AI-Assisted Inverse Design of Two-Dimensional Hybrid Perovskites

Yongxin Lyu

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Materials Science and Engineering

Faculty of Science

June 2025

Abstract

Artificial intelligence (AI) -assisted workflows have transformed materials discovery, enabling rapid exploration of chemical spaces of various material systems. Two-dimensional (2D) hybrid perovskites represent an exciting frontier, and their extraordinary optoelectronic properties can be largely attributed to the versatile choices of organic spacers. However, current efforts to design 2D perovskites rely heavily on trial-and-error and expert intuition approaches, leaving the majority of chemical space unexplored.

This thesis introduces an inverse design workflow specifically designed for Dion-Jacobson perovskites, pivoting on an invertible fingerprint representation for millions of conjugated diammonium organic spacers. A molecular morphing approach was employed to expand the chemical space of organic spacers, which were then evaluated using high-throughput density functional theory (DFT) calculations to determine the energy levels of both the organic and inorganic components in hypothetical perovskite structures. These datasets formed the basis for training various machine learning models, which not only accelerated energy level predictions but also revealed the underlying physical insights between molecular fingerprints and energy levels. Furthermore, a synthesis feasibility screening funnel was developed based on the synthetic accessibility of organic molecules and the formability of 2D structures. Using the above workflow, we inverse-designed new organic spacer candidates with deterministic band alignment between the organic and the inorganic motifs in the 2D hybrid perovskites.

These results highlight the power of integrating invertible, physically meaningful representations into AI-assisted design. By streamlining the property-driven design of synthesizable materials, this framework provides a scalable and efficient pathway for navigating the chemical space of 2D hybrid perovskites. Beyond its immediate applications to perovskites, the methodology demonstrated herein offers a broadly applicable paradigm for the AI-assisted discovery and design of advanced materials, paving the way for future innovations in materials science and technology.

Acknowledgement

Looking back on my PhD journey, I feel incredibly lucky—not just for the experience itself, but for all the amazing people I met along the way and those who helped me get here in the first place.

First and foremost, I would like to express my deepest gratitude to Prof. Tom Wu, my primary supervisor. Thank you for all the time and effort you put into my project—every revision, every piece of feedback, and all the collaborations you helped arrange. This project wouldn't have been possible without you. One of the biggest lessons I've learned from you is the importance of critical thinking—questioning research, challenging assumptions, and pushing for a deeper understanding. I hope to carry that mindset with me throughout my career.

I also want to thank Prof. Jianhua Hao, my MPhil supervisor, especially your support during the beginning of my PhD when I was working remotely. And a special shoutout to Prof. Ran Ding, my former MPhil group member—you were the one who first got me interested in perovskite research, which ultimately led me here.

To my research group members and everyone I met in the School of Materials Science and Engineering and School of Chemistry, thank you for your kindness, encouragement, and all the little moments of support along the way. A special thanks to Alan, for sharing your valuable insights with me. And to the 2024 PGSOC members—thanks for welcoming me into the community.

A massive thank you to Prof. Mira Kim—a mentor, a friend, and the person who introduced me to the PELE community. Your encouragement has continuously pushed me to step outside my comfort zone, and your constant support has meant the world to me. To all the tutors, mentors, and wonderful friends I met through PELE, thank you for making this journey even more meaningful.

Keeping some sense of balance throughout this PhD was crucial, and for that, I have to thank the yoga and Pilates instructors at the UNSW Fitness Centre, as well as the wonderful friends I made there. Yoga became a huge part of my life, giving me some much-needed clarity and calm during this rollercoaster of PhD journey.

On the technical side, I want to acknowledge the national supercomputers, Gadi and

Setonix, along with ResTech at UNSW for providing computing resources, responsive support, and training sessions that helped me sharpen my coding skills.

I would also like to acknowledge the Australian Government Research Training Program (RTP) Scholarship for its financial support, which made this PhD journey possible.

Finally, I want to express my heartfelt gratitude to my family. To my parents, thank you for always supporting my dreams, even though we are separated by thousands of miles—I miss you deeply. And to my cousin, Zefang, I am grateful for our shared reflections on the struggles of PhD life—wishing you all the best in your own journey.

To my husband, Sa, who has tirelessly (but unsuccessfully) tried to convince me that chemistry is superior to materials science. Sharing this PhD journey with you has meant celebrating each other's victories, navigating the struggles together, and always having a teammate through the ups and downs. Thank you for the countless afternoon coffee runs, the nerdy scientific debates, and for making this experience not just bearable, but truly enjoyable. Most of all, thank you for being my rock, my study buddy, and my greatest supporter every step of the way.

Table of Contents

List of Figures

List of Tables

Abbreviations

CBM	Conduction Band Minimum
DFT	Density Functional Theory
DJ	Dion-Jacobson
DL	Deep Learning
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
ML	Machine Learning
PCA	Principle Component Analysis
RMSE	Root Means Square Error
RP	Ruddlesden-Popper
SHAP	SHapley Additive exPlanation
SMARTS	SMiles ARbitrary Target Specification
SMILES	simplified molecular-input line-entry system
SOC	Spin-orbital coupling
STEI	Steric Hindrance Index
SVR	Support Vector Regression
t-SNE	t-distributed tochastic neighbour embedding
VAE	Variational Autoencoder
VBM	Valence Band Maximum

Chapter 1

Introduction

1.1 Background

Recent advances in artificial intelligence (AI) have revolutionized materials discovery, enabling researchers to explore vast chemical and structural spaces with unprecedented efficiency. Traditional experimental and computational methods for materials design are often time-consuming and resource-intensive, making AI-driven approaches particularly attractive. By leveraging machine learning (ML) techniques, researchers can predict material properties, optimize design parameters, and identify promising candidates for various applications[1]. Among these AI-driven strategies, inverse design has emerged as a transformative approach that reverses the conventional forward design process, allowing the direct discovery of materials with targeted properties[2]. This approach employs generative models, optimization algorithms, and invertible material representations to streamline the discovery process across diverse material domains, including inorganic crystals, high-entropy alloys, organic semiconductors, and metal-organic frameworks[3]–[5].

Two-dimensional (2D) hybrid perovskites represent an exciting frontier for AI-assisted inverse design due to their structural tunability and unique optoelectronic properties. Compared to their three-dimensional (3D) counterparts, 2D perovskites offer a signifi-

cantly larger design space owing to the incorporation of organic cation spacers. Among the various 2D perovskite phases, the Dion-Jacobson (DJ) phase has attracted significant attention for its distinctive structural features, including the presence of diammonium organic spacers and the absence of van der Waals gaps. These properties contribute to enhanced charge transport and stability, making DJ-phase perovskites particularly promising for optoelectronic applications such as photovoltaics and light-emitting diodes (LEDs). However, the rational design of organic spacers in DJ-phase perovskites remains a major challenge due to the vast chemical space and the complex interplay between molecular structure and electronic properties[6].

Despite recent progress in AI-assisted workflows for hybrid perovskites, most studies have focused on forward design approaches that rely on exhaustive searches within predefined chemical spaces[7], [8]. These methods often prioritize formability and stability while overlooking critical properties such as energy level alignment, which directly influences charge transport and device performance. Given the quantum-well-like structure of 2D perovskites, where organic and inorganic layers possess distinct electronic properties, understanding and optimizing energy level alignment is crucial for advancing these materials in optoelectronic applications[9]. Addressing this challenge requires a systematic and AI-driven inverse design methodology tailored to hybrid perovskites.

1.2 Research Objectives

The primary objective of this research is to develop an AI-assisted inverse design framework for discovering new organic spacers in DJ-phase hybrid perovskites, with a particular focus on optimizing energy level alignment. The specific goals of this study are:

- To explore AI-driven approaches for materials discovery and evaluate their applicability to 2D hybrid perovskites.
- To investigate inverse design methodologies for identifying organic spacers in 2D perovskites with tailored electronic properties.

- To examine the impact of organic spacers on the electronic structure and synthesize feasibility of DJ-phase hybrid perovskites.
- To inverse design new organic spacer candidates to achieve the targeted energy level alignment of DJ perovskites

This research aims to bridge the gap between AI-assisted design and hybrid perovskite discovery, providing a systematic approach for accelerating materials innovation through inverse design principles.

1.3 Thesis Structure

This thesis is structured into six chapters as follows:

Chapter 1 (Introduction) sets the context for AI-assisted materials discovery, emphasizing the importance of inverse design within hybrid perovskites. It also states the central research objectives and outlines the scope of this work.

Chapter 2 (Literature Review) surveys recent advances in AI-driven materials research and explores various inverse design methodologies. It then examines the structural fundamentals of 2D perovskites—particularly Dion–Jacobson (DJ) perovskites—and highlights why they provide a compelling platform for investigating organic spacer design. Finally, it reviews current AI techniques used for 2D perovskite discovery and underscores the value of pursuing inverse design strategies in this domain.

Chapter 3 (Methodology) presents the overall inverse design framework. It details the development of invertible molecular fingerprints, molecular morphing, high-throughput calculations, machine learning models for energy level prediction, and synthesis feasibility screening.

Chapter 4 (High-Throughput Calculation and Machine Learning Predictions) details the expansion of the chemical space through molecular morphing, visualization of generated

structures, ML model selection and evaluation, and interpretation of structure-property relationships.

Chapter 5 (Synthesis Feasibility Screening and Final Candidate Validation) explains the screening protocol for synthetic accessibility and examines the structural formability of 2D layers. It also focuses on identifying the fingerprint features that correlate with specific energy alignment types and concludes with DFT-based validation of the most promising DJ-phase perovskite candidates.

Chapter 6 (Summary and Outlook) summarizes the key findings of the research and provides an outlook on future directions for AI-assisted hybrid perovskite design.

Chapter 2

Literature Review

This chapter reviews recent advancements and current challenges in two key areas relevant to this work: AI-assisted materials discovery (Section 2.1) and the design of 2D perovskites (Section 2.2). Particular emphasis is placed on inverse design methodologies within AI-driven approaches and their applicability to the discovery and optimization of 2D perovskite materials. Section 2.3 concludes the chapter by summarizing the literature and identifying key research gaps that motivate the development of the inverse design framework presented in this thesis.

2.1 Introduction to AI in Materials Science

As introduced in Chapter 1, AI has become a transformative tool in materials discovery, often referred to as the "fourth paradigm" of science—complementing experimental, theoretical, and computational approaches (Figure 2.1). This section provides an overview of various AI and machine learning (ML) techniques applied in materials science, with a particular focus on their role in accelerating the discovery of novel materials.

The applications of AI in materials science are diverse and rapidly expanding. AI serves as an overarching framework encompassing various concepts, with machine learning (ML)

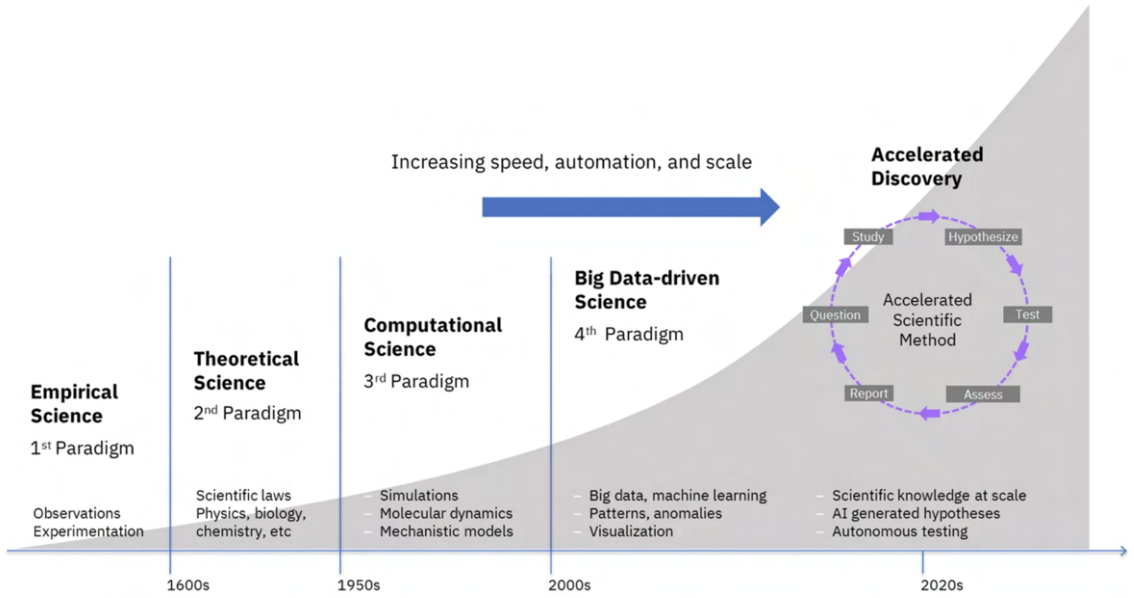


Figure 2.1: The four paradigms of scientific discovery[10].

as a key subset specifically applied to materials science. At its core, ML is utilized to identify complex relationships that are either too intricate or computationally expensive for traditional analytical methods[11]. Key applications include:

1. Learning structure-property relationships: ML models can predict material properties directly from structural information[12]–[14]. For example, deep learning models have been employed to learn the relationship between the structure of organic molecules encoded in fingerprints and their emission characteristics in light-emitting diodes[15]. For inorganic materials, deep learning model has been used to model the relationship between the composition of high-entropy alloys and their thermal expansion coefficients[16].

2. Optimizing synthesis parameters: AI can elucidate reaction mechanisms and optimize experimental conditions for complex chemical reactions[17]–[20]. Supervised ML models have been trained to correlate synthesis parameters with reaction outcomes, enabling the efficient design of experiments and reducing the trial-and-error typically associated with materials synthesis[21].

3. Accelerating computational methods: AI models can reduce the computational cost

of traditional simulation techniques. For instance, active learning models have been developed to perform quantum mechanical calculations using small to medium-sized molecular building blocks instead of individual atoms, as in first-principles methods. These models have demonstrated faster and reliable predictions across diverse material systems and target properties[22].

Given the broad range of AI applications in materials science, this literature review chapter focuses specifically on machine learning for materials discovery. This domain often encompasses one or more of the scenarios discussed above, with the primary goal of identifying novel or previously unexplored materials that exhibit superior properties compared to existing ones. In the following sections, we provide a systematic framework to clarify the various ML methods employed in materials discovery, addressing the diverse terminologies commonly found in the literature. Additionally, we will discuss how these ML methods can be integrated into a comprehensive machine learning workflow and how they can be effectively combined with other data-driven approaches to enhance materials research.

2.1.1 Types of ML methods used for materials discovery

An overview of commonly used ML methods is summarized in Figure 2.2. Within ML, approaches that rely on manually engineered features and relatively simple model architectures are referred to as classical ML methods, with supervised learning being one of the earliest and mostly widely used techniques. The first notable application of supervised learning in materials science dates back to 2010, where a probabilistic model is built to identify the most probable crystal structures of unseen ternary oxide compounds[23]. Classical ML methods remain widely used, particularly when dealing with smaller datasets (typically below thousands of data points) or when model interpretability is a key priority. Meanwhile, deep learning represents a specialized branch of ML that employs multi-layered neural networks capable of automatically extracting features from raw data. Unlike classical ML methods, deep learning does not require manual feature engineering, as it can learn hierarchical representations directly from input data. Deep learning methods have

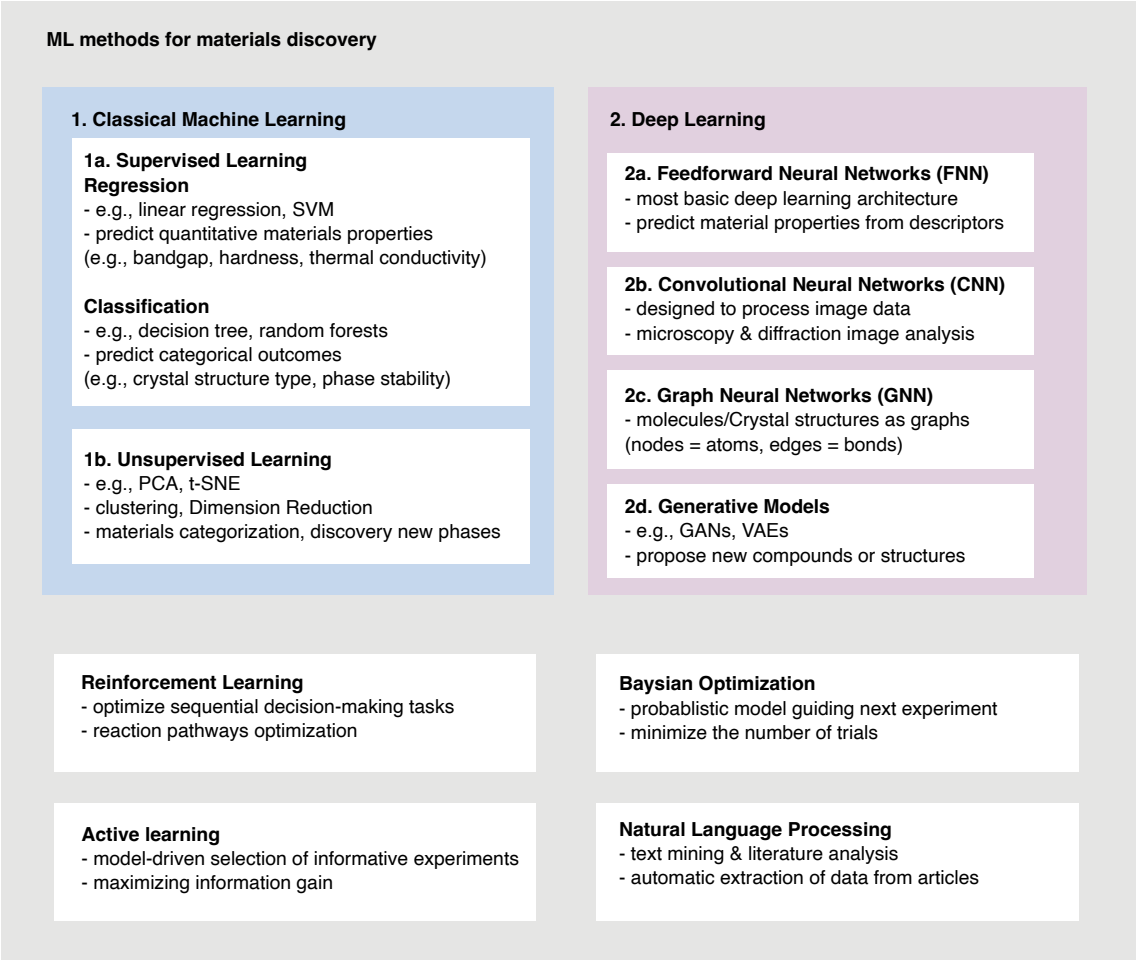


Figure 2.2: Common types of ML methods typically used in materials science.

become increasingly popular in materials science due to their superior performance in tasks involving large and high-dimensional datasets. However, DL methods typically require larger datasets (often exceeding $\sim 10,000$ data points), significant computational resources, and can be less interpretable compared to classical ML approaches.

Classical Machine Learning

The most distinctive types of classical ML methods are supervised and unsupervised learning. In supervised learning, models are trained on labelled datasets, where each data point is associated with a known outcome, such as a material property (for regression tasks) or a class label (for classification tasks).

- Regression models, such as linear regression, random forests, Gaussian process regression, and support vector regression (SVR), are used to map input descriptors (e.g., elemental fractions, lattice parameters) to quantitative material properties such as bandgap, thermodynamic stability, and morphology[24]–[27]. For instance, a random forest regression model has been employed to correlate the elemental composition of adsorption sites with the electrocatalytic adsorption energy of copper-containing intermetallic crystals[28].

- Classification techniques such as logistic regression, support vector machines (SVMs), decision trees are designed to predict discrete labels. These models are commonly applied to tasks such as phase classification, crystal structure identification, and synthesizability prediction[7], [12], [29]. For example, various classification models have been used to distinguish between different oxidation states in metal-organic frameworks (MOFs) based on local environmental features, including metal type, coordination geometry, and chemical environment[12].

Supervised learning approaches excel when reliable labelled data is available, whether from experimental measurements or computational simulations. They are typically more interpretable and computationally efficient compared to many deep learning models, making them attractive for iterative property prediction and materials design workflows.

In unsupervised learning, the objective is to uncover hidden patterns or groupings within unlabelled data without predefined outcomes.

- Clustering algorithms like k-means clustering, hierarchical clustering, and Gaussian mixture models are employed to group materials with similar feature profiles, potentially revealing unexpected trends or novel material families[30], [31]. For example, a study applied hierarchical agglomerative clustering to categorize ternary nitrides into distinct chemical families based on their stability and metastability profiles[31].

- Dimensionality reduction techniques, including principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE), are used to project high-

dimensional materials data into lower-dimensional spaces. This facilitates the visualization of patterns, identification of outliers, and discovery of structure–property relationships that might be obscured in higher dimensions[3], [18], [28].

By revealing latent structures in the data, unsupervised methods can provide valuable insights into structure–property correlations and guide targeted experimental or computational investigations.

Deep Learning

Deep learning (DL) is a specialized branch of ML that employs multi-layered neural networks to automatically learn features from raw data. This approach significantly reduces the need for manual feature engineering, provided that sufficient training data and computational resources are available. DL models excel at capturing complex, non-linear relationships in high-dimensional datasets, making them particularly valuable for a wide range of materials science applications.

Feedforward neural networks (FNNs) represent the most basic form of deep learning architecture. In these models, information flows in a single direction—from input to output—through multiple layers of interconnected nodes (neurons). FNNs are effective for both regression and classification tasks, particularly when datasets are large enough to support the direct learning of representations from raw inputs, such as the chemical structures of organic molecules[20], [32]. In one of the earliest applications of deep learning in materials science, a simple neural network architecture with two hidden layers—the minimum number required to be classified as deep learning—was used to design organic molecules for light-emitting diode (LED) applications[15]. While more advanced deep learning architectures have since been developed, FNNs are often employed as baseline models for benchmarking the performance of more complex networks.

Convolutional neural networks (CNNs) are designed to process grid-like data structures, such as images, by applying convolutional filters that automatically detect spatial hierarchies and local patterns. In materials science, CNNs are widely used for analysing image-

based datasets, including microscopy images (e.g., SEM, TEM) to identify microstructural features or defect distributions and diffraction patterns for automated phase classification and structural analysis[33]–[35]. For example, a CNN architecture with six convolutional layers was developed to classify crystal phases from X-ray diffraction (XRD) patterns automatically[34].

Graph neural networks (GNNs) are particularly well-suited for materials science because they natively process graph-structured data, which naturally represents molecules and crystal structures. In this framework, atoms are represented as nodes, and bonds or interatomic interactions are represented as edges. GNNs have demonstrated remarkable success in predicting a wide range of material properties, including crystal stability, electronic properties, and surface chemistry[36]–[38]. A notable example is GNoMe, a state-of-the-art AI model developed by Google DeepMind, which uses GNNs to predict and discover new crystalline materials. GNoMe has achieved unprecedented scalability, significantly improving the efficiency of materials discovery—reportedly accelerating the process by an order of magnitude[39].

Generative models aim to create new data samples that resemble the distribution of existing data, making them highly effective for inverse materials design. Two widely used generative models in materials science include Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Generative models can propose novel materials with desired properties, such as targeted band gaps, thermal conductivity, or mechanical strength[3], [16], [40]. For instance, MatterGen, a generative AI model developed by Microsoft Research, represents a significant breakthrough in materials science[41]. It enables the design of new inorganic materials with specific target properties, greatly accelerating the exploration of vast compositional spaces.

Other Machine Learning Methods

Active learning is especially valuable when experimental or simulation data are scarce or expensive to obtain. In this method, the model autonomously selects the most informative data points or experiments to label next, thereby optimizing the learning process

with minimal resources. Active learning strategies are often integrated with Gaussian process regression (GPR), due to its inherent uncertainty quantification, or neural networks to target high-uncertainty regions in the materials design space. This approach rapidly converges on optimal candidates while minimizing total experimental cost[16], [28], [34], [39].

In contrast to methods that rely on static datasets, reinforcement learning (RL) focuses on sequential decision-making, where an agent explores chemical or process pathways and receives "rewards" for achieving specific outcomes (e.g., synthesizing a stable or high-performing material). RL is particularly powerful in dynamic laboratory environments, enabling autonomous experimentation through real-time tuning of synthesis parameters and adaptive control of reactors. This exploration-exploitation framework allows RL to discover novel materials and optimize processes beyond human intuition[34], [42]. Bayesian optimization employs probabilistic models to iteratively recommend the next most promising experiments based on current knowledge and uncertainty. By leveraging acquisition functions—such as expected improvement or upper confidence bound—these methods balance exploration and exploitation to efficiently identify parameter sets or compositions with superior properties (e.g., optimized reaction yield or solar cell efficiency)[5], [38], [43].

Natural language processing (NLP) facilitates large-scale text mining of scientific literature, patents, and databases to extract valuable information, including synthesis protocols, property data, and structure-property relationships. Recent advances, particularly transformer-based models, have enhanced the ability to capture complex semantic patterns. When combined with knowledge graph construction, NLP can reveal hidden correlations, accelerate hypothesis generation, and guide experimental design toward previously unexplored materials spaces[34], [44]–[46].

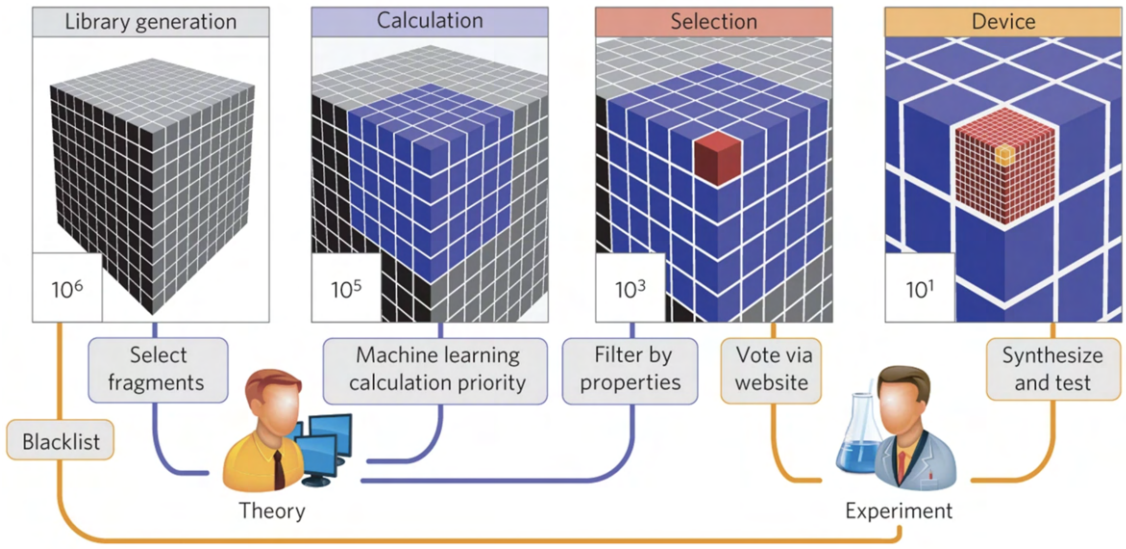


Figure 2.3: A ML-assisted materials discovery workflow in an early OLED study[15].

2.1.2 Incorporation into Materials Discovery Workflows

Building on the discussion of ML methods used in materials discovery, this section explores how ML is seamlessly integrated into materials discovery pipelines. Since ML models require large datasets to learn effectively, they are often combined with high-throughput computational tools and experimental platforms to generate the necessary training data.

On the computational side, large-scale density functional theory (DFT) and molecular dynamics (MD) simulations serve as key sources of labelled datasets for ML model training and fine-tuning. These simulations help establish structure–property relationships, enabling models to make more accurate predictions. Meanwhile, on the experimental side, automated synthesis and characterization platforms generate high-quality data that can be continuously fed back into ML models. This iterative process forms a closed-loop system where predictions guide experiments, and experimental outcomes, in turn, refine the model, accelerating the materials discovery cycle.

One of the earliest examples of ML integration in materials discovery was demonstrated in a 2016 study on organic light-emitting diode (OLED) materials[15]. In this work, a simple

deep learning model with just two hidden layers was incorporated into a materials discovery workflow that included materials generation, high-throughput calculations, virtual screening, and experimental fabrication (Figure 2.3). This streamlined workflow identified thousands of promising candidates, and subsequent experimental validation revealed several materials with exceptionally high quantum efficiency, showcasing the predictive power and practical utility of ML-enhanced discovery pipelines.

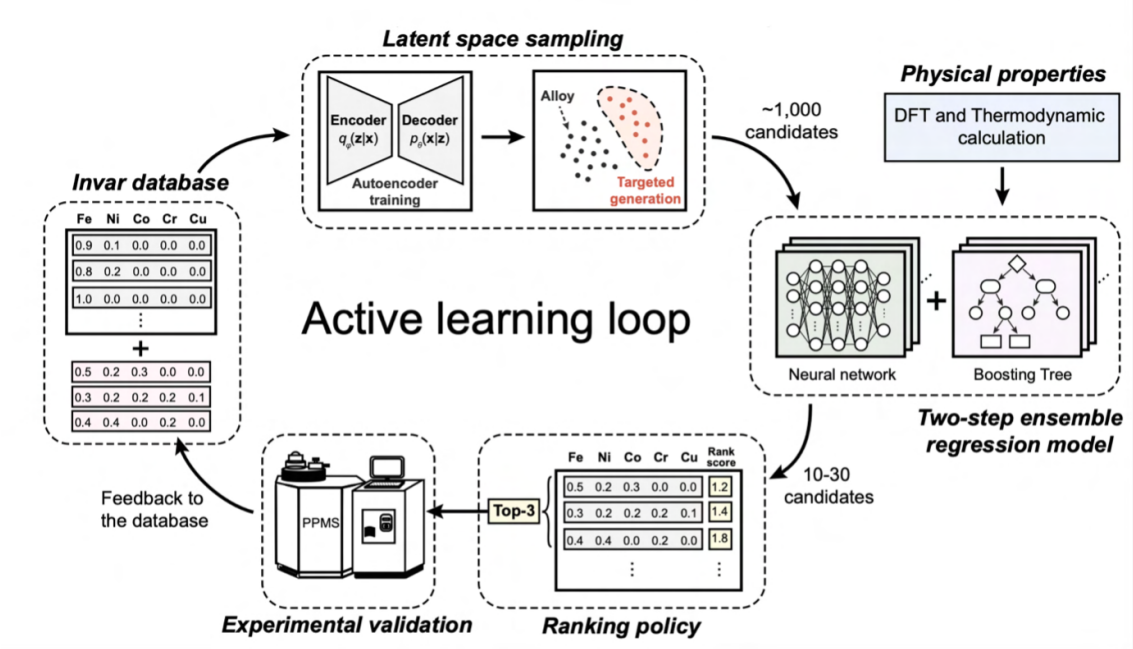


Figure 2.4: An active learning loop applied to high-entropy alloy discovery[16].

A more advanced example is the discovery of high-entropy alloys (HEAs) as shown in Figure 2.4, where an active learning framework was developed, integrating ML with high-throughput DFT calculations, thermodynamic modelling, and experimental synthesis[16]. At the core of this workflow, a generative model proposed new compositions, while a custom two-step regression model predicted their properties. Out of millions of potential compositions, 17 new alloys were synthesized and characterized, leading to the identification of two HEAs with targeted thermal expansion properties.

At the highest level of complexity, fully autonomous materials discovery systems are emerging, integrating multiple AI models into unified closed-loop frameworks. A notable exam-

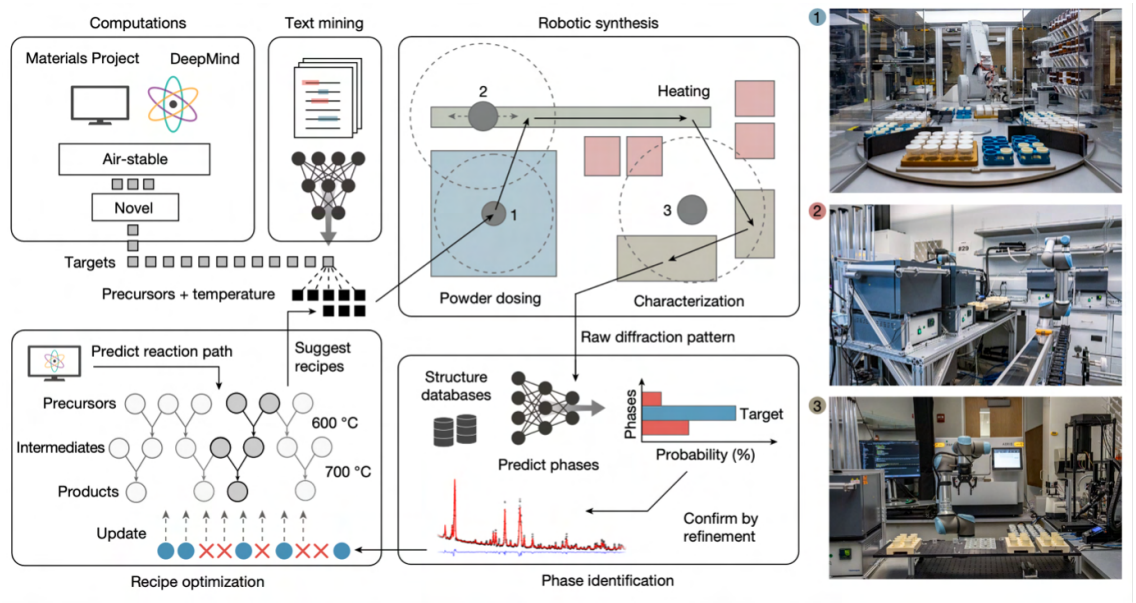


Figure 2.5: Automated materials discovery with the A-Lab platform[34].

ple is the A-Lab platform (Figure 2.5), which employs several ML models to accelerate discovery[34]. First, an NLP model is trained on extensive databases of synthesis procedures extracted from the literature. This is followed by a regression model that predicts optimal synthesis temperatures based on precursor materials. When initial synthesis attempts are unsuccessful, an active learning algorithm iteratively refines the synthesis parameters using experimental feedback. Finally, convolutional neural networks (CNNs) are employed to analyse X-ray diffraction (XRD) patterns of the synthesized materials, aiding in the rapid identification of successful synthesis outcomes.

2.1.3 Inverse Design in Materials Discovery

Building upon the integration of ML into materials discovery workflows, an even more ambitious application is inverse design, often regarded as a significant milestone—or even the "holy grail"—of materials discovery. Rather than adopting the traditional forward or direct-design approach of starting with a known material and then determining its properties, inverse design begins by specifying target properties or functionalities (Figure 2.6). Computational methods are then used to identify, generate, or suggest materials that fulfill

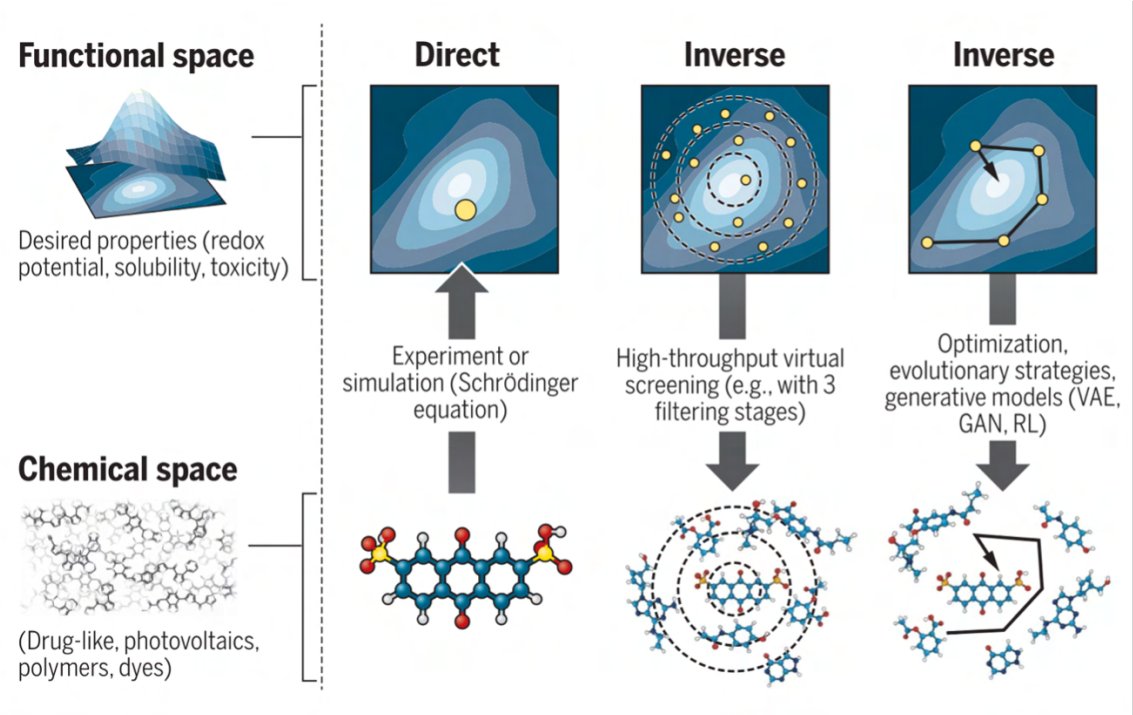


Figure 2.6: Overview of direct and inverse approaches in materials discovery[2].

these specifications. This property-oriented framework is poised to efficiently traverse the vast chemical and structural space, thereby minimizing trial-and-error experimentation and accelerating the discovery of novel materials with precise functionalities[2], [47].

Although the concept of inverse design has been discussed for several decades, the rapid rise of ML in materials science has significantly advanced its practical implementation. Thanks to ML's capacity to model high-dimensional data and capture complex structure–property relationships, inverse design workflows have become more robust and predictive, offering the potential for real-world applicability in diverse materials domains.

Despite its growing popularity, inverse design does not yet have a universally agreed-upon definition within the materials science community. Some methodologies that effectively implement an inverse design workflow do not explicitly label themselves as such. For instance, while high-throughput screening (HTS) is often considered part of forward discovery, some researchers classify it as a subset of inverse design. These terminological distinctions highlight the evolving nature of the field, but the fundamental principle of in-

verse design remains unchanged: instead of starting with a known material and analysing its properties, we begin with a desired property and work backward to identify suitable structures or compositions.

Three primary methodologies are commonly used in inverse design: (1) generative models, (2) iterative design strategies (e.g., active learning, Bayesian optimization, genetic algorithms), and (3) invertible materials representations. Among these, generative models and invertible representations take a fundamentally different approach from iterative techniques. While iterative methods refine candidate materials through sequential optimization loops, generative models and invertible representations aim to construct viable materials from scratch based on predefined target properties. By leveraging these approaches, researchers can efficiently explore vast chemical spaces and propose entirely new material candidates, rather than incrementally improving known ones. This ability to generate novel structures directly from property requirements makes these methods particularly valuable for inverse design. The following sections provide an in-depth overview of generative models and invertible materials representations, examining their respective strengths, limitations, and distinct roles in accelerating materials discovery.

