present and analyze the results of our predictions, highlighting the performance of the model and identifying promising new donor-acceptor pairs. We then discuss the implications of our findings in the field of organic photovoltaics.
- Section 5 concludes the paper, summarizing our key contributions and their significance in advancing the development of efficient organic solar cell materials.

II. RELATED WORKS

The integration of machine learning models with experimental data has emerged as a powerful strategy for discovering new (D/A) combinations in organic photovoltaics. Unlike traditional methods that often rely on theoretical calculations, utilizing experimental data allows researchers to capture the complex relationships between material properties and photovoltaic performance. By compiling extensive datasets from laboratory experiments including parameters such as device efficiency, ML algorithms can identify patterns that inform the selection of promising new material combinations. This data-driven approach not only enhances the predictive accuracy of ML models but also streamlines the discovery process, as it focuses on real-world performance rather than theoretical predictions. For example, recent studies like ([7], [8], [9], [10], [11], [12], [13], [14], [15], [16]) have demonstrated that ML can efficiently analyze experimental results to highlight (D/A) pairs with superior efficiency, ultimately accelerating the development of innovative materials for next-generation OPVs.

In recent works, Kranthiraja et al. [17] developed a random forest model trained on 566 (D/A) pairs to predict organic solar cell power conversion efficiency, demonstrating strong predictive capabilities. However, the dataset notably excluded the Y-series acceptors, limiting the model's applicability to new materials. To overcome this problem Greenstein et al. [18] and Suthar et al. [19] utilized two of the largest experimental datasets to date, comprising respectively 1225 and 1242 (D/A) pairs to predict Voc, Jsc, and PCE using molecular properties derived from Time-Dependent density-functional theory (TD-DFT) and chemical descriptors (Ex: Van Der Waals Surface Area). Their work underscores the importance of comprehensive datasets for optimizing organic photovoltaics.

Our focus is on unique pairs of donor (D) and acceptor (A) molecules, specifically using non-fullerene acceptor (NFA) materials that have demonstrated strong performance in recent studies [8], [11], [20]. Despite advancements, a comprehensive understanding of the relationship between molecular chemical structure, electronic properties, and device performance remains elusive. Our study aims to bridge this gap by exploring the predictive capabilities of only experimental data and molecular descriptors across various material types. Related informations are mentioned in Table I.

III. CONTRIBUTION

A. Dataset

We used the same dataset as Greenstein et al. [18] consisting in 1225 experimental (D/A) pairs devices collected from literature, including key parameters such as Voc, Jsc, Fill Factor (FF) and PCE, along with additional properties like the experimental Frontier Molecular Orbitals (FMO) which are the Highest Occupied Molecular Orbitals (HOMO) and the Lowest Unoccupied Molecular Orbitals (LUMO). To streamline analysis, only distinct (D/A) pairs were retained, focusing on those with the highest PCE for each core structure. The final refined dataset consists of 924 unique (D/A) pairs, encompassing 623 acceptors and 204 donors, along with their respective SMILES (Simplified Molecular Input Line Entry System) codes and electronic properties, providing a diverse range of semiconducting small molecules and polymers.

B. Method:pretreatment and treatment

In our previous work, we implemented a rigorous data cleaning process to ensure the integrity of our dataset by removing duplicate entries and verifying power conversion efficiency values [6]. We extracted over 100 molecular descriptors using RDKit python library which is a well-known package for chemical data handling in ML tasks, consolidating them into a dataset, "data-1," which was further enhanced by adding experimental FMO to create "data-2". For feature selection, we utilized the Minimum Redundancy Maximum Relevance (MRMR) method to identify 20 non-redundant and relevant features that improved predictive accuracy. Subsequently, we applied five supervised machine learning regression models: Support vector Regressor (SVR), Random Forest (RF), Adaptative Boosting Regressor (AdaBoost), Extreme Gradient Boosting regressor (XGBoost) and Gradient Boosting Regressor (GBR), after meticulously tuning their hyperparameters to optimize their performance.

In the present work, we utilized the same model with 125867 unique (D/A) pairs, with 924 pairs already tested in literature, while the rest have not been explored. Our goal was to predict the performance parameters for each pair in each organic photovoltaic device and explore if our model can identify some highly performant (D/A) combinations as mentioned in our methodology which is illustrated in Fig. 1.

IV. RESULTS AND DISCUSSION

A (D/A) pair that was not included in the dataset used to train our ML model is (D18/Y6-Se). The predicted values for the photovoltaic parameters using our ML model (first line in Table II) are consistent with the experimental ones (second line in Table II from [6]). Indeed, the relative uncertainty on the Voc prediction is below 7% while it falls in the 1% range for both Jsc and PCE predicted values. These low relative uncertainties are achieved with the help of the GBR model, an ensemble learning method that combines the predictions of multiple weak learners to improve overall performance and robustness [6].

TABLE I
RESULTS OF NEW PAIRS PREDICTIONS *FROM REF

References	Data	Input Type	Target	ML Models	Validation	Validation Metrics	Results
[17]	566	FMO Weights Firngerprints	PCE	RF	K-Fold CV	r	r=0.85
[19]	1242	CSD + FMO	PCE Jsc Voc	SVR RF ANN GBR	80% Train 20% Test +LOOCV	RMSE r	RMSE=2.004% (PCE) r=0.79 (PCE) r=0.86 (Voc) RMSE = 0.047 V
[18]	503	CSD + FMO	$PCE > 10$	RF	5 Fold CV	RMSE R^2	R^2 = 0.28 RMSE = 1.60%
[6]	924 + 125867 D/A Pairs	CSD + FMO (Exp)	V_{oc} J_{sc} PCE	SVR, , AdaBoost, , RF Xgboost GBR	80% Train 20% Test +LOOCV	RMSE R^2 MAE r	GBR Best Model R^2 = 0.78 (Voc) R^2 = 0.70 (Jsc) R^2 = 0.63 (PCE)

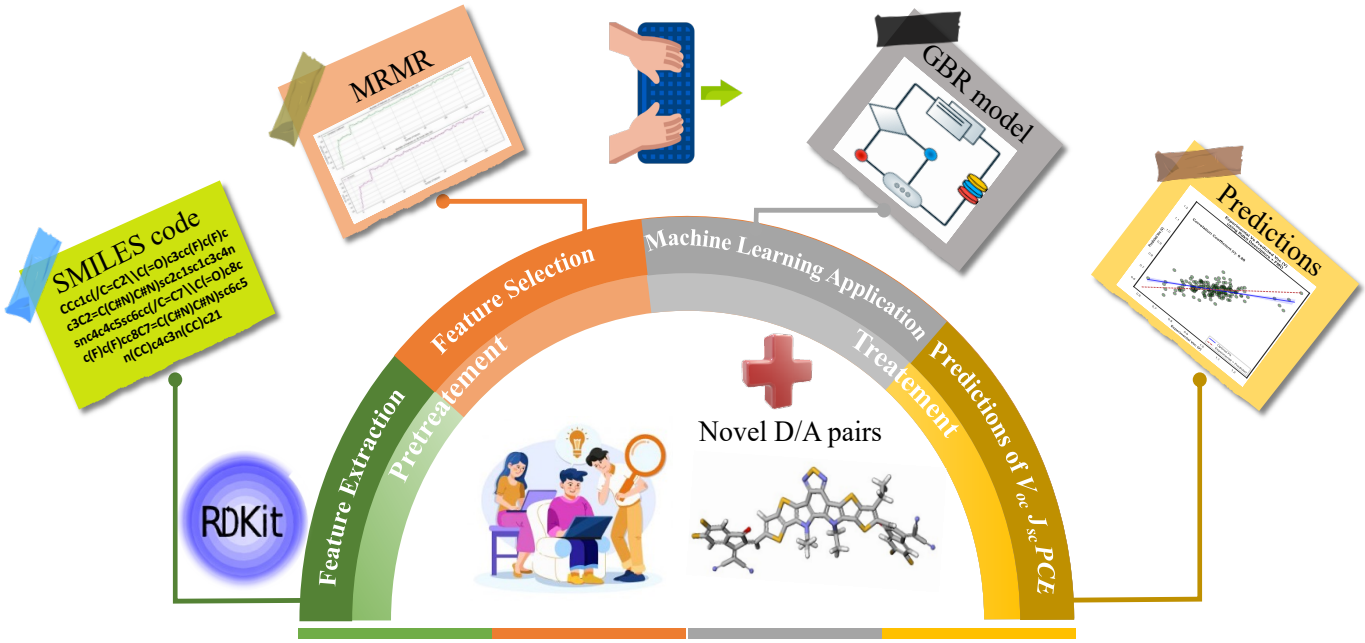


Fig. 1. Schema illustrating the proposed methodology.

This model excels in identifying and prioritizing the most relevant features from our input data, allowing it to predict accurately the photovoltaic parameters of (D/A) pairs not present in the initial dataset.

As shown in Table II, several (D/A) pairs that have not yet been tested in the literature yielded promising predictive values for Voc, Jsc and PCE. The donor material D18 used in binary bulk heterojunction solar cells with the NFA CH1007 and L8-BO achieves predicted PCE values greater than 17% while PBQx-TF blended with CH1007 leads to a predicted PCE value of 16.5%. These novel (D/A) pairs showed high predicted values, suggesting their potential as candidates for high-performance organic photovoltaic (OPV) devices based on NFAs.

All these results not only validates the robustness of our ML model but also highlights its utility in accelerating the discovery of efficient OPV materials pairs.

TABLE II
RESULTS OF NEW PAIRS PREDICTIONS *FROM [6]

Donor	Acceptor	Voc (V)	Jsc (mA/cm^2)	PCE (%)
D18	Y6-Se	0.90	27.5	17.97
*D18	Y6-Se	0.84	27.98	17.7
D18	CH1007	0.88	28.17	17.90
D18	L8-Bo	0.92	25.50	17.44
PBQx-TF	CH1007	0.82	27.13	16.50

V. Conclusion

This study demonstrates the effectiveness of machine learning in predicting the performance of organic photovoltaic devices. By applying our previously developed GBR ML model to a new combinations of donor and acceptor pairs not tested in the literature, we identified several novel pairs with high predicted Voc, Jsc, and PCE values. Notably, the donor D18 paired with acceptors such as CH1007 and L8-BO showed a good photovoltaic potential.

By leveraging machine learning, researchers can significantly reduce the time and resources required to identify high-performing OPV materials, and our approach accelerates the discovery process, allowing for a more efficient path to the development of next-generation solar energy technologies.

Acknowledgment

We would like to express our sincere gratitude to Dr. Olivier BARDAGOT for his invaluable help, his expertise and guidance were instrumental in the completion of this research. Thank you for your unwavering support and encouragement throughout this work.

References

[1] G. J. Moore, O. Bardagot, and N. Banerji, "Advances in organic photovoltaic cells: a comprehensive review of materials, technologies, and performance," Apr. 19, 2023, Royal Society of Chemistry. doi: 10.1039/d3ra01454a.

[2] M. S. A. Kamel, A. Al-jumaili, M. Oelgemöller, and M. V. Jacob, "Inorganic nanoparticles to overcome efficiency inhibitors of organic photovoltaics: An in-depth review," Sep. 01, 2022, Elsevier Ltd. doi: 10.1016/j.rser.2022.112661.

[3] M. C. Scharber and N. S. Sariciftci, "Efficiency of bulk-heterojunction organic solar cells," Dec. 2013. doi: 10.1016/j.progpolymsci.2013.05.001.

[4] Y.-W. Su, S.-C. Lan, and K.-H. Wei, "Organic photovoltaics," Materials Today, vol. 15, no. 12, pp. 554–562, Dec. 2012, doi: 10.1016/S1369-7021(13)70013-0.

[5] G. Kumar and F. C. Chen, "A review on recent progress in organic photovoltaic devices for indoor applications," Aug. 31, 2023, Institute of Physics. doi: 10.1088/1361-6463/acd2e5.

[6] K. Khoussa, L. Boubchir, and P. Leveque, "Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs," 2024. doi: 10.2139/ssrn.4997197.

[7] A. Eibeck et al., "Predicting Power Conversion Efficiency of Organic Photovoltaics: Models and Data Analysis," ACS Omega, vol. 6, no. 37, pp. 23764–23775, Sep. 2021, doi: 10.1021/acsomega.1c02156.

[8] C. Guo et al., "Light-induced quinone conformation of polymer donors toward 19.9% efficiency organic solar cells," Energy Environ Sci, vol. 17, no. 7, pp. 2492–2499, 2024, doi: 10.1039/D4EE00605D.

[9] Q. Liu et al., "18% Efficiency organic solar cells," Sci Bull (Beijing), vol. 65, no. 4, pp. 272–275, Feb. 2020, doi: 10.1016/j.scib.2020.01.001.

[10] A. Mahmood and J. L. Wang, "Machine learning for high performance organic solar cells: Current scenario and future prospects," Jan. 01, 2021, Royal Society of Chemistry. doi: 10.1039/d0ee02838j.

[11] Y. Wang et al., "High-performance nonfullerene polymer solar cells based on a fluorinated wide bandgap copolymer with a high open-circuit voltage of 1.04 v," J Mater Chem A Mater, vol. 5, no. 42, pp. 22180–22185, 2017, doi: 10.1039/c7ta07785h.

[12] J. Cai, X. Chu, K. Xu, H. Li, and J. Wei, "Machine learning-driven new material discovery," Aug. 01, 2020, Royal Society of Chemistry. doi: 10.1039/d0na00388c.

[13] E. Antono et al., "Machine-Learning Guided Quantum Chemical and Molecular Dynamics Calculations to Design Novel Hole-Conducting Organic Materials," Journal of Physical Chemistry A, vol. 124, no. 40, pp. 8330–8340, Oct. 2020, doi: 10.1021/acs.jpca.0c05769.

[14] M.-H. Lee, "A Machine Learning–Based Design Rule for Improved Open-Circuit Voltage in Ternary Organic Solar Cells," Advanced Intelligent Systems, vol. 2, no. 1, Jan. 2020, doi: 10.1002/aisy.201900108.

[15] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," 2018. [Online]. Available: https://www.science.org

[16] N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, and S. P. Russo, "Machine learning property prediction for organic photovoltaic devices," NPJ Comput Mater, vol. 6, no. 1, Dec. 2020, doi: 10.1038/s41524-020-00429-w.

[17] K. Kranthiraja and A. Saeki, "Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells," Adv Funct Mater, vol. 31, no. 23, Jun. 2021, doi: 10.1002/adfm.202011168.

[18] B. L. Greenstein and G. R. Hutchison, "Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms," Journal of Physical Chemistry C, vol. 127, no. 13, pp. 6179–6191, Apr. 2023, doi: 10.1021/acs.jpcc.3c00267.

[19] R. Suthar, T. Abhijith, and S. Karak, "Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors," J Mater Chem A Mater, vol. 11, no. 41, pp. 22248–22258, Sep. 2023, doi: 10.1039/d3ta04603f.

[20] C. Xie et al., "Water-based layer-by-layer processing enables 19% efficient binary organic solar cells with minimized thickness sensitivity," Energy Environ Sci, vol. 17, no. 7, pp. 2441–2452, 2024, doi: 10.1039/D4EE00068D.

Deep Learning Approach for Predicting Efficiency in Organic Photovoltaics from 2D Molecular Images of D/A pairs

Khoukha Khoussa[1], Patrick Lévêque[1,] ,Larbi Boubchir[2]*

K. Khoussa, P. Lévêque

ICube Laboratory, University of Strasbourg-CNRS, 23 rue de Loess, 67037 Strasbourg, France

E-mail: Patrick.leveque@unistra.fr

L. Boubchir

LIASD Laboratory, University of Paris 8, 2 rue de la Liberté, 93526 Saint-Denis, France

Keywords: artificial intelligence, organic photovoltaics performance, deep learning, 2D images of D/A pairs, prediction.

Organic Photovoltaic (OPV) devices have emerged as a promising alternative to conventional solar cells due to their flexibility, lightweight nature, and potential for low-cost production. However, optimizing OPV performance remains a complex challenge, traditionally requiring extensive experimental trials or computational chemistry approaches based on molecular descriptors. To accelerate the development of high-efficiency OPVs, Artificial Intelligence (AI) has been increasingly utilized, particularly machine learning models that rely on chemical descriptors. While these methods have shown success, they are often limited by the quality and completeness of the selected descriptors, potentially overlooking key structural and morphological information. In this work, we propose a novel deep learning framework leveraging Convolutional Neural Networks (CNNs) to predict OPV performance directly from 2D images of donor and acceptor materials. By employing a customized representation of molecular structures, our approach captures spatial and hierarchical patterns that traditional descriptors based ML models may miss. We compare our model's predictive capability to conventional machine learning techniques and demonstrate its potential for improving prediction accuracy and generalization without need to add the Frontier Molecular Orbitals (FMOs) to enhance predictions. Our findings highlight the power of deep learning in accelerating the discovery of efficient organic photovoltaic materials, paving the way for a data-driven approach to materials science and device optimization.

1. Introduction

Photovoltaic (PV) cells, are one of the renewable energy technologies that can provide a nearly infinite supply of electricity, making them a sustainable solution for the long term. However, traditional PV cells are often limited by factors such as high production costs, heavy weight and lack of flexibility that can restrict their applications in certain areas. Organic photovoltaic (OPV) cells offer several advantages, including cost-effectiveness, flexibility, and the potential for integration into lightweight wearable devices, making them an attractive option for a wider range of applications [1], [2], [3], [4], [5], [6], [7]. Nevertheless, OPV cells currently demonstrate lower power conversion efficiency (PCE) around 20% [1], compared to certified PCE of 26.1% for non-concentrated single-crystal silicon solar cells. Further, OPV cells have only estimated operational lifespans around 30 years [2] compared to the certified lifetime of traditional inorganic solar cells [3].

OPV cells are composed of an organic active layer (AL) sandwiched between two electrodes. In efficient OPV cells, the AL is a blend of at least two organic semiconductors forming a bulk heterojunction (BHJ). The usual minimum configuration for BHJ solar cells is a blend of an electron donor (D) and an electron acceptor (A). In the present work, we will focus on (D/A) AL. The AL absorbs light, converts the photons into free charge-carrier and transport them to the electrodes. Optimizing the organic semiconductors in the AL is a complicated task requiring traditionally extensive experimental trials. The different experimental steps are the design of innovative semiconducting materials, their synthesis, the device elaboration and characterization and is therefore a multidisciplinary time-consuming process.

In recent years, data scientists and material experimentalists have shown growing interest in leveraging Artificial Intelligence (AI) for advancements in the Organic Photovoltaic field, aiming to accelerate the development of this technology. By utilizing AI models such as machine learning (ML) and deep learning (DL), key performance metrics can be accurately predicted based on material properties and processing conditions. This AI-driven approach streamlines the optimization of OPV cells, significantly reducing the time required to develop high-performance, cost-effective solar cells [4] [5] [6][7]. Greenstein *et al.* [4], [5] and Suthar *et al.* [12] utilized two of the largest experimental datasets available to date, comprising respectively 1225 and 1242 donor/acceptor (D/A) pairs, along with their molecular properties and their experimental photovoltaic performances. They used Time-Dependent Density Functional Theory (TD-DFT) to calculate the Frontier Molecular Orbital (FMO) properties and used molecular chemical descriptors, to predict key performance indicators including open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), and power conversion efficiency (*PCE*).

Despite the significant advancements in DL and ML, the application of these technologies to predict the performance of OPV cells using images of the chemical structures of (D/A) materials remain relatively rare [8], [9], [10], [11]. Indeed, most existing studies in the field have focused on numerical and textual data for performance prediction. The innovative approach of utilizing image-based inputs for DL models, which can capture the intricate details and spatial features of chemical structures, holds anyway considerable promise. In 2019, Sun *et al.* demonstrates the application of DL to rapidly evaluate the performance of new OPV materials, including a classification task that distinguishes between high and low *PCE* molecules [12] .Using a dataset from the Harvard Clean Energy Project (HCEP), which includes 2.3 million candidate molecules, Sun *et al.* developed a deep neural network model based on ResNet to predict the PCE directly from images of chemical structures. The model

classified molecules into two categories: low-performance (*PCE* ranging from 0 to 4.9%) and high-performance (*PCE* ranging from 5.0 to 9.9%). Initially tested on a small subset of 5000 molecules, the model was further refined using a larger subset of 50,000 molecules, divided into training, testing, and verification sets. It excelled in the classification task, achieving an overall prediction accuracy of 91.02%. The study concludes that deep learning offers a promising approach to the quick evaluation and classification of OPV materials, significantly reducing the time and resources needed for traditional trial-and-error methods, with potential for even greater accuracy through further database expansion. However, while this work represents a significant advancement in using AI models to predict the performance of OPV materials from chemical structures images, the model was developed based on the relatively simple chemical structures of the HCEP dataset. In recent years, more complex D and A semiconductors like Non Fullerene Acceptors (NFA) materials have been developed, achieving power conversion efficiencies approaching 20% [13], [14]. These NFA are not well-represented in the HCEP dataset, which limits the model's applicability for predicting the performance of these high-efficiency OPV materials. Moore *et al.* explored the use of 2D images of chemical structures to predict the frontier molecular orbitals (FMOs) of donor materials, rather than the performance of OPV cells [15]. They employed a convolutional neural network (CNN) with 2D images derived from the HCEP dataset and validated their model using experimental data from HOPV15 [16]. Their model exhibited poorer performance with the more complex experimental data compared to the validation results obtained with the HCEP dataset.

Our work focuses on predicting the main photovoltaic parameters (V_{oc}, J_{sc} and *PCE*) of BHJ (D/A) solar cells using experimental datasets through the innovative use of image data combined with deep learning models. By incorporating image-based analysis, we aim to capture complex molecular features and interactions that might be missed by traditional descriptor-based approaches. Deep learning models excel at extracting and interpreting intricate patterns from images, potentially leading to more accurate predictions and a deeper understanding of OPV cells performance [11], [12], [13], [21].

A schematic representation of an OPV cell is shown in **Figure 1** while a summary of the ML approach to predict OPV device performances and of image-driven AI approach for OPV are presented in **Table 1**.

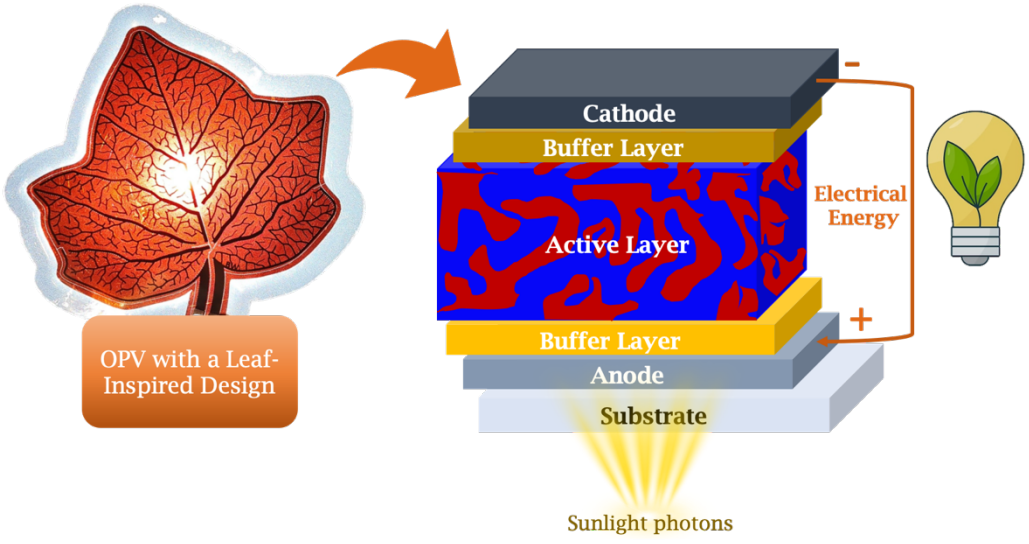


Figure 1. Organic Photovoltaic Device Structure

Table 1: Related works from literature using experimental (Exp) and calculated (Cal) data and our work