Generative models

Generative models including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), have emerged as powerful techniques for inverse design[3], [41], [48]–[50]. In these frameworks, models learn probabilistic distributions over existing materials or molecules. Once trained, they can generate new candidate structures or compositions that are likely to exhibit desired properties, effectively "mapping" targeted characteristics back to plausible chemical formulas or structural motifs. These methods are particularly impactful in organic molecule design (e.g., pharmaceuticals, organic semiconductors), where a dense, high-quality dataset of known molecules can be leveraged to yield innovative new candidates.

One of the earliest notable demonstrations of a generative model in materials discovery was a VAE architecture composed of an encoder and decoder using recurrent neural net-

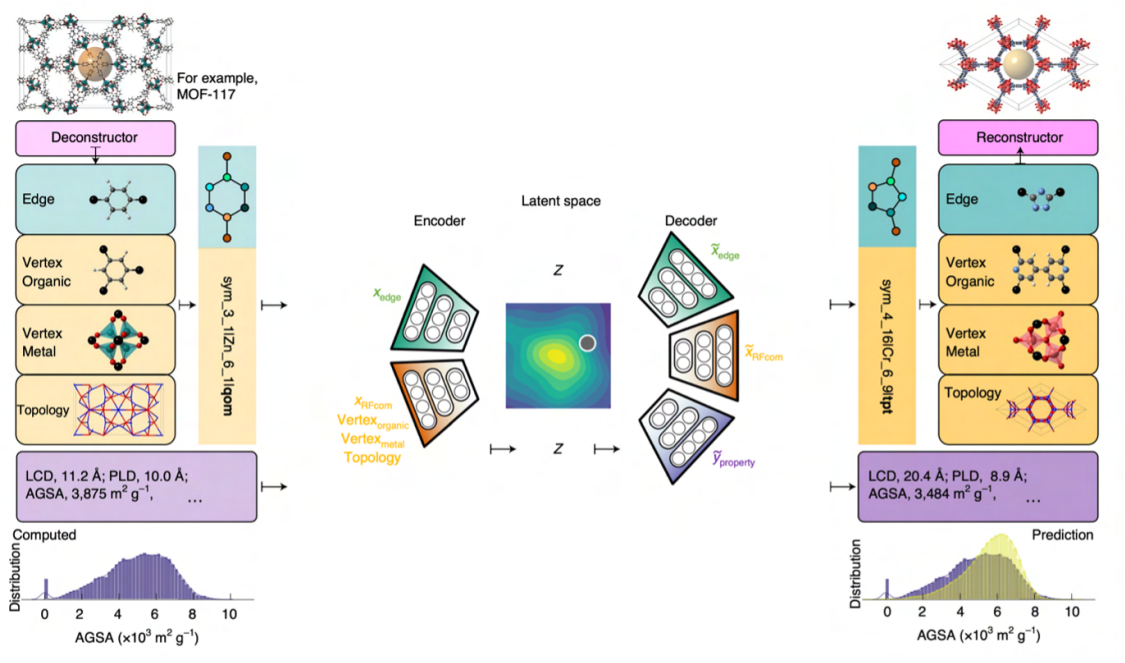


Figure 2.7: Schematic representation of a typical generative model workflow[3].

works, designed to generate organic molecules[51]. Since then, generative models have been applied across various material classes. For example, they have been used to generate novel metal–organic frameworks (MOFs) with exceptional gas separation properties[3]. More recently, MatterGen, a generative model introduced by Microsoft, has demonstrated the ability to generate stable and diverse inorganic materials spanning the entire periodic table. Moreover, it can be fine-tuned to design materials with specific target properties[41].

Despite these successes, the application of generative models remains largely limited to material classes with relatively well-defined structures and abundant training data. Their implementation in highly constrained materials systems, such as the organic spacers in 2D perovskites, is still rare. This limitation may stem from the scarcity of high-quality datasets in these domains, which restricts the model's ability to learn meaningful structure–property relationships. Without sufficient domain knowledge or physics-based constraints integrated into the learning process, generative models struggle to propose viable candidates in these more complex materials spaces. Addressing these challenges will be crucial for expanding the scope of inverse design methodologies to a broader range of

materials.

Invertible Material Representations

A growing area of research focuses on developing bidirectional mappings between material representations and desired property spaces[4], [52]. Unlike many black-box models that learn an inverse mapping implicitly, invertible representations provide a more transparent and structured approach to inverse design. This enables researchers to start in the property domain and systematically work backward to identify candidate structures, making the design process more interpretable and controllable. For molecular systems, several well-established invertible representations exist, including simplified molecular-input line-entry system (SMILES), international Chemical Identifier (INCHI) and molecular graph[53]. These representations allow machine learning models to efficiently encode molecular structures and generate new candidates based on specified properties.

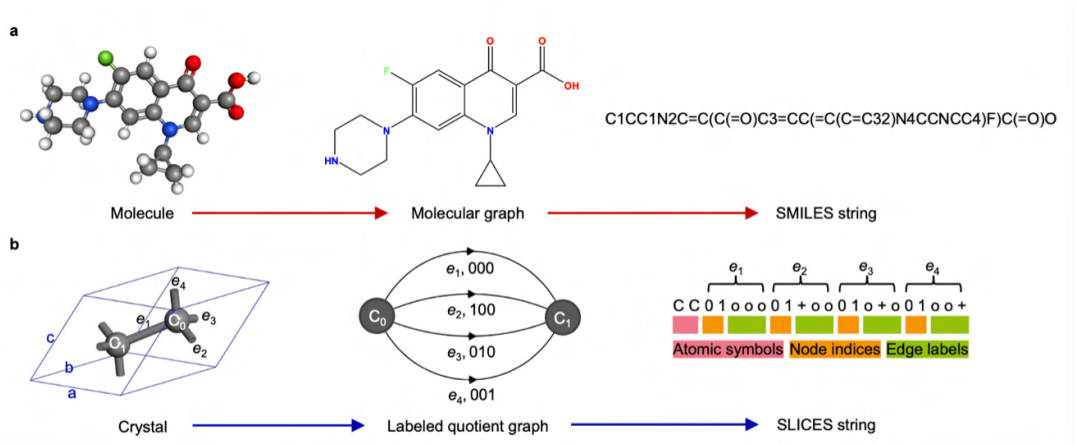


Figure 2.8: Invertible representations in molecular systems and solid-state crystals[54].

For inorganic crystals, the development of invertible representations is more challenging. Unlike molecules, where connectivity can be described in a straightforward manner, solid-state materials require a representation that captures both periodicity and compositional complexity. This has been a longstanding challenge in materials informatics, as achieving invertibility while maintaining generalizability and property-driven design remains difficult. Several representations have been proposed to address this issue, but none have yet become universal standard[4], [54].

Despite these challenges, invertible material representations have significant potential in inverse design, particularly when used in conjunction with machine learning models. By establishing structured and reversible mappings between materials and their properties, these representations provide a powerful framework for rational material generation, facilitating more efficient and targeted discovery.

2.2 2D Hybrid Perovskites

Two-dimensional (2D) hybrid perovskites have emerged as a fascinating class of materials that combine the desirable optoelectronic properties of their three-dimensional counterparts with enhanced chemical and environmental stability. In general, these materials consist of inorganic metal halide layers separated by organic cations, creating a naturally layered architecture. By tuning the thickness of these inorganic layers or altering the organic interlayers, researchers can tailor properties such as bandgap, exciton binding energy, and overall structural stability. Thus, 2D perovskites have attracted attention for applications ranging from solar cells and light-emitting diodes to photodetectors and field-effect transistors[55].

Unlike 3D perovskites, where small organic cations (e.g., methylammonium or formamidinium) fit within the perovskite lattice, 2D perovskites incorporate larger organic spacer molecules that enforce a layered structure and introduce new functionalities. This dimensional reduction results in strong quantum confinement effects, leading to distinct optical and electronic behaviours compared to their 3D analogy. In the following subsections, we discuss the structural variations of 2D perovskites, focusing on key phase families, and explore their fundamental electronic properties.

2.2.1 Structural and Electronic Fundamentals

Structural fundamentals

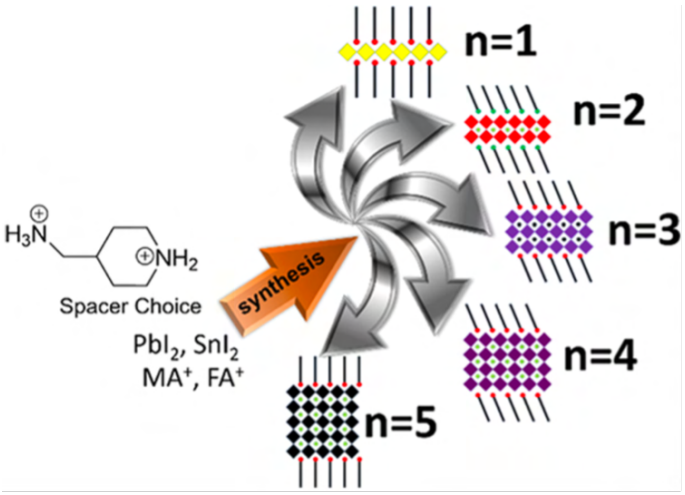


Figure 2.9: Design strategy for organic spacers in 2D perovskites[6].

2D perovskites exhibit rich structural diversity, primarily arising from variations in both the inorganic metal-halide framework and the organic spacers[56]. The inorganic layers are composed of corner-sharing metal halide octahedra, which can stack to form structures ranging from single-layered (strictly 2D perovskites, n = 1) to multi-layered (quasi-2D perovskites, n > 1)[57]. As the number of inorganic layers approaches infinity (n $\rightarrow \infty$), the structure converges to that of a 3D perovskite. Additionally, the metal (e.g., Pb, Sn) and halide (e.g., I, Br, Cl) compositions in the inorganic framework can be tuned, with large chemical space available[58]. Meanwhile, the organic spacer cations play a pivotal role in determining the exact structural phase (Figure 2.10), giving rise to three main families of 2D perovskites: Ruddlesden–Popper (RP), Dion–Jacobson (DJ), and Alternating Cation-Interlayer (ACI).

The **RP phase** represents the most extensively studied family of 2D perovskites and can be described by the general formula $(A')_2A_{n-1}M_nX_{3n+1}$. Here, A′ is a bulky monovalent organic spacer cation, A is a smaller monovalent cation (as seen in 3D perovskites), M is a metal cation (e.g., Pb^{2+}, Sn^{2+}), and X is a halide anion. Two layers of organic spacers intercalate between the inorganic slabs, creating a van der Waals gap. While many RP-phase perovskites display an in-plane octahedral shift of (1/2, 1/2), more complex tilts and distortions can occur with bulky or flexible organic spacers. Despite potential

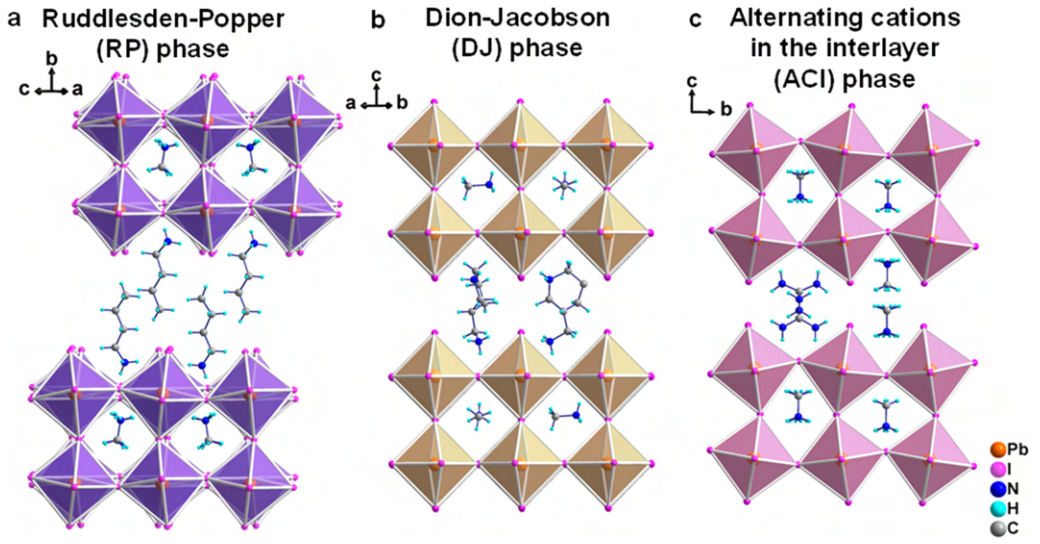


Figure 2.10: Schematic illustration of different structural phases of 2D perovskites[6].

drawbacks—such as restricted out-of-plane transport due to the van der Waals gap—RP perovskites are valued for their strong excitonic effects, tunable bandgaps, and suitability for light-emitting and photodetection applications[55], [59].

The **DJ phase**, described by the formula A'$A_{n-1}M_nX_{3n+1}$, where A' is typically a divalent organic spacer cation (e.g., diammonium compounds) with two ammonium tethering groups anchor to the inorganic layers. In contrast to the RP phase, DJ perovskites have only a single organic spacer layer between inorganic slabs, thus eliminating the van der Waals gap and often reducing the interlayer distance[60]. This configuration improves out-of-plane electronic coupling and charge transport while strengthening hydrogen-bonding interactions—leading to excellent structural stability. Such attributes render DJ perovskites promising candidates for high-efficiency solar cells and transistors[61].

The **ACI phase** is a relatively new structural motif where two different organic cations alternate between the inorganic layers, with the first example demonstrated in 2017[62]. Often, the organic spacer cations are relatively small (comparable to methylammonium), resulting in shorter interlayer distances and stronger interlayer coupling[63]. This structural motif can yield a reduced bandgap and enhanced charge transport, and there have been demonstrations of improved solar-cell efficiencies over comparable RP and DJ phases[64],

[65]. However, because of stricter size constraints on the organic cations, the chemical space for ACI-phase perovskites is comparatively smaller.

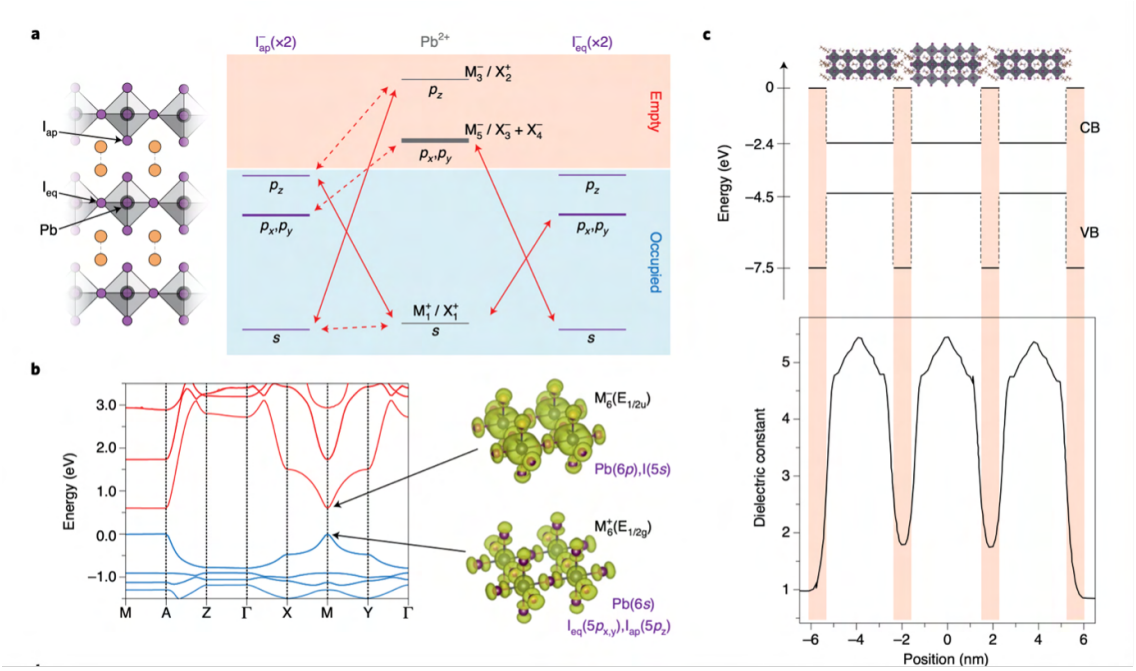


Figure 2.11: Electronic properties of 2D perovskites[66].

Electronic Fundamentals

The electronic structure of 2D perovskites is now relatively well understood, with their band edges predominantly governed by the inorganic layers and their structural and dielectric environment modulated by the organic spacers (Figure 2.11).

Generally, the bandgap widens as the thickness of the inorganic layer decreases (i.e., as n decreases). For example, monolayer (n = 1) perovskites typically exhibit bandgaps $\sim$ 2.4 eV, while quasi-2D structures with larger n values approach the bandgap of the corresponding 3D perovskite ($\sim$ 1.5 eV for $MAPbI_3$)[67]. This tunability has been leveraged in studies to induce an "energy funneling" effect, where charge carriers move from wider-bandgap (lower-n) to narrower-bandgap (higher-n) regions, enhancing the efficiency of light-emitting diodes[68]. Beyond structural thickness, the bandgap can also be finely tuned by altering the metal cation (M = Pb^{2+}, Sn^{2+}) or halide anion (X = I^-, Br^-, Cl^-)[69].

A defining feature that distinguishes 2D from 3D perovskites lies in the strong quantum confinement effects inherent to their layered structure[70]. In 2D perovskites, the inorganic slabs act as quantum wells, while the organic layers serve as wide-bandgap barriers[56]. This configuration leads to notably large exciton binding energies, ranging from 100 to 500 meV—substantially higher than the 10–50 meV typical of 3D perovskites—thus endowing 2D systems with pronounced excitonic behaviour advantageous for optoelectronic applications. However, the layered structure also induces highly anisotropic charge transport. In-plane transport benefits from robust orbital overlap between the metal and halide, often rivalling that in 3D perovskites (particularly for larger n values), whereas out-of-plane transport is hindered by the insulating organic layers acting as barriers to carrier motion[66].

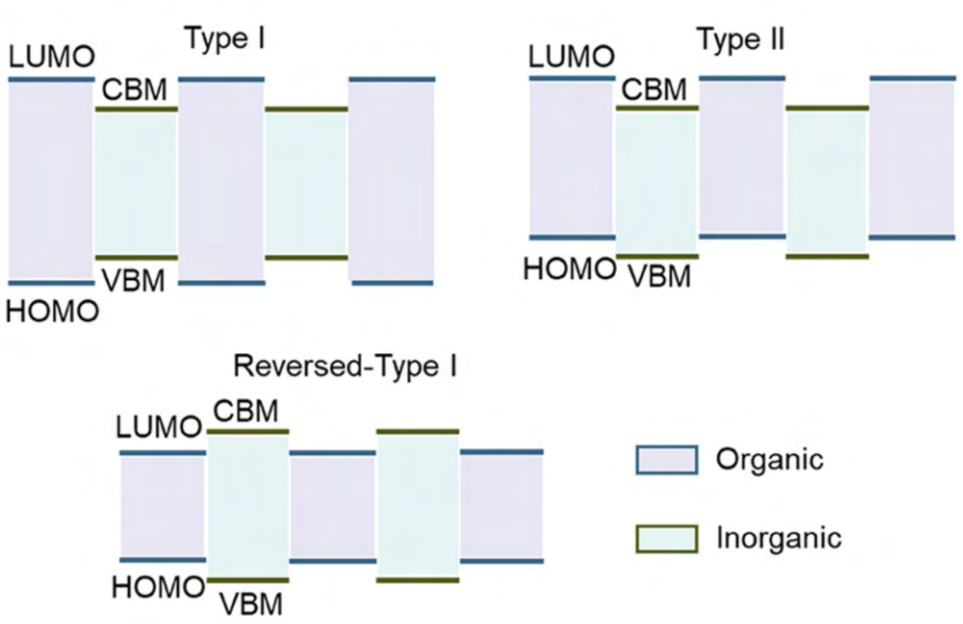


Figure 2.12: Schematics of the quantum well effect in 2D perovskites[71].

One of the most actively explored strategies for improving 2D perovskite performance is tuning their quantum well structure through composition engineering. Energy-level alignments in these systems are generally categorized as type-I (straddling gap) or type-II (staggered gap), with four possible configurations. Most 2D perovskites naturally adopting a type-I configuration with organic component acts as the insulating barrier (Figure 2.12).

Research has explored various approaches to modulating energy level alignment, including altering the thickness and composition of the inorganic layer and tailoring the organic spacer design[72]–[74]. Among these strategies, organic spacer engineering has emerged as the most promising, demonstrating the ability to achieve all four possible types of energy level alignment.

Despite these advancements, the tuning of energy level alignment remains limited, and many organic spacers have yet to be explored for further optimization. In the following section, we introduce design strategies for organic spacers to enhance their impact on perovskite properties.

2.2.2 Design Strategies for Organic Spacers

As mentioned above, the organic spacer cations in 2D perovskites play a pivotal role in determining their structural, electronic, and environmental stability properties. Rational design of these spacers enables the tuning of perovskite properties to optimize performance in optoelectronic devices.

The early discovery of organic spacers for 2D perovskites was largely serendipitous, driven by the limited availability of organic cations known to incorporate into the perovskite framework. Initial studies primarily focused on simple alkylammonium spacers, such as butylammonium (BA) and propylammonium (PA), with research efforts centred on varying spacer length or functional groups to improve film morphology and device performance[75]–[77].

In recent years, spacer engineering has evolved significantly, shifting towards the deliberate design of organic cations with tailored functionalities. This includes the exploration of conjugated organic spacers, typically featuring aromatic systems such as benzene or thiophene rings[78], [79]. The introduction of conjugation has been shown to enhanced interactions between organic spacers, improved charge transport, and enable tunable optoelectronic properties[80]. Additionally, spacer modification strategies now encompass

functional group engineering—such as side-chain substitutions, fluorination, and positional adjustments of ammonium tethering groups—to finely control interlayer interactions, defect passivation, and environmental stability.

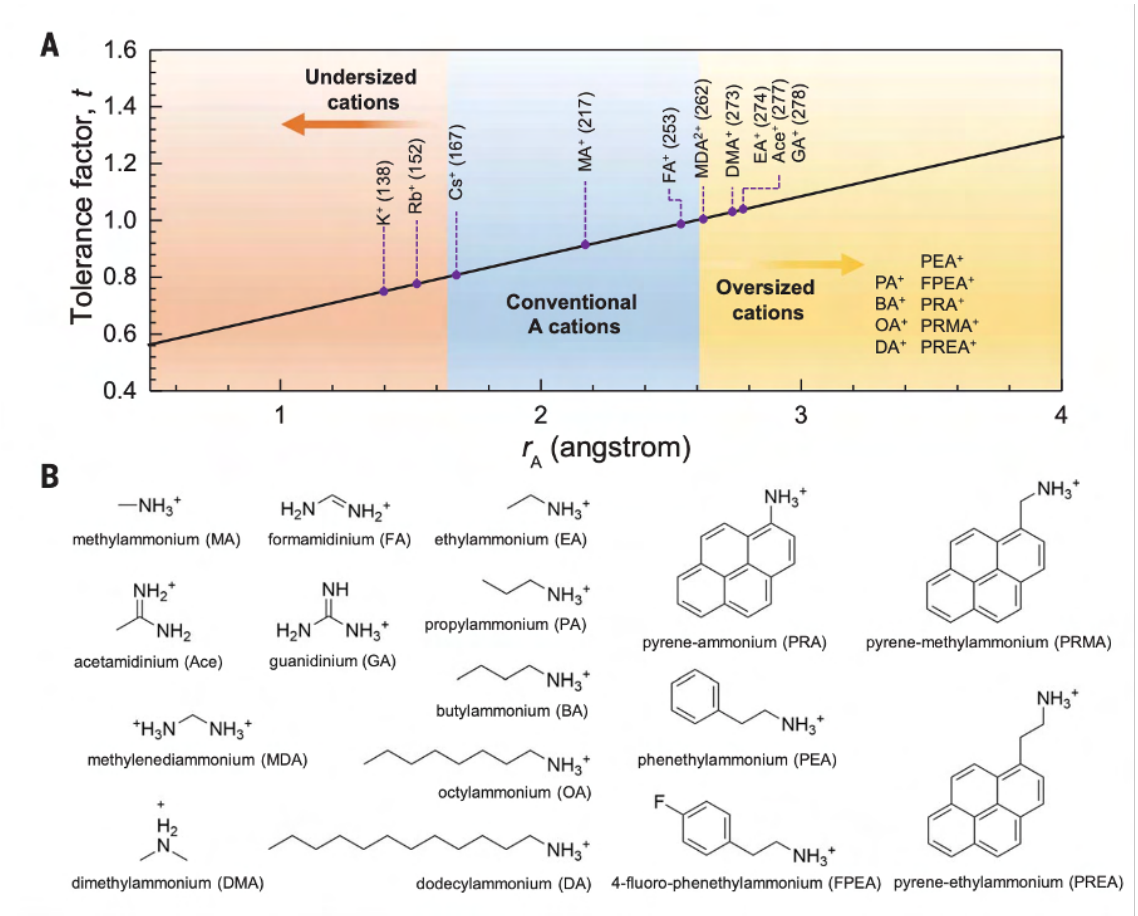


Figure 2.13: Molecular size constraints for organic spacers in 2D perovskites[81].

To maintain the perovskite framework, organic spacers must satisfy certain structural and chemical criteria. From a steric perspective, the size and shape of the organic spacer play a crucial role in determining whether it can fit within the inorganic framework and stabilize the 2D perovskite structure (Figure 2.13). While there is no strict limitation on the length of the organic spacer—studies have successfully incorporated linear alkyl organic spacers with up to 18 carbon atoms into 2D perovskite framework[82]–[84]—the cross-section area must remain within a certain threshold to avoid exceeding the available space within the 2D perovskite lattice[85]. In particular, the cross-section width of the organic spacer, often

approximated by the diameter of its conjugated backbone, should be smaller than width of the MX_6 octahedral unit[72]. If the spacer is too bulky, studies have shown that it can lead to different structural dimensions or even disrupt the formation of a stable 2D lattice, leading to 0D or 1D perovskite[86], [87].

Another critical consideration in designing organic spacers is the shape and configuration of the ammonium head group, which strongly influences hydrogen bonding. In 2D perovskites, this ammonium head typically fits into the cavity formed by the inorganic octahedral network, creating hydrogen bonds that stabilize the overall structure[85], [88]. The halide ions in the inorganic framework serve as hydrogen-bond acceptors, necessitating that the organic spacer contains a suitable hydrogen bond donor—typically an electron-deficient nitrogen bearing at least one hydrogen. Primary ammonium groups (i.e., NH_3^+) are widely used because they offer relatively strong hydrogen bonding, though secondary (NH_2^+) or tertiary ammonium group (NH^+) can also participate in hydrogen bond, albeit with generally lower bonding strength. Excessive steric hindrance around the nitrogen can impede its insertion into the inorganic pocket and weaken these critical bonds, potentially destabilizing the 2D perovskite. Typically, the hydrogen bond distance must remain below $\sim 3.0 - 3.5$Å to provide sufficient structural stabilization[88].

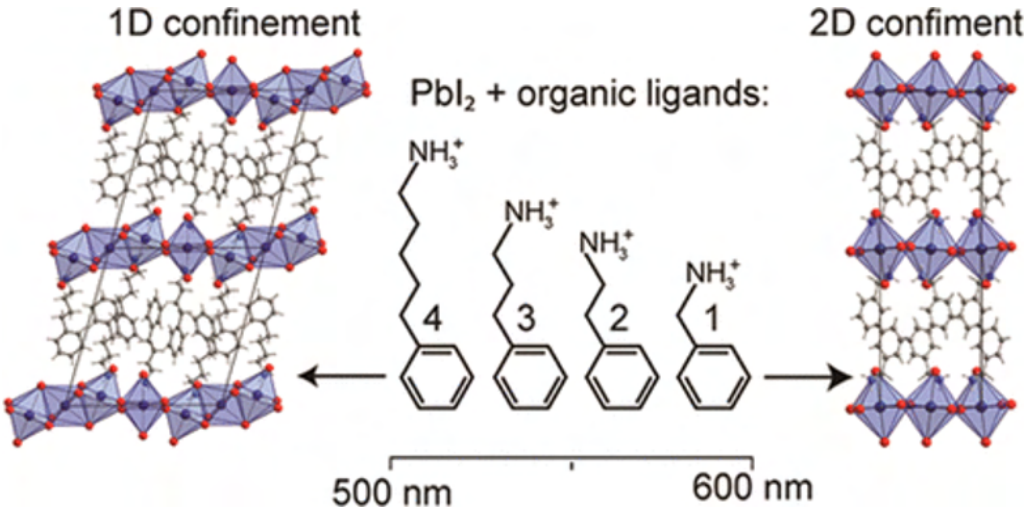


Figure 2.14: Modulation of 2D perovskites formability by tuning linker length[89].

Recent work has demonstrated various strategies for modulating hydrogen bonding. In one

study, researchers investigated organic spacers in which the ammonium tethering group was placed at different positions along the molecule, altering the distance between the ammonium head and the inorganic framework[90]. The spacer with the weakest hydrogen bonding formed a one-dimensional (1D) perovskite, whereas the two stronger-bonding spacers yielded 2D structures. Another approach involved varying the length of linker segment between the ammonium head and the main body of the spacer, thereby increasing conformational flexibility and enhancing hydrogen-bond interactions with the inorganic lattice (Figure 2.14)[89], [91]. Additionally, functional group engineering—such as fluorination—has been shown to further strengthen hydrogen bonds, likely due to the high electronegativity and induced dipole moment of fluorine atoms[92].

Beyond the structural considerations necessary for forming a 2D perovskite, organic spacer engineering also provides opportunities to tailor electronic properties. In many cases, the organic spacer does not directly participate in the electronic structure of the inorganic framework, allowing it to act primarily as a structural template. For instance, shorter organic spacers can reduce the interlayer distance and thereby enhance out-of-plane charge transport[93]. Additionally, adjustments to the inorganic octahedral tilting can further modulate the perovskite bandgap[79]. Some conjugated organic cations—particularly those containing aromatic rings—are known to reduce energy barriers for charge transport relative to their alkyl-based counterparts[94], [95].

Modifying the quantum well structure offers another route for electronic tuning. Introducing organic spacers with extended conjugation can shift the spacer's frontier orbitals, potentially inverting the quantum-well alignment from type I to type II. Achieving this often requires complex spacer architectures with extended π-conjugation and functional-group modifications. In one study, organic cations featuring a π-conjugated pyrene backbone with varying linker lengths were introduced onto the perovskite surface. These cations contributed electronically to the surface band edges and influenced carrier dynamics, ultimately improving solar cell efficiency[96]. In another series of studies (Figure 2.15), oligothiophene-based organic spacers were incorporated into DJ-phase perovskites, with both DFT calculations and experimental results confirming that increasing the number

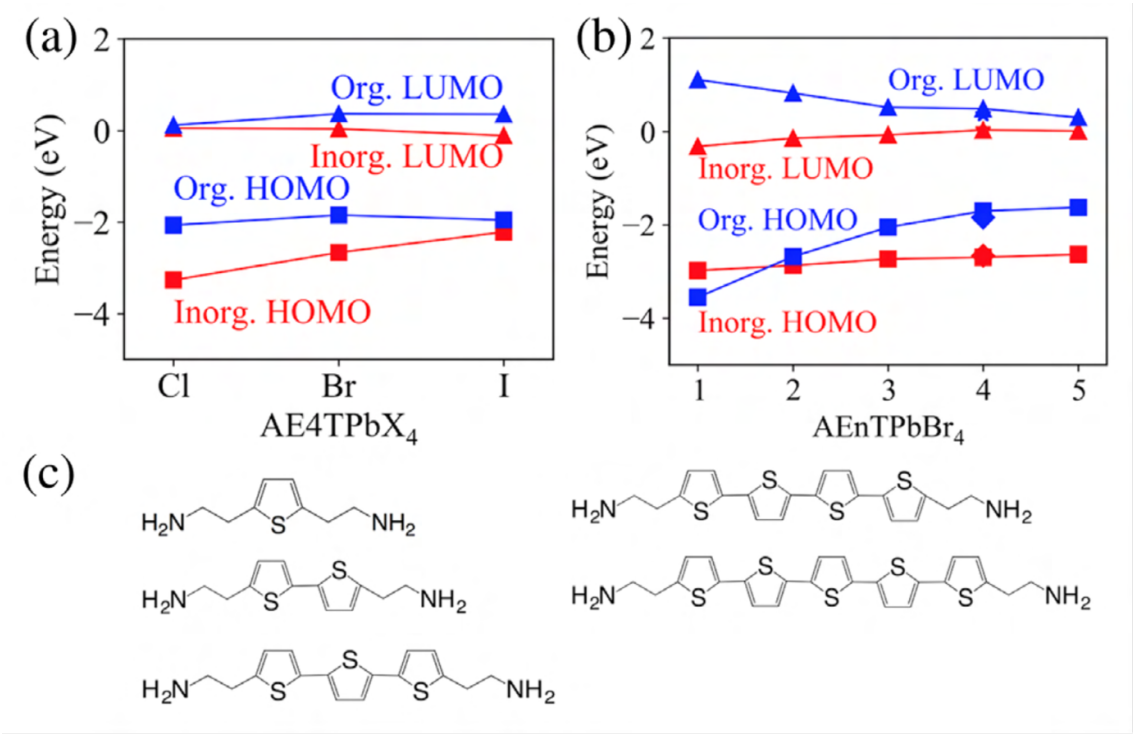


Figure 2.15: Engineering the energy level alignment in 2D perovskites by designing organic spacers[9].

of aromatic rings effectively tuned the energy-level alignment, leading to the realization of type-II alignment[9], [97]. Additionally, research on RP-phase perovskites has explored highly conjugated organic spacers, demonstrating that modifications to functional groups can achieve all four types of energy-level alignment. For instance, the introduction of electron-withdrawing units enabled type-II alignment, where the organic LUMO and inorganic VBM define the band edges. Meanwhile, incorporating a small bandgap unit facilitated a reversed type-I alignment[72].

Despite the established design strategies for organic spacers in 2D perovskites, the optimization process remains highly complex. The intricate structural features of organic spacers, along with their interactions with the inorganic framework, introduce a vast number of variables that are challenging to fully understand through conventional trial-and-error methods or human intuition alone. To address this complexity, AI-driven approaches have emerged as powerful tools for accelerating material discovery and optimization. The

following section explores recent advancements in AI-driven methodologies for 2D perovskites, with a particular emphasis on organic spacer design.

2.2.3 AI-Driven Approaches for 2D perovskites

ML methods have been extensively applied to the design of 3D perovskites, yet their use in guiding the discovery and optimization of 2D perovskites remains in a comparatively early stage. In 3D perovskites, a key focus has traditionally been an optimizing the inorganic frameworks—often represented by elemental compositions that are relatively straightforward for computational methods to handle[26], [98], [99]. By contrast, 2D perovskites incorporate both inorganic layers and organic spacers. The design challenge therefore shifts prominently to selecting and engineering the organic spacer, which must meet specific geometric (e.g., size constraints) and chemical (e.g., functional group compatibility) requirements.

Given that ML-assisted discovery of 2D perovskites is still in its early stages, relevant studies in this field remain scarce. Therefore, to provide a broader perspective, below we review ML application in three closely related research areas: organic spacer design in 2D perovskites, hybrid materials interface design, and organic materials design.

The virtually unbounded chemical space of possible organic spacers has driven researchers to explore ML and other data-driven strategies for 2D perovskites (Figure 2.16). An early study has used ML models trained on 86 reported organic spacers in lead-based 2D perovskites to derive design rules for predicting the perovskite dimensionality of five new organic spacers[100]. Recent approaches have expanded the scope of spacer exploration considerably. For instance, Wu et al. utilized a ML model trained on 80 high throughput synthesized lead-free double perovskites to evaluate the synthesis feasibility of 8,460 organic spacers from PubChem[7]. In another study, molecular dynamics simulations on over ten-thousand hypothetical organic spacers were used as training data to select six new ligands for perovskite synthesis[8].

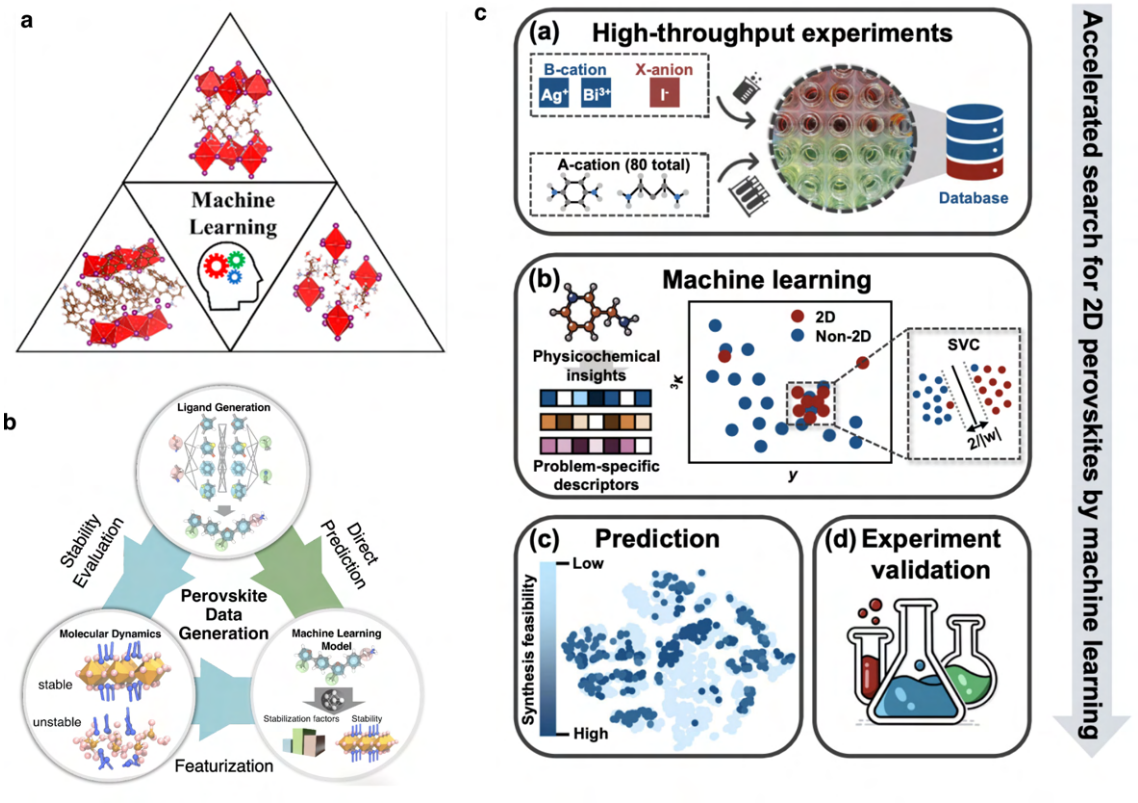


Figure 2.16: Machine learning workflows for 2D perovskite design[7], [8], [100].

Beyond 2D perovskites, similar ML workflows have been applied to other hybrid materials interfaces involving small molecules, particularly in passivation materials for perovskite solar cells[27], [101], [102]. In these studies, ML workflows typically follow a three-step approach. First, the physiochemical descriptors are selected based on domain knowledge. Second, the descriptors are computed from the organic molecule structure. Finally, these descriptors are used as input features for supervised machine learning models to predict key properties. For example, a recent study used various regression models to predict the power conversion efficiency (PCE) of perovskite solar cells passivated by ammonium salts, using a set of physiochemical descriptors[101].

Compared to hybrid perovskites, ML-assisted design of organic materials is a relatively mature field, with well-established methods for molecular representation and advanced machine learning models[5], [24], [32], [35]. The most critical aspect of ML workflows in

organic materials is how molecules are represented. Common molecular representations include:

(1) SMILES (Simplified Molecular Input Line Entry System) – An invertible molecular representation widely used in text-based ML models.

(2) Molecular fingerprints – Such as Morgan fingerprints or Extended Connectivity Fingerprints (ECFPs), which have been extensively applied to polymer design, peptide engineering, and organic emitter discovery.

For instance, a recent study used a 2048-bit Morgan fingerprint to represent heat-resistant polymers, combined with a feed-forward neural network to down-select promising candidates from a virtual library[32]. These representation techniques, combined with advanced ML models, have proven effective for guiding molecular design. However, directly applying these techniques to 2D perovskites is not straightforward due to the added complexity of the organic–inorganic interface.

A major limitation of the current ML approaches in 2D perovskites is their reliance on forward design workflows, which require exhaustive brute-force screening of chemical space. Since organic spacers must be predefined before descriptors can be calculated, the approach is not invertible, preventing direct generation of new molecular structures from target properties. Moreover, descriptor-based representations often fail to fully capture the molecular complexity of organic spacers. Such brute-force screening can be computationally expensive and may miss promising regions of chemical space if the initial library is not sufficiently diverse. Therefore, inverse design—where desired properties are defined a priori, and algorithms propose candidate molecules or structures—holds the promise of accelerating materials discovery for 2D perovskites.

Furthermore, much of the pioneering ML work on 2D perovskites has been centred on questions of formability and stability, a critical gap remains in the application of AI-assisted workflow to predict physical properties of 2D perovskites. Energy level alignment, a key property controlling the spatial distribution and transfer of charge carriers and excitations

in semiconducting materials and their interfaces, directly impacts the performance of optoelectronic devices. Different from well-studied elemental and compound semiconductors, organic and inorganic components in hybrid perovskites are heterogeneous with separate energetics, forming quantum-well-like structures[9]. Although 2D perovskites have been investigated using traditional workflows, such as the Edisonian approach[72] and high-throughput calculations[103], [104], systematic exploration of the energy level alignment through AI-assisted approaches is still in its early stages[105], presenting a significant opportunity for advancement.

2.3 Summary and Research Gaps

In the preceding sections, we introduced the two main pillars of this thesis: AI-assisted materials discovery—with particular attention to inverse design methodologies—and the design challenges posed by 2D perovskites, especially in the context of organic spacers. Their intersection is the central focus of this work.

From the perspective of 2D perovskite design, the vast chemical space of organic spacers necessitates the use of data-driven and machine learning techniques. These materials exhibit highly intricate structure–property relationships, with many potentially relevant features that are difficult to fully grasp through traditional approaches. To tackle this complexity, we focus on DJ-phase n=1 Pb–I-based perovskites as a prototype system for studying structure–property relationships. Insights gained from this prototype can then be generalized to other 2D perovskite phases and alternative inorganic frameworks. We choose energy level alignment as our target property, as it is crucial for optoelectronic applications yet remains insufficiently understood.

From the viewpoint of AI-assisted inverse materials design, this 2D perovskite system presents an equally compelling challenge. Unlike more extensively studied materials with large, well-curated datasets, 2D perovskites are relatively data-scarce and uniquely hybrid in nature, requiring careful consideration of both organic and inorganic components.

While inverse design methodologies have shown promise for inorganic materials, organic molecules, and polymers, 2D perovskites introduce additional constraints—such as spacer size limitations and the complex organic–inorganic interface—that necessitate novel AI-driven approaches.

Moreover, the insights gained from this research could extend beyond 2D perovskites to a broader class of hybrid materials, where the intricate interplay between organic and inorganic components defines material properties. As emerging materials fields often suffer from data scarcity, developing machine learning strategies tailored to these challenges is essential. By leveraging AI in such contexts, we aim to contribute to a generalizable framework for hybrid material discovery, enabling data-driven innovation in materials science.

This thesis addresses several critical research gaps in the field of AI-driven 2D perovskite design:

- Lack of large, high-quality datasets for 2D perovskites – Unlike well-established materials, 2D perovskites suffer from data scarcity. Existing datasets are often small, inconsistent, or lack standardization, limiting the ability of machine learning models to generalize effectively.
- Limited understanding of structure–property relationships – A quantitative and predictive understanding of how organic spacer chemistry influences electronic properties and synthesizability remains underdeveloped. The complexity of organic–inorganic interactions make it challenging to establish clear design rules.
- Challenges in inverse design for 2D perovskites – Existing inverse design models do not fully accommodate the unique constraints of hybrid materials, such as the size limitations of organic spacers and the chemical constraints associated with functional group compatibility.
- Challenges in AI-assisted approaches for hybrid materials – The absence of a standardized ML workflow for hybrid materials poses a significant barrier. Feature

selection remains underdeveloped, and machine learning models struggle to encode key chemical and structural descriptors necessary for accurately modelling organic–inorganic interactions.

Addressing these challenges requires robust, AI-driven frameworks tailored to 2D perovskite design—particularly those that incorporate domain expertise, and feature property-driven design. The present thesis aims to bridge some of these gaps by exploring inverse design workflows that couple molecular design with inorganic framework constraints, ultimately accelerating the discovery of high-performance 2D perovskite materials.

Chapter 3

Methodology

This chapter outlines the methodology developed for the inverse design of DJ-phase 2D perovskites. Section 3.1 provides an overview of the overall design framework, which integrates chemical space establishment, property prediction, and synthesis feasibility. Section 3.2 introduces the core component of the workflow: an invertible and interpretable molecular fingerprint tailored for organic spacer design. The subsequent sections detail the key pipelines of the framework, including high-throughput calculations (Section 3.3), machine learning model development and evaluation (Section 3.4), and the two-step synthesis feasibility screening process (Section 3.5).

3.1 Overview of the inverse design workflow

The AI-assisted inverse design workflow is illustrated in Figure 3.1. This workflow hinges on a unique 12-digit fingerprint representation scheme to navigate the chemical space of organic spacers, integrating DFT calculations, machine learning, and synthesis feasibility screening. First, hypothetical candidates are generated using a molecular morphing approach and selected for DFT calculation. Second, the DFT data are used to train interpretable machine learning models, accelerating property predictions and revealing

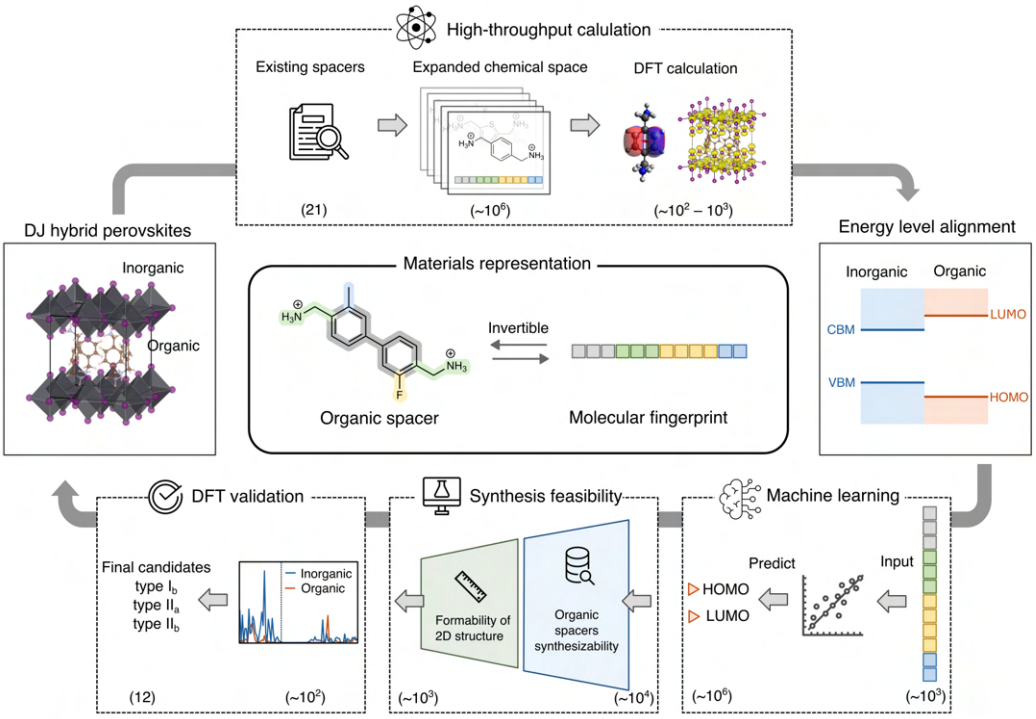


Figure 3.1: AI-assisted inverse design workflow for discovering DJ perovskites with targeted energetics and synthesis feasibility.

structure-property relationships. Third, synthesis feasibility is assessed based on the synthetic accessibility of organic spacers and their potential to form stable 2D structures. Finally, these 2D perovskites undergo DFT validation to confirm their energy level alignment, leading to a selection of recommended candidates.

This workflow was designed based on the unique nature of 2D hybrid perovskites and the targeted property of band alignment. It begins with chemical space expansion using a molecular morphing approach. To realize an invertible representation of conjugated diammonium organic spacers, they are encoded into a compact 12-digit fingerprint vector. Based on the physical insights obtained on 21 existing spacers reported for DJ perovskites, we generated the fingerprints of approximately 4×10^6 hypothetical spacers with complexity comparable to the reported ones. High-throughput density functional theory (DFT) calculations were then used to evaluate the energy levels of the corresponding hybrid per-

ovskites within a designated subset of the chemical space, which were used as the training data. Next, various regression models were trained using fingerprints as input features and organic frontier levels as target property, aiming to extract insights on the structure-property relationship. The hypothetical spacers were then down selected using a two-step synthesis feasibility screening funnel based on their availability in the PubChem database and multiple reported formability descriptors specific to forming 2D perovskite structures. Lastly, feasible candidates for targeted energy level alignment types are validated using DFT calculation. By integrating these components, the workflow facilitates inverse design of DJ perovskites with rarely explored Ib, IIa and IIb band alignment types.

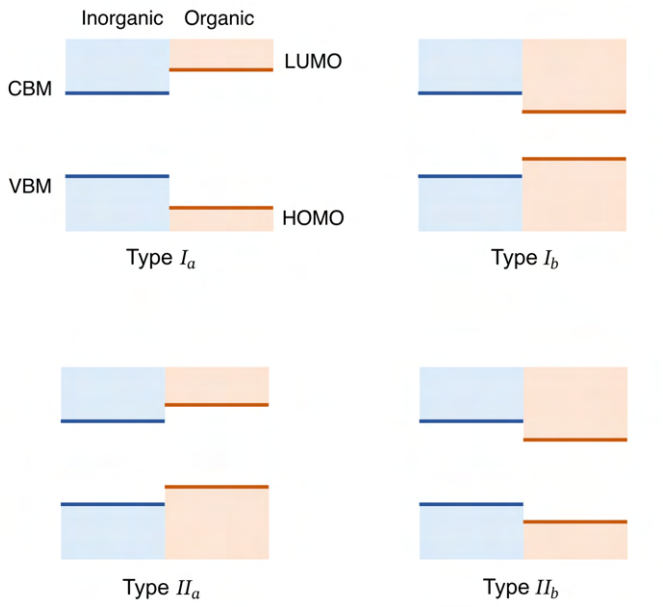


Figure 3.2: Schematic representation of energy level alignment types in 2D perovskites.

The classification of the energy level alignment types is shown in Figure 3.2. In type I alignment, both low-energy electrons and holes are localized in the same component: type Ia in the inorganic layers and type Ib in the organic layers. In type II alignments, electrons and holes are separated between different components: in type IIa, electrons are localized in the inorganic layers and holes in the organic layers, whereas in type IIb, the reverse occurs. It should be noted that the designations “a” and “b” are sometimes

interchanged in the literature depending on the component being emphasized[9]. In other studies, only type I and type II are referenced without further categorization[72], and type Ib in our context is occasionally described as "reversed type I"[106]. Notably, alignment type Ia is referred to as type Ib in some literature[9], [69]; here, we defined type Ia as the configuration where inorganic states serving as the band edges.

While the components of this workflow—database generation, high-throughput calculations, machine learning, and DFT validation—are common to AI-assisted materials discovery[29], [107], the distinctive feature here is the integration of an invertible materials representation. Invertibility is a key attribute for materials representations in inverse design[2], ensuring two-way conversion between molecular structure and their representation. This type of invertible representation has been applied to some materials systems[3], [4], but this is the first implementation in the context of hybrid materials. The absence of a versatile scheme of organic spacer representation has confined 2D perovskite research to forward design approaches, limiting the exploration of available chemical space. As we will show in this thesis, the workflow developed herein overcomes these limitations, facilitating the energy level alignment prediction. In addition, we expect that this fingerprint-based workflow will be generalized to investigate the correlation of other material properties with organic motifs in a wide range of hybrid material systems.

3.2 Invertible Molecular Fingerprints

Figure 3.3 depicts an overview of our fingerprinting scheme, which is based on the specific attributes of conjugated organic cations in 2D DJ perovskites, comprising two key components: molecular fragmentation and functional group encoding. Organic spacers are first fragmented into their building blocks (backbone, tethering ammonium, side chain, and substitutions) and then these building blocks are further encoded into a 12-digit fingerprint.

Fragmentation

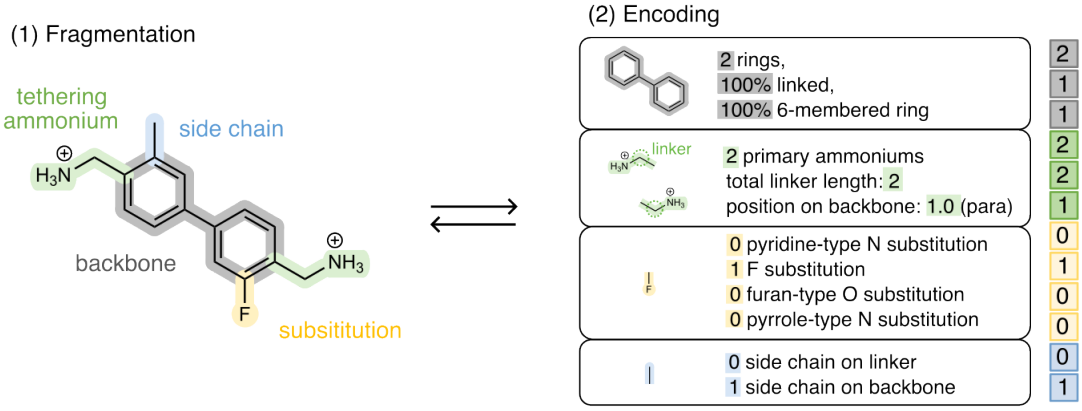


Figure 3.3: Invertible molecular fingerprint representation for organic spacers in DJ perovskites.

The fragmentation is designed considering the structural motifs shared by reported conjugated diammonium spacers, as shown in Figure 3.4. The DJ-phase organic spacers explored in this work are assumed to consist of four fragments:

(1) a conjugated backbone of aromatic rings;

(2) two tethering ammonium groups that anchor the spacer to the inorganic framework;

(3) optional heteroatom substitutions;

(4) optional side chains.

In this work, we limit the heteroatom substitutions to fluorine (F), oxygen (O), and nitrogen (N) on aromatic rings (benzene and thiophene). This simplification is based on two primary considerations: (1) their widespread use in semiconducting organic spacers within 2D perovskite systems, and (2) the need to preserve a chemically interpretable and synthetically accessible design space. Although including additional heteroatoms such as chlorine (Cl), bromine (Br), or phosphorus (P) could enhance electronic diversity, doing so would substantially expand the chemical space, increase fingerprint complexity, and introduce greater uncertainty in terms of synthetic feasibility and structural compatibility with the perovskite lattice.

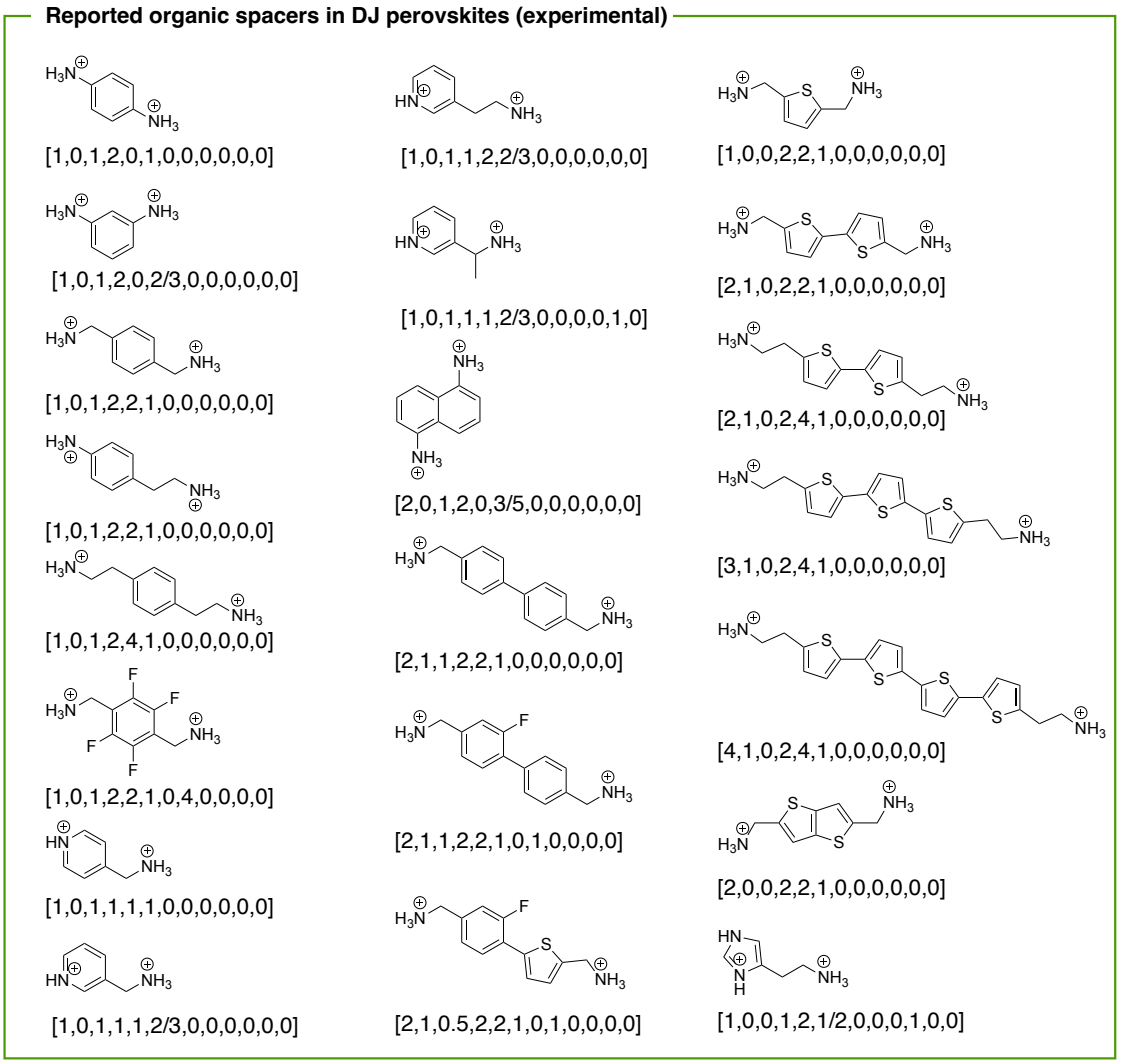


Figure 3.4: Reported organic spacers included in this study with their molecular fingerprint.

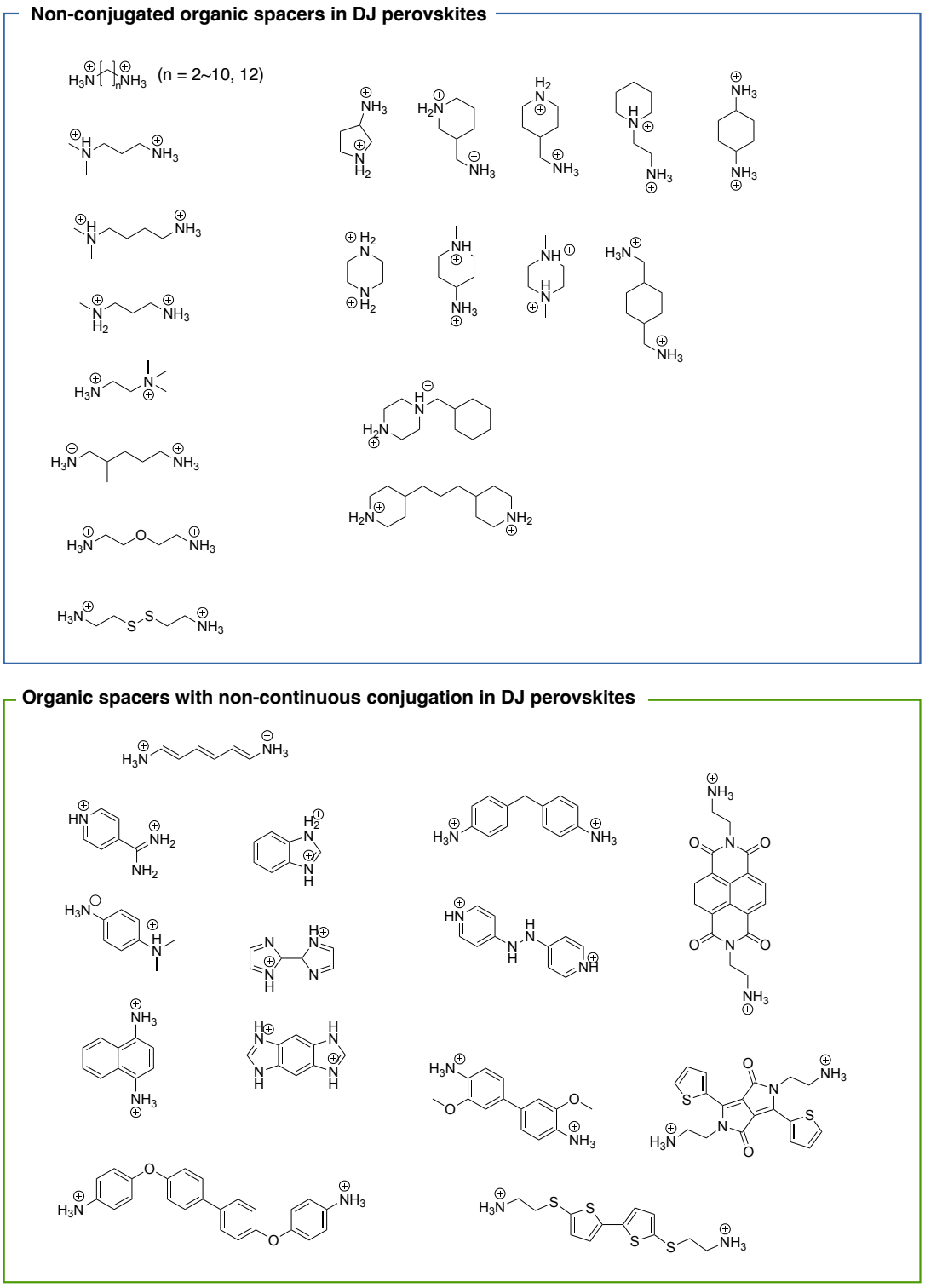
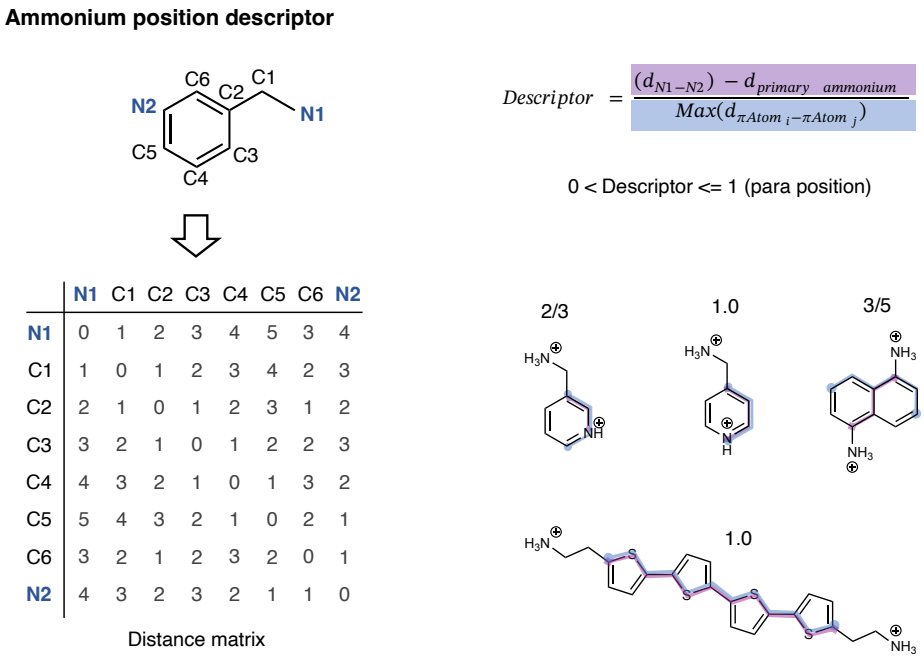


Figure 3.5: Organic spacers excluded from the scope of this study.

These structural constraints significantly narrow the chemical space from a potentially immense size (estimated at $\sim 10^{60}$ molecules for small organic molecules, as recognized in the context of drug discovery[108]) to a much smaller subspace of organic spacers.

We should note that the resulting chemical space is not exhaustive, leaving out some spacers, for example ones with alkyl backbones or non-continuous conjugation (Figure 3.5). This fingerprinting scheme leads to a chemically relevant and computationally manageable set of organic cations (vide infra), giving rise to 2D DJ perovskite candidates with tailored properties. We primarily focused on semiconducting π-conjugated molecules due to their high relevance to optoelectronic applications of 2D perovskites and rich chemical diversity.

Encoding



	N1	C1	C2	C3	C4	C5	C6	N2
N1	0	1	2	3	4	5	3	4
C1	1	0	1	2	3	4	2	3
C2	2	1	0	1	2	3	1	2
C3	3	2	1	0	1	2	2	3
C4	4	3	2	1	0	1	3	2
C5	5	4	3	2	1	0	2	1
C6	3	2	1	2	3	2	0	1
N2	4	3	2	3	2	1	1	0

Figure 3.6: Illustration of the ammonium position descriptor.

The encoding component of the scheme translates molecular structure into a fingerprint vector containing 12 customized descriptors, each representing a specific structural feature. Eleven descriptors are obtained by counting functional groups, while a unique “ammonium

position" descriptor is derived from a distance matrix (Figure 3.6). The main principle is to choose a minimal number of descriptors to reduce computational cost while these descriptors must be sufficient to describe the organic spacers relevant to DJ perovskites. As we will show later in the ML results in Chapter 4, there is minimal overlap between the descriptors, and they capture essential features for energy level prediction.

For high-throughput purposes, all organic spacer structures in this study were stored in the Simplified Molecular Input Line Entry System (SMILES) format, a widely used textual representation of molecular structure. Molecular fingerprinting was performed to extract key structural and chemical descriptors automatically using the RDKit library in Python. The workflow took the SMILES representation of the organic spacer as input and returned a set of 12 organic descriptors, categorized as follows:

1. Conjugated backbone descriptors: number of rings; percentage of ring linkages; percentage of six-membered rings
2. Tethering ammonium descriptors: number of primary ammonium groups (NH_3^+); linker length (distance between ammonium groups and backbone); and ammonium position on the backbone
3. Heteroatom substitution descriptors: number of nitrogen atoms (pyridine-type); number of fluorine atoms, number of oxygen atoms (furan-type); number of nitrogen atoms (pyrrole-type)
4. Side chain descriptors: number of side chain attached to linkers; number of side chains attached to the backbone

Descriptors were computed using SMARTS (SMiles ARbitrary Target Specification) patterns, enabling the identification and counting of specific functional groups. A new descriptor, the ammonium position on the backbone, was developed specifically for this work. This descriptor quantifies the relative position of tethering ammonium groups on the conjugated backbone using a distance matrix approach, as depicted in Figure 3.6.

The ammonium position descriptor is derived from the molecular skeleton using a distance matrix. This descriptor is calculated as the ratio of the maximum distance along the conjugated backbone to the distance between the tethering ammonium group and the backbone. Representative organic spacers with their corresponding ammonium position descriptors are shown in Figure 3.6.

We should note that the molecule-fingerprint correspondence is not exclusive, in other words, some molecular isomers share the same fingerprint. Although additional descriptors, or longer fingerprints (e.g., heteroatom substitution position, and side chain position) could offer more structural detail, we found such features have minimal impact on electronic properties, making the current fingerprinting scheme sufficient for predicting new DJ perovskites with all four band alignment types.

Figure 3.7 illustrates the relationship between a fingerprint and its corresponding organic spacer(s). The upper panel shows a one-to-one mapping, where a fingerprint corresponds to a single organic spacer. In contrast, the lower panel shows a one-to-many mapping, where multiple organic spacers (isomers) share the same fingerprint, including the example shown in Figure 3.3. These isomers may differ in structural features such as the position of heteroatom substitution, side chain placement, or the linker length between tethering ammonium groups. While such variations may affect to a certain extent the chemical and physical properties of the cations, the the energy levels are almost the same due to the shared molecular backbone. However, these isomers may have different levels of synthesis feasibility, which warrants elucidation based on future detailed analysis and experimental efforts. At this stage, no additional screening is applied to these isomers. All molecules corresponding to a given fingerprint are retained in the dataset to preserve the full chemical diversity for downstream analysis.

In previous AI-assisted 2D perovskite discovery efforts, organic spacers are typically represented using physiochemical descriptors[7], [8], but an effective molecular representation scheme that can explicitly capture the molecular structure has not been established. In the myriad research fields involving organic molecules, the structural variations are often encoded using digits (e.g., fingerprints), strings (e.g., SMILES), or graph-based methods[2].

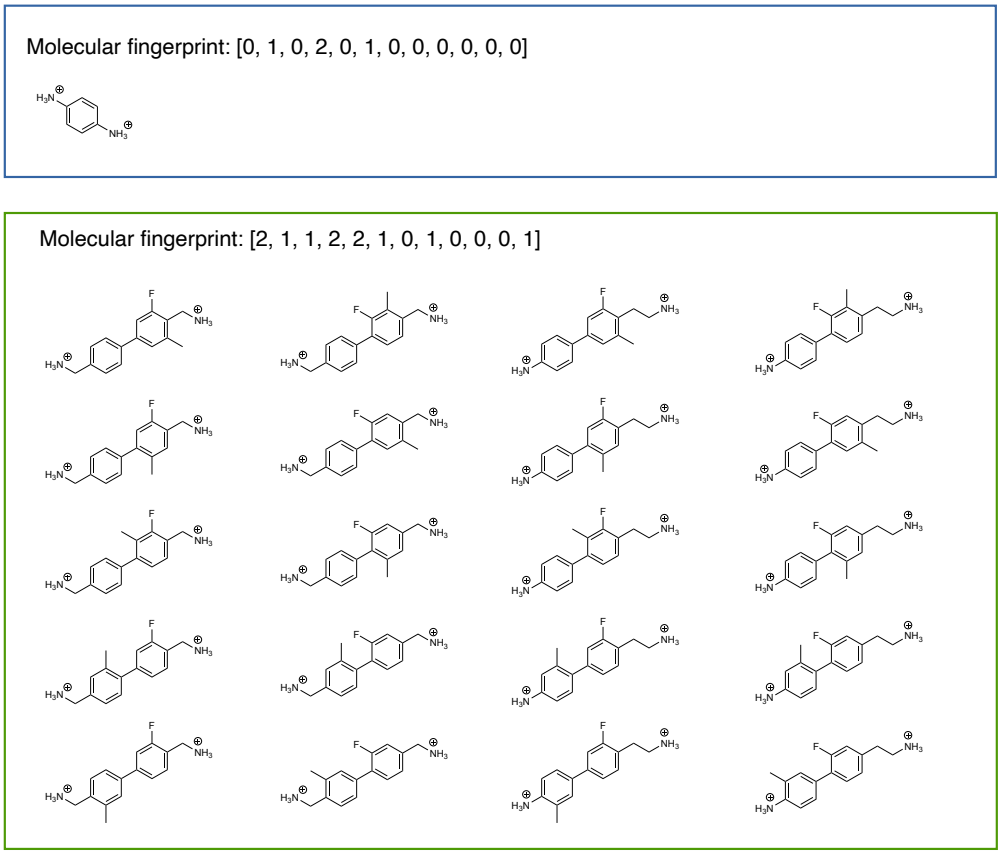


Figure 3.7: An illustration of one-to-one (top) and one-to-multiple (bottom) mappings between a molecular fingerprint and its corresponding organic spacer(s).

Among these, fingerprinting methods—such as the widely adopted but non-invertible 2048-digit Morgan fingerprint—have demonstrated their efficiency in AI-assisted workflow[32], [43]. In contrast, our 12-digit fingerprint scheme has been tailored according to the specific attributes of 2D hybrid perovskites, offering several advantages. First, it is efficient, with minimal redundancy and overlap between descriptors, ensuring a compact representation that captures structural variation most relevant to DJ perovskites. Second, it is interpretable, enabling human experts to extract meaningful insights into the encoded structural variations. Finally, it is invertible, allowing direct mapping back to the molecular structure by both human experts and machines, which is essential for inverse design.

3.3 High-throughput calculation

Molecular morphing

Molecular morphing was implemented as a systematic method to explore chemical space by generating variants of molecular structure. The process used reaction SMARTS patterns implemented in the RDKit library to iteratively apply predefined chemical transformations. This approach performs stepwise modifications on a starting molecular structure, generating new variants while adhering to defined chemical constraints.

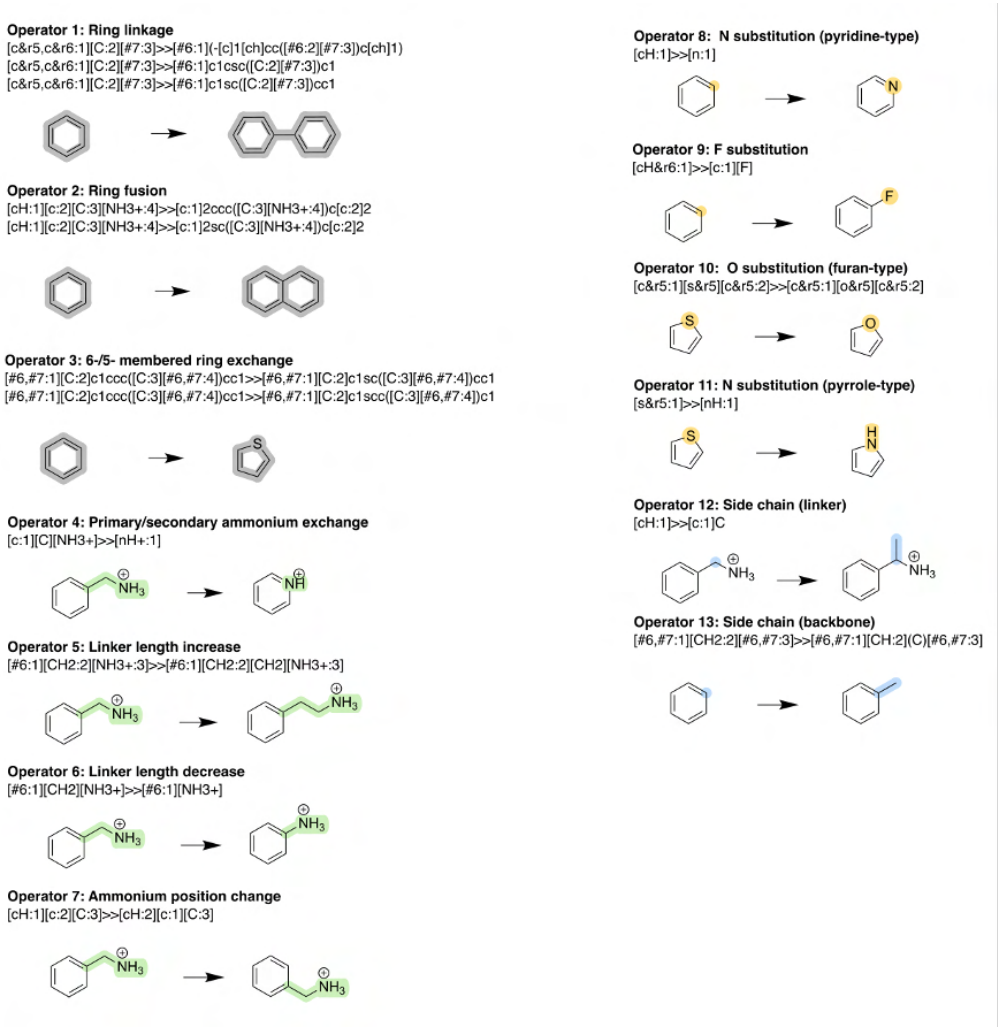


Figure 3.8: List of molecular morphing operators used in this study for generation of hypothetical organic spacers.

The molecular morphing process began with PDMA, a well-characterized prototype molecule, represented in SMILES format. As shown in Figure 3.8, a set of 13 morphing operators, encoded as 17 unique SMARTS patterns, was defined to ensure that each transformation adhered to established chemical constraints. These morphing operation include: increasing number of rings, substituting heteroatoms on aromatic rings, modifying linker lengths, etc.

Each operator was applied iteratively to the starting molecule to generate new molecular structures. The newly generated molecules were stored as SMILES strings, ensuring compatibility with downstream fingerprinting and modelling workflows.

PDMA was chosen as the starting molecule due to its structural simplicity and extensive study in the literature. Figure 3.9 illustrates the distribution of existing spacers across generations when different candidates are selected as G_0 molecules. PDMA (top right) was selected because it results in most existing spacers appearing in early generations, demonstrating its structural simplicity and suitability for easy transformation into other molecular structures via morphing operations.