References	Dataset	Exp or Cal (HCEP)	Input Type	Target	ML models	Validation	Validation Metrics	Results
[12]	2.3 million	HCEP	2D images of donor	*PCE*	ResNet	Train:70% Test: 20% Validation : 10%	Accuracy (classificatio)	91.02 %
[15]	2.3 million+ HOPV15	HCEP + Exp	2D images of donor	FMO of Donor	CNN	Train test split	R^2	Exp : R^2 = 0.55 (HOMO) R^2 = 0.63 (LUMO) cal : R^2 = 0.99 (HOMO) R^2 = 0.99 (LUMO)
[6]	1242	Exp	CSD+FMO	*PCE* V_{oc} J_{sc}	SVR, RF, ANN, GBR	80% Train 20% test +LOOCV	RMSE /r	2.004 % /0.79 (*PCE*) 0.047 V/0.86 (V_{oc})
[4]	503	Exp	CSD +FMO (TDs-DFT)	*PCE* >10	RF	5 Fold CV	R^2 RMSE	R^2 = 0.28 RMSE = 1.60%
[6]	924	Exp	CSD+FMO (Exp)	V_{oc} J_{sc} *PCE*	SVR,RF, GBR,Xg boost,	80% Train 20% test +LOOCV +New data	RMSE MAE R^2 r	(*PCE %*) with FMO RMSE=2.38 MAE =1.84 $R^{2.}$=0.63 r= 0.79
Our Work	707	Exp	2D images of D/A pairs	V_{oc} J_{sc} *PCE*	CNN	80% Train 20% test	RMSE MAE R^2 r	(*PCE %*) without FMO RMSE=2.53 MAE =1.87 $R^{2.}$=0.54 r= 0.73

2. Contribution

2.1. Dataset

Our dataset has been collected from experimental data on 1,225 non-fullerene acceptor (NFA)/donor devices reported in scientific literature [4]. This dataset encompasses key performance metrics such as open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), fill factor (*FF*), and power conversion efficiency (*PCE*). It should be noted that only three parameters are necessary to calculate all photovoltaic parameters, explaining why *FF* has been omitted. To ensure data reliability and uniqueness, duplicate (D/A) pairs were refined by selecting only the highest *PCE* values for each core structure. This curation resulted in a final dataset of 999 unique donor-acceptor (D/A) pairs, comprising 605 distinct acceptors and 214 donors, each represented by their Simplified Molecular Input Line Entry System (SMILES) notation. After excluding pairs lacking reported frontier molecular orbital (FMO) values, the dataset was further reduced to 924 unique combinations with corresponding electronic properties. The *PCE* values in this dataset range from 0.01% to 18.77%, with the majority falling between 6% and 14%. Analysis of the most frequently utilized molecules identified PBDB-T (also known as PCE12), PBDB-T-2F (PM6), PTB7-Th (PCE10), and P3HT as the top donor materials and the most common acceptors were ITIC, IT-4F, Y6, and IDIC. As shown in **Figure 2** we can see the distribution of key photovoltaic parameters in the dataset, where each histogram illustrates the frequency of values for the respective parameter, with an overlaid kernel density estimation (KDE) curve to depict the underlying probability distribution. The *PCE* distribution exhibits a right-skewed shape, the V_{oc} one follows a near-normal pattern, peaking around 0.9 V and the J_{sc} distribution displays a peak around 15-20 mA/cm² and shows a moderately right-skewed trend.

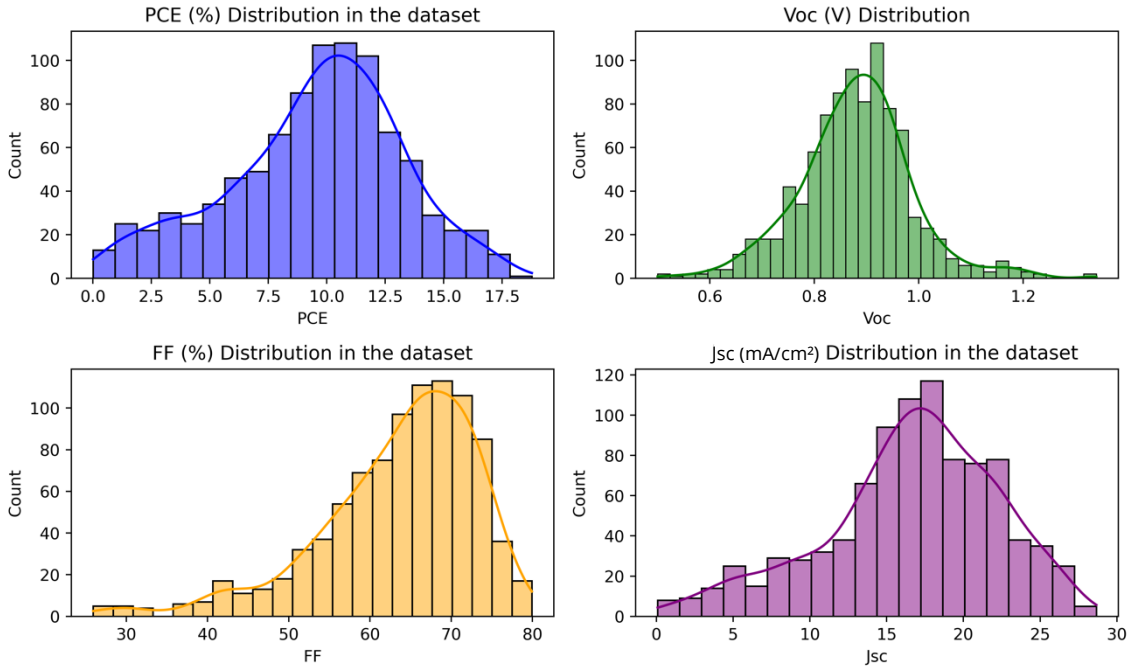


Figure 2. ***PCE*, V_{oc}, J_{sc} and *FF* values distribution in the dataset**

2.2. Data Preparation

The SMILES (Simplified Molecular Input Line Entry System) codes for donor and acceptor materials were used to remove invalid entries. We then used RDKit, a widely used cheminformatics toolkit, to convert the cleaned SMILES codes into molecular objects. Further, we generated 2D coordinates for each molecule to create planar visual representations. In our representation, each atom is depicted as a circle with a radius proportional to its calculated atomic radius, providing a more scalable and visually informative representation of atom sizes. The bonds are represented by differeng colors depending of their nature (black for single bonds, red for double bonds and cyan for triple bonds). Two representations were tested, one with short side chains and another one with real side chains in order to estimate the increase in calculation time using real chains. To prepare the images to be used as inputs into our Convolutional Neural Network (CNN), we determined the maximum X and Y coordinates from the 2D molecular representations. All images were then resized to a uniform resolution of 80 pixels per dimension (80 PDI), with final dimensions of (259 x 480) pixels. This standardization ensures that the input images have consistent size and aspect ratio, which is critical for effective CNN training (See **Figure 3**). The final dataset consists of images of donor-acceptor pairs, each visualized with our custom style to enhance clarity and differentiation of atom and bond types.

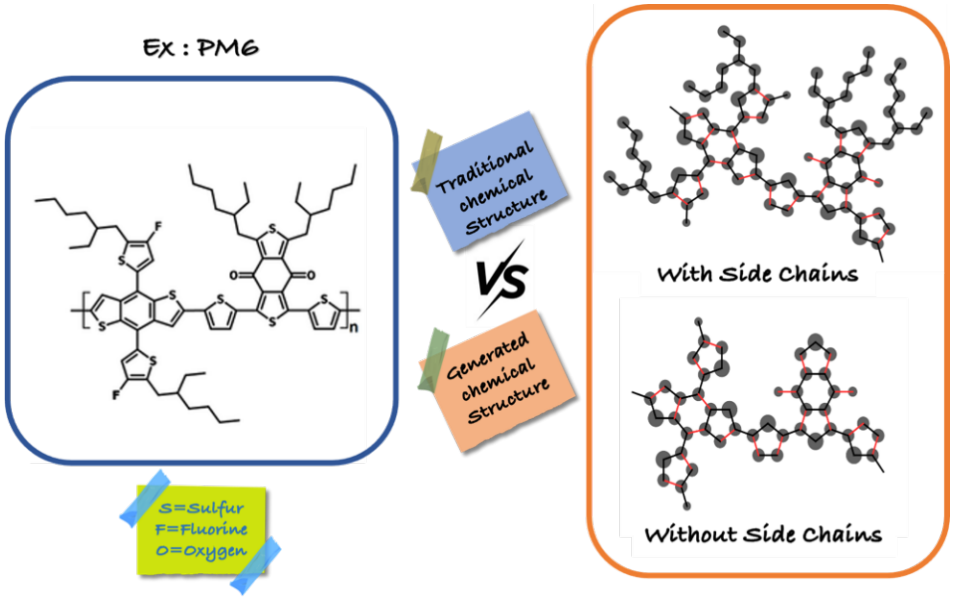


Figure 3. Generated 2D images example

2.3. Model Training

In the present work, we explored two approaches to predict the V_{oc}, J_{sc}, and *PCE* values of organic photovoltaic devices using convolutional neural network. The core data used for all approaches consisted of 2D images representing the donor and acceptor materials. In the first approach, we utilized 2D CNN to directly extract features from the donor and acceptor images. The architecture of the CNN comprised nine layers: five 2D convolutional layers interspersed with four max-pooling layers. After feature extraction from both donor and acceptor images, the features were concatenated into a single feature vector. This combined vector was then fed into fully connected layers to predict the V_{oc}, J_{sc}, and *PCE* values of the OPVs. This approach focused on leveraging the spatial information from the images to calculate the predictions. In a second approach, we enhanced the CNN-based model by incorporating the experimental FMOs or Frontier Molecular Orbitals of both donor and acceptor materials into the feature vector. The features extracted from the donor and acceptor images using the same CNN architecture as in the first approach were concatenated with the FMOs values before the prediction stage. This enriched feature vector was then used to predict the V_{oc}, J_{sc} and the *PCE* values. This approach is aiming to improve prediction as it is well known that the V_{oc} mainly depends on FMOs[17][18]. Both approaches are described in **Algorithm 1** and **Figure 4**.

Algorithm 1 : The proposed Approach for predicting photovoltaic parameters using 2D images of D/A pairs and CNN.

```
Input:
   Donor image D
   Acceptor image A
   Experimental FMOs values: F_Donor, F_Acceptor

Output:
   Predicted Voc, Jsc, and PCE
Begin
        Begin
          // Step 0: Hyperparameter Tuning
          Set Learning_Rate, Epochs based on tuning

          // Step 1: Feature Extraction
          Features_D ← CNN_ExtractFeatures(D)
          Features_A ← CNN_ExtractFeatures(A)

          // Step 2: Feature Fusion
          Combined_Features ← Concatenate(Features_D, Features_A)
```

```
        // Step 3: Prediction
        Output_Vector ← FullyConnectedNetwork(Combined_Features)

        // Step 4: Return predicted values
        Return Voc, Jsc, PCE from Output_Vector
    End
    Begin
        // Step 0: Hyperparameter Tuning
        Set Learning_Rate, Epochs based on tuning

        // Step 1: Feature Extraction
        Features_D ← CNN_ExtractFeatures(D)
        Features_A ← CNN_ExtractFeatures(A)

        // Step 2: FMOs Processing
        FMOs_Vector ← Concatenate(F_Donor, F_Acceptor)

        // Step 3: Feature Fusion
        Combined_Features ← Concatenate(Features_D, Features_A, FMOs_Vector)

        // Step 4: Prediction
        Output_Vector ← FullyConnectedNetwork(Combined_Features)

        // Step 5: Return predicted values
        Return Voc, Jsc, PCE from Output_Vector
    End
End
```