Frontier level calculation of organic spacers

The 3D molecular structures of organic spacers were generated from SMILES string using the RDKit library, which efficiently converts the molecular graph representations into 3D coordinates. Conformational sampling was conducted assuming the isolated, gas-phase configurations of the spacers, independent of their incorporation into the 2D perovskite structure.

DFT calculations were performed using Gaussian 09 package with the B3LYP functional and 6-31G** basis set to calculate the energy levels of the HOMO and LUMO.

Hybrid Perovskite Structure Generation

The hybrid perovskite structures were constructed by inserting the organic spacers into a PbI_4-based inorganic framework. Initial conformations of the organic spacers were visu-

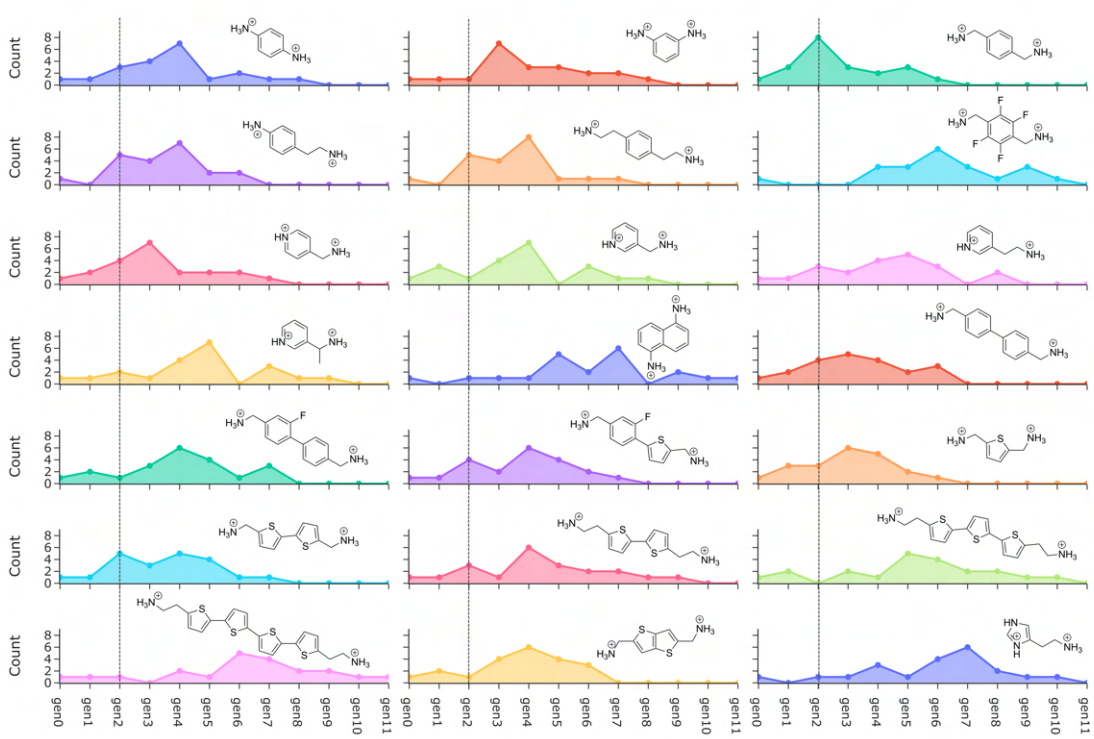


Figure 3.9: Rationale for selection of PDMA as the generation G_0 organic spacer.

alized in Avogadro and adjusted according to literature-reported configurations to reflect their likely conformations within the 2D perovskites. Specific modifications include:

(1) Ensuring the linearity of linker groups to reduce aggregation effects observed in isolated forms.

(2) Constraining rotations of linked aromatic rings to reflect dihedral angles typically found in hybrid perovskites, which are smaller than those in the isolated gas phase. For instance, benzene-benzene linkages were assigned a dihedral angle of 20°; thiophene-thiophene and thiophene-benzene linkages were constrained to 0°.

A $2 \times 2 \times 1$ supercell of PbI_6 octahedra was employed (each unit cell containing four organic spacers and four PbI_4 units), starting with an ideal cubic configuration in inorganic layers. Organic spacers are aligned along the lattice c direction, with two ammonium tethering groups intercalated within the cavities formed by the PbI_6 octahedra. Organic

spacers are arranged in herringbone (out-of-phase) configuration. This configuration was chosen as it is frequently observed in experimental studies, with minimal influence on the electronic structure compared to other configurations. Interlayer distances between neighbouring PbI_4 layers were modulated according to spacer length, ensuring realistic structural representation. The framework was built programmatically using a combination of RDKit and pymatgen.

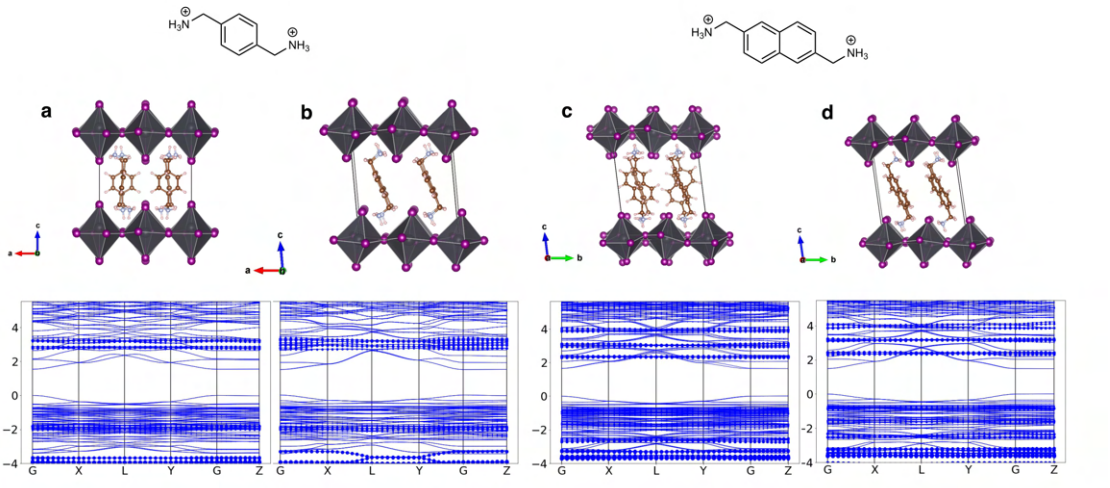


Figure 3.10: Crystal structures and band structures of DJ perovskites with different organic spacer packing arrangements. Two example spacers are shown. Panels **a** and **c** illustrate out-of-phase (herringbone) packing, while **b** and **d** show in-phase packing.

Figure 3.10 shows a comparison between different packing pattern of organic spacers and their influence on the energy level in 2D perovskites. Figure 3.10a and c shows the out-of-phase (herringbone) arrangement of organic spacers. This configuration is commonly observed in reported 2D perovskites, and our DFT calculation indicate this type of packing leads to minimal dispersion of molecular orbitals. Figure 3.10b and d shows the in-phase arrangement, a configuration typically found in oligomer thiophene-based spacers. Our DFT calculation indicate that the molecular orbitals exhibit sizable dispersion (~ 0.7 eV as observed in this study), consistent with previous reports attributing this behaviour to electronic coupling among tightly packed adjacent organic spacers[109]. Notably, one study indicates that out-of-phase configurations are energetically favoured[110]. For consistency, we adopt the out-of-phase packing pattern across all structures. We anticipate that this

in-plane dispersion of molecular orbitals will not affect our proposed final candidates for energy level alignment type Ib, IIa, and IIb, as this dispersion primarily broadens the organic frontier levels without shifting their centres.

Perovskite Structure Relaxation

DFT-based structure relaxations were performed using the Vienna Ab initio Simulation Package (VASP) with the Perdew-Burke-Ernzerhof (PBE) functional and projector augmented wave (PAW) pseudopotentials. Grimme's DFT-D3 dispersion correction with zero damping was included to account for van der Waals interactions critical to layered perovskites. Relaxation was conducted in two steps:

(1) A preliminary relaxation with a loose reciprocal density of 64 (resulting in k-point grids such as $1 \times 1 \times 1$ or $1 \times 1 \times 2$, depending on the lattice parameters along the c-axis).

(2) A tighter relaxation with a reciprocal density of 300 (resulting in k-point grids such as $3 \times 3 \times 2$, $3 \times 3 \times 3$, or $3 \times 3 \times 4$). Convergence criteria required an energy difference per atom below 5×10^{-6} eV.

Electronic Structure Calculations

To investigate the electronic properties of 2D perovskites, spin-orbit coupling (SOC) was included due to the significant relativistic effects in Pb-based compounds. The following workflow addressed the known limitations of commonly used functionals in predicting bandgap values and ensure alignment with experimental data:

1. Band structure shape identification: The band structure across the Brillouin zone was initially calculated using the PBE+SOC functional. This functional effectively captures the qualitative shape of the band structure but is known to underestimate the bandgap for perovskite materials. The results provided a foundational map of the electronic band dispersion.

2. Accurate bandgap calculation at representative points: To obtain accurate bandgap values, the HSE06+SOC hybrid functional was applied at critical points in the

Brillouin zone (Γ and Z). This method corrects the bandgap underestimation of PBE+SOC, ensuring quantitative agreement with experimental values. For structures with interlayer distances below 5 Å, calculations were performed at both Γ and Z points to capture the dispersion of inorganic bands from Γ to Z. For larger interlayer distances (greater than 5 Å), calculations were focused on the Γ point only, as the dispersion becomes negligible. A Hartree-Fock exchange (HF) percentage of 40% was used in the HSE06+SOC calculations, yielding bandgap values with a deviation of only 0.05 eV from experimental results.

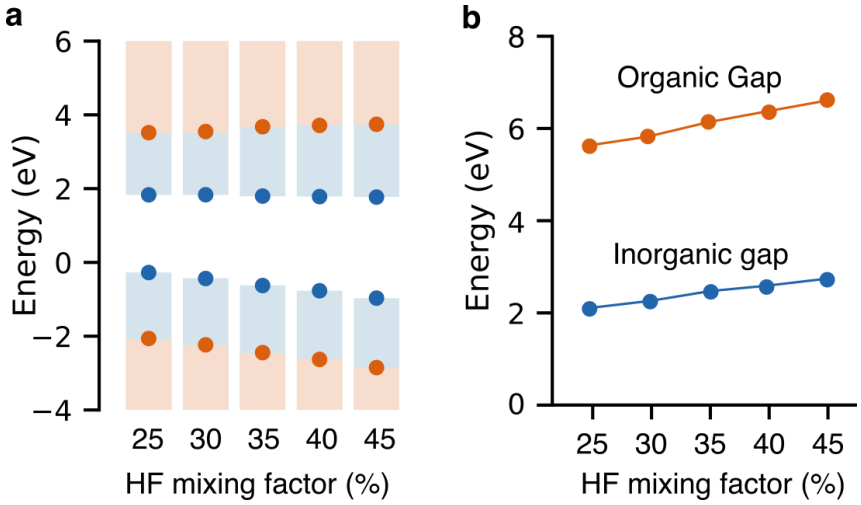


Figure 3.11: Effect of Hartree–Fock (HF) mixing factor on the calculated energy levels of organic and inorganic components in DJ-phase perovskites with the G_0 molecule (PDMA). **a** Variation in energy level alignment with increasing HF mixing. **b** Evolution of the organic and inorganic energy gaps as a function of HF mixing factor.

Figure 3.11 shows the impact of mixing factor on the organic frontier levels in DJ perovskite. A range of mixing factors (25% to 45%) has been employed in earlier studies to match experimental bandgap values. In this work, we select a mixing factor of 40%, as it yields the smallest average error of 0.05 eV. The mixing factor impacts the organic levels slightly more than the inorganic levels, though both follow very similar trends. This suggests that the energy level alignment type is unlikely to vary significantly for most proposed DJ perovskites.

Table 3.1 and 3.2 shows the bandgap of several existing DJ perovskites with different mixing factors, confirming the mixing factor of 40% yield the smallest error compared

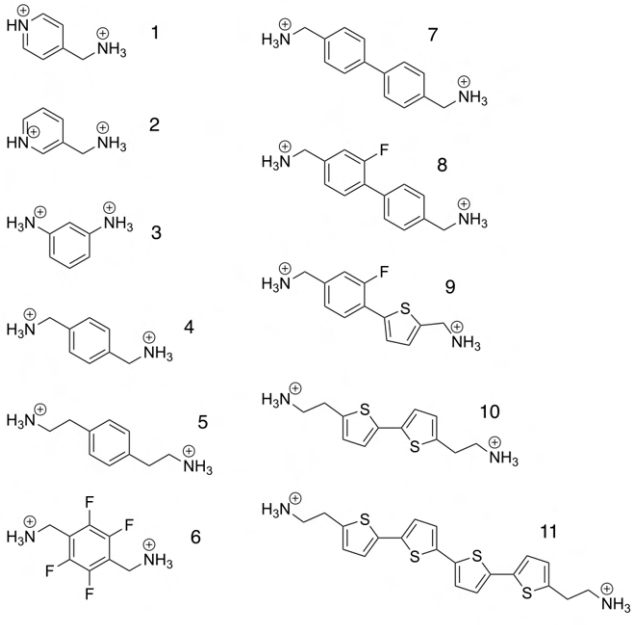


Figure 3.12: Reported organic spacers for which experimental bandgap values have been measured in the corresponding DJ perovskites.

Table 3.1: Comparison of bandgap values calculated using HSE + SOC, HF=40% and reported experimental values. Compound IDs correspond to the organic spacers listed in Figure 3.12.

Compound	**Calculation (eV)**	**Experiment (eV)**
1	2.42	-[79]
2	2.26	2.34[79]
3	2.39	2.42[111]
4	2.56	2.44[91], 2.42[112]
5	2.42	2.46[91], 2.43[113]
6	2.57	2.58[113]
7	2.38	2.52[114]
8	2.56	2.42[114]
9	2.56	2.39[114]
10	2.79	-[110]
11	2.38	2.38[97]

to experimental value.

3. Refined band structure across the Brillouin zone: To extend the high accuracy of HSE06+SOC calculations across the entire Brillouin zone without incurring the prohibitive computational cost, a scissoring technique was applied. The band structure

Table 3.2: Effect of HF mixing factor on the bandgap values of DJ perovskites. Compound IDs correspond to the organic spacers listed in Figure 3.12.

	Compound 2 (eV)	**Compound 4 (eV)**
Experiment	2.34[79]	2.44[91], 2.42[112]
25%	1.82	2.12
30%	1.97	2.26
35%	2.11	2.41
40%	2.26	2.56
45%	2.42	2.71

obtained from PBE+SOC was adjusted using the band edges calculated at the Γ and Z points with HSE06+SOC. The scissoring operation aligned the PBE+SOC-derived band structure with the inorganic band edge and organic frontier orbital levels predicted by HSE06+SOC, providing a consistent, high-fidelity representation of the electronic structure.

High-throughput framework

The high-throughput DFT calculations for both organic spacers and hybrid perovskite structures were conducted on the Gadi supercomputer, utilizing a workflow based on the Materials Project (MP) input parameter template. The MP standards are widely recognized as the minimal benchmark in the field of DFT calculations, ensuring reproducibility and consistency across studies. These parameters are particularly suitable for capturing the general structural and electronic properties of a wide range of materials.

To meet the specific demands of hybrid perovskite systems, which are characterized by strong spin-orbit coupling, van der Waals interactions, and low-symmetry structures, we further refined and tightened the computational parameters. Key refinements included:

- Energy Convergence Criterion: The MP default for electronic energy convergence (EDIFF) is 5×10^{-5} eV per calculation. For our study, we set a stricter threshold of 5×10^{-6} eV per atom to ensure reliable energy differences for low-symmetry perovskite structures.

- Plane-Wave Cutoff Energy: While the MP standard cutoff energy is typically 520 eV, we adjusted this to 480 eV, as testing showed that this value-maintained accuracy for perovskite systems while optimizing computational efficiency.

- Reciprocal Space Sampling: For initial structural relaxation, we used the default MP reciprocal density of 64 (resulting in k-point grids such as $1 \times 1 \times 1$ or $1 \times 1 \times 2$, depending on the c-axis lattice parameter). For final relaxation and electronic property calculations, we increased the reciprocal density to 300, corresponding to dense k-point grids such as $3 \times 3 \times 3$, $3 \times 3 \times 2$, or $3 \times 3 \times 4$. This stricter k-point sampling ensured accurate modelling of structural distortions and electronic properties in the layered perovskites.

- Dispersion Corrections: To account for van der Waals interactions in layered systems, Grimme's DFT-D3 dispersion correction with zero damping was included. This refinement is critical for accurately describing interlayer interactions in 2D perovskites.

The high-throughput framework was automated using the pymatgen library to generate VASP input files and parse outputs, enabling systematic and efficient exploration of $\sim 3,000$ organic spacers and ~ 400 hybrid perovskite structures. By building on the minimal standards established by the Materials Project and incorporating additional refinements specific to perovskites, we ensured that our computational results met the highest standards of accuracy and reliability for this material system.

3.4 Machine Learning

All machine learning tasks in this thesis were conducted using the Scikit-learn library in Python. This included data preprocessing, dimensionality reduction, training of regression and classification models, and extraction of feature coefficients for model interpretation.

Dimensional reduction for chemical space visualization

To visualize the chemical space of organic spacers, we employed unsupervised learning techniques for dimensionality reduction. We compared Principal Component Analysis (PCA), a linear dimensionality reduction method, with t-distributed stochastic neighbour embedding (t-SNE), a nonlinear technique designed to capture complex high-dimensional data structures in a lower-dimensional space. While PCA provided an initial overview, the resulting plots exhibited significant overlap between data points, limiting its ability to distinguish between structurally diverse spacers. Consequently, we selected t-SNE for its superior capability in preserving local and global relationships within the data.

Each spacer was represented numerically using a 12-dimensional fingerprint vector that encodes key structural and chemical features relevant to our analysis. The dataset, comprising $\sim 20,000$ fingerprints corresponding to $\sim 4 \times 10^6$ spacers generated across generations G_0 to G_6, was subjected to t-SNE analysis. We utilized a perplexity of 40 to balance the consideration of local and global data structures, optimizing the algorithm's sensitivity to both densely and sparsely populated regions of the chemical space.

The output of this process was a set of two-dimensional coordinates that effectively represent the high-dimensional chemical space of the spacers, enabling clear visualization of structural similarities and differences across spacer generations.

Data Collection and Input Features

The dataset comprised high-throughput computational data on 3,239 organic molecules across generations G_0 to G_4 with varying structural and electronic properties. The target properties for prediction were the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies of isolated organic spacers.

The input features were 12-digit organic fingerprints, which are highly relevant to the target properties due to their comprehensive representation of molecular descriptors. These fingerprints capture the chemical, electronic, and structural characteristics of the organic molecules, providing a robust basis for predictive modelling. Correlation analysis revealed minimal redundancy among the features, negating the need for additional feature selection

methods.

To ensure comparability across features, all input data were normalized to have zero mean and unit variance using the StandardScaler module in the Scikit-learn library. This normalization step mitigates bias arising from differences in feature scales, thereby optimizing model performance.

Model Training and Validation

The data was partitioned into training and test sets using an 80:20 random split, ensuring an unbiased evaluation of model performance. The training data was further subjected to five-fold cross-validation to ensure robustness and to avoid overfitting.

Table 3.3: Hyperparameters for various regression ML models for HOMO and LUMO predictions.

Method	**HOMO**	**LUMO**
Linear regression	no hyperparameter	no hyperparameter
Lasso regression	$\alpha = 0.001$	$\alpha = 0.001$
Ridge regression	$\alpha = 1.0$	$\alpha = 5.0$
Elastic net regression	$\alpha = 0.001$	$\alpha = 0.001$
SVM (kernel = linear)	$C = 20, \epsilon = 0.1$	$C = 10, \epsilon = 0.1$
SVM (kernel = rbf)	$C = 10, \epsilon = 0.1$	$C = 10, \epsilon = 0.1$
SVM (kernel = poly)	$C = 1, \epsilon = 0.1$	$C = 1, \epsilon = 0.1$
K neighbours regressor	$n\_neighbours = 7$	$n\_neighbours = 7$
Random forest regressor	$n\_estimators = 100$	$n\_estimators = 100$

We evaluated the performance of several machine learning models implemented in the Scikit-learn library. Grid search with cross-validation was used to identify optimal hyperparameters for each model, as summarized in Table 3.3.

All models with optimal hypterparameters were evaluated using 15-fold cross-validation (cv=15) implemented via Scikit-learn's cross_validate function to ensure consistent and reliable estimation of fitting error.

Performance metrics

The models were evaluated based on two key metrics:

(1) R^2 Score. Quantifying the proportion of variance explained by the model, with higher values indicating better fit.

$$R^2 = 1 - \frac{SS_{\text{RES}}}{SS_{\text{TOT}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.1}$$

(2) Root Mean Squared Error (RMSE): Providing an absolute measure of predictive accuracy, with lower values indicating smaller residual errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{3.2}$$

Both metrics were calculated for the training and test datasets. The results demonstrated excellent predictive performance for both linear (~ 0.95) and non-linear models (~ 0.99). This level of performance suggests that the relationships in the data are well captured by the chosen models, obviating the need for more advanced techniques, such as deep learning, for this study.

LASSO regressions were selected as the optimal model due to their decent predictive accuracy and interpretability. The importance of each input feature, both normalized coefficient and unnormalized coefficient, was analysed to verify the contribution of the organic fingerprint descriptors to the target predictions.

SHAP value analysis.

To interpret the feature importance and contribution of individual organic descriptors in predicting HOMO/LUMO, we employed SHapley Additive exPlanations (SHAP) values. SHAP values provide a game-theoretic approach to quantify each feature's impact on the model's output.

Unlike the conventional approach of using the average feature value as the reference point, we calibrated all SHAP values using a baseline molecule from Generation 0. This calibra-

tion allowed us to directly compare feature contributions relative to a chemically meaningful reference, facilitating more insightful interpretations.

The SHAP values were computed and visualized using the SHAP library in Python.

3.5 Synthesis feasibility screening

PubChem existence

The presence of an organic spacer in the PubChem database was used as a proxy for synthetic accessibility. PubChem is a comprehensive chemical information repository, widely used to assess the availability and feasibility of molecular synthesis.

The neural form of organic spacers is converted to SMILES format to ensure compatibility with PubChem's search algorithms. The neutral SMILES string was used to retrieve chemical information via the pubchempy library, which interacts with the PubChem API. Key identifiers, including the compound identifier (CID) and International Union of Pure and Applied Chemistry (IUPAC) name, were extracted for each molecule.

Due to limitations on the number of requests allowed by the PubChem API, this process was computationally slower than other components of the high-throughput pipeline. Therefore, synthetic accessibility screening was limited to the generations G_0-G_4 molecules, totalling approximately $\sim 10^4$ candidates.

2D structure formability

The formability of a 2D perovskite structure was evaluated based on the spatial and chemical properties of hydrogen-donor nitrogen atoms within the organic spacers. Four descriptors were used to quantify the formability of the candidate spacers:

1. Steric hindrance index (STEI): This descriptor measures the steric hindrance around a target nitrogen atom, calculated as the inverse sum of the cubed distances to all

other atoms in the molecule:

$$\mathrm{STEI}_{N_i} = \sum_{j=1}^{n} \frac{1}{\left(d_{N_i-\mathrm{Atom}_j}\right)^3} \tag{3.3}$$

A larger STEI indicates a higher density of nearby atoms, increasing steric hindrance and potentially reducing the likelihood of forming hydrogen bonds with the inorganic framework.

2. Eccentricity: Eccentricity quantifies the molecular shape with respect to a nitrogen atom, measuring the longest distance between the nitrogen and any other atom in the molecule:

$$\mathrm{Eccentricity}_{N_i} = \max\left(d_{N_i-\mathrm{Atom}_j}\right) \tag{3.4}$$

Larger eccentricity values correspond to more elongated molecules, which are favorable for forming 2D perovskite layers.

3. Number of rotatable bonds (NumRot): This descriptor reflects the flexibility of the ammonium group tethered to the nitrogen atom, calculated as the minimum distance from the nitrogen to the nearest atom in the conjugated backbone:

$$\mathrm{Num\_Rot}_{N_i} = \min\left(d_{N_i-\mathrm{RingAtom}_j}\right) \tag{3.5}$$

A higher number of rotatable bonds improves flexibility, facilitating the anchoring of the organic spacer to the inorganic framework.

4. N-N pair distance (Dis_{NN}): This descriptor measures the spatial separation between two nitrogen atoms in a molecule:

$$\mathrm{Dis}_{N_i-N_j} = \frac{1}{\left(d_{N_i-N_j}\right)^2} \tag{3.6}$$

A smaller value indicates a larger distance, reducing repulsive interactions and enhancing structural stability in 2D perovskite layers.

All descriptors were computed from the distance matrix of organic spacers, which is obtained using RDKit library in python. Briefly, all hydrogen-donor nitrogen atoms were

identified in the distance matrix, and their corresponding descriptors were calculated using the equation above.

A boundary-based approach was applied to screen out unsuitable organic spacers. Decision boundaries were defined using the properties of known, successfully synthesized spacers from the literature. Molecules failing to meet the criteria for any of the four descriptors were excluded from further consideration. This systematic screening ensured that only candidates with favourable synthetic accessibility and structural formability proceeded to the subsequent stages of analysis.

Chapter 4

High-Throughput Calculation and Machine Learning Predictions

In this chapter, we describe the generation of hypothetical organic spacers and the subsequent prediction of their properties using a combined high-throughput DFT approach and machine learning techniques. Section 4.1 outlines the systematic expansion of the organic spacer library through molecular morphing operations. Section 4.2 presents general physical insights derived from high-throughput DFT calculations. Sections 4.3 and 4.4 discuss the selection and evaluation of machine learning models, highlighting their predictive performance and their role in extracting structure-property relationships.

4.1 Molecular generation for chemical space expansion

4.1.1 Morphing operation

To systematically explore a broad and diverse chemical space, we employed a deterministic molecular morphing approach rather than stochastic molecular generation methods[103], [115]. This ensures controlled diversification while maintaining the chemical interpretabil-

	Number of existing organic spacers	Expanded chemical space	Pubchem existence
G_0	1	1	1
G_1	3	15	13
G_2	8	236	94
G_3	3	2,987	411
G_4	2	35,495	1,367
G_5	3	401,932	-
G_6	1	4,446,434	-

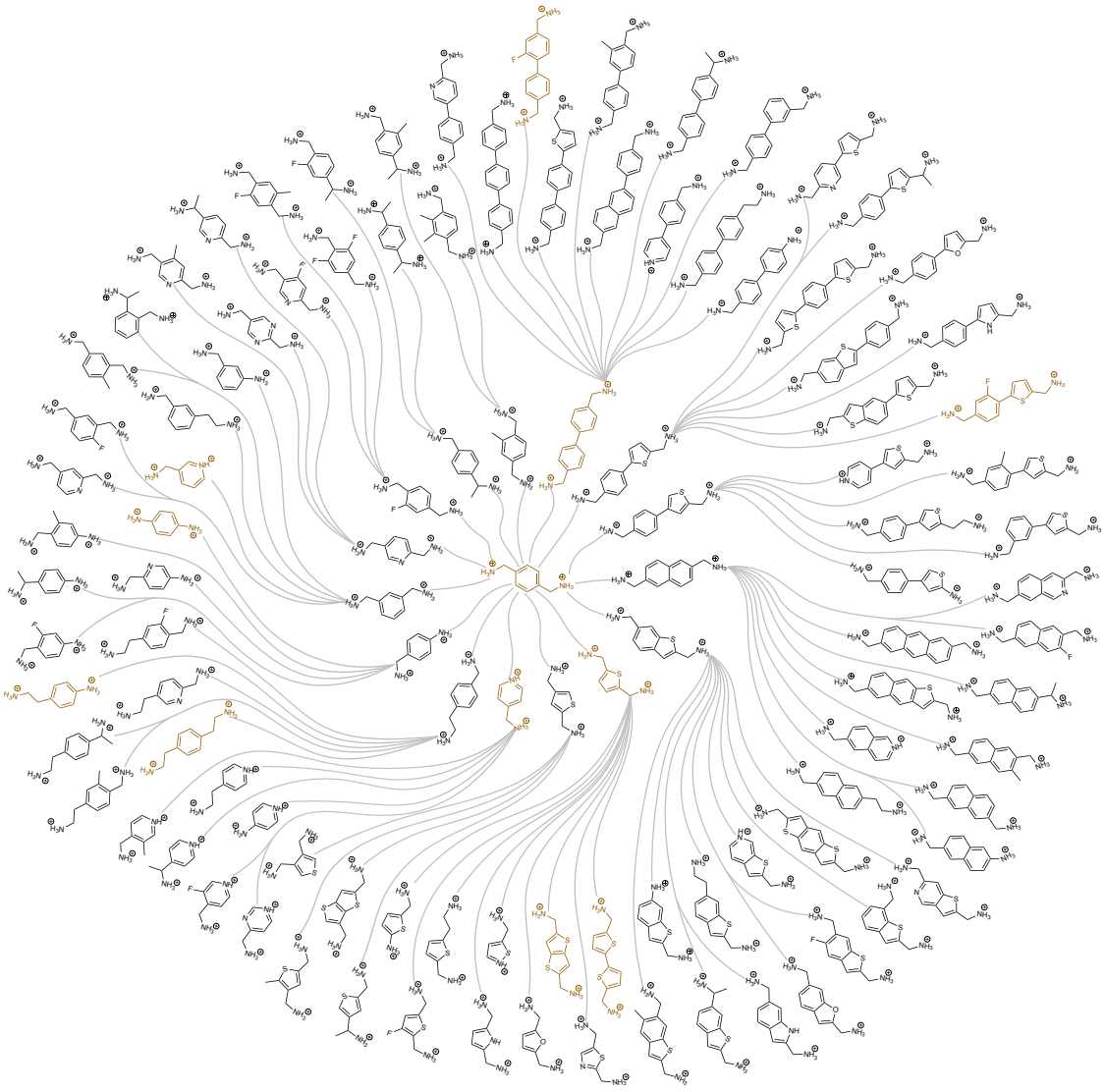


Figure 4.1: Scaffold tree plot illustrating the organic spacer generation process.

ity of the generated spacers. As detailed in Chapter 3, each spacer is represented by a 12-digit fingerprint vector, which encodes key molecular features such as π-conjugation, ammonium tethering, heteroatom substitution, and side-chain modifications. All the morphing operators are associated with organic desciptors in the molecular fingerprint (Table 4.1). This approach enables the enumeration of structurally diverse yet chemically meaningful molecular variations, ensuring a comprehensive and uniform coverage of the chemical space.

We initiate the morphing process from PDMA (Generation 0, G_0), a well-characterized organic spacer[112]. We iteratively apply 13 distinct morphing operators to introduce incremental modifications, systematically generating higher-order generations ($G_1 - G_6$). The molecular generation process is illustrated in Figure 4.1. All organic spacers in G_0 (centre core), G_1(first circle) are displayed. For G_2 (second and third circle), only representative structures with unique molecular fingerprints to maintain clarity. Experimentally reported molecules within $G_0 - G_2$ are highlighted, while the inset table quantifies the exponential expansion of the chemical space, which increases from an initial set of 21 reported organic spacers to millions of hypothetical spacers within $G_0 - G_6$.

Table 4.1: List of organic descriptors and their associated morphing operators.

Number	**Morphing operation**	**Organic descriptor**
1	Benzene-thiophene ring exchange	five-membered ring
2	Ring linkage	ring linkage
3	Ring fusion	ring fusion
4	Primary secondary amine exchange	number of primary amine
5	Linker length increase	Linker length
6	Linker length decrease	Linker length
7	Linker position change	linker position
8	Hetero-nitrogen substitution	hetero-nitrogen
9	Fluorination	fluorination
10	Furan exchange	furan
11	Pyrrole exchange	pyrrole
12	Side chain on backbone	no. side chain on backbone
13	Side chain on linker	no. side chain on linker

The morphing operations yield progressively complex sets of organic spacers. For example, to incorporate five-membered ring backbones, we utilize a ring contraction operator that

transforms benzene into thiophene, thereby diversifying the core molecular framework. This systematic expansion extends beyond commonly studied phenyl- and thiophene-containing families, encompassing a broader spectrum of heteroatom-substituted structures (e.g., F, O, N) and various side-chain modifications.

The 21 experimentally reported organic spacers were captured within generations $G_0 - G_6$, and with comparable complexity in these generations, we enumerated 21,306 fingerprints, corresponding to 4,887,100 hypothetical organic spacers. To assess the chemical feasibility of the generated molecules, the neutral forms of the hypothetical spacers were cross-referenced against the PubChem database. Within generations $G_0 - G_4$, where computational feasibility allowed exhaustive searches, $\sim 10^3$ spacers were identified in PubChem, confirming that a subset of our generated structures aligns with known chemical compounds. This validates the chemical plausibility of the generated molecular space.

4.1.2 Visualization of generated chemical space

To provide an intuitive understanding of the distribution and structural diversity of the generated organic spacers, we employ a dimensionality reduction technique to visualize the chemical space. Since the molecular fingerprints exist in a 12-dimensional space, we utilize t-distributed stochastic neighbour embedding (t-SNE)[28], an unsupervised learning algorithm that transforms high-dimensional data into a two-dimensional representation while preserving the relative distances between structurally similar molecules.

The resulting t-SNE map is shown in Figure 4.2, representing $\sim 2 \times 10^4$ unique fingerprints corresponding to $\sim 4 \times 10^6$ organic spacers across generations $G_0 - G_6$. Experimentally reported spacers are highlighted, with representative molecules from G_0 and G_6 explicitly labelled. The spatial organization of the t-SNE clusters reflects the underlying structural similarities among different spacers: closely grouped molecules share similar fingerprint features, while larger inter-cluster distances correspond to structurally dissimilar spacers.

Notably, among all reported spacers, the highest-generation example (G_6), AE4T[97], is

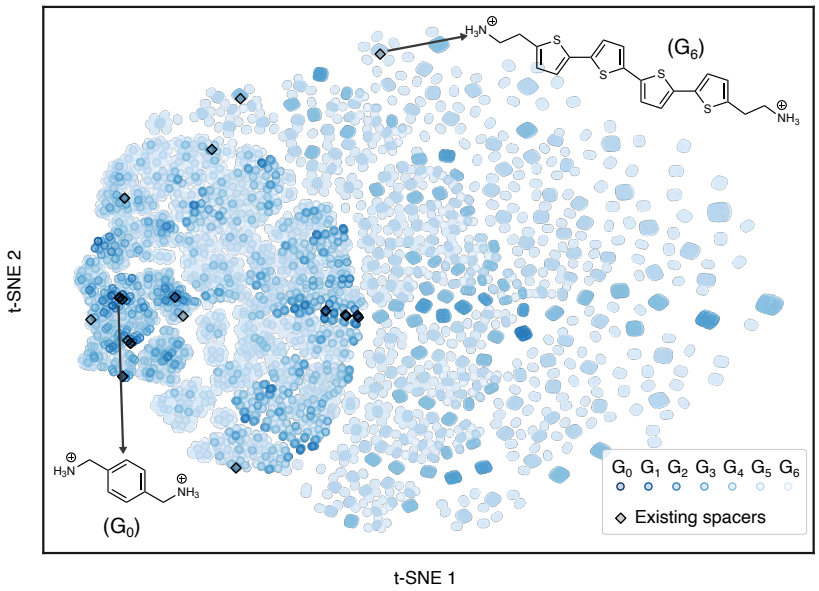


Figure 4.2: t-SNE representation of the generated chemical space containing the hypothetical organic spacers.

distinctly separated from other spacers, reflecting its more complex structure. This visualization demonstrates that our generative workflow comprehensively covers the chemical space surrounding experimentally known spacers, including AE4T. Compared to traditional molecular datasets compiled from pre-existing databases, our approach ensures a more uniform and representative sampling, avoiding biases toward extreme or rare structures. By maintaining a balanced distribution of generated molecules, we enable reliable high-throughput calculations and robust machine learning predictions. As we will demonstrate later in this chapter, this curated dataset serves as a reliable foundation for training machine learning models, enhancing their ability to predict electronic properties with high accuracy.

4.1.3 Descriptor-based visualization and cluster analysis

To further examine the relationships between molecular descriptors and structural similarity, we present Figure 4.3, which depicts the same t-SNE map but color-coded according

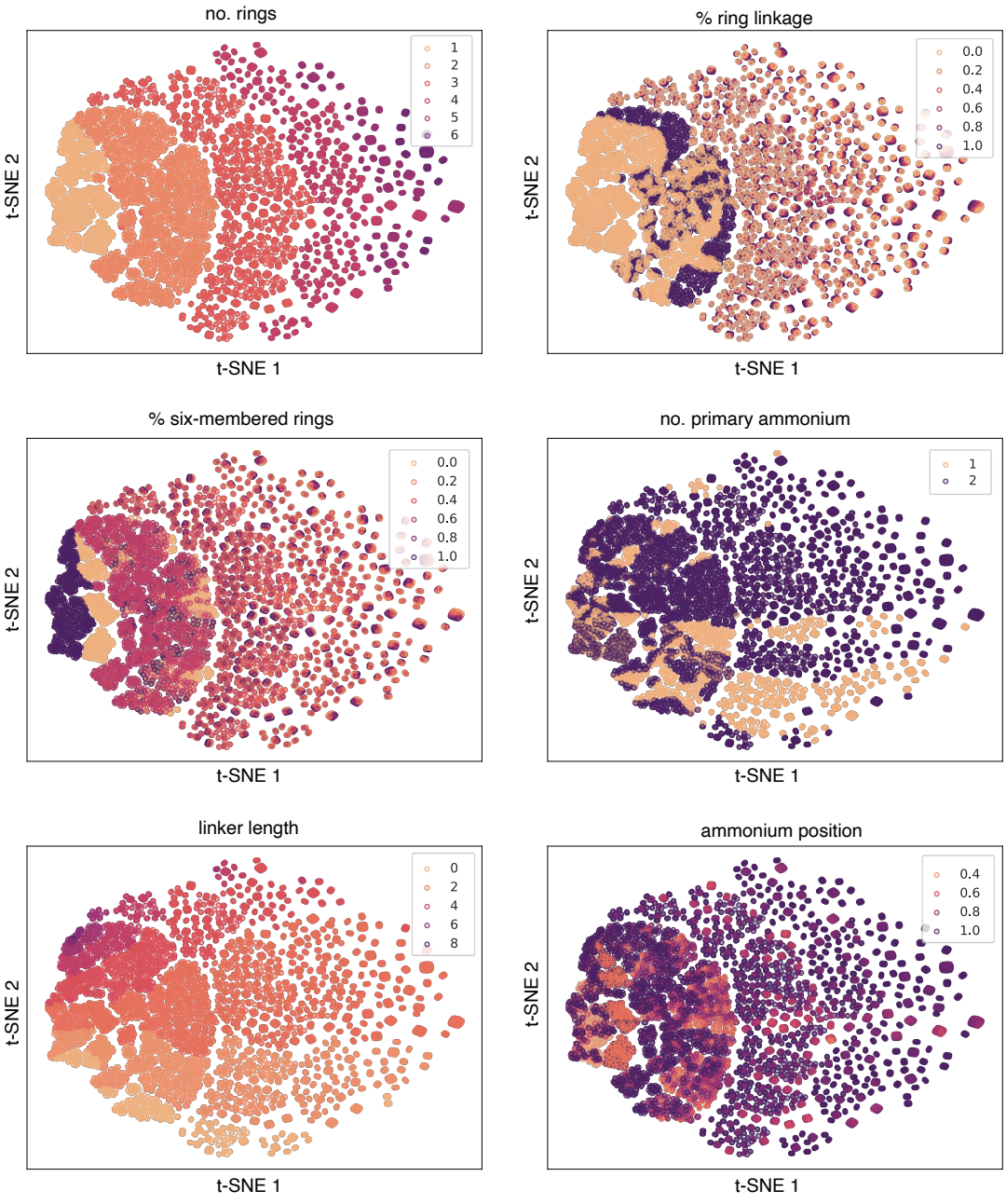


Figure 4.3: Visualization of the chemical space with respect to organic descriptors in molecular fingerprint (Part 1).

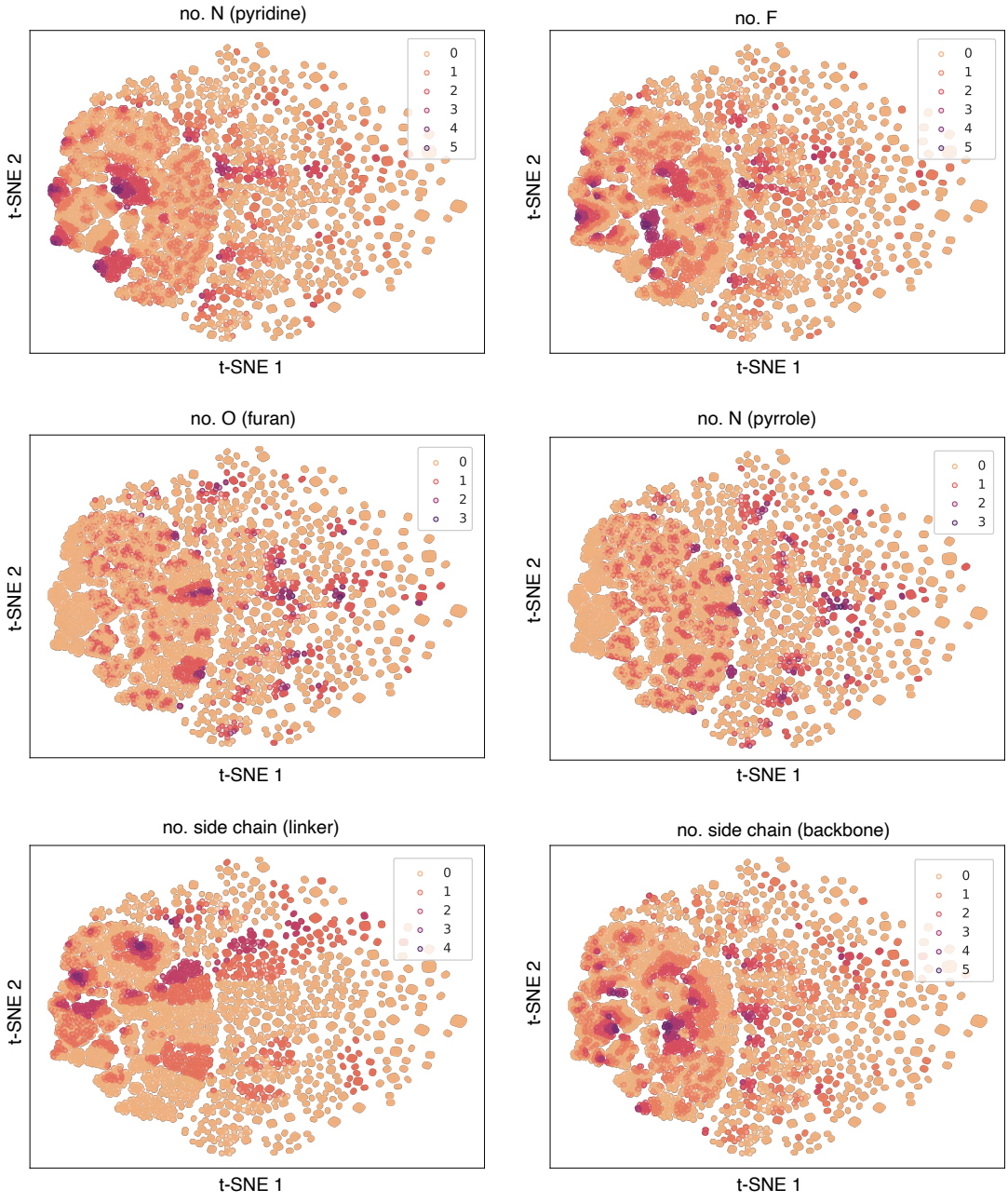


Figure 4.3: Visualization of the chemical space with respect to organic descriptors in molecular fingerprint (Part 2, continued).

to different molecular descriptors derived from the fingerprint vector. This visualization highlights key trends:

1. Molecular spacers with similar descriptors cluster together, confirming that the fingerprint representation effectively captures structural relationships.

2. The non-linear characteristic of t-SNE is well observed and it makes it easier to understand how t-SNE work. Note that the non-linearity organization of data points reflects the fundamental difference between t-SNE and principal component analysis (PCA). While PCA, a linear dimensionality reduction technique, is widely used in materials informatics, we found that it resulted in excessive overlap between structurally distinct molecules. By contrast, t-SNE preserves local structures more effectively, providing better differentiation of organic spacers.

A deeper analysis of specific clusters within the t-SNE map is presented in Figure 4.4, focusing on the existing organic spacers. The full distribution of the generated organic spacers is shown in Figure 4.4a. A clear gradient is observed from left to right, reflecting increasing number of aromatic rings. Spacers such as PDMA and ThDMA, which contain a single aromatic ring and are widely used in 2D perovskites, cluster on the left side of the plot. On the other hand, highly conjugated organic spacers, such as the oligothiophene-based spacers (AE2T, AE3T, AE4T) are found progressively further to the right.

Figure 4.4 b,c provide focused views of subspaces containing spacers with a single aromatic ring. In **b**, molecules are colored by ring type, highlighting distinct clustering between five-membered ring and six-membered ring as conjugated backbones. In **c**, the same subset is colored by linker length. A spatial gradient is visible, with shorter linkers (e.g., PDA, linker length = 0) clustering in the top and longer linkers (e.g., PDEA, linker length = 4) spreading toward the bottom. This demonstrates that the fingerprint-based t-SNE projection effectively captures meaningful structural and functional diversity among known spacers.

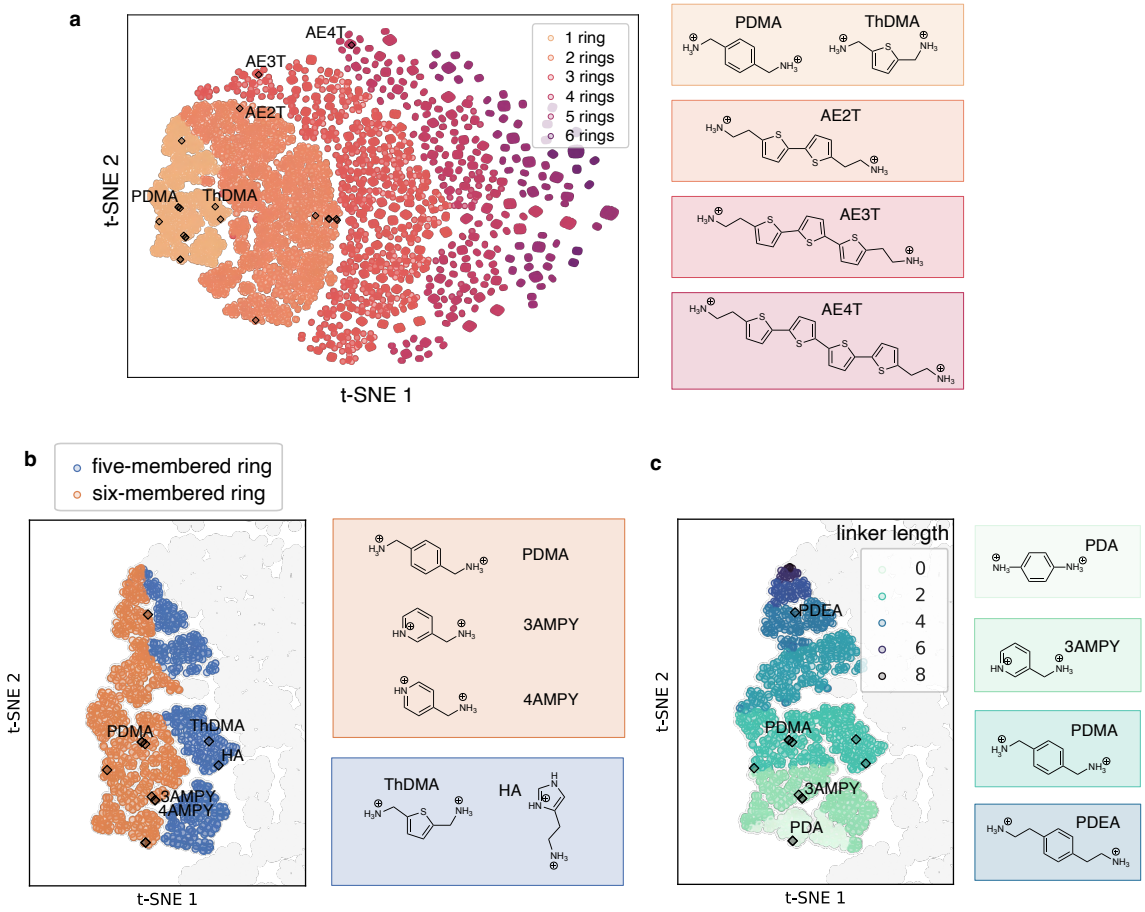


Figure 4.4: Chemical space visualization of existing spacers. **a** Complete chemical space. **b,c** Enlarged views highlighting spacers containing a single aromatic ring.

4.2 Influence of organic spacer structure on energy levels

4.2.1 DFT calculation of DJ perovskites

To assess the impact of organic spacer selection on the electronic properties of DJ perovskites, we performed a detailed high-throughput DFT analysis on 261 DJ-phase perovskites. This dataset includes both experimentally reported spacers and hypothetical spacers from generations $G_0 - G_2$ of our expanded chemical space. Model crystal structures were constructed by inserting organic spacers between the PbI_4 layers, with each unit cell containing four diammonium spacers and four PbI_4 units (see Chapter 3 for computational details).

Organic spacer packing configurations

In computational studies of DJ perovskites, two primary packing arrangements of organic spacers are typically assumed: in-phase and out-of-phase configurations. Unlike the inorganic sublattice, which undergoes significant structural distortions (such as octahedral tilting and bond-angle variations), the organic spacers exhibit minimal changes in packing pattern upon structural relaxation. This behaviour suggests that interactions between organic spacers and the inorganic layers, as well as interactions among the organic spacers themselves, are relatively weak, occurring primarily through hydrogen bonding and van der Waals interactions. Consequently, both in-phase and out-of-phase arrangements lead to well-converged structures, indicating that the energy difference between these configurations is small. This suggests a shallow energy landscape with respect to the relative alignment of organic spacers.

Despite the energetic similarity between different packing arrangements, our analysis revealed that the packing configuration has minimal influence on the energy level alignment. Therefore, to align with experimentally observed structures, all organic spacers were assign to arranged in herringbone (out-of-phase) configurations[110]. The optimized geometries were computed at the DFT-PBE level.

Bandgap calibration and energy level alignment

The energy level alignment between the organic frontier molecular orbitals and the inorganic band edges was subsequently determined using HSE + SOC calculations. A key parameter influencing the computed bandgap values is the mixing factor, which defines the fraction of exact exchange in the hybrid functional. Previous studies on 2D perovskites have used mixing factors ranging from 0.25 to 0.45 to match experimental bandgaps. Our calculations confirm that while the choice of mixing factor significantly alters absolute bandgap values, its impact on energy level alignment trends is minimal, as both organic and inorganic components exhibit a similar response to changes in the mixing factor. To ensure consistency with experimental data, we benchmarked our DFT results against available experimental bandgap values, finding that a mixing factor of 0.4 provided the best agreement (see Chapter 3 for further discussion).

Energy level trends in DJ perovskites

Our analysis revealed that the majority of DJ perovskites (18 out of 21 experimentally reported structures) exhibit type Ia energy level alignment, characterized by low-energy electrons and holes localized in inorganic layers. The remaining three structures exhibit type IIa alignment. The primary factor dictating this variation in energy level alignment is the frontier molecular orbitals levels of the organic spacers, which exhibit a broad energy variation of $\sim$ 6.1 eV. In contrast, the inorganic band edges remain relatively invariant, with a variation of only $\sim$ 0.9 eV (see Figure 4.5). This observation aligns with the common approximation cited in the literature that inorganic energy levels of 2D perovskites can be assumed unchanged with different organic spacers[9], [72].

Furthermore, Figure 4.5 illustrates how our expanded chemical space exploration has significantly extended the range of organic frontier levels compared to previous studies. The generated DJ perovskites—including $G_0 - G_2$ and inverse designed final candidates (introduced later in Chapter 5)—exhibit an expanded energy distribution of organic frontier levels, covering a much wider range than the reported organic spacers, underscoring the effectiveness of our chemical space exploration approach.

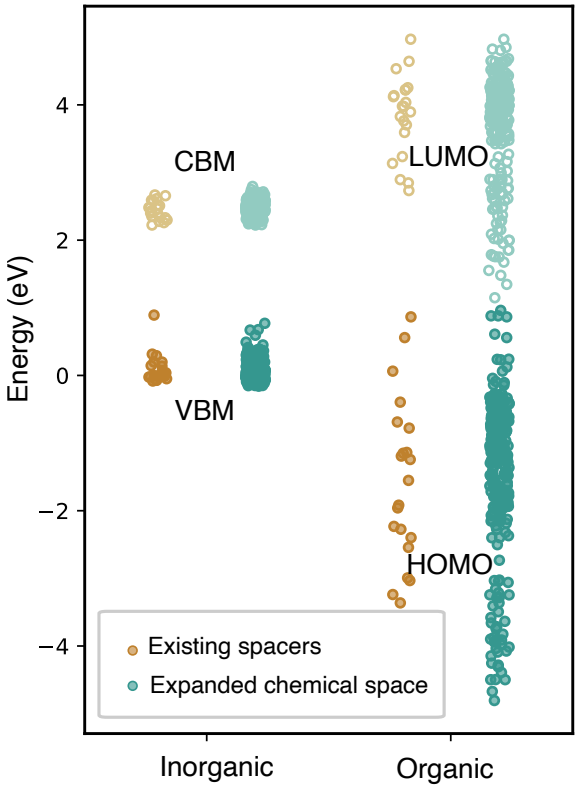


Figure 4.5: Energy level alignment in DJ perovskites with existing spacers and hypothetical organic spacers.

4.2.2 Four factors governing energy level alignment

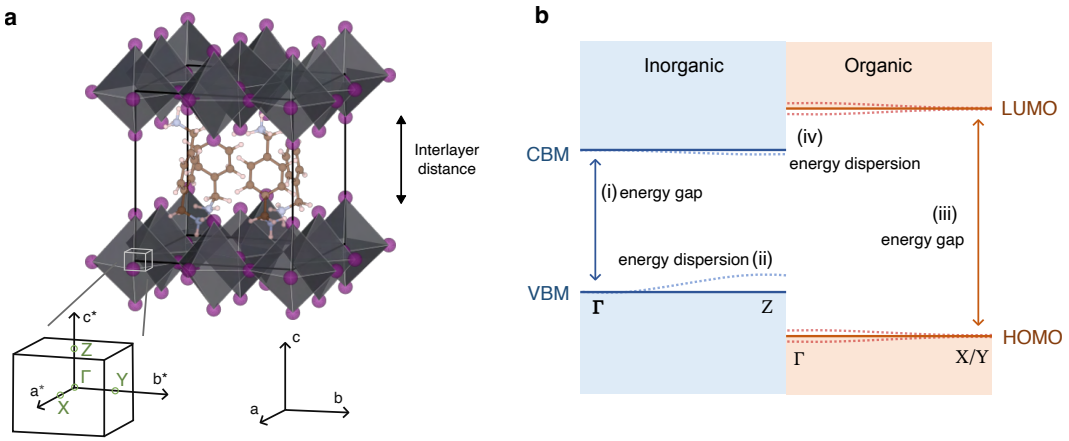


Figure 4.6: Four factors affecting the energy level alignment in DJ perovskites.

To further understand the structure-property relationships governing energy level alignment in DJ perovskites, we analysed key electronic band structure trends across all structures studied. Four dominant factors were identified, as illustrated in Figure 4.6:

(i) Energy gap at the inorganic band edge (Γ point)

(ii) Inorganic band dispersion along the stacking direction (Γ to Z)

(iii) Energy gap between the organic frontier orbitals (HOMO-LUMO levels)

(iii) Frontier orbital dispersion of the organic spacers

Electronic structure of inorganic component

The inorganic layers in DJ perovskites typically form direct bandgap semiconductors with their valence band maximum (VBM) and conduction band minimum (CBM) located at the Γ point in the Brillouin zone. However, in cases where interlayer coupling is significant, the bandgap shifts to the Z point.

The band dispersion is strongly anisotropic, exhibiting strong dispersion in the $\Gamma - X/Y$ direction (in-plane directions in real space), while the dispersion in the $\Gamma - Z$ direction

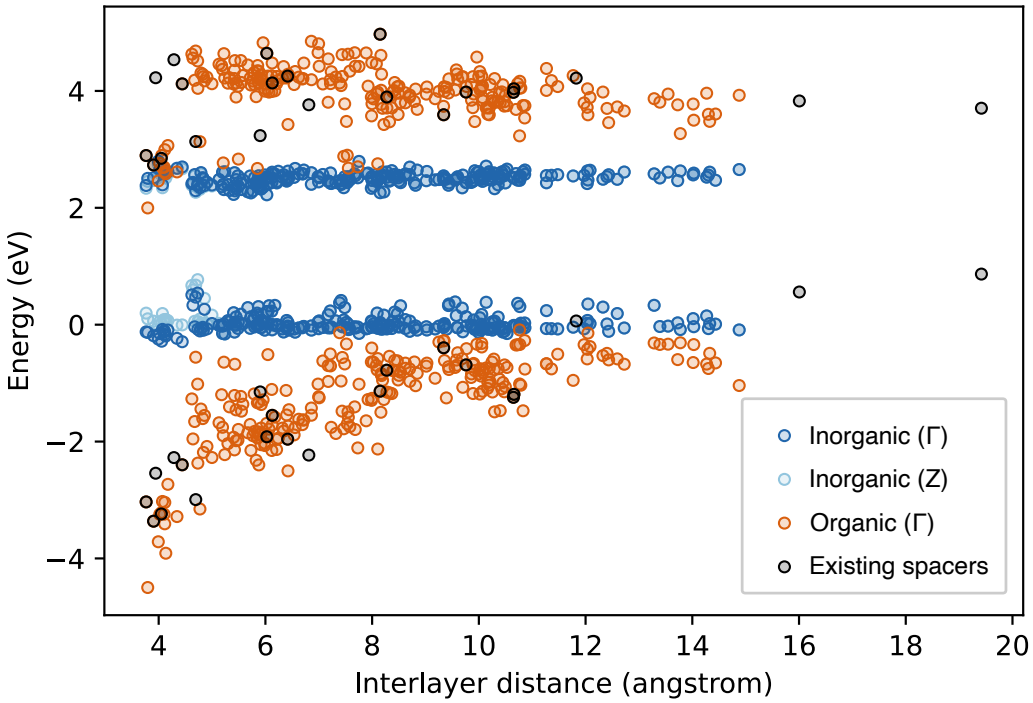


Figure 4.7: Energy level alignment in calculated DJ perovskites plotted against interlayer distance.

(stacking direction) depends on interlayer interactions. This trend of interlayer interaction is illustrated in Figure 4.7, where the energy level alignment is plotted against the interlayer distance. Interlayer coupling, characterized by the energy difference between the Γ and Z point of the inorganic band edge, becomes significant only when the interlayer distance is relatively small.

To further quantify the structure-property relationship between inorganic framework geometry and energy levels, we examined two key structural factors (Figure 4.8):

- Factor (i) in Figure 4.6 is mainly affected by octahedral tilting and distortion. As shown in Figure 4.8a, variations in the Pb-I-Pb bond angle and I-Pb-I internal distortion significantly impact the inorganic energy gap at the Γ point, leading to energy shifts of approximately 1 eV. These distortions arise due to hydrogen bonding interactions between the PbI_6 octahedra and the organic cations.
- Factor (ii) in Figure 4.6 is mainly affected by interlayer distance. As shown in Figure

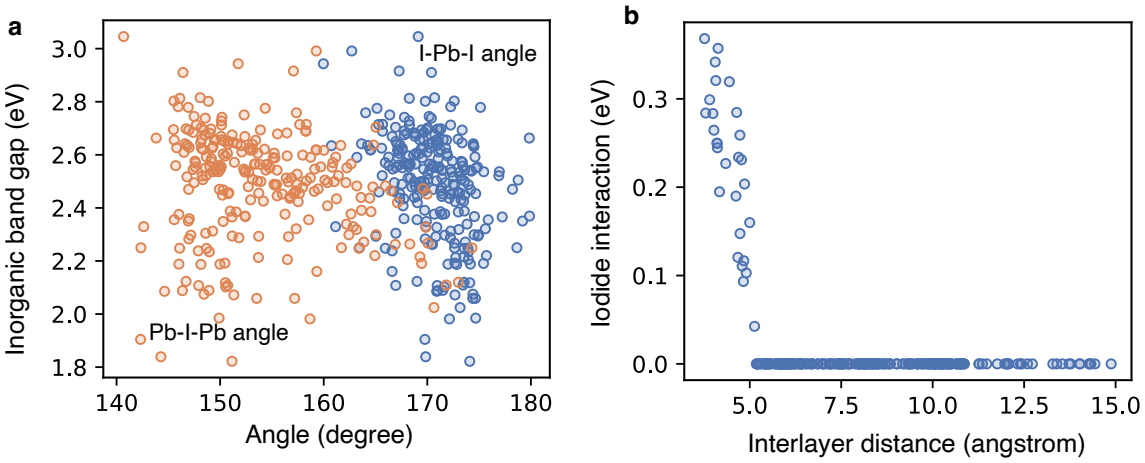


Figure 4.8: Indirect influence of organic spacers on inorganic band edges.

4.8b, when the interlayer distance decreases below 5.0 Å, iodide-iodide interactions enhance $\Gamma - Z$ energy dispersion, also inducing ~1 eV energy shifts. This behaviour is commonly observed in DJ-phase perovskites with short organic spacers and is consistent with trends reported in ACI-phase perovskites[63], [79].

Electronic structure of the organic component

Unlike the band-like behaviour of the inorganic layers, the organic frontier orbitals (HOMO and LUMO) remain localized and discrete, with minimal dispersion, closely resembling their isolated molecular forms. This occurs because:

- There is not interlayer interaction between organic spacers in adjacent layers, resulting in zero dispersion along the $\Gamma - Z$ direction.
- The in-plane ($\Gamma - X/Y$) dispersion is also minimal due to the herringbone packing of organic spacers, which restricts electronic interactions between adjacent organic units. More detailed discussions on the relationship between packing pattern and in-plane dispersion can be found in Chapter 3.

4.2.3 Simplified modelling of organic frontier levels

Our analysis confirms that the primary influence of organic spacers on DJ perovskite energy levels lies in their HOMO and LUMO levels, which is primarily a consequence of weak bonding interactions between organic cations and the inorganic framework[6], [81].

However, the high computational cost of DFT calculations for large DJ perovskite unit cells limits our ability to simulate thousands of structures. To address this, we propose an efficient approximation: Organic frontier energy levels in hybrid perovskites can be estimated using calculations on isolated cations.

We computed frontier levels of isolated organic spacers using the B3LYP functional in Gaussian. Figure 4.9 compares the frontier orbital energies obtained from hybrid perovskite calculations and isolated organic cations. While absolute energy values differ because of functional choices, basis sets, and the chemical environment, a near-linear relationship emerges between the two calculations:

- Figure 4.9a shows that HOMO levels exhibit a linear correlation between hybrid perovskites and their isolated cation counterparts.
- Figure 4.9b highlights a similar trend in LUMO levels, with noticeable dependence on the ring count of the organic spacer.
- Figure 4.9c indicates that indicates that the HOMO–LUMO gap remains consistent between the two methods.

By leveraging these correlations, we can predict organic frontier levels without the need for fully relaxed DJ perovskite calculations for each candidate spacer. This significantly reduces the computational cost, enabling us to scale from hundreds to thousands of hypothetical structures—an essential step for generating robust datasets to train machine learning models in the subsequent sections.

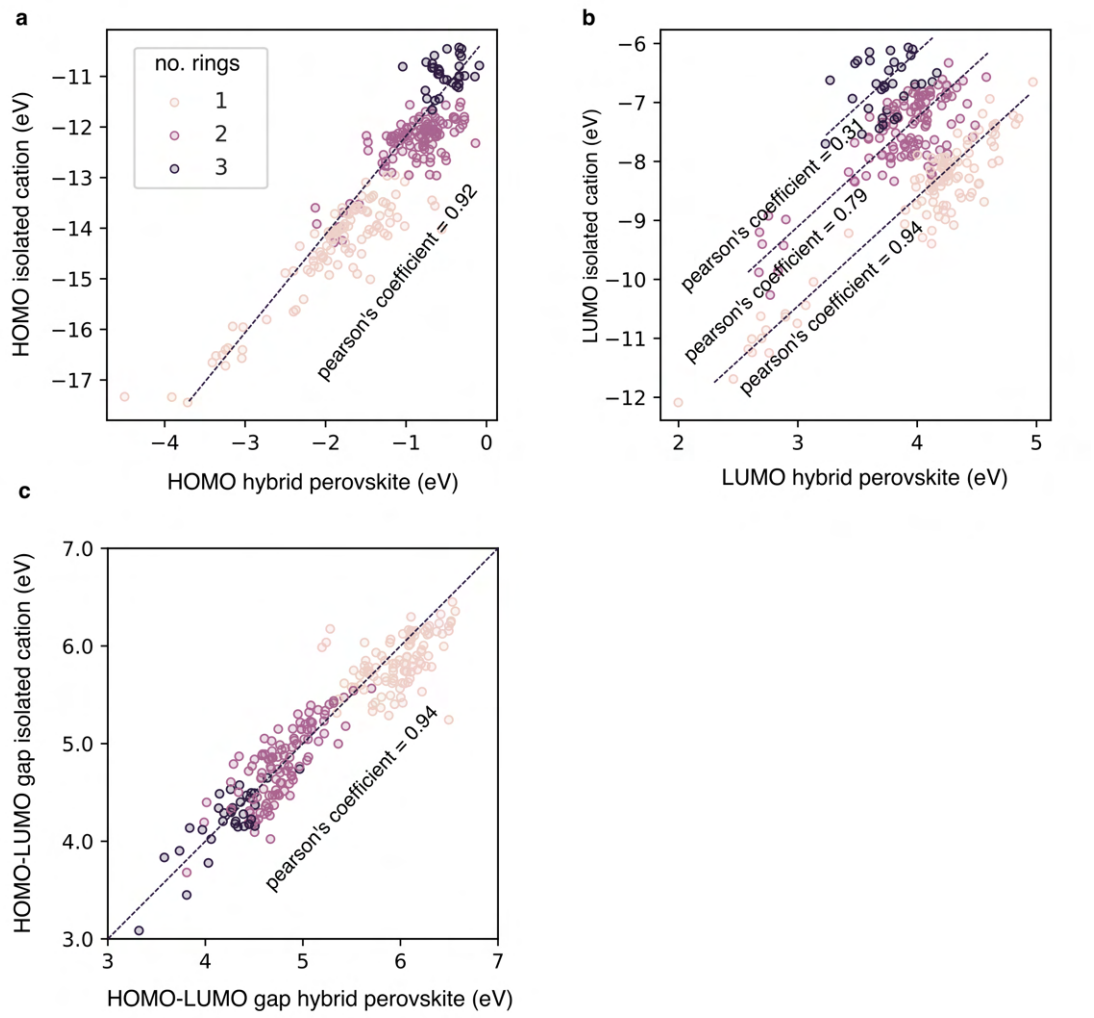


Figure 4.9: Correlation between organic frontier levels in hybrid perovskites and their isolated molecular forms.

4.3 Selection and evaluation of machine learning models

In the previous sections, we established how organic spacers affect the energy level alignment in DJ perovskites, and we identified key structural descriptors that can serve as input features for predictive modelling. Here, we describe how ML is employed to capture the structure–property relationships between molecular fingerprints and their frontier energy levels, and facilitate the rapid identification of promising spacer candidates.

4.3.1 Molecular fingerprint as input feature

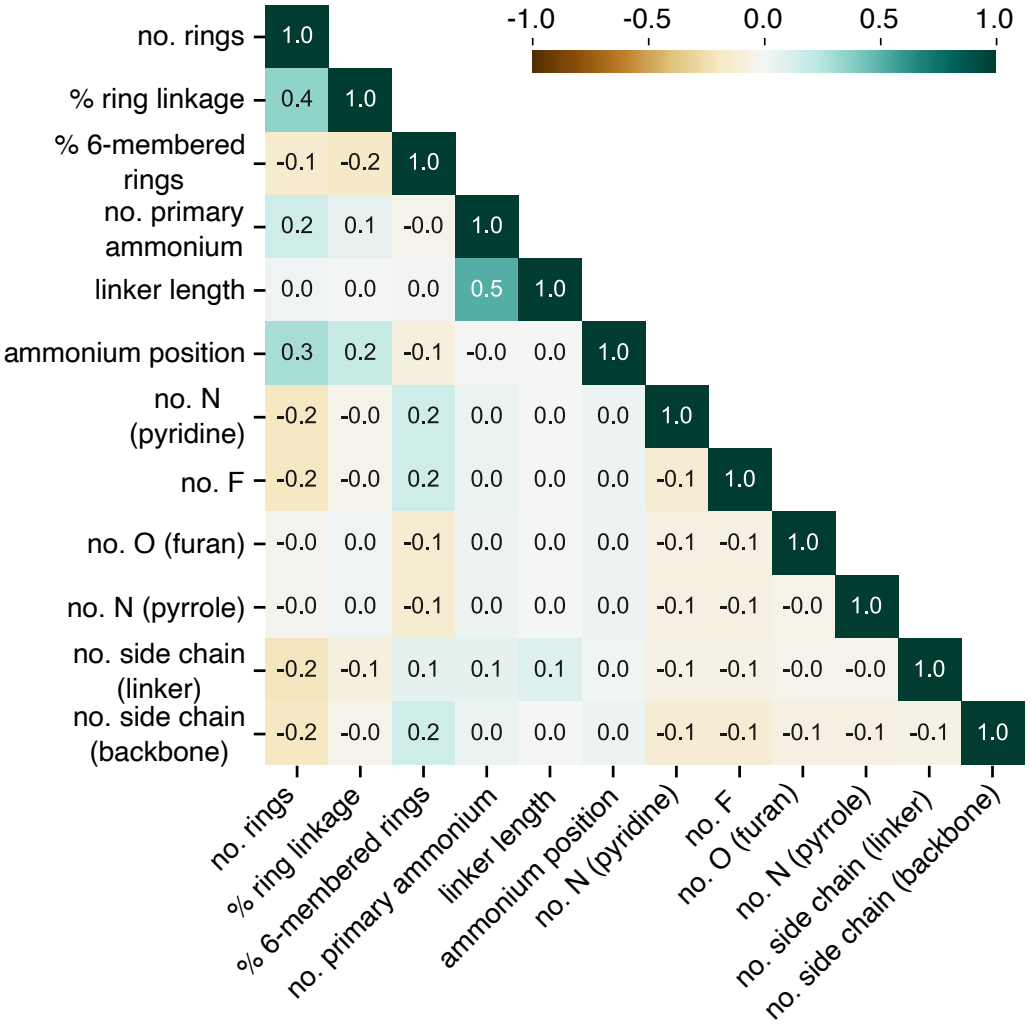


Figure 4.10: Correlation matrix of organic descriptors in the molecular fingerprint.

To represent each organic spacer, we use a 12-digit molecular fingerprint as the input feature for our machine learning pipeline. This fingerprint does not require additional feature selection, as it was specifically designed to minimize redundancy among descriptors. Indeed, Pearson's correlation coefficients between fingerprint descriptors are all below 0.5 (see Figure 4.10), confirming low inter-feature correlation and justifying the direct inclusion of all 12 features in model training. This is a significant advantage compared to previous studies that rely on diverse chemical descriptors and require feature selection to reduce multicollinearity[7], [8].

Our machine learning dataset consists of 3,239 organic spacers spanning generations $G_0 - G_3$, with HOMO/LUMO values from high-throughput calculations serving as the target properties. This dataset allows us to explore a broad range of structures and electronic characteristics, setting the stage for robust model training.

4.3.2 Comparison of different models

Because our goal is to predict HOMO and LUMO energies from known labels (i.e., computed frontier levels), this is a supervised regression problem. We trained separate machine learning models for HOMO and LUMO predictions respectively, with the dataset split into training and testing sets (80: 20 ratios). All models were evaluated using 15-fold cross-validation to ensure consistent and reliable estimation of fitting error. A variety of linear and non-linear regression methods commonly used in materials science literature were benchmarked, including:

- Linear models: linear regression, LASSO regression, Ridge regression, Elastic net regression, Support vector regression with a linear kernel.
- Non-linear models: Random Forest, K nearest neighbour, Support Vector Regression with radial basis function (rbf) kernel or polynomial kernel.

Figure 4.11 summarizes the R^2 scores (ranging from 0 to 1, with higher values indicating better predictions) for each model. Across both HOMO and LUMO predictions, nonlinear

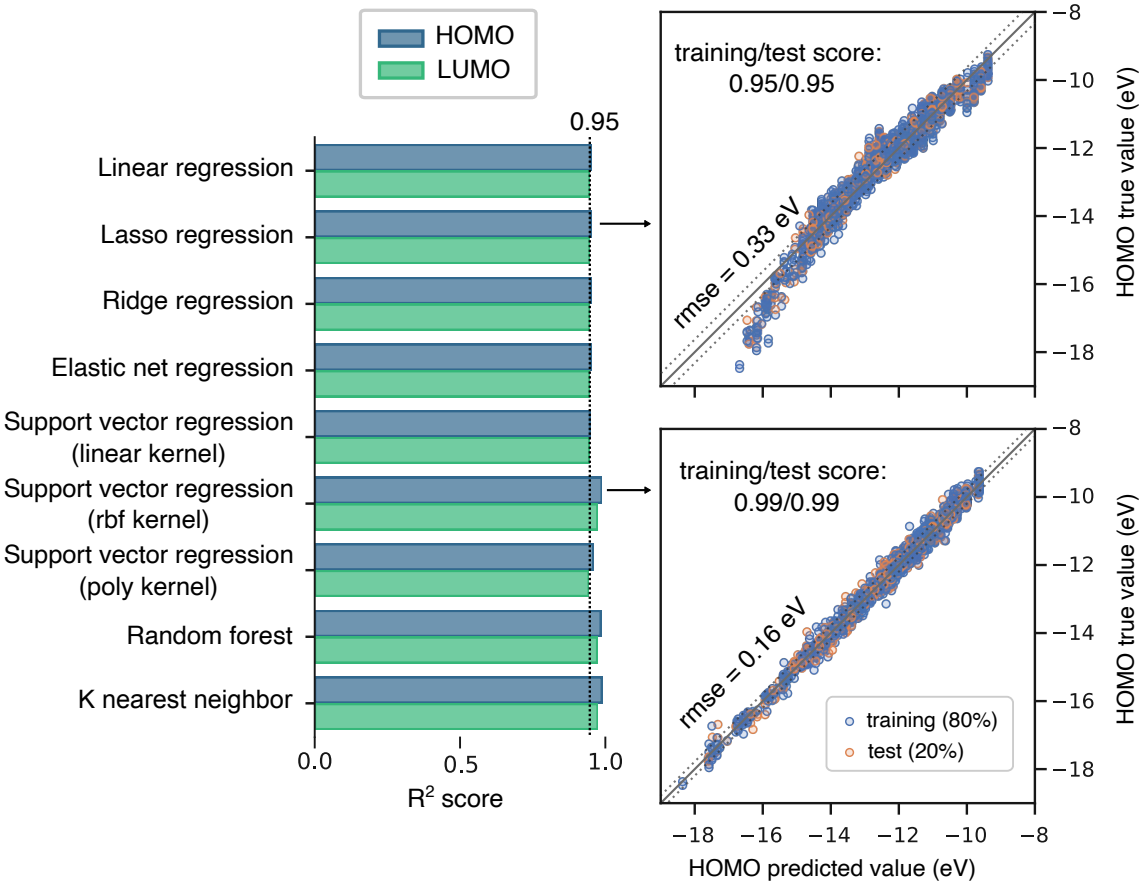


Figure 4.11: Summary of ML model performance for HOMO/LUMO prediction.

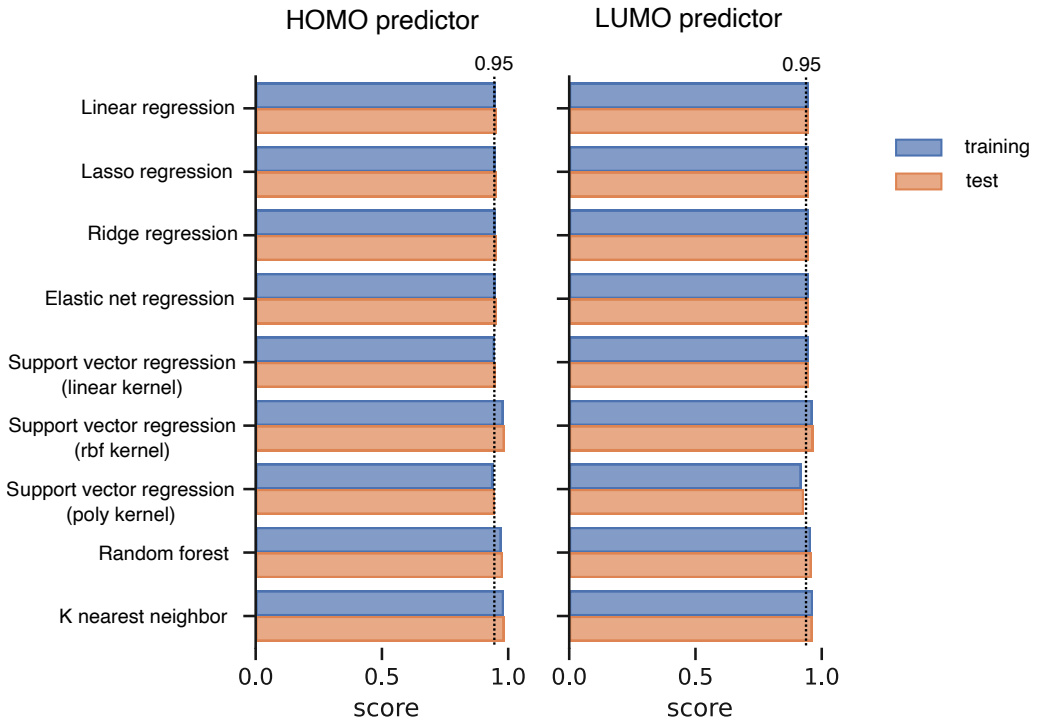


Figure 4.12: Comparison of training/test scores of various ML models for HOMO/LUMO prediction.

models achieve slightly higher R^2 scores (around 0.99/0.97) than linear models (around 0.95/0.95). Notably, no overfitting was detected, as evidenced by comparable performance on both training and test sets (Figure 4.12).