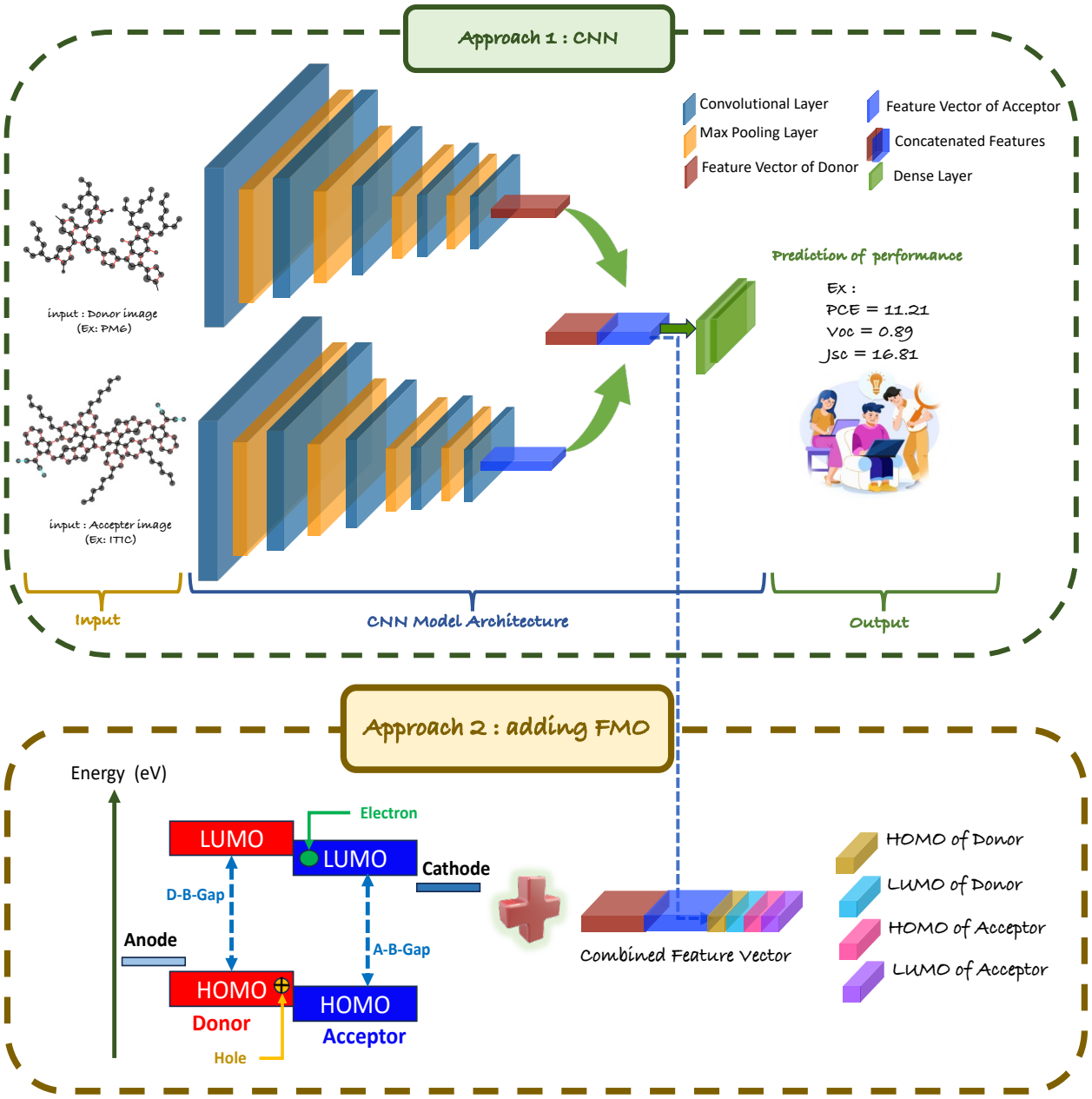


Figure 4. Our approach

3. Results and Discussion

In our first approach, to optimize model performance, we fine-tuned hyperparameters such as the learning rate and kernel size, selecting the optimal configuration that yielded the best predictive accuracy for V_{oc}, J_{sc} and *PCE* predictions. We trained our model using only CPUs, with each epoch taking approximately 2 to 5 seconds, eliminating the need for a GPU and making the process cost-effective. The maximum RAM usage was 16 GB. We experimented with 50, 100, 200, and 300 epochs, but found that even with just 20 epochs, our model achieved promising results. In the second approach, we aimed to enhance the predictive capability of the model by incorporating additional chemical information, providing a more comprehensive representation of the material properties to improve prediction.

As a results, we achieved remarkable results using only 2D images of donor and acceptor materials without incorporating FMO values. When compared to our previous work [6], which relied on traditional machine learning models using chemical descriptors extracted from SMILES representations, the CNN-based model demonstrated comparable predictive performance. This suggests that deep learning can effectively extract and learn relevant molecular features directly from image data, achieving approximately similar results to models explicitly utilizing FMO descriptors. Specifically, our CNN model attained, for the *PCE* prediction, an RMSE of 2.53 (%), an r of 0.73, an MAE of 1.87 (%), and an R^2 of 0.54, which closely matched the results obtained when FMO features were included. Including the FMOs improved, as expected, the V_{oc} prediction but the effect on the *PCE* prediction is rather limited. These findings indicate that deep learning possesses the capability to autonomously identify and represent key molecular features necessary for accurate power conversion efficiency prediction. Further, it was not necessary to include a feature selection part (using for instance Minimum Redundancy Maximum Relevance) as the CNN itself can select relevant features and predict the key parameters of OPV performance. The fact that our model achieved high predictive accuracy solely based on image data underscores the potential of CNNs as powerful tools for material property prediction without requiring predefined chemical descriptors. All results are illustrated in the **Table 2** and **Figure 5**.

Table 2. Results of our two approaches : using only CNN and using CNN with FMO.

Performance Parameters	***PCE* (%)**				***J_{sc}* (mA/cm²)**				***V_{oc}* (V)**			
Model evaluation metrics	R^2	MAE (%)	RMSE (%)	r	R^2	MAE (mA/cm²)	RMSE (mA/cm²)	r	R^2	MAE (V)	RMSE (V)	r
Approach 1 (Only CNN)	0.54	1.87	2.53	0.73	0.57	2.63	3.37	0.78	0.66	0.05	0.06	0.84
Approach 2 (CNN With FMO)	0.55	1.89	2.52	0.73	0.58	2.57	3.36	0.79	0.69	0.04	0.05	0.86

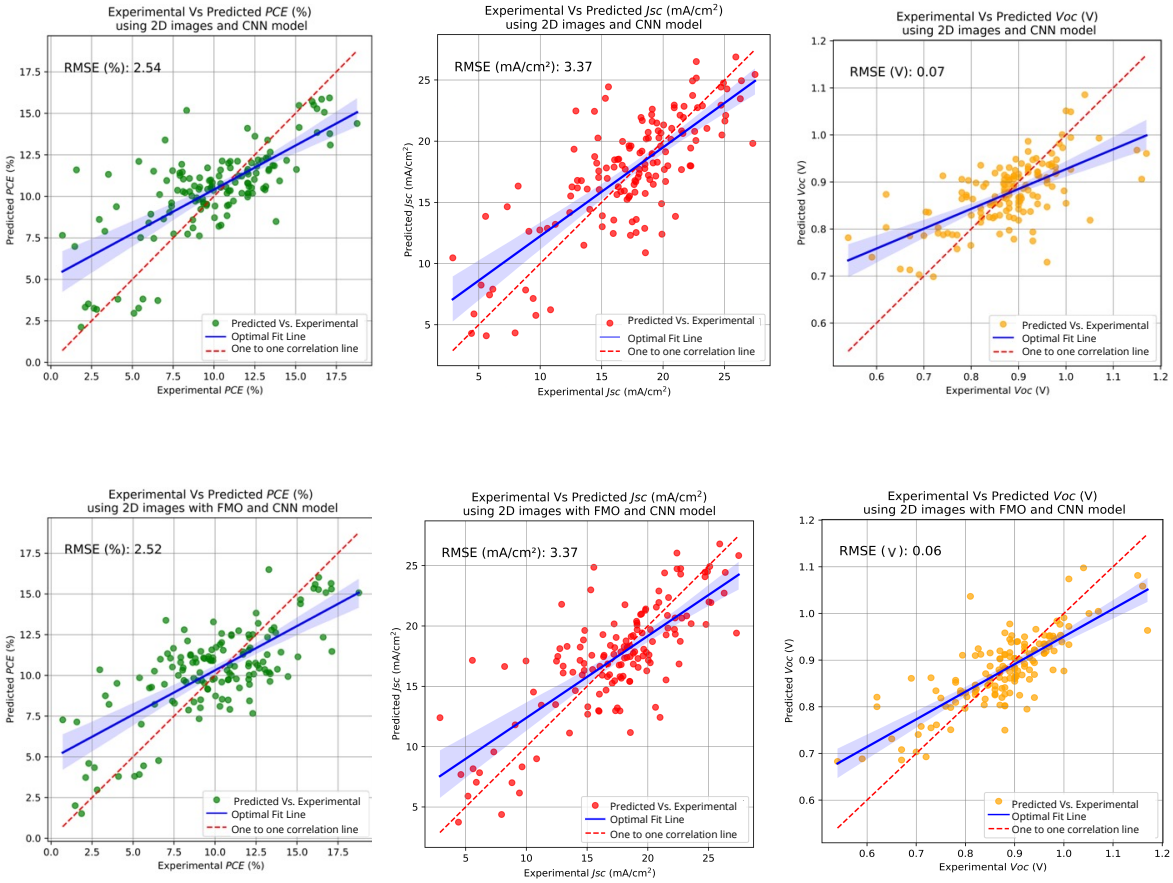


Figure 5. Results of prediction for the main photovoltaic parameters using only the CNN model (top line) or the CNN model and the FMOs (bottom line).

4. Conclusion

In this work, we demonstrated the effectiveness of deep learning, specifically Convolutional Neural Networks, in predicting the power conversion efficiency of organic photovoltaic devices using only 2D images of donor and acceptor materials. Our approach eliminates the need for handcrafted chemical descriptors, showing that CNNs can autonomously extract relevant molecular features for accurate performance prediction.

By comparing different modeling strategies including purely image-based CNN predictions and CNNs enhanced with frontier molecular orbital (FMO) values we found that CNNs alone could achieve competitive accuracy. Our results indicate that deep learning is a viable alternative to conventional descriptor-based methods, offering a more automated and scalable approach for OPV material discovery. Further, our model used only CPUs with good results on only 20 epochs taking a total time lower than 100 seconds. The maximum RAM usage was 16 GB making the process cost-effective. These findings highlight the potential of deep learning in accelerating OPV research, reducing dependence on predefined molecular descriptors, and paving the way for AI-driven materials science.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author. It includes a definition of Frontier Molecular Orbitals (HOMO and LUMO), and an explanation of donor and acceptor materials in organic solar cells. It also describes the Convolutional Neural Network (CNN) architecture used in this study, including the role and impact of key hyperparameters such as learning rate, number of epochs, and batch size..etc.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author.

ORCID

Pr. Larbi Boubchir: 0000-0002-5668-6801

Dr. Patrick Lévêque: 0000-0002-6927-9025

Khoukha KHOUSSA : https://orcid.org/0009-0006-4021-841X

Received: ((will be filled in by the editorial staff))

Revised: ((will be filled in by the editorial staff))

Published online: ((will be filled in by the editorial staff)

5. References

[1] C. Guo *et al.*, "Light-induced quinone conformation of polymer donors toward 19.9% efficiency organic solar cells," *Energy Environ Sci*, vol. 17, no. 7, pp. 2492–2499, 2024, doi: 10.1039/D4EE00605D.

[2] Y. Li *et al.*, "Non-fullerene acceptor organic photovoltaics with intrinsic operational lifetimes over 30 years," *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-25718-w.

[3] W. B. Tarique and A. Uddin, "A review of progress and challenges in the research developments on organic solar cells," Aug. 15, 2023, *Elsevier Ltd*. doi: 10.1016/j.mssp.2023.107541.

[4] B. L. Greenstein and G. R. Hutchison, "Screening Efficient Tandem Organic Solar Cells with Machine Learning and Genetic Algorithms," *Journal of Physical Chemistry C*, vol. 127, no. 13, pp. 6179–6191, Apr. 2023, doi: 10.1021/acs.jpcc.3c00267.

[5] R. Suthar, T. Abhijith, and S. Karak, "Machine-learning-guided prediction of photovoltaic performance of non-fullerene organic solar cells using novel molecular and structural descriptors," *J Mater Chem A Mater*, vol. 11, no. 41, pp. 22248–22258, Sep. 2023, doi: 10.1039/d3ta04603f.