Despite their slightly lower performance for low-energy frontier levels, linear models capture the overall trend effectively. This difference becomes clearer in Figure 4.13 (HOMO) and Figure 4.14 (LUMO), which plot predicted vs. true energy values alongside model scores and root mean squared errors (RMSE). While nonlinear models better capture the extreme regions of energy, linear models show no significant deviations from the overall relationship.

Since our primary objective was to classify energy level alignment types (e.g., Type Ia, Type IIa, etc.) rather than predict absolute HOMO/LUMO values, a slight accuracy gap at the extremes is not critical. By optimizing the decision boundary, linear models can serve as a highly interpretable alternative without a significant loss in performance. Consequently, linear models are chosen for subsequent analysis and feature interpretation.

4.4 Interpretation of structure-property relationships

In this section, we explore various approaches to interpreting the machine learning model. We demonstrate how different model interpretation methods can yield slightly different feature importance rankings while still preserving the same underlying physical meaning.

4.4.1 Feature coefficient

In many traditional machine learning approaches—especially linear models—the model parameters (often referred to as coefficients or weights) provide a direct way to interpret how each input feature influences the predicted outcome. In our case, these coefficients offer insight into how various organic descriptors (encoded in the 12-digit fingerprint) affect the HOMO/LUMO energy levels of the organic spacers.

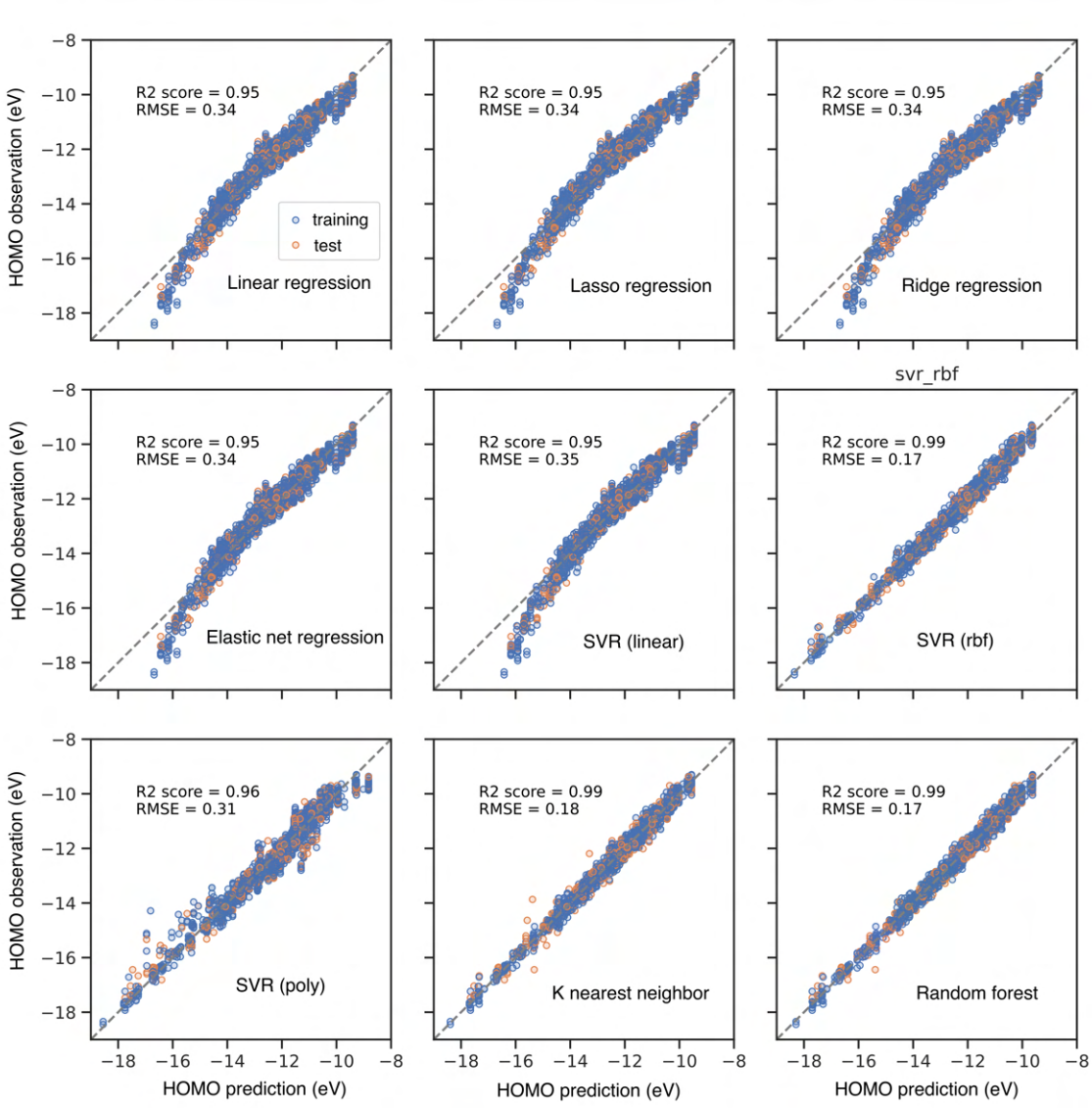


Figure 4.13: Predicted vs. true values for HOMO level across various ML models.

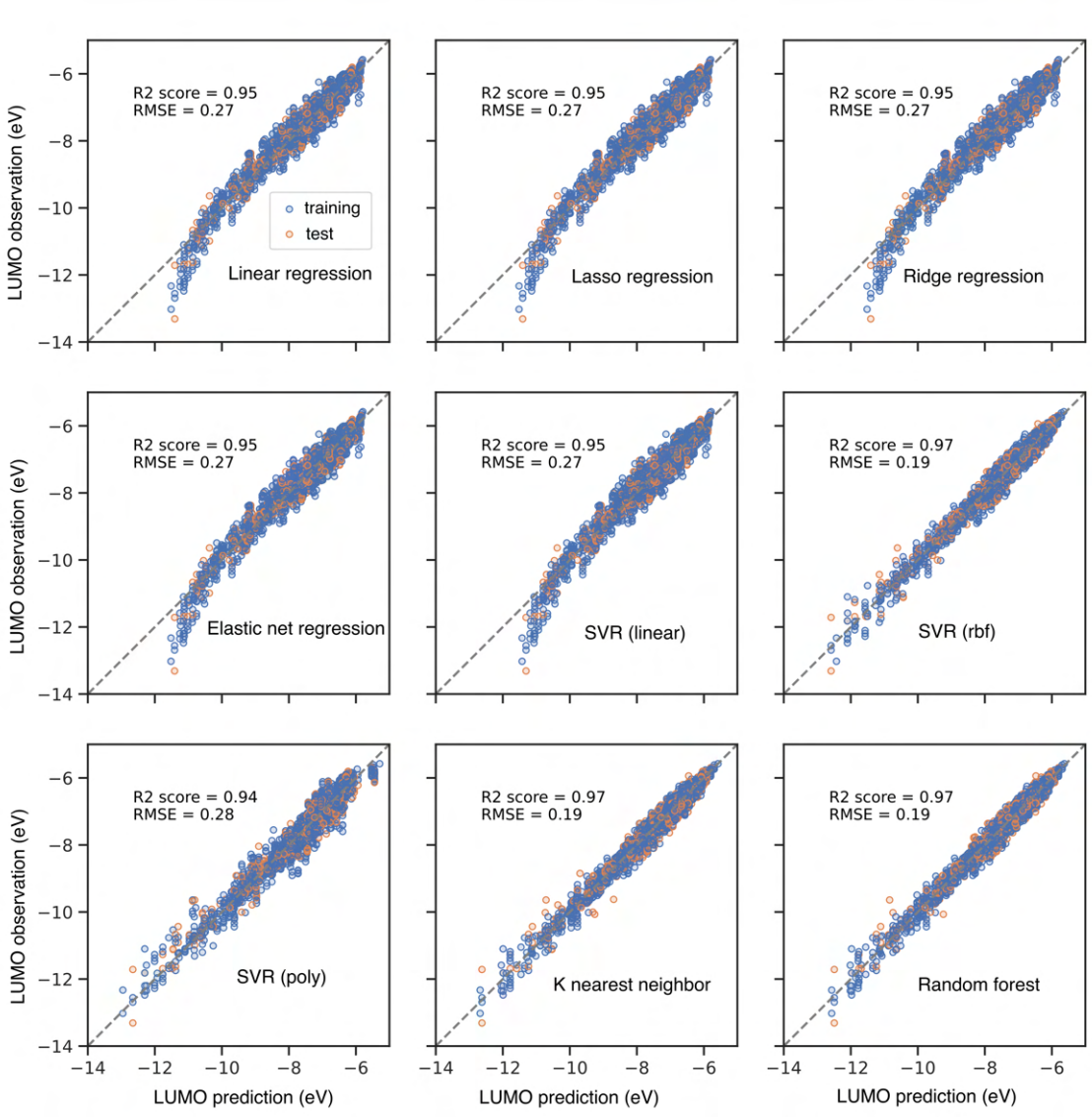


Figure 4.14: Predicted vs. true values for LUMO level across various ML models.

Since our data undergo feature normalization (standardization) before training, two types of coefficients are relevant: normalized feature coefficient (directly reported by the trained model), and unnormalized feature coefficient (raw coefficient rescaled to the original units).

Normalized feature coefficient

During model training, each descriptor is standardized by subtracting its mean and dividing by its standard deviation. As a result, the coefficients reported by the model reflect the effect of a one-standard-deviation change in a descriptor on the predicted HOMO or LUMO. Larger absolute coefficients indicate greater importance, while the sign (positive/negative) denotes whether the feature increases or decreases the predicted value.

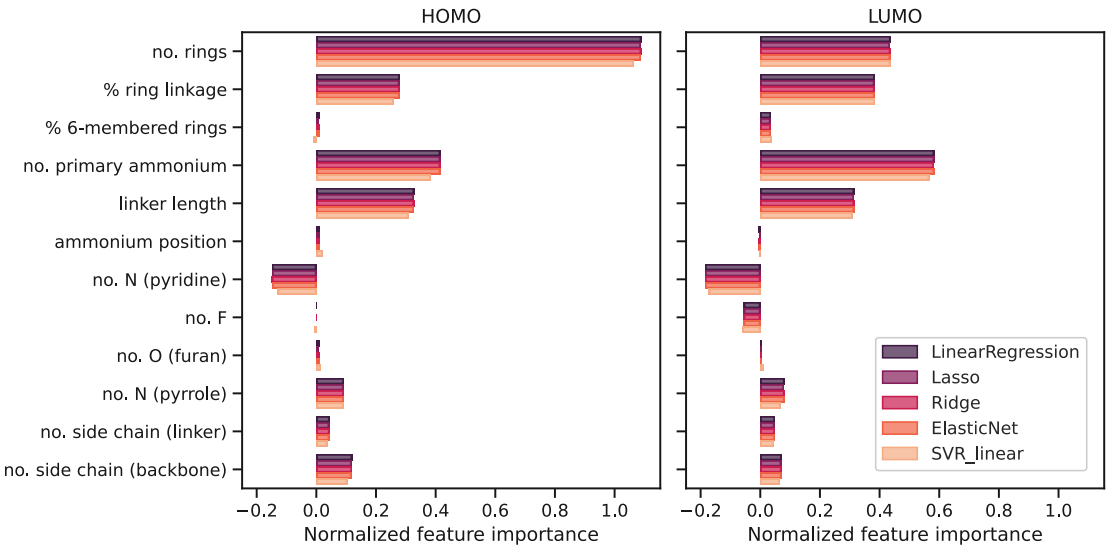


Figure 4.15: Normalized feature coefficients of linear ML models used in this work.

Figure 4.15 compares the normalized feature coefficients from various linear models—including Linear Regression, LASSO, Ridge, Elastic Net, and Linear SVR. All models yield similar coefficients, suggesting that regularization (L1, L2, or a combination) does not drastically alter the identification of the most influential features. Moreover, the consistency in coefficients across models indicates minimal overfitting.

In general, features related to the conjugated backbone (e.g., no. rings) and tethering ammonium groups (e.g., no. primary ammonium) strongly influence both HOMO and

LUMO predictions. Notably, most features affect HOMO and LUMO in a comparable manner, implying that variations in the HOMO–LUMO gap primarily arise from a few key descriptors. For HOMO, all features except "no. pyridine-type nitrogens" contribute positively, with number of rings exerting the largest effect. For LUMO, pyridine-type nitrogen and fluorine substitution exert negative contributions, whereas number of rings plays a slightly smaller role compared to HOMO.

Given its L1 regularization, LASSO provides a convenient means of highlighting key features by favouring sparse solutions. Therefore, LASSO regression serves as our representative linear model for subsequent interpretation and prediction, although the insights remain applicable to all linear models.

Unnormalized feature coefficient

While normalized coefficients gauge the effect of a one-standard-deviation change, the unnormalized coefficients express how a one-unit increase in each descriptor affects the HOMO or LUMO in real (eV) units. Consequently, unnormalized coefficients are often more intuitive in a materials science context, where absolute energy shifts matter.

As an example, the LASSO-based HOMO predictor can be written as:

$$\begin{aligned} HOMO = {} & 1.33\,x_1 + 0.61\,x_2 + 0.05\,x_3 + 1.33\,x_4 + 0.52\,x_5 + 0.18\,x_6 - 0.30\,x_7 + 0.00\,x_8 \\ & + 0.06\,x_9 + 0.44\,x_{10} + 0.11\,x_{11} + 0.24\,x_{12} - 19.25 \end{aligned} \tag{4.1}$$

Here, $x_1 \ldots x_{12}$ represent the 12 molecular descriptors, and the coefficients specify how changes in each descriptor shift the predicted HOMO (in eV). Notably, x_1 and x_4—which corresponded to no. rings and no. primary ammonium groups—dominate, indicating their strong contribution to the HOMO level.

A similar expression applies to LUMO, and Figure 4.16 visualizes the absolute values of the unnormalized coefficient for both HOMO and LUMO in radar plot. Overall, nor-

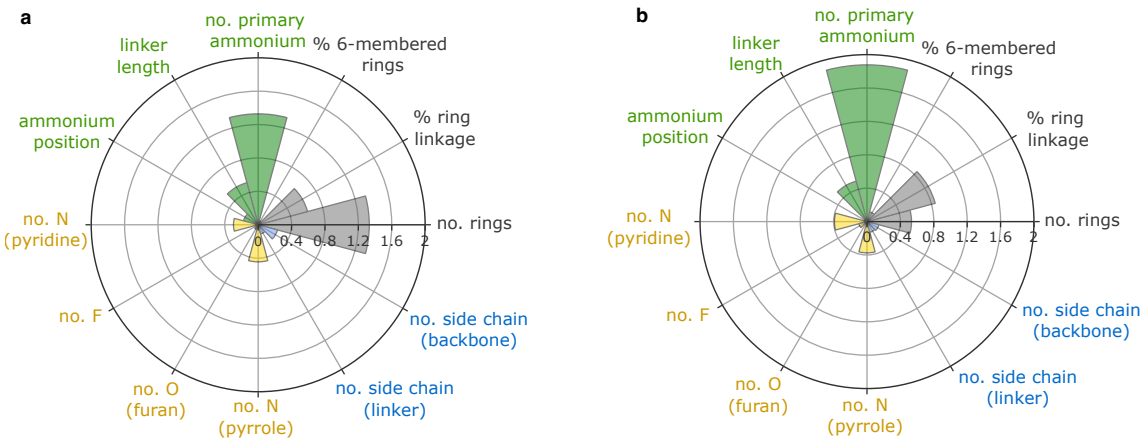


Figure 4.16: Unnormalized feature coefficients (absolute value) from Lasso regression model.

malized and unnormalized coefficients are consistent; however, descriptors with broader numerical ranges (e.g., number of rings, which varies from 1 to 4) show relatively smaller unnormalized coefficients, whereas features with narrow ranges (e.g., number of primary ammonium groups, which varies from 1 to 2) appear more prominent.

Despite these differences, the physical insights remain the same: conjugation (number of rings) and tethering ammonium groups play crucial roles in frontier orbital energies. The next subsection delves into a comprehensive interpretation of how each feature influences HOMO/LUMO levels.

4.4.2 SHAP value analysis

SHAP (SHapley Additive exPlanations) is a popular method for explaining the output of complex machine learning models. At its core, SHAP leverages concepts from cooperative game theory—specifically Shapley values—to attribute the contribution of each feature to a model's prediction. To further interpret how individual features influence predictions, we employ SHAP analysis, which assigns each descriptor a contribution (positive or negative) to the final HOMO or LUMO prediction.

Instead of using the average predicted value (the default SHAP baseline), we use the pre-

dicted value of G_0 molecule, PDMA, as reference molecule. This makes the SHAP values more physically meaningful because each feature's contribution is interpreted relative to the G0 molecule rather than to a broad average. Concretely, for each data sample, our model produces 12 SHAP values (one per feature). Summing all 12 SHAP values plus the baseline prediction (the prediction for G_0) typically yields the model's actual prediction for that sample.

Global feature importance

Figure 4.17 shows a SHAP summary plot (i.e., beeswarm plot), which provides a global view of feature importance and how each feature's impact varies across samples. The 12 features appear on the y-axis in order of their overall influence, with the most impactful features at the top. Each dot in a row represents a SHAP value for one data point, and the x-axis indicates the magnitude and direction of the feature's contribution (left for a negative shift from G_0, right for a positive shift). Because most features (e.g., no. rings) take discrete values, their SHAP values often cluster in distinct groups rather than forming a continuous distribution. A wider spread of SHAP values in a feature's row means that feature has a more variable effect across different molecules. For example, "number of rings" can shift the predicted HOMO by as much as 4 eV relative to G_0.

The summary plot highlights the influence of key features. Consistent with our analysis of feature coefficient in previous subsection, the features related to the conjugated backbone and tethering ammonium groups being the most significant. Among these, the number of aromatic rings in the conjugated backbone emerges as a critical factor, directly influencing the degree of conjugation—a well-established design rule in organic semiconductors[116] that has also found application in 2D perovskites[9], [72]. In addition to conjugation, the analysis underscores the significance of electron richness, another foundational principle in the design of organic semiconductors[116]. For tethering ammonium groups, the electron-rich alkyl groups associated with primary ammonium can raise the frontier levels by increasing the linker length or the number of primary ammonium groups. The effect of heteroatom substitution varies depending on the electronic nature of the substituent. For example, pyridine-type nitrogen, being electron-withdrawing, lowers both HOMO and

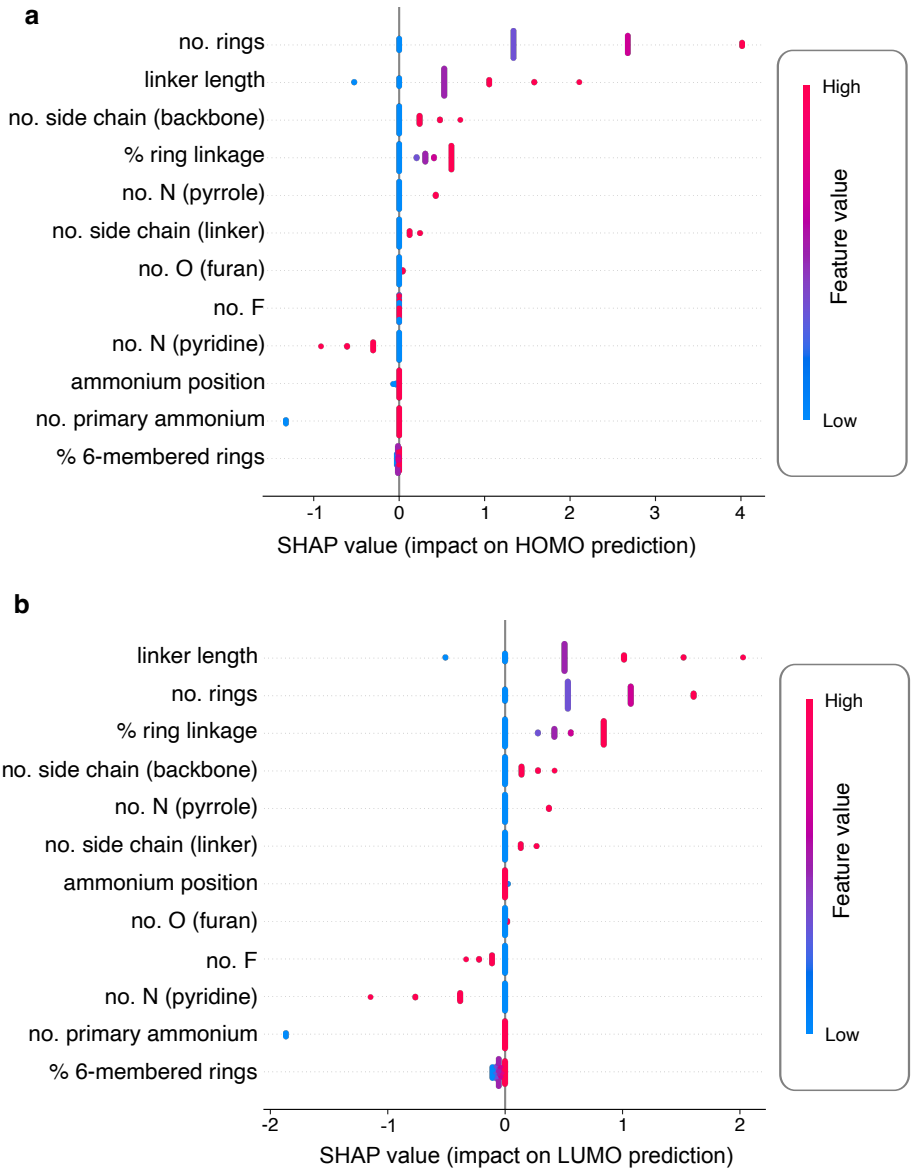


Figure 4.17: SHAP value analysis of HOMO and LUMO predictor.

LUMO, while pyrrole-type nitrogen, being electron-donating, raises both levels. Interestingly, fluorination—widely used to enhance stability in 2D perovskite spacers due to the large dipole moment induced by its electron-withdrawing ability[114], [117]—showed a relatively minor influence on the frontier levels in this study. This limited effect may stem from fluorine substitution not directly participating in the conjugated π-system. While highly electronegative, fluorine's influence remains localized, resulting in minimal perturbation to the frontier orbitals.

Compared to feature coefficient discussed in previous section, the linker length emerges as the most influential factor in LUMO prediction. This is because all the data is calibrated by baseline value of G_0 molecules, instead of the mean value. This shows that change in linker length can produce the largest deviation from LUMO of G_0 molecule.

Individual feature importance of representative molecule

After examining the global summary, we now present visualizations of SHAP values for individual samples. This is achieved using waterfall plots, which decompose a single prediction into its component feature contributions, illustrating how each feature influences the model's output above or below the baseline (which, in this case, is the G_0 molecule). Two categories of organic spacers are analyzed below, molecules with relatively high HOMO, and molecules with relatively low LUMO.

Figure 4.18 presents SHAP value analysis for four organic spacers with relatively high HOMO levels, which are later validated to achieve Type IIa alignment. The SHAP values are calibrated using the G_0 molecule as a baseline (horizontal axis on the left, where HOMO = -13.998 eV). Each feature's SHAP value is represented by a bar extending to the right (positive contribution, colored in pink) or to the left (negative contribution, colored in blue). The bars are arranged in order of their magnitude of impact, allowing for a visual step-by-step decomposition from the baseline to the final predicted value.

The primary driving factors for a higher HOMO level vary across molecules. For instance, in the first molecule, the most significant contributor is the number of rings, followed by

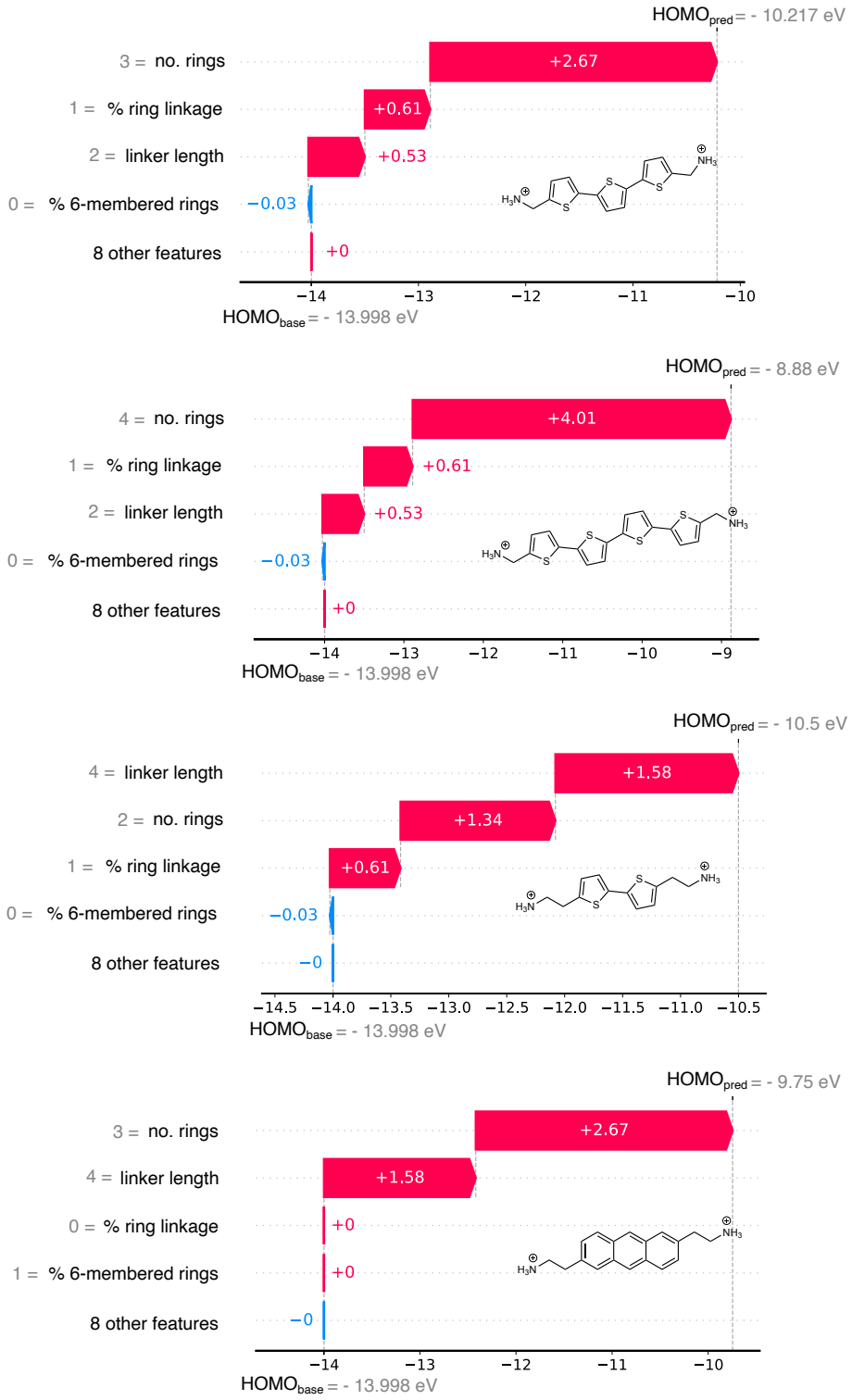


Figure 4.18: SHAP value analysis of representative organic spacers in type IIa.

the percentage of ring linkage. In the third molecule, the linker length is the dominant feature, followed by the number of rings.

Figure 4.19 illustrates the SHAP analysis of eight organic spacers predicted to achieve Type IIb alignment, characterized by relatively low LUMO levels.

The results indicate that the most significant factor driving a lower LUMO level across these molecules is the reduction in the number of primary ammonium groups.

These findings demonstrate the predictive capability of the interpretable machine learning model, which allows for the estimation of organic frontier energy levels—and by extension, the energy level alignment of DJ perovskites—for any organic spacer given its fingerprint representation. This capability accelerates the discovery process by enabling the rapid identification of promising candidates with targeted energy level alignment types.

4.5 Chapter summary

This chapter presented a combined high-throughput DFT and machine learning workflow for exploring the chemical space of organic spacers in 2D perovskites. We first generated a diverse library of hypothetical organic cations through systematic morphing operations. High-throughput DFT calculations were then performed to evaluate key physical properties across this expanded chemical space. Machine learning models were trained and benchmarked, achieving high predictive accuracy. Importantly, these models enabled rapid property prediction for large molecular sets and revealed meaningful structure–property relationships. The outcomes from this chapter establish a robust foundation for the inverse design and screening of novel spacer candidates in the following chapters.

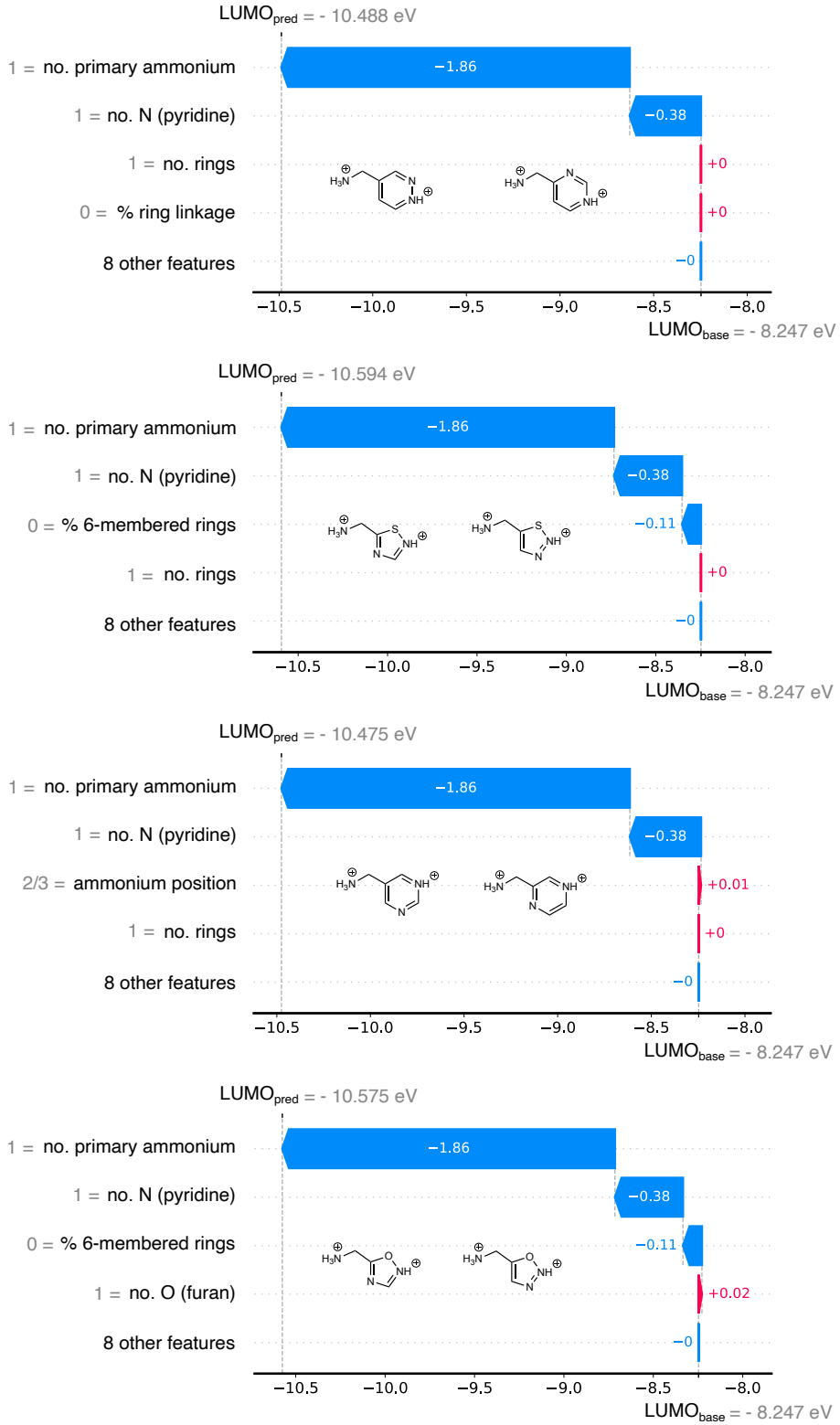


Figure 4.19: SHAP value analysis of representative organic spacers in type IIb.

Chapter 5

Synthesis Feasibility Screening and Final Candidate Validation

One of the fundamental challenges in AI-assisted materials discovery is ensuring that computationally predicted materials are experimentally realizable. Section 5.1 introduces a two-step screening approach designed to evaluate the practical synthesis feasibility of 2D perovskites. Sections 5.2 present the inverse design of final candidates, integrating both targeted energy level alignment and synthesis feasibility constraints to identify experimentally viable DJ-phase perovskites.

5.1 Synthesis feasibility screening

5.1.1 Rationale and challenges

Synthesis feasibility is a key bottleneck in AI-assisted materials discovery, acting as the bridge between theoretical predictions and experimental realization[118], [119]. While previous studies have investigated the synthetic feasibility of RP perovskites[7], [8], no systematic approaches have been applied to DJ perovskites. This lack of established

protocols arises due to two main challenges:

1. Limited exploration of diammonium spacers: Unlike RP perovskites, which utilize a wide range of monoammonium cations, DJ perovskites require diammonium cations, whose synthetic pathways remain relatively underexplored.

2. Scarcity of negative data: The absence of well-documented failed synthesis attempts (e.g., cases forming 1D or 0D phases instead of 2D structures) makes it difficult to train predictive models using traditional machine learning approaches[99], [120].

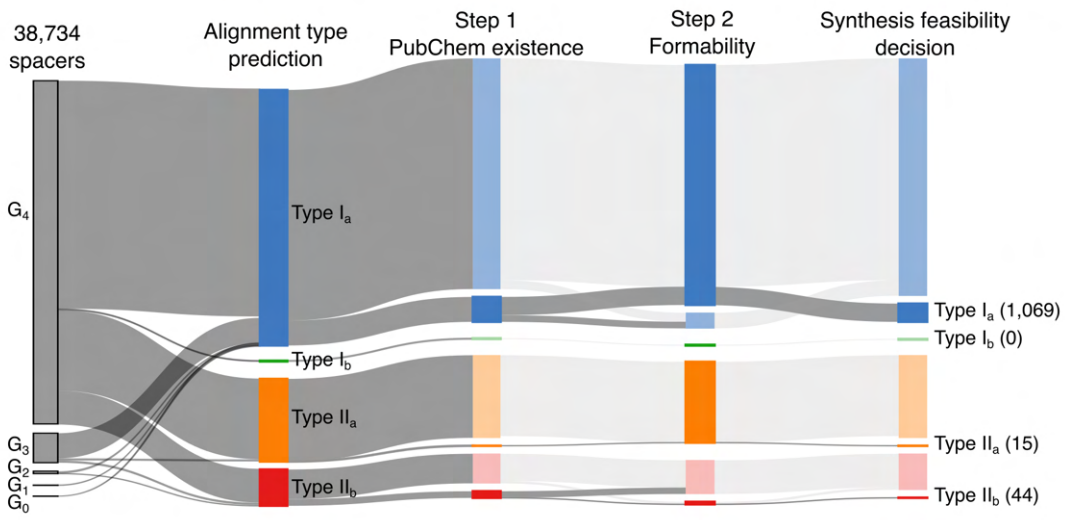


Figure 5.1: Summary of synthesis feasibility screening result.

To address this challenge, we developed a two-step computational screening approach, mimicking the experimentalist's approach to 2D perovskite synthesis:

Step 1: Synthetic accessibility of organic spacers—determines whether the generated organic spacers are practically synthesizable using established chemical routes.

Step 2: Formability of 2D DJ perovskite structures—evaluate whether a given organic spacer is likely to form a stable 2D DJ perovskite phase rather than collapsing into 1D or 0D structures.

This workflow is illustrated in Figure 5.1, with screening results shown for generations $G_0 - G_4$. The following subsections provide detailed analyses of each screening step.

5.1.2 Step 1: Synthetic accessibility of organic spacers

Using PubChem as a proxy for practical synthesizability

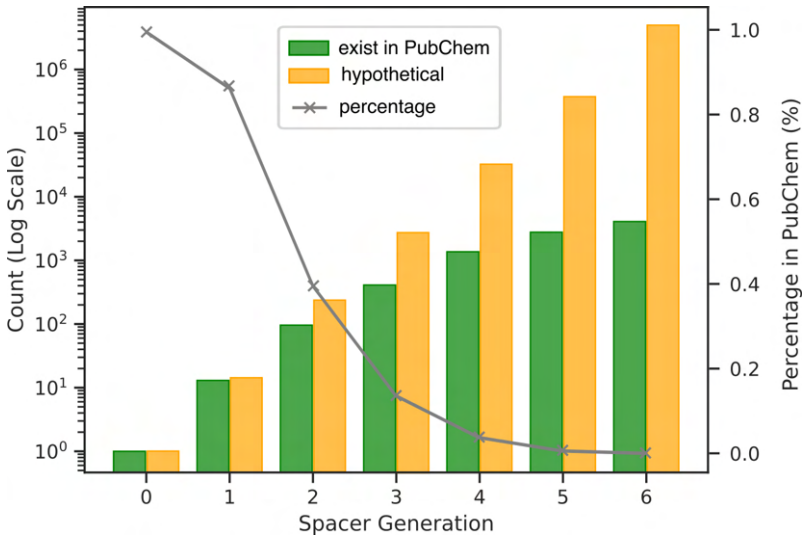


Figure 5.2: Number of generated organic spacers vs. existing spacers in G0-G4.

Rather than calculating a theoretical synthesis feasibility score, which is common in organic chemistry, we use PubChem presence as a proxy for practical synthesizability. This approach is well-established in 2D perovskite literature[7], as molecules listed in PubChem are generally commercially available or synthetically documented.