[6] K. Khoussa, L. Boubchir, and P. Leveque, "Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency from the Molecular Chemical Structure of (Donor/Acceptor) Pairs," 2024. doi: 10.2139/ssrn.4997197.

[7] K. Khoussa, P. Leveque, and L. Boubchir, "On the use of Machine Learning to Discover Novel Donor-Acceptor Pairs For Organic Photovoltaic Devices," in *2024 IEEE International*

Conference on Big Data (BigData), IEEE, Dec. 2024, pp. 4659–4662. doi: 10.1109/BigData62323.2024.10825948.

[8] A. Mahmood and J. L. Wang, "Machine learning for high performance organic solar cells: Current scenario and future prospects," Jan. 01, 2021, *Royal Society of Chemistry*. doi: 10.1039/d0ee02838j.

[9] G. Kumar and F. C. Chen, "A review on recent progress in organic photovoltaic devices for indoor applications," Aug. 31, 2023, *Institute of Physics*. doi: 10.1088/1361-6463/acd2e5.

[10] G. J. Moore, O. Bardagot, and N. Banerji, "Advances in organic photovoltaic cells: a comprehensive review of materials, technologies, and performance," Apr. 19, 2023, *Royal Society of Chemistry*. doi: 10.1039/d3ra01454a.

[11] S. Bhatti *et al.*, "Machine learning for accelerating the discovery of high performance low-cost solar cells: a systematic review," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.13893

[12] W. Sun *et al.*, "The Use of Deep Learning to Fast Evaluate Organic Photovoltaic Materials," *Adv Theory Simul*, vol. 2, no. 1, Jan. 2019, doi: 10.1002/adts.201800116.

[13] C. Guo *et al.*, "Light-induced quinone conformation of polymer donors toward 19.9% efficiency organic solar cells," *Energy Environ Sci*, vol. 17, no. 7, pp. 2492–2499, 2024, doi: 10.1039/D4EE00605D.

[14] C. Xie *et al.*, "Water-based layer-by-layer processing enables 19% efficient binary organic solar cells with minimized thickness sensitivity," *Energy Environ Sci*, vol. 17, no. 7, pp. 2441–2452, 2024, doi: 10.1039/D4EE00068D.

[15] G. J. Moore, O. Bardagot, and N. Banerji, "Deep Transfer Learning: A Fast and Accurate Tool to Predict the Energy Levels of Donor Molecules for Organic Photovoltaics," *Adv Theory Simul*, vol. 5, no. 5, May 2022, doi: 10.1002/adts.202100511.

[16] S. A. Lopez *et al.*, "The Harvard organic photovoltaic dataset," *Sci Data*, vol. 3, Sep. 2016, doi: 10.1038/sdata.2016.86.

[17] M. C. Scharber and N. S. Sariciftci, "Efficiency of bulk-heterojunction organic solar cells," Dec. 2013. doi: 10.1016/j.progpolymsci.2013.05.001.

[18] M. C. Scharber *et al.*, "Design rules for donors in bulk-heterojunction solar cells - Towards 10 % energy-conversion efficiency," *Advanced Materials*, vol. 18, no. 6, pp. 789–794, Mar. 2006, doi: 10.1002/adma.200501717.

TOC:

This study highlights the potential of deep learning, particularly Convolutional Neural Networks (CNNs), for predicting the photovoltaic performance of organic solar cells. By leveraging 2D images representing donor/acceptor molecular pairs, the model accurately estimates key performance indicators proving that this image-based approach offers a fast and efficient alternative to traditional computational methods, enabling high-throughput screening of material combinations and accelerating the discovery of high-performance organic photovoltaic materials.

Khoukha Khoussa, Patrick Lévêque*, Larbi Boubchir

Deep Learning Approach for Predicting Efficiency in Organic Photovoltaics from 2D Molecular Images of D/A pairs

Supporting Information

Deep Learning Approach for Predicting Efficiency in Organic Photovoltaics from 2D Molecular Images of D/A pairs

Khoukha Khoussa[1], Patrick Lévêque[1,] ,Larbi Boubchir[2]*

Contents

1 Frontier Molecular Orbitals (FMOs): refer to the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO) of a semiconducting molecule. These orbitals are critical to determine the molecules light absorption range, exciton dissociation driving force, charge transfer efficiency and overall device efficiency. FMOs play a central role in balancing the efficiency and stability of OPV devices [1].

2 Donor and Acceptor Materials: in organic photovoltaics are the two key components responsible for light absorption and charge generation within the active layer of a solar cell.

- The **donor material** can absorb sunlight photons and generate excitons (bound of electron-hole pairs) and it may donate free electrons to the acceptor material.
- The **acceptor material** can absorb sunlight photons and generate excitons. It may donate free holes to the donor material.

Efficient energy level alignment between the donor's and the acceptor's FMOs is essential for effective exciton dissociation and charge transfer. As a result, the proper combination of donor and acceptor materials directly influences the power conversion efficiency and operational stability of OPV devices [2].

3 **CNN (Convolutional Neural Network)** is a type of deep learning model used for both classification and prediction tasks based on image inputs. To achieve optimal performance for a specific task, it is important to select apprpriate hyperparameters, as outlined in Table S1 [3].

Table S1. CNN hyper-parameters

Term	Meaning
Learning Rate	Controls how fast the model updates its weights during training
Epochs	One full pass through the entire training dataset and more epochs allows the model to better learn from the input data
Batch Size	Number of samples processed at once before weights are updated, it impacts memory usage and model stability during training
Loss Function	Measures the error between predicted and actual values (e.g: MSE = Mean Squared Error)
Convolutional Layer	Extracts features (Feature Map) from the input data
Pooling Layer	Reduces spatial size of data to focus on key features that influence the predictions so it will speed up computation and avoids overfitting
Activation Function (e.g: ReLU, sigmoind, Linear..)	It introduces non-linearity, allowing the network to learn complex relationships in the data, and helps tailor the model for either classification or regression tasks
Dropout	Randomly disables neurons during training to prevent overfitting

References :

[1] J. Yu, N. Q. Su, and W. Yang, "Describing Chemical Reactivity with Frontier Molecular Orbitalets," *JACS Au*, vol. 2, no. 6, pp. 1383–1394, Jun. 2022, doi: 10.1021/jacsau.2c00085.

[2] K. Müllen and W. Pisula, "Donor-acceptor polymers," Aug. 05, 2015, *American Chemical Society*. doi: 10.1021/jacs.5b07015.

[3] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," Mar. 01, 2021, *Elsevier B.V.* doi: 10.1016/j.isprsjprs.2020.12.010.

Khoukha KHOUSSA

Application of Artificial Intelligence to Accelerate the Development of the Active Layer of Organic Photovoltaic Cells

Résumé

Face aux enjeux liés au changement climatique et à l'épuisement des ressources fossiles, le développement de sources d'énergie renouvelable constitue une priorité. Les cellules photovoltaïques organiques (OPVs) représentent une alternative prometteuse grâce à leur flexibilité, leur légèreté et leur potentiel de fabrication à faible coût. Toutefois, leur rendement et leur stabilité restent des défis majeurs pour une adoption à grande échelle. Dans ce contexte, les avancées en intelligence artificielle (IA), notamment en apprentissage automatique (machine learning (ML)) et en apprentissage profond (deep learning (DL)), offrent des perspectives inédites pour la découverte et l'optimisation de matériaux organique des OPVs. Cette thèse explore l'application de techniques d'IA à la prédiction des performances des OPVs à partir des structures chimiques des matériaux organique (paires Donneur/Accepteur). Nous nous sommes concentrés sur les cellules constituées d'un donneur (ou D) et d'un accepteur (ou A) d'électrons en hétérojonction volumique (Bulk Heterojunction, BHJ). Trois contributions principales sont proposées : La prédiction des performances des OPVs à partir de descripteurs moléculaires extraits des représentations SMILES des matériaux donneurs et accepteurs via l'apprentissage automatique, l'entraînement de modèle de machine learning sur deux bases de données : l'une contenant uniquement des accepteurs de type fullerène (FA) et l'autre incluant également d'autres non dérivés de fullerène (NFA) et enfin le développement d'une approche d'apprentissage profond exploitant les images 2D des structures chimiques des paires donneur/NFA. Ces travaux illustrent le potentiel de l'IA pour accélérer la conception rationnelle de matériaux performants, contribuant ainsi au progrès des technologies solaires organiques durables.

Mots clée : Dispositifs photovoltaïques organiques (OPVs), Intelligence artificielle (IA), Paires donneur/accepteur (D/A), Structure chimique, Apprentissage automatique, Apprentissage profond , Indicateurs de performance des OPVs, Prédictions

Summary

Faced with the challenges of climate change and the depletion of fossil fuels, the development of renewable energy sources is a priority. Organic photovoltaic cells (OPVs) represent a promising alternative thanks to their flexibility, light weight and low-cost manufacturing potential. However, their efficiency and stability remain major challenges for large-scale adoption. In this context, advances in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), offer novel perspectives for the discovery and optimization of organic materials for OPVs. This thesis explores the application of AI techniques to the prediction of OPV performance based on the chemical structures of organic materials (Donor/Acceptor pairs). We focus on cells consisting of a Bulk Heterojunction (BHJ) of electron donor (or D) and electron acceptor (or A) molecules. Three main contributions are proposed: prediction of OPV performance from molecular descriptors extracted from SMILES representations of donor and acceptor materials via machine learning, machine learning model training on two databases: one containing only fullerene-type acceptors (FA) and the other also including other non-fullerene derivatives (NFA) and finally the development of a deep learning approach exploiting 2D images of the chemical structures of donor/NFA pairs. This work illustrates the potential of AI to accelerate the rational design of high-performance materials, thus contributing to the progress of sustainable organic solar technologies.

Keywords: Organic PhotoVoltaic devices (OPVs), Artificial Intelligence (AI), Donor/Acceptor Pairs (D/A), Chemical structure, Machine Learning, Deep Learning, OPVs Performance Metrics, Predictions.

UNIVERSITE DE STRASBOURG

ECOLE DOCTORALE MATHEMATIQUES SCIENCES DE L'INFORMATION ET DE L'INGENIEUR

RESUME DE LA THESE DE DOCTORAT

Discipline : Informatique

Spécialité : Informatique

Présentée par :
Khoukha KHOUSSA

Titre : Application of Artificial Intelligence to Accelerate the Development of the Active Layer of Organic Photovoltaic Cells

Unité de Recherche : ICube (Le laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (UMR7357))

Directeur de Thèse :
Dr. Patrick LEVEQUE

Co-Directeur de Thèse :
Pr. Larbi BOUBCHIR

Localisation : ICube, Université de Strasbourg-CNRS, 23 rue de Loess, 67037 Strasbourg, France

Thèse confidentielle : X NON ☐ OUI

Introduction générale :

Face au changement climatique et à la raréfaction des ressources fossiles, le développement de sources d'énergie durables est devenu une priorité majeure. Les cellules photovoltaïques organiques (OPVs) constituent une alternative prometteuse pour la conversion de l'énergie solaire grâce à leur souplesse, leur faible poids et leur potentiel de fabrication à coût réduit [1]. Cependant, leur efficacité et leur stabilité restent limitées, freinant leur déploiement à grande échelle [2]. Parallèlement, l'essor de l'intelligence artificielle (IA), notamment du « machine learning » (ML) et du « deep learning » (DL), ouvre de nouvelles perspectives pour la découverte et l'optimisation de matériaux [3], [4], [5], [6], [7]. Ces approches, capables d'exploiter de vastes jeux de données et de révéler des relations complexes, offrent des outils puissants pour accélérer les avancées dans le domaine des OPVs [8], [9]. Dans cette thèse, nous nous sommes intéressés à l'application de techniques d'IA pour la prédiction des performances des cellules photovoltaïques organiques en se basant sur la structure chimique des matériaux de la couche active. Nous nous sommes focalisés sur des couches actives (voir figure 1) contenant seulement deux matériaux actifs : un semiconducteur donneur d'électrons (ou D) et un semiconducteur accepteur d'électrons (ou A) mélangés en hétérojonction volumique (*Bulk HeteroJunction*, BHJ) avec leurs orbitales moléculaire frontière. Nous avons proposé trois contributions majeures :

I. La prédiction des performances des cellules OPV à partir de descripteurs chimiques extraits des représentations SMILES (*Simplified Molecular Input Line Entry System*) des matériaux D et A, via l'apprentissage automatique.
II. L'élargissement de la base de données utilisée en entraînant les modèles de ML à partir d'une base de données contenant des accepteurs d'électrons de type dérivés de fullerène (*Fullerene Acceptors*, FA) puis à partir d'une autre qui contient les 2 types d'accepteurs FA et NFA.
III. Une approche d'apprentissage profond basée sur les images 2D de la structure chimique des paires (D/NFA).

Ces travaux visent à démontrer le potentiel de l'IA pour rationaliser et accélérer la conception de matériaux performants pour les OPVs, contribuant ainsi au développement de sources d'énergie plus durables.

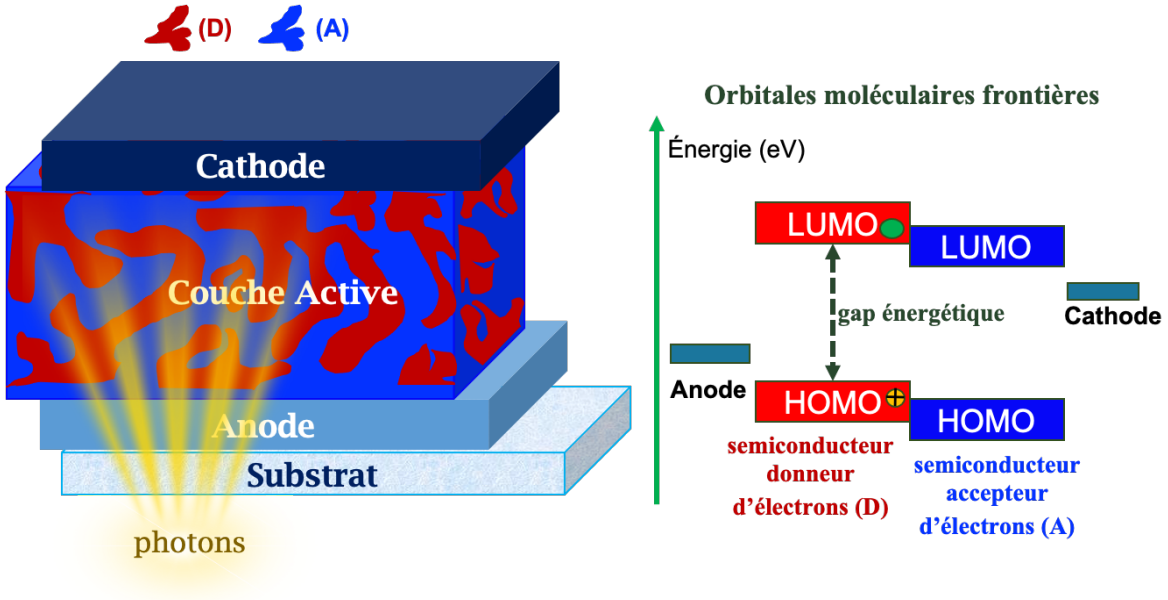


Figure 1. Structure des cellules solaire organique (OPVs)

I. Prédiction des performances à partir de descripteurs chimiques par l'apprentissage automatique.

Positionnement :

Dans cette étude, nous avons développé et évalué une approche d'apprentissage automatique (voir figure 2) visant à prédire, à partir de paires donneur/accepteur non-dérivés de fullerène (D/NFA), les principales métriques de performance des cellules photovoltaïques : la densité de courant de court-circuit ou J_{sc} (*short-circuit current-density* ayant pour unité les mA/cm^2), la tension en circuit ouvert ou V_{oc} (*open-circuit voltage* qui s'exprime en *V*) et l'efficacité de conversion ou *PCE* (*Power Conversion Efficiency* en *%*). Ce travail a pour objectif de réduire significativement le coût et le temps de développement par rapport aux méthodes empiriques classiques. Notre contribution se distingue par l'utilisation conjointe de descripteurs chimiques expérimentaux, extraits directement des représentations SMILES pour des paires uniques et par une évaluation rigoureuse de plusieurs modèles de Machine Learning (ML) optimisés via l'ajustement d'hyperparamètres. De plus, nous avons constitué un jeu de données enrichi, générant toutes les combinaisons possibles de paires (D/NFA), y compris celles encore inexplorées dans la littérature, ce qui nous a permis d'identifier de nouvelles combinaisons prometteuses présentant un fort potentiel en termes de performances photovoltaïques.

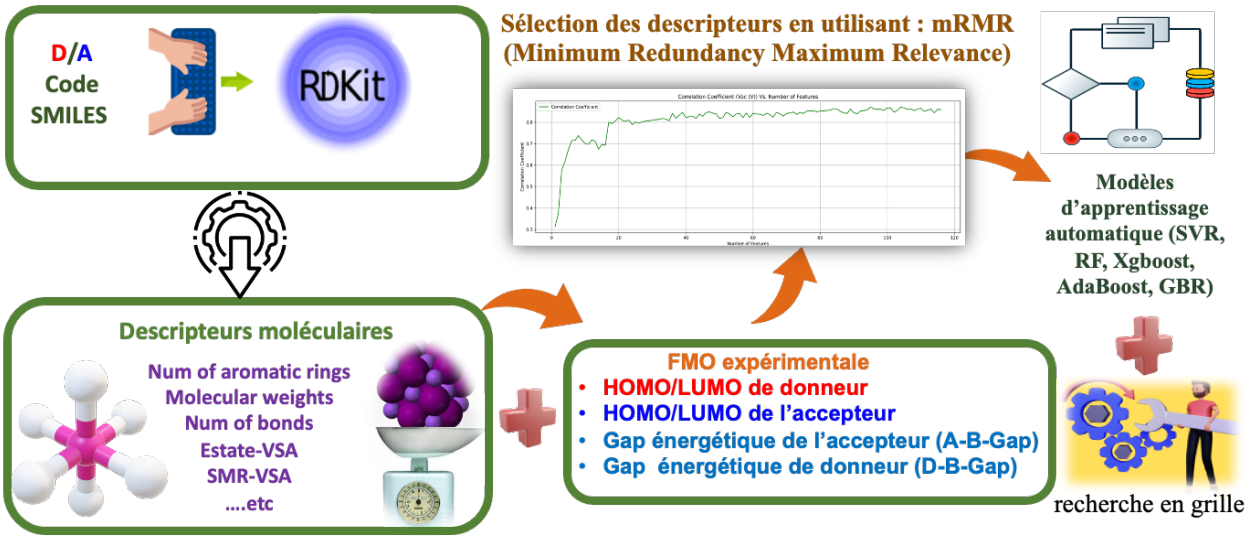


Figure 2. Notre approche d'apprentissage automatique.

Données utilisées et prétraitement :

Un jeu de données riche de 1225 paires (D/NFA) issues de la littérature scientifique a été compilé [10], contenant les principales métriques photovoltaïques (V_{oc}, J_{sc}, *PCE*). Après un important travail de nettoyage et de standardisation incluant la vérification de la cohérence des *PCE*, l'unification des noms commerciaux et la sélection des meilleures performances par paire le jeu final comporte 924 paires uniques annotées avec les codes SMILES (data_1) et, si disponibles, les niveaux énergétiques frontières expérimentaux (FMO pour *Frontier Molecular Orbitals*) (data_2). Plus de 100 descripteurs chimiques ont été extraits à l'aide de la bibliothèque RDKit en utilisant les code SMILES. Le jeu de données final est homogène, riche en informations chimiques et prêt à être utilisé comme entrée pour les modèles d'apprentissage automatique.

Approche méthodologique :

Dans cette phase de l'étude, nous avons mis en œuvre cinq algorithmes d'apprentissage automatique pour prédire les performances photovoltaïques à partir des descripteurs chimiques extraits précédemment : *Random Forest* (RF), *Support Vector Regression* (SVR), *Gradient Boosting Regressor* (GBR), *XGBoost* et *AdaBoost*. Afin d'optimiser les performances de chaque modèle, une recherche systématique des hyperparamètres a été réalisée à l'aide de la méthode *Grid Search*. Par ailleurs, une étape de sélection de variables a été intégrée au processus à l'aide de la méthode mRMR (*Minimum Redundancy Maximum Relevance*), ce qui nous a permis d'identifier un sous-ensemble optimal de 20 descripteurs offrant un très bon compromis entre complexité et précision. L'efficacité des modèles a été évaluée à travers deux approches de validation : une division 80% - 20% du jeu de données en ensembles respectivement d'entraînement et de test, ainsi qu'une validation croisée *Leave-One-Out* (LOOCV), garantissant une évaluation rigoureuse et robuste de la capacité prédictive des modèles développés. Dans la continuité de ce travail, nous avons exploité

les combinaisons de paires (D/NFA) générées à partir de notre base de données initiale pour les entrainer en utilisant le modèle *Gradient Boosting Regressor* (GBR). L'objectif était d'identifier de nouvelles combinaisons potentiellement performantes, encore non explorées dans la littérature offrant des performances théoriquement supérieures à celles observées jusqu'à présent, ouvrant ainsi la voie à de futures expérimentations ciblées.

Résultats et discussion :

Les meilleurs résultats ont été obtenus avec le modèle *Gradient Boosting Regressor*. Avec celui-ci, comme le montre la **Figure 1** pour la prédiction de V_{oc} à partir des seuls descripteurs (data_1), le modèle a atteint un R^2 de 0,68, un coefficient de corrélation r de 0,83 et un RMSE de 0,052 V. L'ajout des FMOs expérimentales (data_2) a permis d'améliorer ces performances ce qui est logique pour V_{oc} avec un R^2 de 0,78, un r de 0.88 et un RMSE de 0,045 V, des valeurs qui sont comparable avec l'état de l'art. Des résultats similaires ont été obtenus pour la prédiction de J_{sc} avec un R^2 de 0,67 (data_1) et 0,70 (data_2), un r de respectivement 0,82 et 0,83. Pour J_{sc} la valeur de RMSE est de 3,33 mA/cm² (data_1) contre 3,19 mA/cm² (data_2). Pour le *PCE*, les prédictions ont donné un R^2 de 0,61, un r de 0,78 et une RMSE de 3,3 % avec data_1 et un R^2 de 0,63, un r de 0,79, et un RMSE de 2,3 % avec data_2. Le LOOCV a confirmé la robustesse des modèles avec des RMSE faibles pour V_{oc} : 0,042 V (data_1) et 0,044 V (data_2). Enfin, une comparaison directe avec les travaux de Greenstein *et al.* [10] sur des dispositifs à PCE > 10 % a mis en évidence la supériorité de notre modèle GBR, obtenant un R^2 de 0,46 et une RMSE de 1,38 %, contre respectivement 0,28 et 1,6 % pour le modèle RF de Greenstein *et al.* [10]. Cette comparaison directe démontre la pertinence et l'efficacité de notre approche pour la prédiction des performances des OPVs. Nous avons également évalué notre modèle sur un ensemble de données non incluses dans le jeu initial et avons observé qu'il maintenait de bonnes performances prédictives. Par exemple, la paire (D18/Y6-Se) était absente des données d'entraînement mais a été étudiée expérimentalement par la suite. Les prédictions de notre modèle GBR sur cette paire prédit V_{oc} avec une incertitude relative inférieure à 7 % et J_{sc} et *PCE* avec moins de 1 % d'erreur relative. De plus, notre modèle a permis d'identifier d'autres paires prometteuses, comme (D18/CH1007) et (PBQx-TF/CH1007) avec des *PCEs* prédits supérieurs à 16 %, suggérant leur fort potentiel pour de futures applications en photovoltaïque organique à base de NFAs. Tous ces résultats démontrent la fiabilité de notre modèle, même pour des combinaisons inédites et la capacité de généralisation de notre modèle. En résumé, notre approche est fiable et peut aider à accélérer le développement des matériaux utilisés dans la couche active des OPVs.