In our expanded chemical space (G_0-G_4), we find that 4.9% of generated spacers are present in PubChem. The fraction of synthesizable molecules decreases progressively from G_0 to G_4 (Figure 5.2), reflecting the increasing structural complexity of higher-generation molecules. This trend is expected since earlier generation spacers are structurally simpler and more likely to resemble known compounds, while higher-generation spacers, derived through iterative molecular morphing, tend to be chemically novel.

Synthetic accessibility across energy level alignment types

To determine whether synthesis feasibility is correlated with the energy level alignment type, we analysed the PubChem presence of spacers in G_0-G_4 with different ML-predicted energy level alignments:

- Type Ib spacers present the greatest synthetic challenges, as none of them are found in PubChem.
- Type IIa spacers are also rare, with only 0.1% appearing in PubChem.
- Type IIb spacers are the most readily accessible, with 17.5% present in PubChem.

These findings suggest that certain energy alignment types are inherently more difficult to synthesize, posing additional challenges in the inverse design process.

Structural factors affecting synthetic accessibility

To further understand why certain molecular structures are more synthetically accessible, we trained a classification model to predict whether a given organic spacer appears in PubChem. This model follows a logistic regression framework, where:

- Input Features: Molecular fingerprint descriptors.
- Target Property: Binary classification (exists in PubChem = 1, not in PubChem = 0).

We present the key performance metrics of the logistic regression model in Figure 5.3. The confusion matrix (Figure 5.3a) compares the model's predictions with the actual labels, showing an overall accuracy of 93%, meaning that 93% of the total predictions were correct.

To further evaluate the model's discriminative ability, we examine the Receiver Operating Characteristic (ROC) curve (Figure 5.3b) and its corresponding Area Under the Curve (AUC) score. In general, a model with no discriminative power (random guessing) produces a diagonal ROC curve from (0,0) to (1,1), with an AUC of 0.5. In contrast, a perfect

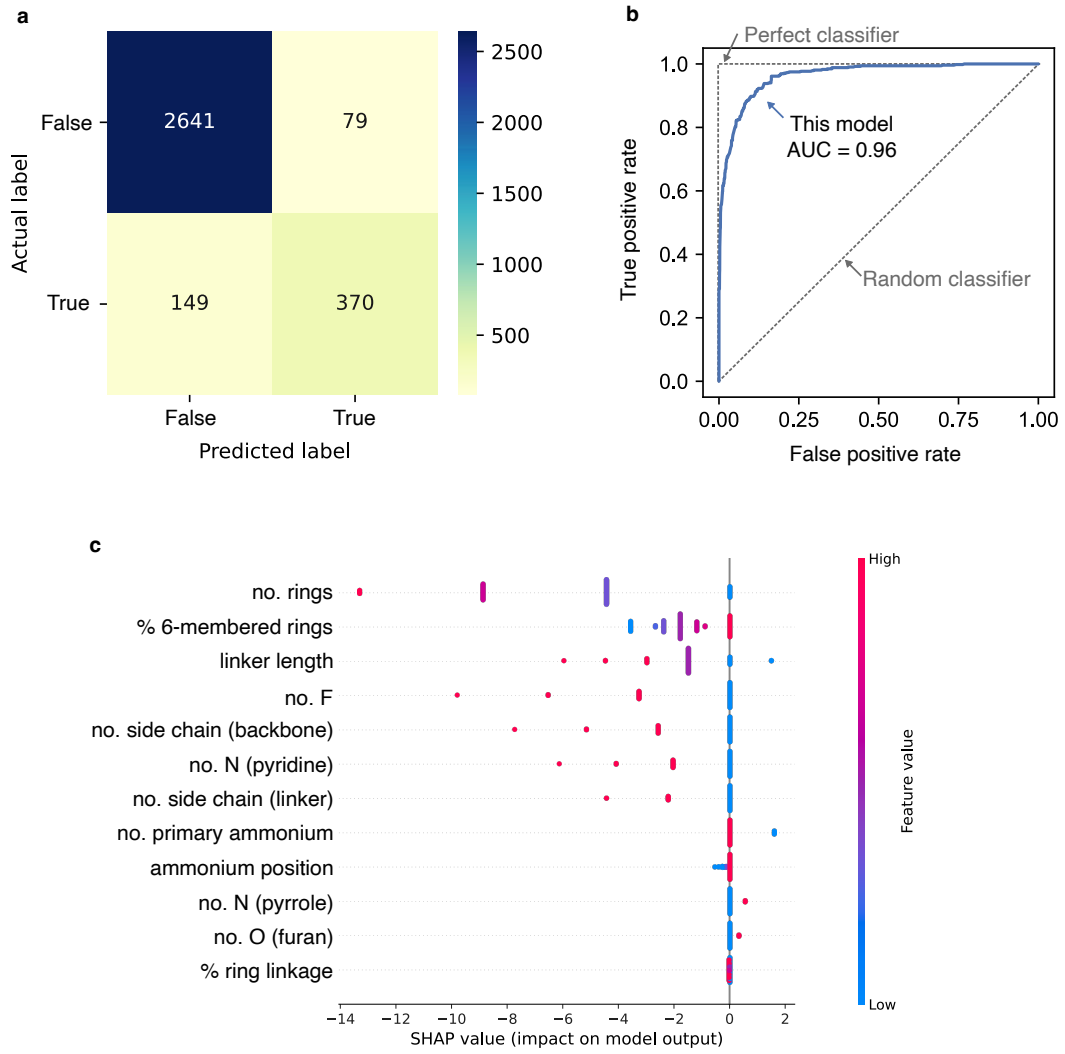


Figure 5.3: Logistic regression analysis of the relationship between fingerprints and PubChem existence.

classifier would have a curve that rises sharply to (0,1) and extends to (1,1), corresponding to an AUC of 1.0. Our model achieves an AUC of 0.96, with the ROC curve closely approaching the top-left corner, indicating strong predictive performance and a high degree of reliability in capturing the relationship between molecular fingerprints and synthesis feasibility.

Using SHAP value analysis, we identified key molecular features that influence synthetic accessibility (Figure 5.3c). The most significant descriptors include:

- Number of aromatic rings: Molecules with more rings are generally less likely to be found in PubChem;
- Side-chain modifications: Certain branched or bulky substituents reduce synthetic accessibility;
- Heteroatom substitutions, especially fluorination and pyridine type nitrogen negatively impact synthesizability.

These findings suggest that molecular complexity—particularly higher ring counts, side chains and heteroatom substitution—tends to reduce synthetic feasibility. This aligns with general organic synthesis trends, where molecules with multiple fused rings and electronegative substitutions are often more difficult to synthesize.

By integrating PubChem data into our feasibility screening, we ensure that our selected candidates remain practically synthesizable. Although most of the identified $G_0 - G_4$ spacers have not yet been explored for 2D perovskites, their presence in PubChem suggests that their chemical synthesis pathways are well established, making them promising candidates for experimental validation.

5.1.3 Step 2: 2D structure formability analysis

Following the synthetic accessibility screening, the next step evaluates the formability of 2D DJ perovskite structures. A key determinant of perovskite formability is the hydrogen-

bonding interaction between the organic spacer and the inorganic framework. It has been established in the field of 2D perovskite that hydrogen bonding plays a crucial role in stabilizing 2D perovskite structures.

Hydrogen bonding as a formability criterion

To establish hydrogen bonding potential, we examine the interaction between donor and acceptor atoms. For a hydrogen bond to form, two conditions must be met:

1. A hydrogen donor atom – A hydrogen atom covalently bonded to an electronegative element (e.g., nitrogen in ammonium groups).
2. A hydrogen acceptor atom – A highly electronegative element, capable of accepting a hydrogen bond (e.g., halide anions (I^-, Br^-, Cl^-) in the perovskite framework).

In 2D perovskites, the halide atoms in the inorganic layers serve as hydrogen acceptors, while hydrogen-donor groups originate from the organic spacers, typically from nitrogen atom. However, not all nitrogen atoms in organic spacers are capable of forming hydrogen bonds.

Figure 5.4 categorizes the four nitrogen types present in the organic spacers examined in this study. Among these, only three can act as hydrogen donors:

(1) Primary ammonium ($-NH_3^+$)

(2) Protonated pyridine-type nitrogen ($-NH^+-$)

(3) Pyrrole-type nitrogen ($-NH-$)

Conversely, unprotonated pyridine-type nitrogen ($-N-$) cannot serve as a hydrogen donor, as it lacks a covalently bonded hydrogen.

A few representative organic spacers containing these nitrogen types are illustrated in Figure 5.4. For instance, in the last example, the molecule contains two pyridine-type nitrogen atoms. In its neutral form, both nitrogen atoms are equivalent, with no hydrogen

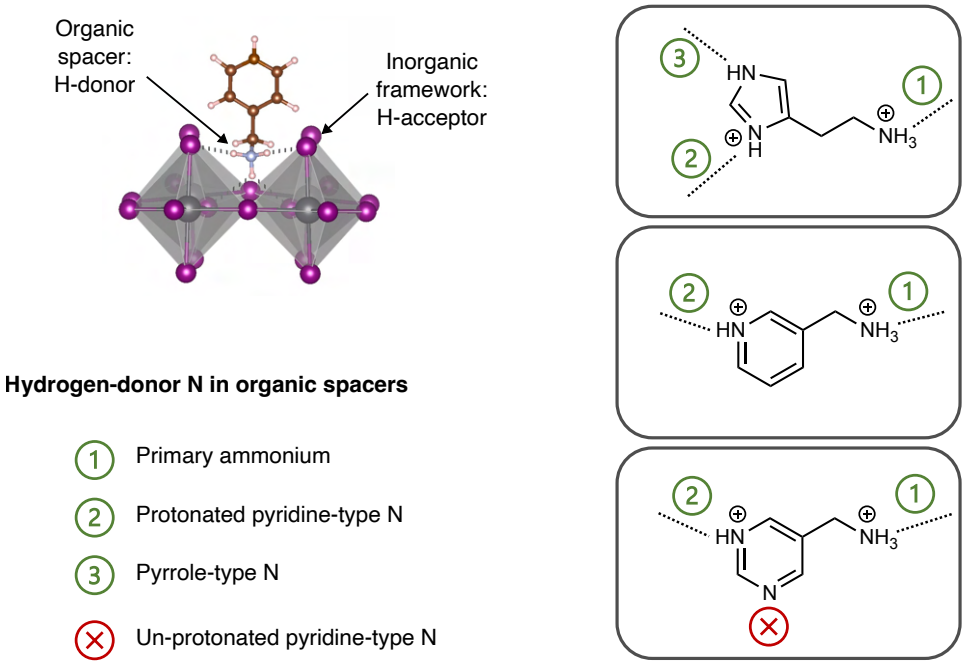


Figure 5.4: Hydrogen-donor nitrogen for hydrogen bond formation in 2D perovskite.

bonding capability due to the absence of a covalently bonded hydrogen. However, in its charged form, if one of the nitrogen atoms becomes protonated, the additional hydrogen enables hydrogen bond formation with the inorganic framework, thereby influencing 2D structure formability.

To ensure accurate descriptor selection, our formability analysis exclusively considers hydrogen-donor nitrogen atoms while excluding non-donor types (e.g., unprotonated pyridine-type nitrogen).

Formability descriptors

The ability of a hydrogen-donor nitrogen to form a hydrogen bond depends not only on its presence but also on whether it can reach the inorganic framework's halide atoms. This interaction is influenced by the topological and steric properties of the organic spacer. To quantify these effects, we adopt four key formability descriptors, previously established in RP perovskite formability studies[7], [100].

- steric hindrance index (STEI) – measures spatial constraints around the hydrogen donor.
- eccentricity – evaluates molecular shape (ratio between height and width)
- nitrogen-nitrogen pair distance (Dis_{NN})—Assess the spatial separation of tethering ammonium groups.
- the number of rotatable bonds in the spacer's tail—Reflects molecular flexibility, influencing hydrogen bond formation.

Each descriptor is computed using distance matrix calculations and is anchored to one or more nitrogen atoms (Figure 5.5). Among them, Dis_{NN} evaluates two nitrogen atoms, whereas the other three descriptors focus on a single nitrogen centre.

Formability decision framework

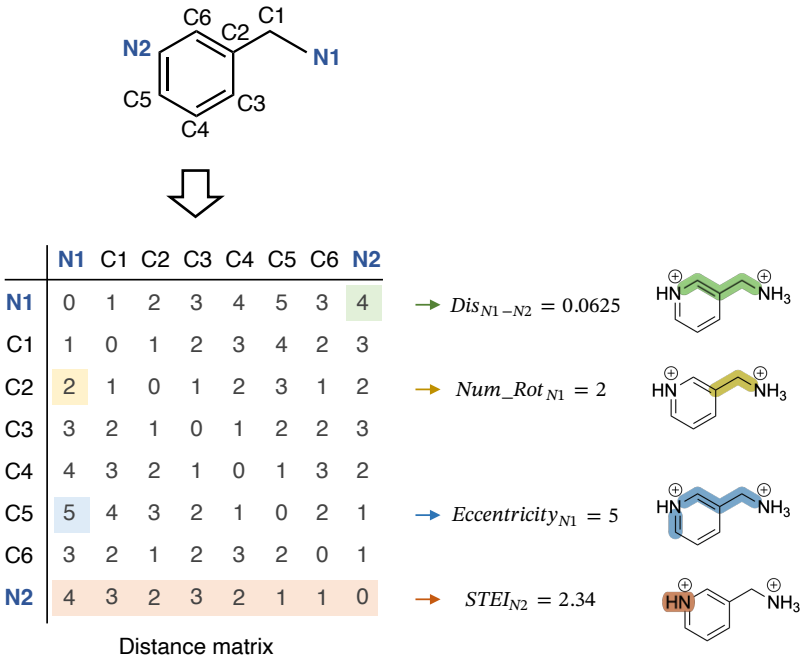


Figure 5.5: Calculation of formability descriptors for organic spacers.

To access the formability of our hypothetical organic spacers, we employed a boundary-based approach, defining thresholds for each descriptor based on their physical interpretation and experimental reported positive data.

Since prior studies suggest a linear relationship between formability and these descriptors, threshold values are determined based on the minimum or maximum range observed in experimentally validated spacers. For example, in the case of STEI, previous studies and organic chemistry principles suggest that higher steric hindrance impedes 2D perovskite formation. Thus, we set an upper limit for STEI to define formability boundaries.

For an organic spacer to be classified as formable, it must satisfy the boundary conditions for all four formability descriptors. Compared to machine learning classification methods commonly used in similar studies, our method addresses the unique challenge for DJ perovskites: the limitation of highly imbalanced dataset (dominated by positive data) and high correlations among descriptors which can lead to multicollinearity and reduced model

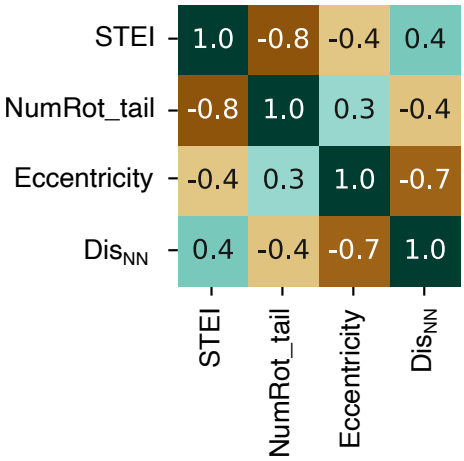


Figure 5.6: Calculation of formability descriptors for organic spacers.

accuracy. As shown by the Pearson correlation analysis of formability descriptors in Figure 5.6, the four formability descriptors exhibit strong correlations with each other (above 0.7 for multiple descriptors), limiting their independent utility in machine learning models. Our approach applies stricter, more interpretable criteria by evaluating descriptors individually rather than collectively, enhancing its robustness for formability prediction.

Formability screening result

Applying the formability criteria to generations $G_0 - G_4$, we find that 7.4% of organic spacers fail the screening (Figure 5.1). However, the impact varies across energy level alignment types:

- Type IIb spacers are the most affected (74% excluded).
- Type Ib and Type IIa spacers are largely unaffected, with no exclusions.

The influence of the four formability descriptors on final screening decisions is also unevenly distributed. Among them, STEI exhibits the strongest impact on formability constraints. A closer examination reveals that STEI alone accounts for 68.1% of exclusions among Type IIb spacers (Figure 5.7). This suggests that steric effects play a dominant role in restricting the formability of these organic spacers, further reinforcing the importance of spatial accessibility in hydrogen bonding interactions within 2D DJ perovskite structures.

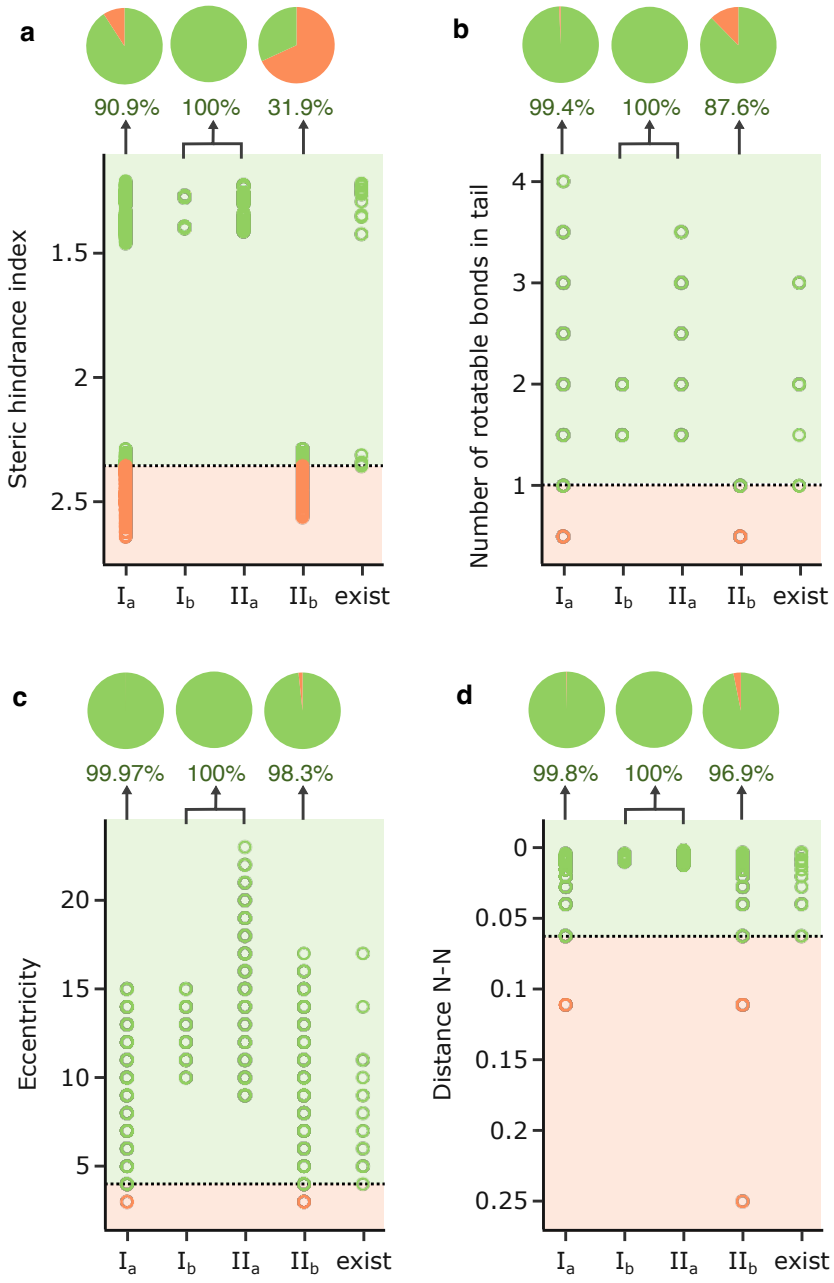


Figure 5.7: Analysis of influence of the formability descriptors on the final decision of formability.

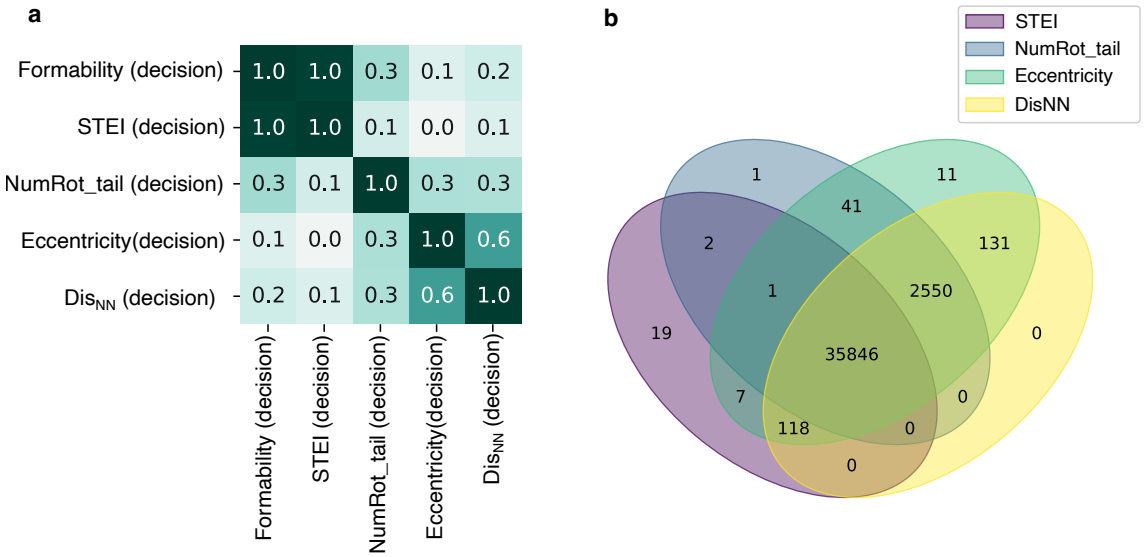


Figure 5.8: Relationship between the decision of four formability descriptors. **a** Pearson correlation coefficients between formability decisions and molecular descriptors. **b** Venn diagram illustrating the overlap in decisions among the four formability descriptors.

To better understand the relationship between descriptors and formability decisions, Figure 5.8a presents a Pearson correlation analysis between the decision of individual descriptors and the final decision. The near-perfect correlation (close to 1.0) between STEI and formability confirms steric hindrance as the most critical determinant.

Additionally, Figure 5.8b provides a Venn diagram illustrating the overlap between the decision among the four descriptors. The shared area among all four descriptors represents the organic spacers that satisfy the final formability decision criteria. Non-overlapping areas indicate that each formability descriptor serves a slightly different role in screening candidates. The largest number of candidates outside the STEI circle underscores that STEI is the most effective screening descriptor.

Relationship between fingerprint and formability screening

To further investigate the structural factors influencing 2D perovskite formability, we analysed the relationship between molecular fingerprints and formability decisions. As shown in Figure 5.9, Pearson's correlation coefficient reveals that key structural features affecting formability include linker length and the number of primary ammonium groups.

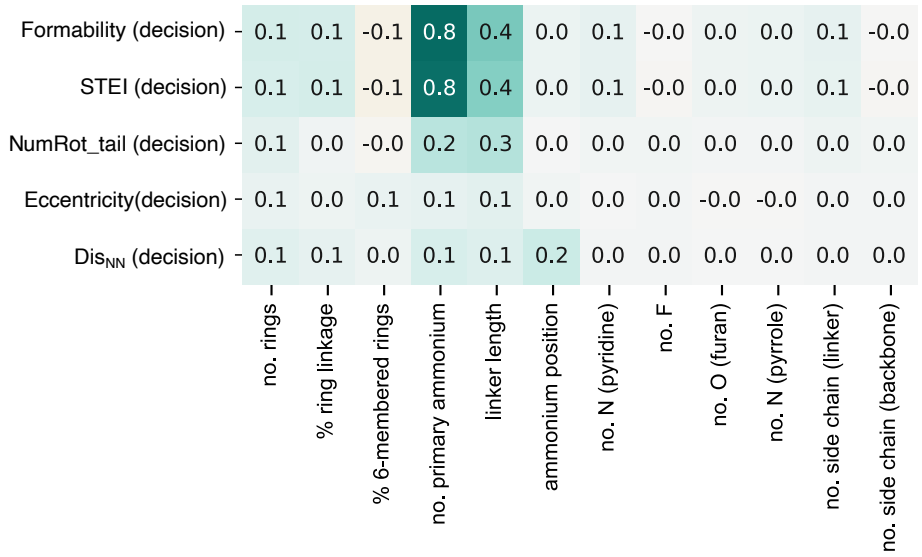


Figure 5.9: Correlation between formability descriptor decision and molecular fingerprint.

This align with established understanding in this field[6], emphasizing that the tethering ammonium group plays a crucial role in the formation of 2D perovskite structure.

The influence of the tethering ammonium group is primarily exerted through its impact on the STEI, which determines whether the ammonium donor can effectively engage in hydrogen bonding with the inorganic framework. This effect explains why Type IIb spacers are disproportionately affected by formability screening—compared to Type IIa and Type Ib spacers, Type IIb spacers exhibit shorter linker lengths and fewer primary ammonium groups, leading to increased steric hindrance and reduced hydrogen bonding potential.

Although our analysis confirms that the tethering ammonium group is the dominant factor in formability, we also observe weaker correlations between formability and other structural descriptors. To explore these additional dependencies, we analysed representative groups of organic spacers to illustrate that formability is governed by complex interactions between multiple structural variables.

As shown in Figure 5.10, formability in this set of organic spacers is determined by a combination of factors: tethering ammonium position, linker length, number of primary ammoniums, and the percentage of six-membered rings.

Among these descriptors, the rotatable bond in the tail descriptor is primarily influenced by linker length, particularly when there is a single primary ammonium, and the linker length is zero. In contrast, STEI involves a more complex interplay of factors, including the number of primary ammoniums, linker length, and ammonium position. While steric hindrance is often associated with the number of primary ammoniums, our findings reveal that this factor alone is insufficient to disqualify an organic spacer. Instead, the spatial environment of secondary ammoniums, particularly the position of the tethering ammonium on the ring, is a key determinant of the STEI boundary.

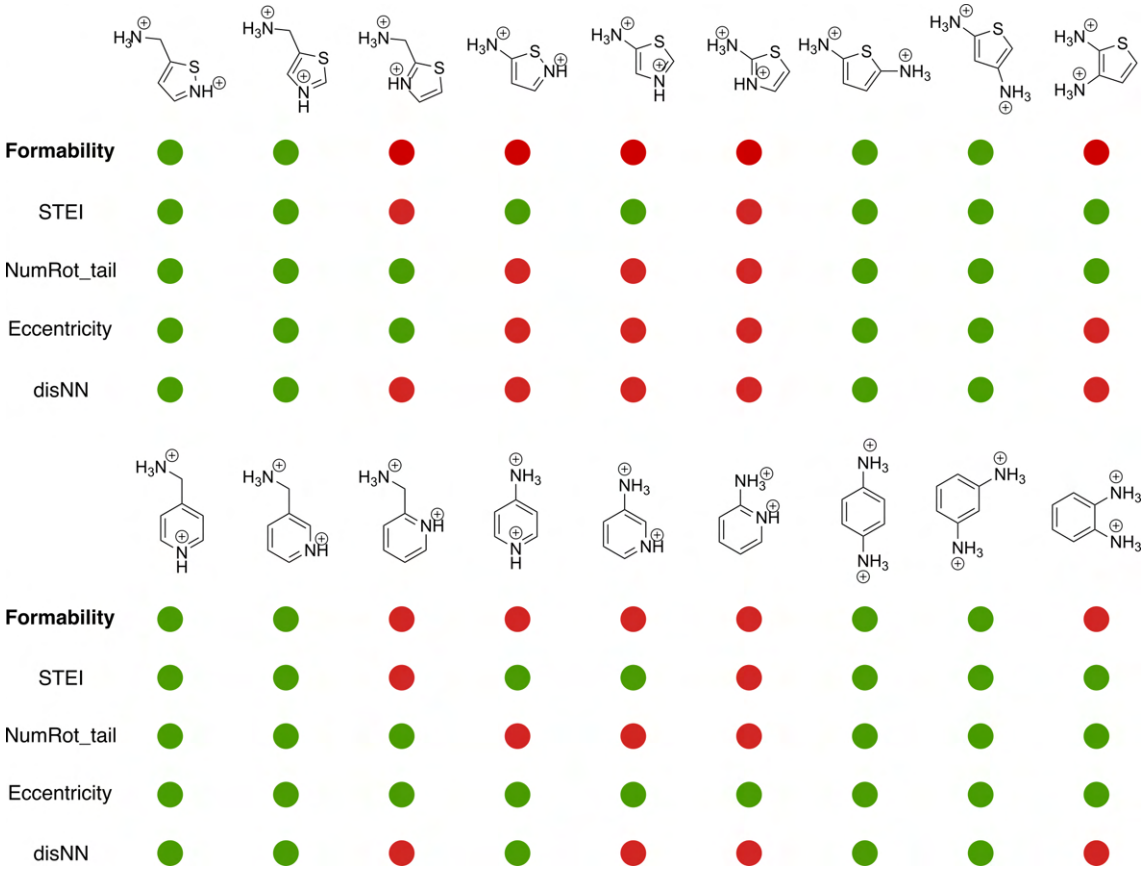


Figure 5.10: Examination of similar organic spacers near the formability decision boundary.

5.1.4 Synthesis feasibility screening summary

Our synthesis feasibility screening reveals distinct synthesis feasibility challenges for DJ perovskites with type Ib, IIa, and IIb alignments. For type Ib and IIa organic spacers, the primary bottleneck lies in their synthetic accessibility, whereas for type IIb spacers, the main challenge is achieving the formability of the 2D structure.

While this two-step screening process effectively narrows the range of DJ perovskite candidates, limitations remain when compared to real experimental synthesis. First, while PubChem provides a practical and high-throughput filter, certain organic spacers not listed in its database may still be accessible through complex synthetic routes, as demonstrated in organic photovoltaic research[116]. Second, the formability descriptors rely on boundaries derived from reported positive data, leaving unexplored regions in the parameter space. Refining these boundaries as new DJ-phase spacers are reported could further expand the pool of viable candidates. Finally, the screening process does not fully capture certain experimental considerations critical to practical synthesis. For conjugated organic spacers, solubility stands out as a significant factor. Specifically, increasing the number of rings to three or more can lead to solubility issues[71], which are particularly relevant for type Ib and IIa spacers. This challenge may be mitigated by structural modifications, such as incorporating short alkyl side chains to disrupt the planarity of the conjugated backbone, a strategy commonly employed in organic photovoltaics[116], [121]. Additionally, key experimental parameters—such as solvent choice, precursor ratios, temperature, and pH—are not accounted for in our method. These factors can influence whether the DJ phase forms or if alternative phases (e.g., 1D, 0D, or RP phase) are favoured with the same organic spacer[100].

5.2 Inverse design of DJ perovskites with targeted energy level alignment

In this section, we apply an inverse design strategy for selecting organic spacers that achieve three specific energy-level alignment types relevant to DJ perovskite applications: Type IIa, Type IIb, and Type Ib. This approach leverages the invertible molecular fingerprint representation, allowing us to map from desired energy level alignments back to potential organic spacer structures. First, we identify unique fingerprint characteristics based on ML-predicted energy level alignment and synthesis feasibility analysis. These fingerprints are then inverted to reconstruct the corresponding organic spacer structures within the expanded chemical space. Since the fingerprint criteria follow well-defined boundary conditions, this method enables a systematic and exhaustive exploration of the chemical space for targeted alignment types.

5.2.1 Rationale for inverse design framework

The primary objectives of our inverse design approach are:

(1) Exploring uncharted regions of the DJ perovskite energy landscape, focusing specifically on Type IIa, Type IIb, and Type Ib alignments.

(2) Ensuring synthetic feasibility, so that the identified organic spacers are relevant for experimental validation.

Limitations of the forward design approach

Up to this point, our computational design strategy has relied on a forward design approach, employing high-throughput screening to evaluate $\sim 10^4$ hypothetical organic spacers. This has successfully led to experimentally viable Type IIa and Type IIb candidates. However, two critical challenges remain:

(1) Absence of Type Ib candidates: Despite screening a large chemical space, no organic

spacers satisfying Type Ib energy alignment have been identified. This is primarily due to synthetic challenges, as many potential Type Ib spacers are systematically filtered out during feasibility screening.

(2) Limited Exploration of the Chemical Space. The screening process has been constrained to lower-generation organic spacers ($G_0 - G_4$), totaling $\sim 10^4$ candidates. However, the number of possible organic spacers increases exponentially in later generations. This means that a significant portion of the design space remains unexplored, potentially missing optimal organic spacers that exist beyond the current screening limits.

Transitioning to an inverse design approach

To overcome the limitations of forward design, we adopt an inverse design approach, leveraging our invertible molecular fingerprint representation—a key feature of our forward design workflow. This enables us to reverse-engineer organic spacer structures directly from target energy level alignments while ensuring synthetic feasibility. Unlike forward design, which requires exhaustive screening, this inverse approach allows us to map directly from desired properties (energy level alignment and synthesis feasibility) to molecular fingerprints, and subsequently to organic spacer structures. The inverse design process consists of the following steps:

(1) Identify molecular fingerprints associated with targeted energy level alignments and synthetic feasibility constraints.

(2) Reconstruct organic spacers by inverting the selected fingerprints. Obtain synthesizable candidate through synthesis feasibility screening.

(3) Validate the energy level alignment of the designed DJ perovskites through DFT calculations.

By implementing this target-driven approach, we bypass the computational bottlenecks of exhaustive forward screening, enabling a systematic and efficient exploration of the chemical space to identify optimal organic spacers.

5.2.2 Constructing fingerprint for targeted alignment types

Previous chapters have established a strong correlation between molecular fingerprints and energy level alignment types. By analysing statistical trends in the fingerprint data from $G_0 - G_4$ organic spacers, we identify distinct molecular features that are closely associated with specific alignment types.

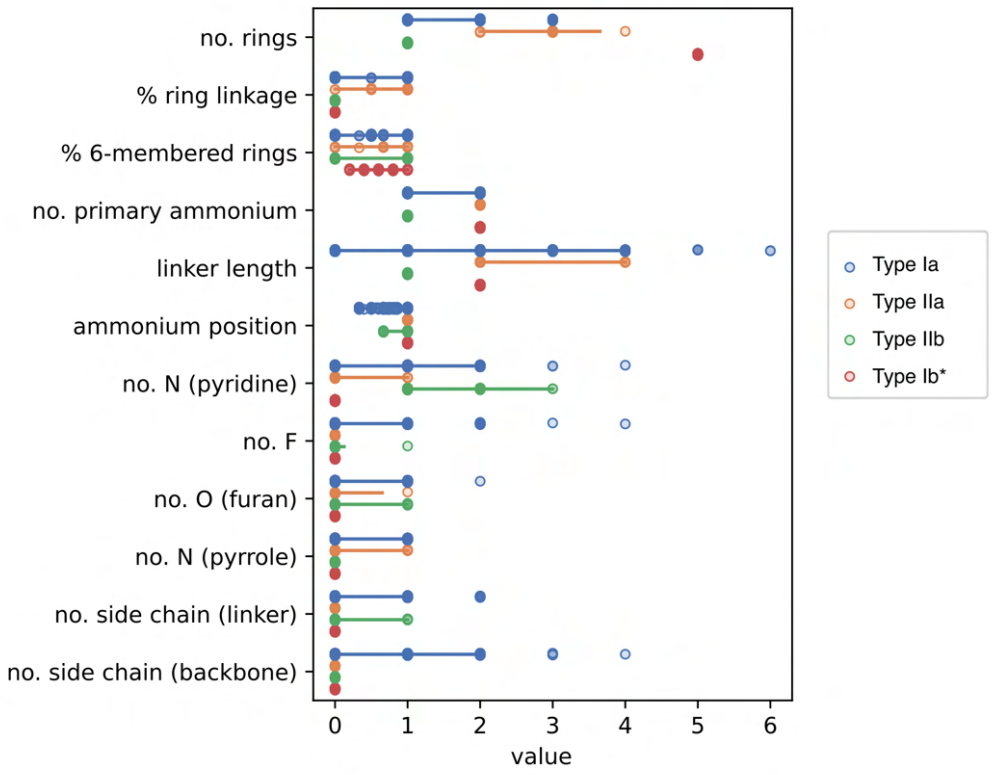


Figure 5.11: Distribution of organic fingerprints associated with different energy level alignment types.

Figure 5.11 presents the distribution of 12 molecular descriptors in organic spacers that pass synthetic feasibility screening, categorized by their respective alignment types. The statistics of the 12 organic descriptors for qualified organic spacers from $G_0 - G_4$ are depicted. The dots indicate the range, while the bars represent the 95% confidence interval of each descriptor, color-coded according to their alignment type (Ia, IIa, IIb). For Type Ib spacers, no candidates were identified in $G_0 - G_4$, primarily due to synthetic accessibility

constraints. To gain insight into their characteristic features, we considered organic spacers that theoretically satisfy Type Ib energy alignment without enforcing synthetic feasibility constraints.

We identified distinct molecular fingerprint characteristics associated with different energy level alignments. Type Ia spacers exhibit a broader range of descriptor values, which corresponds to their higher prevalence in the organic spacer dataset. In contrast, Type IIa, IIb, and Ib spacers display more defined clustering patterns, suggesting that specific molecular features play a critical role in determining their alignment behaviour.

Among these, the number of rings, primary ammonium groups, and linker length emerge as the key distinguishing factors. Notably, the number of rings differs significantly across alignment types: Type Ib = 5; Type IIa = 2–4; Type IIb = 1.

These trends provide valuable insights into the molecular fingerprint characteristics most likely to yield qualified organic spacers for targeted energy level alignments.

Based on statistical probability analysis, we define boundary conditions for fingerprint values that are most likely to correspond to targeted energy level alignment types (Figure 5.12). The fundamental principle in setting these criteria is to capture the range indicated by the confidence interval, ensuring that the selected fingerprint space is broad enough to include promising candidates while remaining computationally manageable.

For example, Type IIa spacers encompass a large number of candidates if a single fingerprint set is used. To refine the selection and improve specificity, we introduce additional fingerprint criteria based on the conjugated backbone structure. Specifically, we classify Type IIa spacers into two distinct subgroups: oligothiophene-like organic spacers, characterized by linked 5-membered rings; and acene-like organic spacers, featuring fused 6-membered rings.

By establishing fingerprint-based selection criteria, we define a finite and exhaustible chemical search space (as demonstrated below), enabling a systematic and targeted search for viable organic spacers that satisfy both energy alignment and synthetic feasibility con-

	Type Ib	Type IIa		Type IIb
		oligothiophene-based	acene-based	
no. rings	[5, 7]	[2, 6]	[2, 4]	1
% ring linkage	0	1	0	0
% 6-membered rings	1	0	1	[0, 1]
no. primary ammonium	2	2	2	1
linker length	[0, 2]	[0, 6]	[0, 6]	[0, 3]
ammonium position	[0.5, 1]	1	1	(0, 1]
no. N (pyridine)	0	0	0	[0, 4]
no. F	0	0	0	[0, 2]
no. O (furan)	0	0	0	[0, 1]
no. N (pyrrole)	0	[0, 2]	0	[0, 1]
no. side chain (linker)	0	0	0	[0, 2]
no. side chain (backbone)	0	0	0	0
Number of molecules	3	12	2	58

Figure 5.12: Fingerprint criteria for targeted energy level alignment type.

straints.