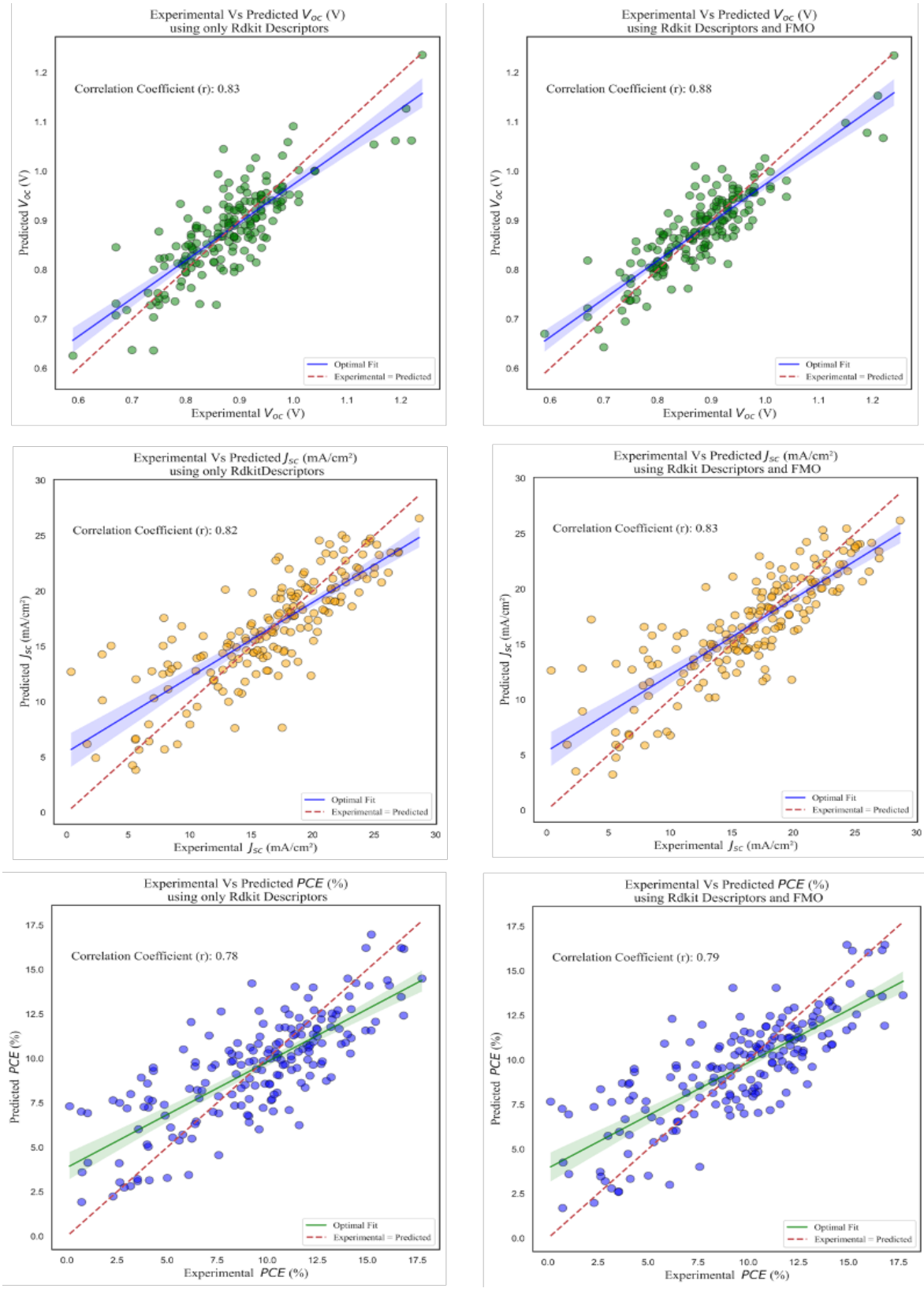


Figure 1. Valeurs expérimentales comparées aux valeurs prédites de V_{oc} (haut), J_{sc} (milieu) et *PCE* (bas) obtenues à l'aide du modèle GBR, utilisant les descripteurs chimiques des paires (D/A) (à gauche) et les descripteurs chimiques combinés aux FMOs des paires (D/A) (à droite).

II. Entraînement de modèle ML à partir d'une base de données contenant des accepteurs de type dérivés de fullerènes (FA) et à partir d'une autre qui contient les 2 types d'accepteurs FA et NFA.

Positionnement :

Dans ce travail, nous adoptons une approche complémentaire en nous concentrant d'abord sur les combinaisons donneur/accepteur dérivés de fullerènes (D/FA). Cette étude se déroule en deux étapes : premièrement, un modèle, le GBR, est entraîné uniquement avec des données (D/FA). Cela permet d'analyser l'influence spécifique des accepteurs dérivés de fullerènes sur l'efficacité des OPVs à partir d'un jeu de données dédié. Deuxièmement, les données (D/FA) sont combinées au jeu de données plus large (D/NFA), formant ainsi un ensemble incluant les deux types d'accepteurs. Le modèle GBR est alors réentraîné sur ce nouvel ensemble pour examiner si l'intégration des accepteurs dérivés de fullerènes améliore la précision, la généralisation du modèle, ou modifie l'importance des descripteurs moléculaires clés. Contrairement aux travaux antérieurs, notre contribution apporte une meilleure diversité des données en utilisant un jeu de données expérimental combinant à la fois NFA et FA, ces derniers étant particulièrement pertinents puisqu'ils étaient les seuls accepteurs utilisés avant l'apparition plus récente des NFAs. Cette double approche vise à révéler la capacité prédictive autonome des systèmes (D/FA) ainsi que leur contribution lorsqu'ils sont intégrés à des données plus diversifiées, offrant ainsi de nouvelles perspectives pour la sélection des matériaux et la modélisation des performances des OPV.

Données utilisées et prétraitement :

Dans cette étude, nous avons utilisé un jeu de données open source comprenant 249 paires (D/FA) expérimentales pour des systèmes photovoltaïques organiques qui après le prétraitement (suppression des données manquantes et ajouts des FMOs expérimentales si disponible) a donné 193 paires uniques, issues de la littérature scientifique publiée entre 2013 et 2017 et compilées par Padula en 2019 [11]. Ce jeu de données inclut principalement des cellules solaires BHJ ainsi que quelques dispositifs en bicouches et contient des métriques photovoltaïques expérimentales (V_{oc}, J_{sc} et *PCE*). Le jeu de données se concentre exclusivement sur des accepteurs dérivés de fullerènes (C60, PC61BM, PC71BM) reflétant la pratique expérimentale, courante avant l'introduction des NFAs, consistant à varier les donneurs tout en fixant l'accepteur. L'analyse montre une prédominance de PC71BM, utilisé dans près de la moitié des cas, due à ses propriétés optiques favorables par rapport à PC61BM qui absorbe très peu dans le visible. Les paramètres photovoltaïques présentent une large gamme de valeurs, avec des efficacités de conversion généralement modestes (0 à 7 %). Afin d'accroître la diversité, les données sur les accepteurs dérivé de fullerènes ont été combinées avec celles sur les accepteurs non dérivés de fullerènes, formant un

ensemble étendu de 1111 paires. Cette combinaison révèle que les dispositifs à accepteurs dérivés de fullerènes affichent en moyenne des performances inférieures à celles des NFAs, conformément à l'efficacité supérieure de ces derniers pour l'absorption dans les longueurs d'onde du spectre solaire.

Approche méthodologique :

Pour décrire les structures chimiques, plus de 100 descripteurs moléculaires pertinents ont été extraits à l'aide de RDKit en suivant la même méthodologie que celle utilisée dans nos précédents travaux. Ces descripteurs couvrent un large spectre de propriétés physico-chimiques, structurelles et électroniques essentielles pour prédire les performances des dispositifs photovoltaïques organiques. Pour la prédiction, nous avons utilisé le *Gradient Boosting Regressor* (GBR), une méthode d'apprentissage robuste, reconnue pour sa précision et sa résistance au surapprentissage comme remarqué lors de la première partie de nos travaux (I). Le modèle a été optimisé via une recherche exhaustive des hyperparamètres (*Grid Search*) adaptée pour maximiser la performance prédictive sur l'ensemble des paramètres cibles.

Résultats et discussion :

Les performances prédictives du modèle GBR ont été évaluées via une séparation *train/test* (80%/20%), avec deux stratégies d'entraînement : l'une utilisant uniquement le jeu de données FA, l'autre combinant les jeux FA et NFA pour augmenter la diversité chimique. Le modèle entraîné uniquement sur les données FA a montré une précision compétitive, notamment pour *PCE* avec un RMSE de 1,51 % et un coefficient de corrélation (r) de 0,69, comparable aux résultats de Padula *et al.* [11]. Pour V_{oc}, RMSE était de 0,07 V avec un r de 0,62, tandis que pour J_{sc}, le modèle a atteint un r de 0,72 et un RMSE de 2,36 mA/cm^2. Toutefois, la taille limitée et la faible diversité chimique du jeu FA restreignent la généralisation du modèle. En intégrant les données NFA dans un jeu combiné, le modèle réentraîné montre une nette amélioration sur toutes les cibles prédictives : pour *PCE*, RMSE est de 2,27 % avec un r accru à 0,85 alors que pour V_{oc}, RMSE reste stable à 0,07 V tandis que r passe à 0,83, pour J_{sc}, le modèle atteint une valeur de RMSE de 3,10 mA/cm² et un r de 0,88. Ces résultats démontrent que la diversité accrue des données améliore la robustesse et la fiabilité des prédictions, malgré une légère augmentation des erreurs liée à une plus grande variabilité des données photovoltaïques. Ce compromis souligne l'importance de la diversité chimique pour optimiser les modèles prédictifs pour le photovoltaïque organique.

III. Approche d'apprentissage profond basée sur les images 2D des paires (D/NFA).

Positionnement :

Dans un autre travail montré en figure 3, nous avons exploré une approche alternative basée sur l'utilisation d'images 2D représentant les structures moléculaires des D et des NFA en remplacement des descripteurs classiques extraits par calculs ou outils dédiés. Cette méthode exploite les réseaux de neurones convolutifs (CNN), une architecture de *Deep Learning* (DL) particulièrement adaptée à l'analyse d'images, afin d'apprendre directement les caractéristiques pertinentes pour la prédiction des performances photovoltaïques à partir des données expérimentales seules. Cette stratégie présente l'avantage de ne pas nécessiter l'utilisation préalable des orbitales frontalières moléculaires (FMO) ni d'une étape de sélection de descripteurs, puisque le CNN réalise automatiquement l'extraction des caractéristiques importantes au cours de l'entraînement. Ainsi, cette approche simplifie le « pipeline » de modélisation tout en explorant la capacité du DL à améliorer la précision prédictive sans pour autant dépendre de l'apport empirique de données supplémentaires comme les FMOs.

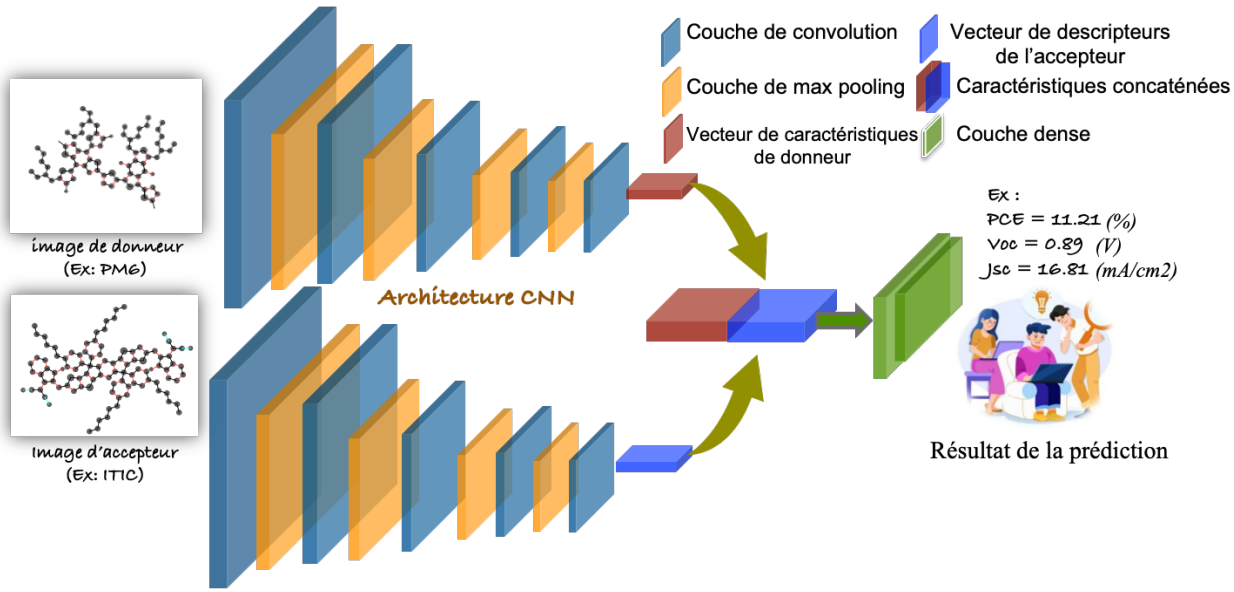


Figure 3. Notre approche de l'apprentissage profond

Données utilisées et prétraitement :

Le jeu de données utilisé dans cette étude est identique à celui exploité dans nos travaux précédents. Cependant, contrairement aux approches antérieures basées sur l'extraction de descripteurs moléculaires classiques, nous avons généré des représentations visuelles 2D des molécules à partir de leurs codes SMILES. À cet effet, chaque chaîne SMILES a été convertie en un objet de type

molécule via RDKit. Nous avons ensuite produit des coordonnées 2D planaires pour chaque atome des molécules afin de créer des images standardisées. Pour aller au-delà des visualisations symboliques classiques de RDKit, nous avons développé un rendu personnalisé des structures moléculaires. Chaque atome est ainsi représenté par un cercle dont le rayon est proportionnel au rayon atomique, permettant de mieux percevoir la diversité chimique des atomes constituant les molécules. Les liaisons chimiques sont codées en couleur selon leur type : noir pour une simple liaison, rouge pour une double liaison et cyan pour une triple liaison. Par ailleurs, afin d'optimiser le coût de calcul et évaluer l'impact des chaînes latérales solubilisantes, deux versions d'images ont été créées : une incluant les chaînes latérales complètes, et une autre ne représentant que le squelette conjugué principal. Enfin, les images ont été normalisées à une résolution uniforme de 80 pixels par dimension avec des dimensions finales de (259 × 480) pixels, garantissant la cohérence nécessaire pour l'entraînement de réseaux de neurones convolutifs (CNN). Ce prétraitement aboutit à un jeu de données visuel prêt pour une analyse DL et offrant une nouvelle manière d'exploiter les données expérimentales sans recourir à l'extraction de descripteurs.

Approche méthodologique :

Chaque image est traitée séparément par un CNN identique, permettant d'extraire automatiquement des caractéristiques structurales et électroniques, sans recourir à des descripteurs chimiques manuels. Les vecteurs issus des deux CNN sont ensuite fusionnés et passés dans des couches entièrement connectées pour estimer les métriques de performance. L'ensemble du modèle a été entraîné uniquement sur des CPU avec seulement 16 Go de RAM, démontrant une approche légère et accessible. Une seconde variante a été testée en combinant les images avec les valeurs expérimentales des FMOs en les ajoutant au vecteur 1D qui contient toutes les descripteurs extraits à partir des images normalisées des molécules D et A. Cet ajout a pour but d'évaluer l'apport de descripteurs numériques sur les performances du modèle.

Résultats et discussion :

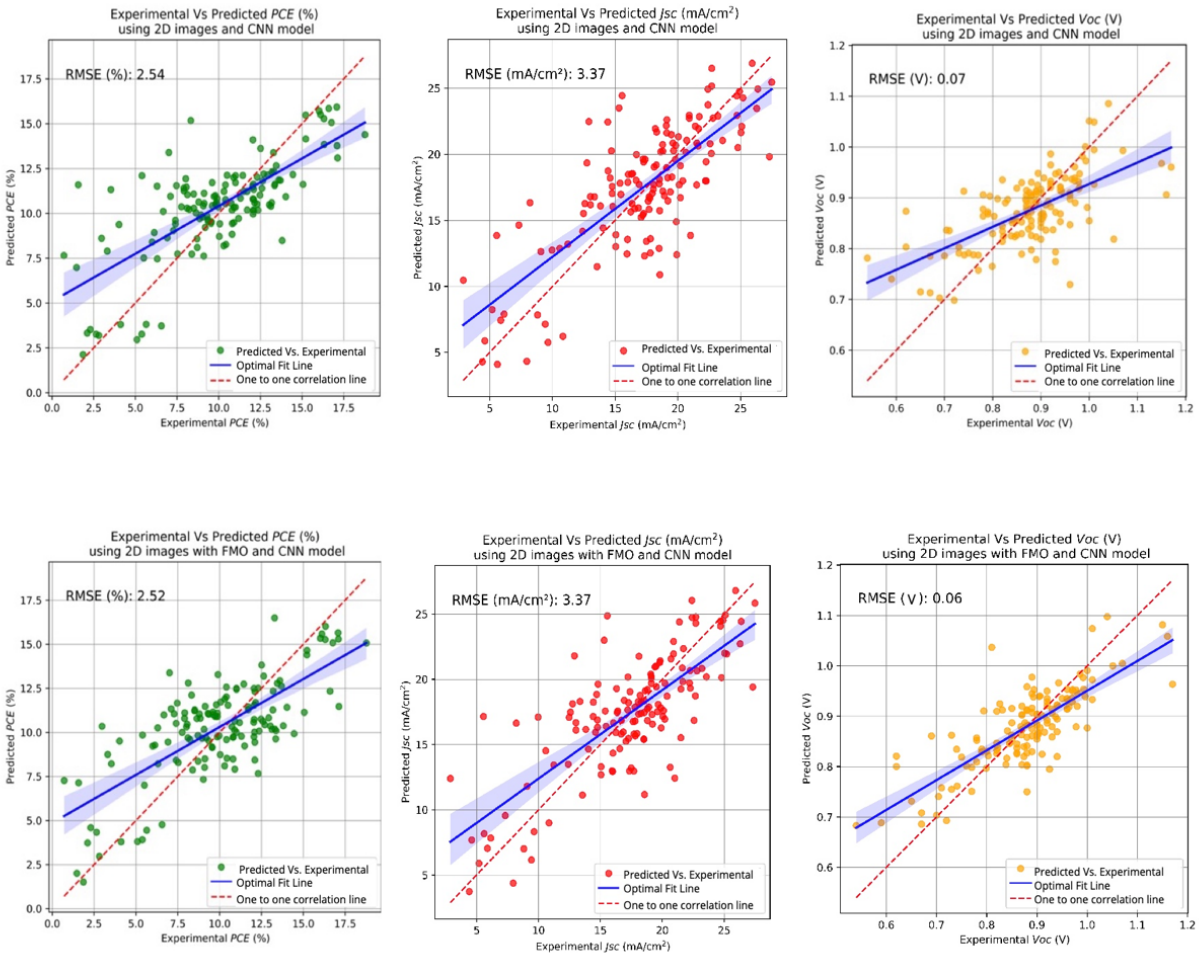


Figure 2. Résultats de la prédiction : en utilisant des images 2D et un réseau de neurones convolutif (haut) et en utilisant des images 2D associées aux FMO et un CNN (bas).

L'optimisation du modèle CNN a été réalisée par un ajustement ciblé des hyperparamètres, notamment le taux d'apprentissage (*Learning Rate*, LR) et la taille du noyau (*Kernel Size*). Nos résultats démontrent une approche efficace et peu coûteuse. Même avec seulement 20 époques, le modèle a montré une bonne capacité de convergence. En se basant uniquement sur des images 2D des molécules, le CNN a atteint des performances comparables, voire supérieures, à celles obtenues dans nos travaux précédents utilisant des descripteurs chimiques. Comme montre la **Figure 2**, pour la prédiction du *PCE*, le modèle a obtenu un RMSE de 2,53 %, un coefficient de corrélation r de 0,73 et un R^2 de 0,54. L'ajout des valeurs des FMOs a légèrement amélioré la précision des prédictions de V_{oc}, mais a eu peu d'effet sur la précision des prédictions de *PCE*.

Ces résultats confirment que les CNN peuvent extraire automatiquement les caractéristiques pertinentes à partir des images, sans nécessiter de sélection de descripteurs, ni de techniques de réduction de dimension comme mRMR. Cette approche montre ainsi le potentiel des réseaux convolutifs pour la prédiction des performances photovoltaïques à partir de simples représentations visuelles des molécules, ouvrant la voie à un criblage plus rapide et plus accessible des matériaux.

Conclusion :

Ces recherches démontrent l'apport de l'intelligence artificielle dans la prédiction des performances des cellules photovoltaïques organiques, en s'appuyant sur des descripteurs chimiques ou des images 2D des semiconducteurs organiques de la couche active. Elles ouvrent des perspectives prometteuses pour la conception accélérée de nouveaux matériaux organiques semiconducteurs utilisés dans les OPVs.

Nos travaux :

- Khoukha KHOUSSA, Yves-André CHAPUIS, and Nicolas LACHICHE. "Optimisation de matériaux et dispositifs pour l'énergie à partir de concepts d'intelligence artificielle pour small data." *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2023)*. Strasbourg, France. hal-04944011. (Publié)

- Khoukha KHOUSSA, Patrick LEVEQUE, and Larbi BOUBCHIR. "On the use of Machine Learning to Discover Novel Donor-Acceptor Pairs For Organic Photovoltaic Devices." *2024 IEEE International Conference on Big Data (BigData)*. **15** (2024). pp. 4659-4662. doi : 10.1109/BigData62323.2024.10825948 (Publié)

- Khoukha KHOUSSA, Larbi BOUBCHIR and Patrick LEVEQUE. " Artificial Intelligence in Organic Photovoltaics: Predicting Power Conversion Efficiency From the Molecular Chemical Structure of (Donor/Acceptor) Pairs," *Engineering Reports* 7, no. 8 (2025): e70334, doi: 10.1002/eng2.70334 (Publié)
- Khoukha KHOUSSA, Patrick LEVEQUE, and Larbi BOUBCHIR. "Deep Learning Approach for Predicting Efficiency in Organic Photovoltaics from 2D Molecular Images of D/A pairs". 2025 (submitted to *Advanced theory and simulation,* Under revision)

Références :

[1] Y.-W. Su, S.-C. Lan, and K.-H. Wei, *Mater. Today*, **15** (2012), p. 554–562. doi: 10.1016/S1369-7021(13)70013-0.

[2] W. B. Tarique, and A. Uddin, *Mater. Sci. Semicond.* Process. **163** (2023), 107541. doi: 10.1016/j.mssp.2023.107541.

[3] S. Bhatti, H. U. Manzoor, B. Michel, R. S. Bonilla, R. Abrams, A. Zoha, S. Hussain, and R. Ghannam, *arXiv*, (2022), 2212.13893. Available: http://arxiv.org/abs/2212.13893.

[4] Y. Zhao, R. J. Mulder, S. Houshyar, and T. C. Le, *Polym. Chem.,* **14** (2023), p. 4213–4236. doi: 10.1039/d3py00395g.

[5] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, and A. Tropsha, *Chem. Soc. Rev*., **49** (2020), 3525–3564. doi: 10.1039/d0cs00098a.

[6] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, *WIREs Comput. Mol. Sci*., 12 (2022), e1608. doi: 10.1002/wcms.1608.

[7] G. Kumar, and F. C. Chen, *J. Phys. D: Appl. Phys*., **56** (2023), 503001. doi: 10.1088/1361-6463/acd2e5.

[8] M. Jeong, J. F. Joung, J. Hwang, M. Han, C. W. Koh, D. H. Choi, and S. Park, *NPJ Comput. Mater*., 8 (2022), 206. doi: 10.1038/s41524-022-00834-3

[9] G. J. Moore, O. Bardagot, and N. Banerji, *Adv. Theory Simul*., **5** (2022), 2100511. doi: 10.1002/adts.202100511.

[10] B. L. Greenstein and G. R. Hutchison, *J. Phys. Chem. C*, **127** (2023), 6179–6191. doi: 10.1021/acs.jpcc.3c00267.

[11] D. Padula, J. D. Simpson, and A. Troisi, *Mater. Horiz*., **6** (2019), 343–349. doi: 10.1039/c8mh01135d.