Exhaustive search of chemical space

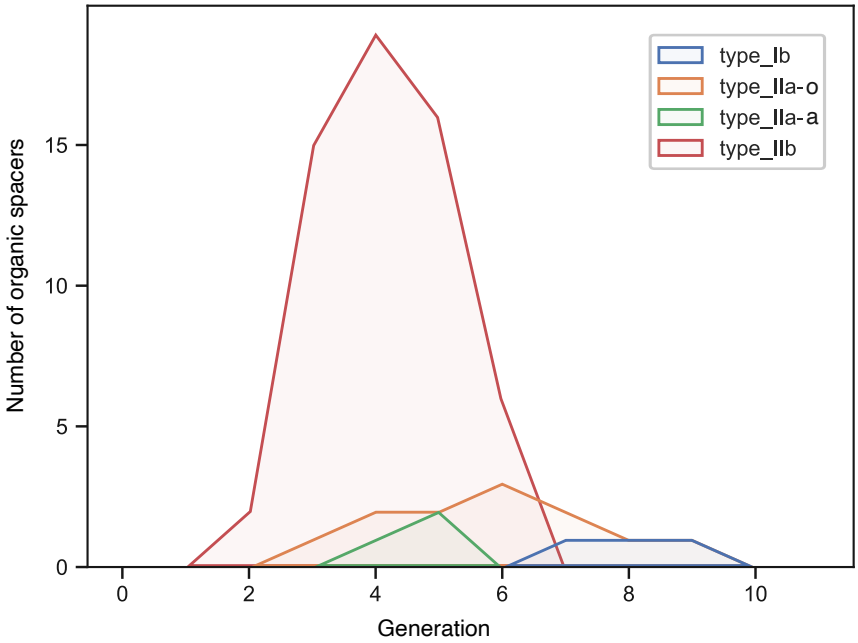


Figure 5.13: Organic spacer counts for each energy alignment type across generations $G_0 - G_{11}$

Using the fingerprint criteria, we systematically map the chemical space to identify potential organic spacers that meet the desired energy alignment and synthesis feasibility constraints. As shown in Figure 5.13, the number of viable organic spacers for each fingerprint criterion follows a single-peak distribution across generations: Starts at zero in early generations ($G_0 - G_1$).; Peaks at an intermediate generation ($G_5 - G_9$); Diminishes to zero by G_{11}, marking a natural endpoint where no additional spacers satisfy the criteria. The decline at G_{10} and G_{11} suggests that beyond these generations, no additional chemically meaningful spacers are likely to exist within the defined fingerprint constraints.

This approach enables us to overcome the limitations of the enumerable chemical space ($G_0 - G_6$, approximately 10^6 spacers, with the number expected to increase exponentially in later generations) and conduct an exhaustive search across the entire chemical space within the defined fingerprint constraints. While viable spacers may exist beyond this

subregion, our analysis suggests that it represents the most promising region for identifying candidates efficiently while maintaining an affordable computational cost.

Our search identified three type Ib organic spacers in $G_7 - G_9$, 14 type IIa candidates in $G_3 - G_9$, and 58 type IIb candidates in $G_2 - G_6$.

5.2.3 Mapping fingerprint to organic spacers structures

Type Ib

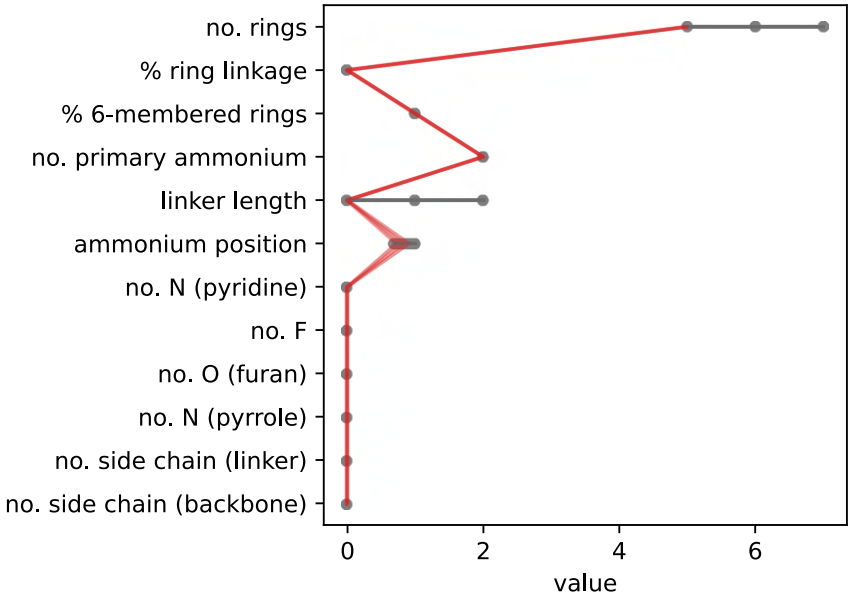


Figure 5.14: The explored fingerprint range and vs. fingerprint range of type Ib final organic spacer candidate.

For Type Ib alignment, the identified fingerprint range corresponds to 347 organic spacers. However, only three of these spacers pass the synthesis feasibility screening, primarily due to synthetic accessibility constraints.

The fingerprint range of the selected organic spacers is shown in Figure 5.14. All three share highly similar fingerprint characteristics, with the only variation being ammonium position. Their key structural features include:

- A conjugated backbone consisting of five fused 6-membered rings.

- Two ammonium groups serving as tethering groups.

- No heteroatom substitutions (e.g., nitrogen, oxygen, or fluorine).

- No side chains attached to the backbone.

As illustrated by the explored range of organic fingerprints, organic spacers with additional variations—such as an increased number of rings, longer linker lengths, or alternative linker positions—were excluded from the final selection, primarily due to their absence from the PubChem database, indicating limited synthetic accessibility.

The structures of the three identified organic spacers are presented in Figure 5.15.

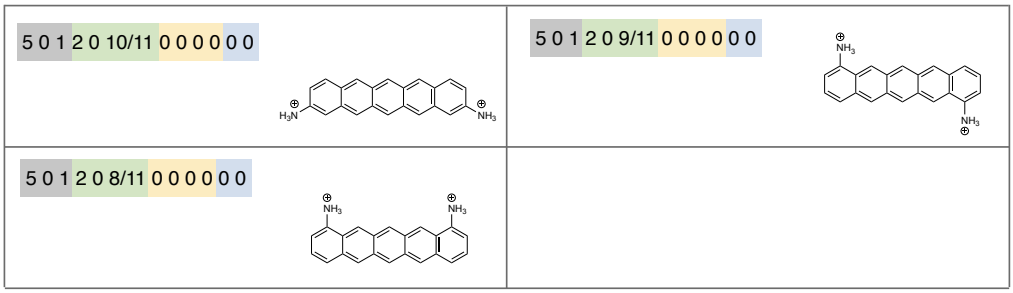


Figure 5.15: Inverse designed candidates for type Ib alignment.

Type IIa

For Type IIa spacers, the identified fingerprint set corresponds to 720 organic spacers for oligothiophene-like structures and 88 organic spacers for acene-like structures. After applying synthesis feasibility screening, only 14 candidates remained as final selections.

The fingerprint range of the selected final candidates is shown in Figure 5.16. These candidates primarily exhibit the following key structural characteristics:

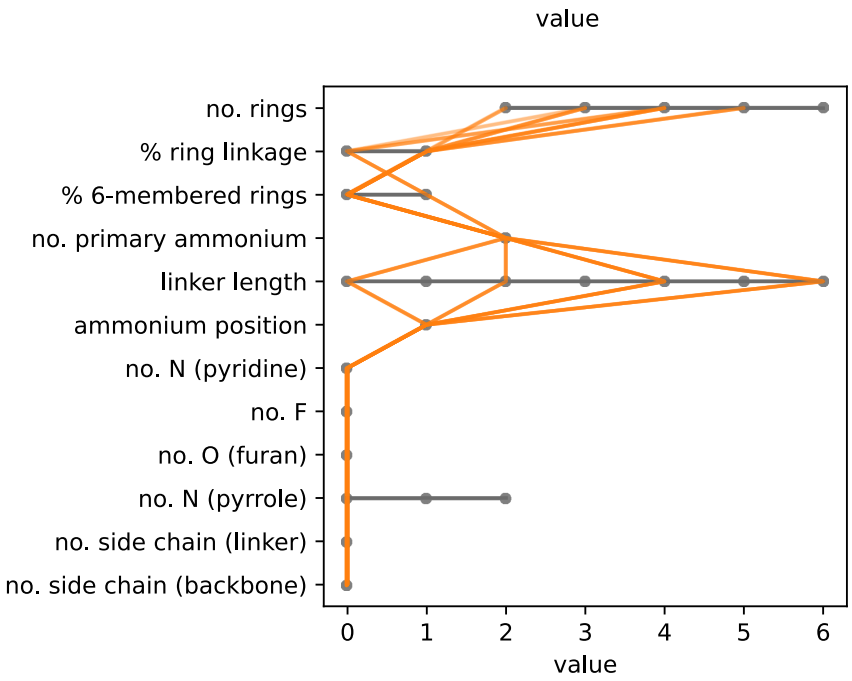


Figure 5.16: The explored fingerprint range and vs. fingerprint range of type IIa final organic spacer candidate.

- Conjugated Backbone: Multiple linked thiophene rings (oligothiophene-like) or fused benzene rings (acene-like).

- Tethering ammonium groups: Two primary ammonium groups with varying linker lengths.

- No heteroatom substitutions (e.g., nitrogen, oxygen, fluorine).

- No side chains attached to the backbone.

The organic spacer structures are presented in Figure 5.17. Notably, for each fingerprint, only one organic spacer successfully passed synthesis feasibility screening, while all other isomers were filtered out.

For instance, considering AE4T as an example, although its isomers exhibit similar electronic properties, their synthetic feasibility—specifically, synthetic accessibility—varies significantly. The excluded isomers were primarily filtered out due to: Uneven linker

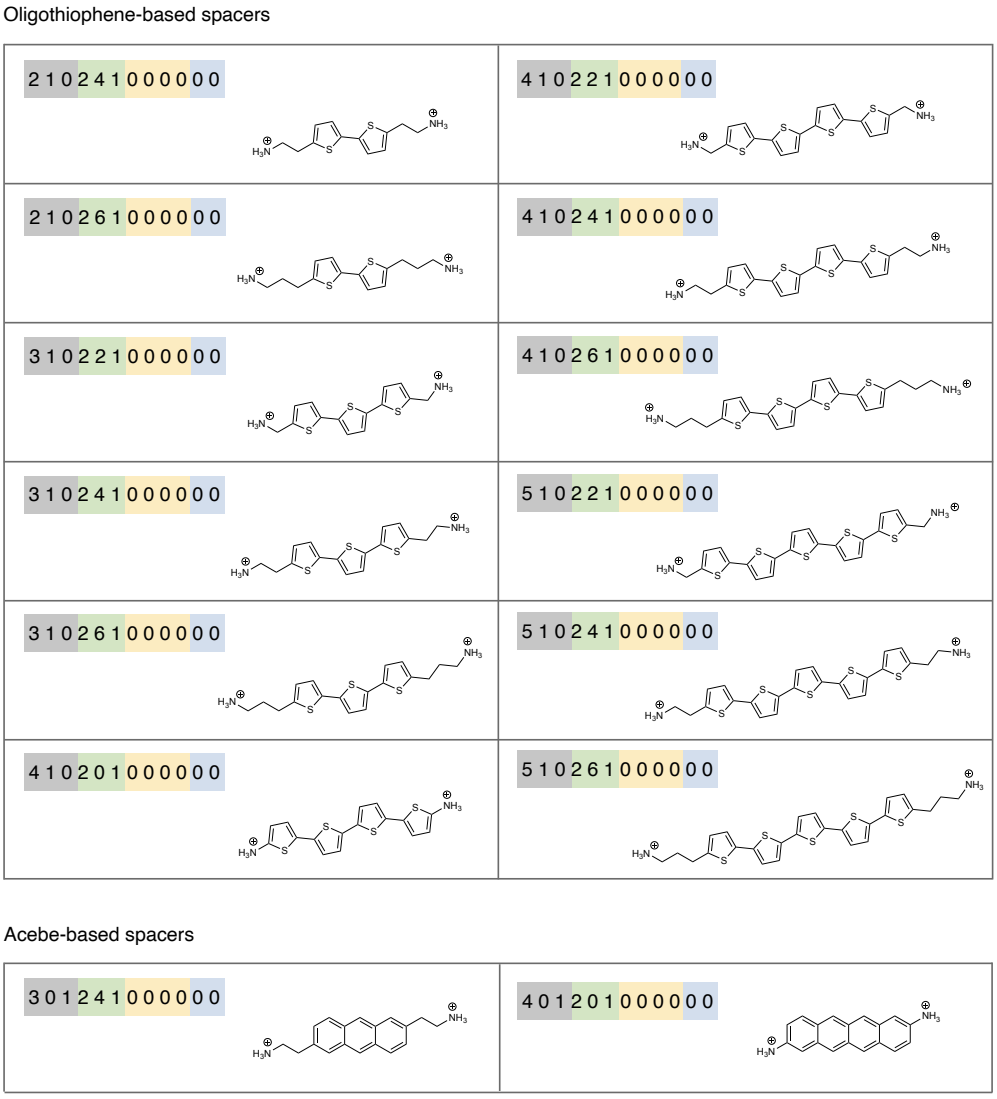


Figure 5.17: Inversed designed candidates for type IIa alignment.

lengths on the two primary ammonium groups, and alternative ring linkage patterns that deviated from the feasible synthetic routes.

These findings highlight the importance of synthesis feasibility constraints in determining the practical viability of Type IIa organic spacers, beyond their electronic properties alone.

Type IIb

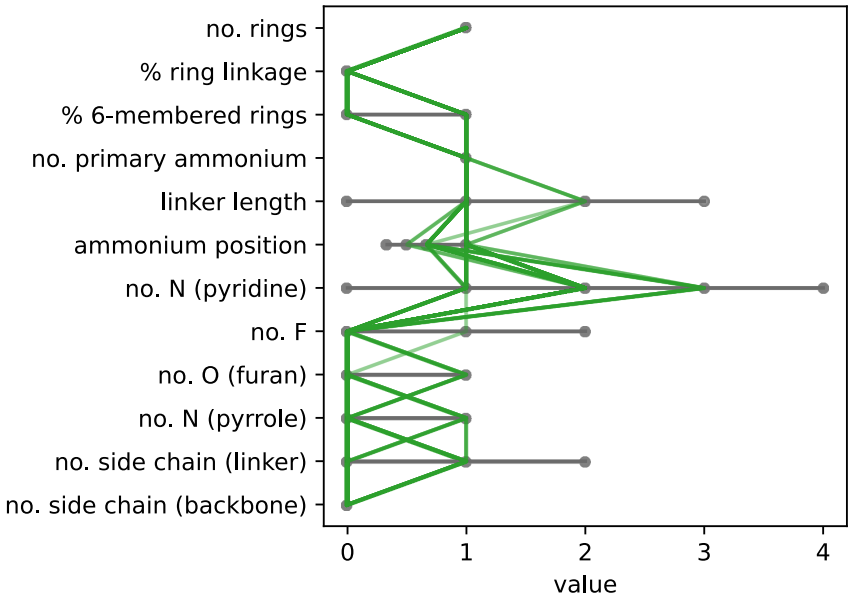


Figure 5.18: The explored fingerprint range and vs. fingerprint range of type IIb final organic spacer candidate.

For Type IIb spacers, the identified fingerprint set corresponds to 823 organic spacers. After applying synthesis feasibility screening, 58 candidates remained as final selections.

The fingerprint range of the final synthesizable candidates is shown in Figure 5.18. Compared to other alignment types, the distribution suggests that Type IIb spacers exhibit less variation in backbone and tethering group features, while showing greater diversity in heteroatom substitutions and side-chain modifications. These candidates primary exhibit the following key structural characteristics:

- Conjugated backbone: Typically one ring, either 5-membered or 6-membered.

- Tethering ammonium group: One primary ammonium group, mostly with a single carbon in the linker.

- Heteroatom substitutions: Presence of one or more pyridine-type nitrogen substitutions.

- Side-chain variations: Broader diversity compared to other alignment types.

The organic spacer structures are presented in Figure 5.19. Interestingly, we observed that multiple Type IIb organic spacers often share identical fingerprints.

5.2.4 DFT validation of designed DJ perovskites

We constructed DJ perovskite structures based on the designed organic spacers and performed DFT calculations to validate their energy level alignments. The DFT validation results, summarized in Figure 5.20, compare all final candidates with experimentally reported DJ perovskites. The newly identified organic spacers significantly expand the energy level alignment landscape, covering a much broader range than previously reported structures. Notably, most of the inverse-designed structures exhibit the targeted energy level alignment, demonstrating the effectiveness of our AI-assisted inverse design workflow.

In the following sections, we discuss the discovered organic spacers for each energy level alignment type, along with key molecular design insights. We show that all selected organic spacers originate from well-established research fields:

- Type Ib and Type IIa candidates are primarily derived from organic electronics and optoelectronic research [122], [123], including acene- and oligothiophene-based molecules.
- Type IIb candidates are predominantly associated with medicinal chemistry, featuring structures such as diazine-based molecules.

Type Ib energy level alignment

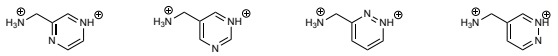
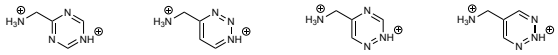
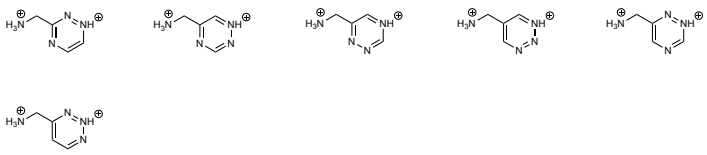
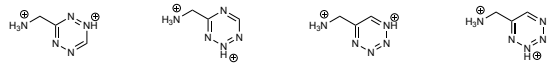
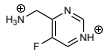
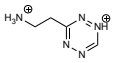
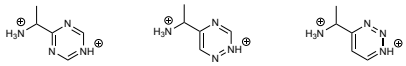
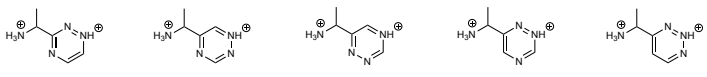
Fingerprint	Organic spacers
1 0 1 1 1 1 1 0 0 0 0 0	
1 0 1 1 1 2/3 1 0 0 0 0 0	
1 0 1 1 1 1 2 0 0 0 0 0	
1 0 1 1 1 2/3 2 0 0 0 0 0	
1 0 1 1 1 1 3 0 0 0 0 0	
1 0 1 1 1 2/3 3 0 0 0 0 0	
1 0 1 1 1 1 1 1 0 0 0 0	
1 0 1 1 2 2/3 3 0 0 0 0 0	
1 0 1 1 1 1 2 0 0 0 1 0	
1 0 1 1 1 2/3 2 0 0 0 1 0	

Figure 5.19: Inverse designed candidates for type IIb alignment (Part 1).

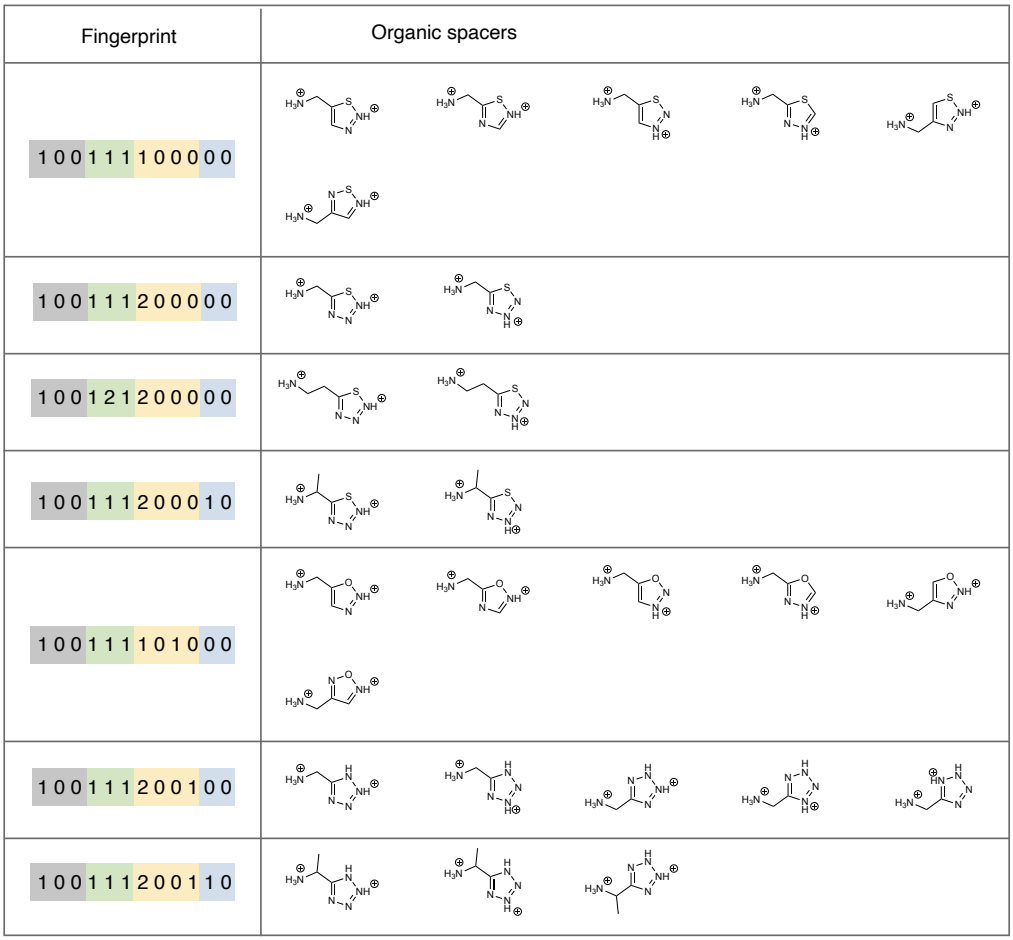


Figure 5.19: Inverse designed candidates for type IIb alignment (Part 2, continued).

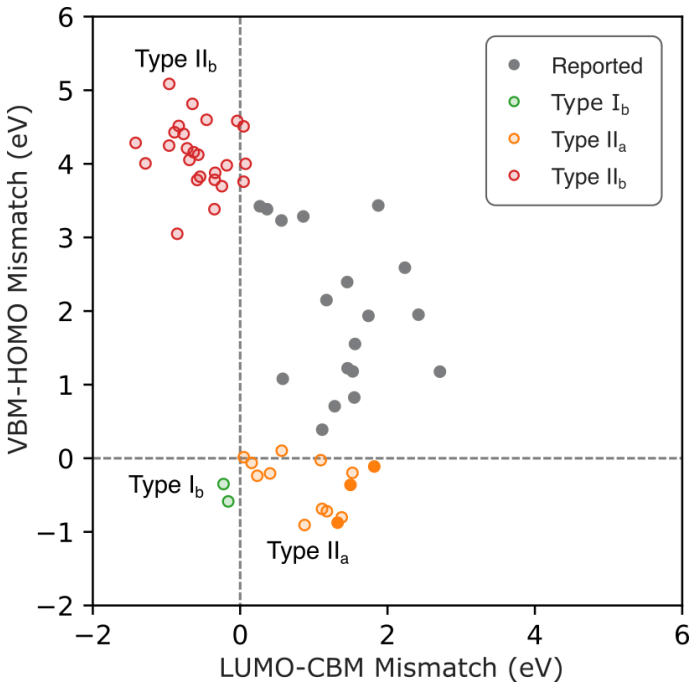


Figure 5.20: Scatter plot depicting the predicted DJ perovskites with targeted alignment types (Ib, IIa, and IIb) alongside previously reported structures.

Designing type Ib spacers proved the most challenging due to the need for a highly conjugated backbone with small HOMO-LUMO gap. Our analysis revealed that only acene-based spacers with at least five linearly fused benzene rings can achieve the required small HOMO-LUMO gap (< 2.3 eV, below the inorganic bandgap). Other conjugated backbones, such as benzene (linked) or thiophene (either linked or fusion) with a comparable number of rings, are ineffective.

Acene-based materials, extensively studied in organic electronics[123], exhibit a progressively narrowing HOMO-LUMO gap as the number of rings increases.

Both two identified type Ib spacers feature a pentacene backbone with two ammonium tethering groups. While higher acene derivatives (e.g., hexacene, heptacene) could theoretically achieve even smaller HOMO-LUMO gaps and guarantee type Ib alignment, they were absent from the PubChem database, likely due to the limited chemical stability of higher acenes.

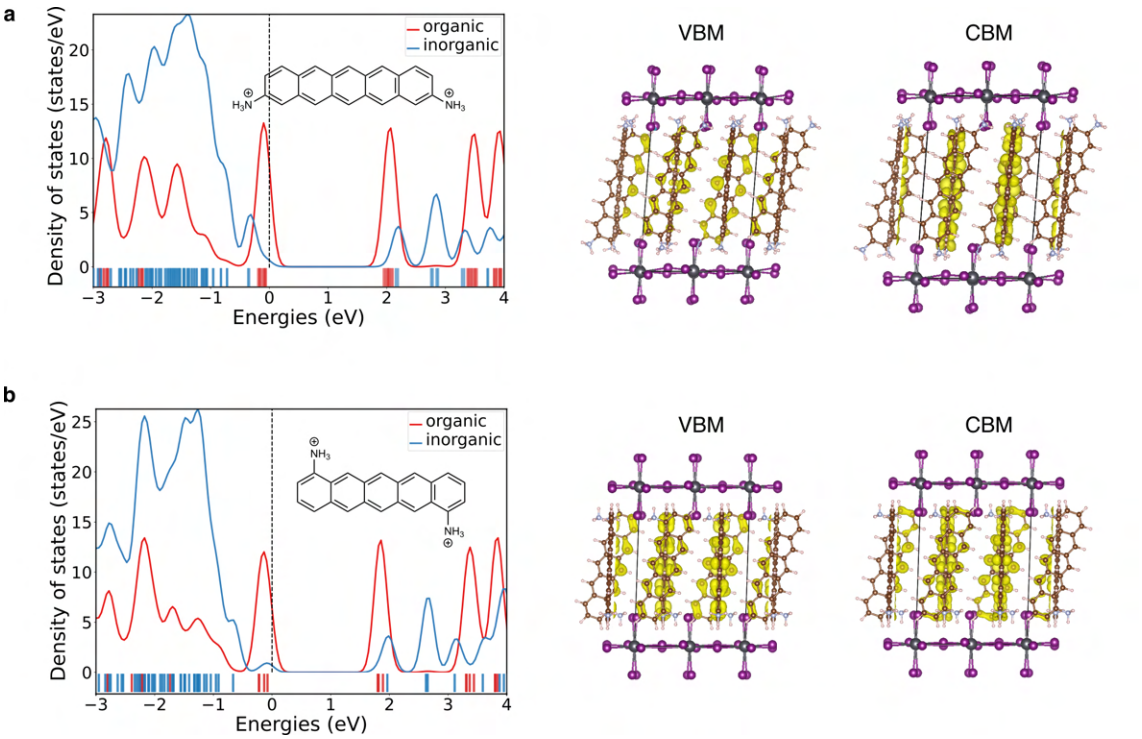


Figure 5.21: Electronic structure of final candidate selections for type Ib DJ perovskites.

The electronic structure of the two proposed DJ perovskites exhibiting Type Ib alignment is shown in Figure 5.21. The left panels display the projected density of states (PDOS), illustrating the electronic contributions from the organic (red) and inorganic (blue) components, with the Fermi level set to 0 eV (dashed line). The right panels show the charge density distributions of the band edge states (VBM and CBM). The yellow isosurfaces indicate the spatial distribution of electronic states, confirming that both the VBM and CBM of the DJ perovskite are primarily contributed by the organic component. Moreover, the HOMO and LUMO of the organic spacers are predominantly localized along the pentacene backbone, where the π-electron density is most concentrated and delocalized.

Comparison with reported 2D perovskites

To date, only one diammonium organic spacer featuring type Ib alignment has been theoretically designed in DJ lead-iodide perovskites[109]. This spacer, which also features a pentacene backbone with two methylammonium tethering groups, was identified in our

inverse design (appear in G_4), but was excluded in the subsequent PubChem filtering step.

The only experimentally synthesized 2D lead-iodide perovskite spacer with type Ib alignment belong to RP phase[72].In that study, Type Ib alignment was inferred based on photoluminescence (PL) emission from the organic component and qualitative energy level estimations using a series of approximations. However, relying on optical properties alone to deduce electronic alignment is questionable, as optical gaps do not always directly correspond to electronic energy levels. Moreover, their DFT calculations suggested a mixed alignment between Type IIa and Type Ib, with a mixing factor of 0.25.

Our calculations (Figure 5.22), performed using the same perovskite structure as in their study, indicate that the alignment is actually Type IIa, regardless of whether we apply the systematic mixing factor of 0.4 used throughout our study or the 0.25 mixing factor employed in the previous work. The discrepancy in computational results is likely due to differences in software implementations and methodological choices.

While direct comparisons across different simulation methodologies and experimental techniques remain challenging in 2D perovskite research[97], the relative trends within our calculations—conducted with a consistent set of computational parameters—are internally reliable. These findings suggest that our proposed pentacene-based organic spacers provide a more precise and controllable approach for achieving Type Ib energy level alignment in 2D perovskites.

Type IIa energy level alignment

Achieving Type IIa alignment typically requires extended conjugation through an increased number of aromatic rings to raise the HOMO energy level. Our inverse design approach identified two major families of organic spacers that satisfy these criteria: acene-based molecules with fewer rings than pentacene and oligothiophene-based molecules. It is well established in organic chemistry that both families exhibit a progressively narrower HOMO-LUMO gap as the ring count increases.

In the context of organic spacers, these molecules rely on the same principle—extending

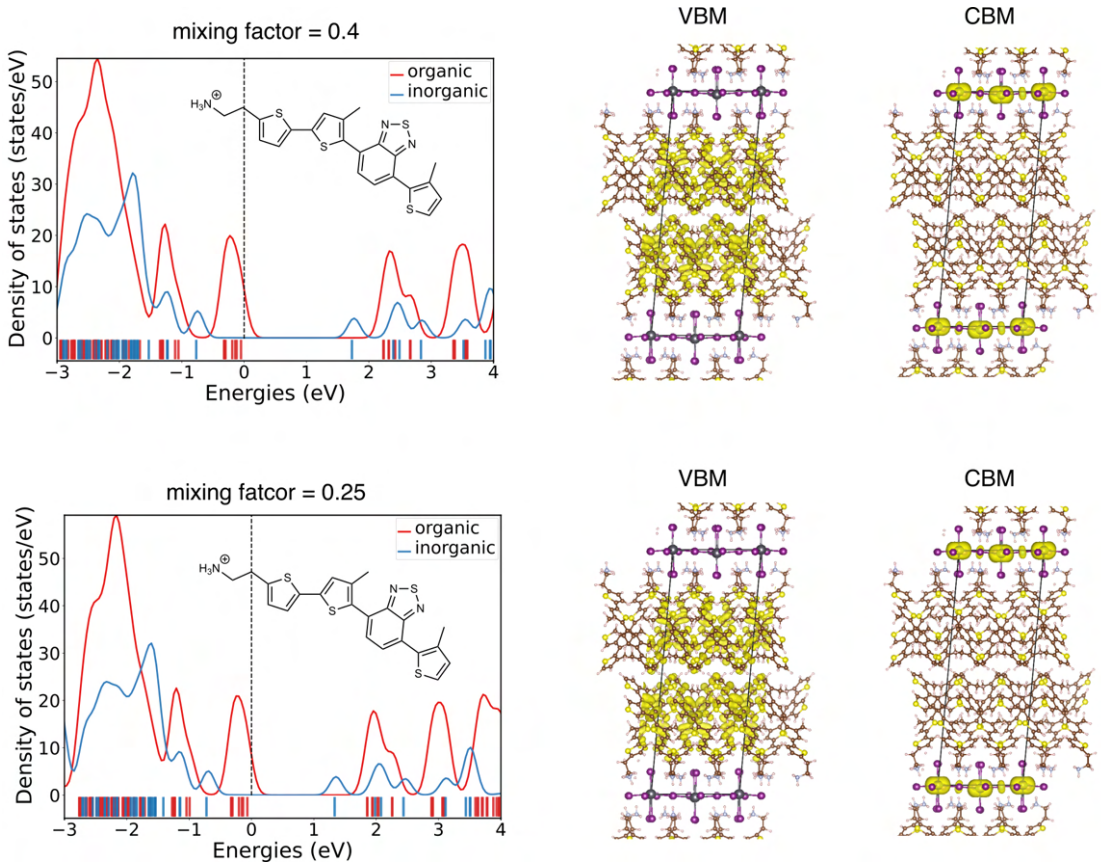


Figure 5.22: Energy level alignment of the previously reported RP-phase spacer claimed to exhibit type Ib alignment.

conjugation to elevate HOMO energy levels and reduce LUMO levels. Additionally, our analysis indicates that increasing the linker length in the tethering ammonium group raises both HOMO and LUMO energy levels, further influencing energy alignment.

The electronic structures of the 12 DJ perovskites structures exhibiting type IIa alignment are shown in Figure 5.23. The VBM is primarily contributed by the organic component, while the CBM remains dominated by the inorganic framework.

Within the oligothiophene family, all identified organic spacers feature linked thiophene rings with variable ring counts and different linker lengths. This family represents one of the earliest explored organic spacers in 2D perovskites, and three of these spacers (Figure 5.23a,c,e) have already been reported in the literature [9], [97].

Only one acene-based spacer (Figure 5.23l) was confirmed to exhibit Type IIa alignment, featuring anthracene with an ethylammonium tethering group. This molecule has been extensively studied in a theoretical work [109], where it was reported to exhibit a mixed Type IIa/Ia alignment. The slight mismatch with our findings is likely due to differences in structural details; however, the overall trend remains consistent between our study and previous research.

Type IIb energy level alignment

Type IIb spacers typically require a single primary ammonium group and pyridine-type nitrogen substitutions at multiple positions on aromatic rings to effectively lower the LUMO level. These organic spacers frequently feature nitrogen-substituted ring systems as their conjugated backbone, which are well-established in medicinal chemistry and related fields. For example, our study identified several organic spacers containing six-membered aromatic diazines, including pyrazine, pyridazine, and pyrimidine. Two of these spacers were previously predicted in theoretical studies with similar results [104]; however, to date, no experimental studies have been conducted on DJ perovskites featuring Type IIb alignment.

Compared to the only known spacer reported in the RP phase, our identified spacers

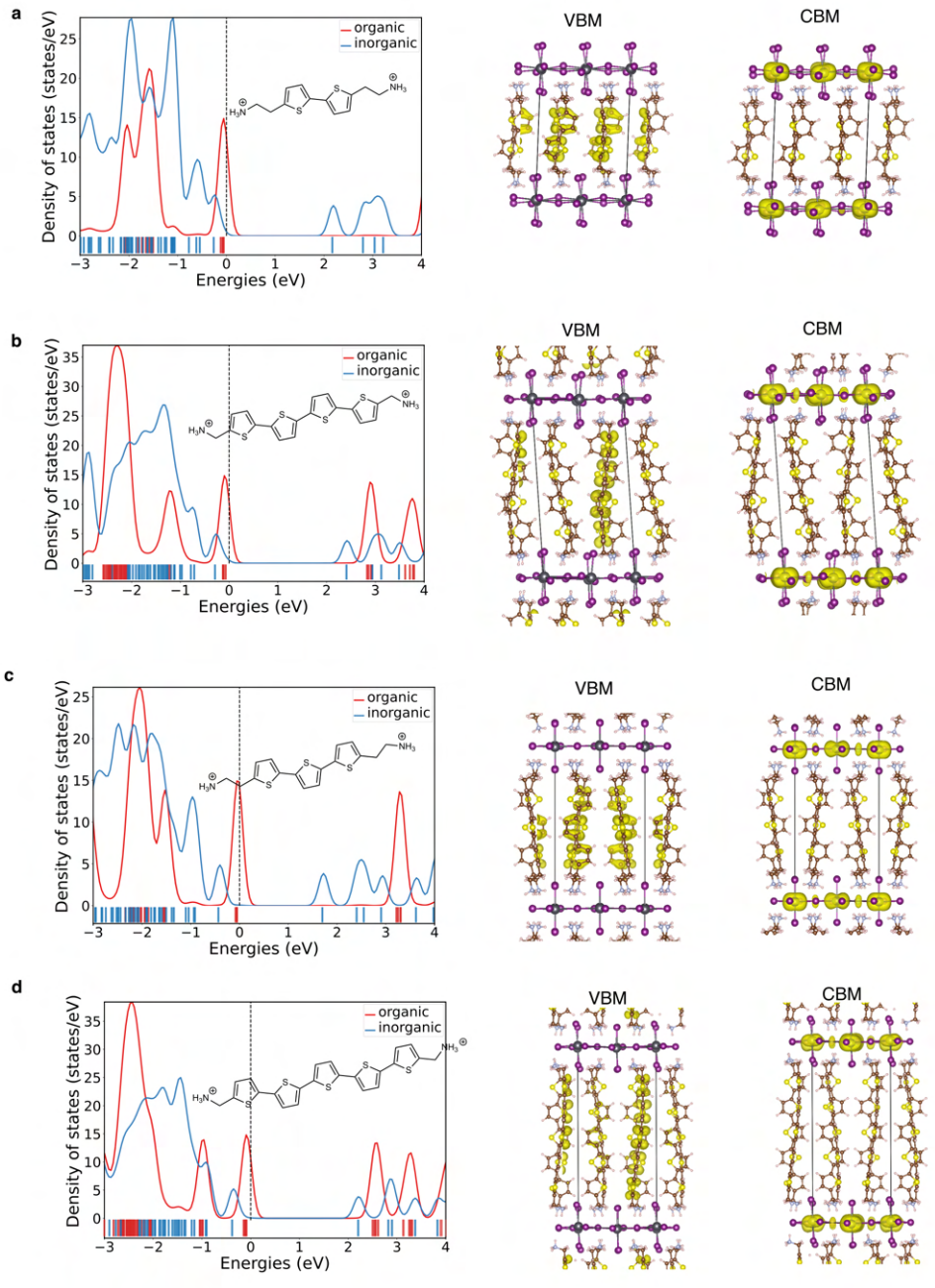


Figure 5.23: Electronic structure of inverse-designed type IIa DJ perovskites (Part 1).

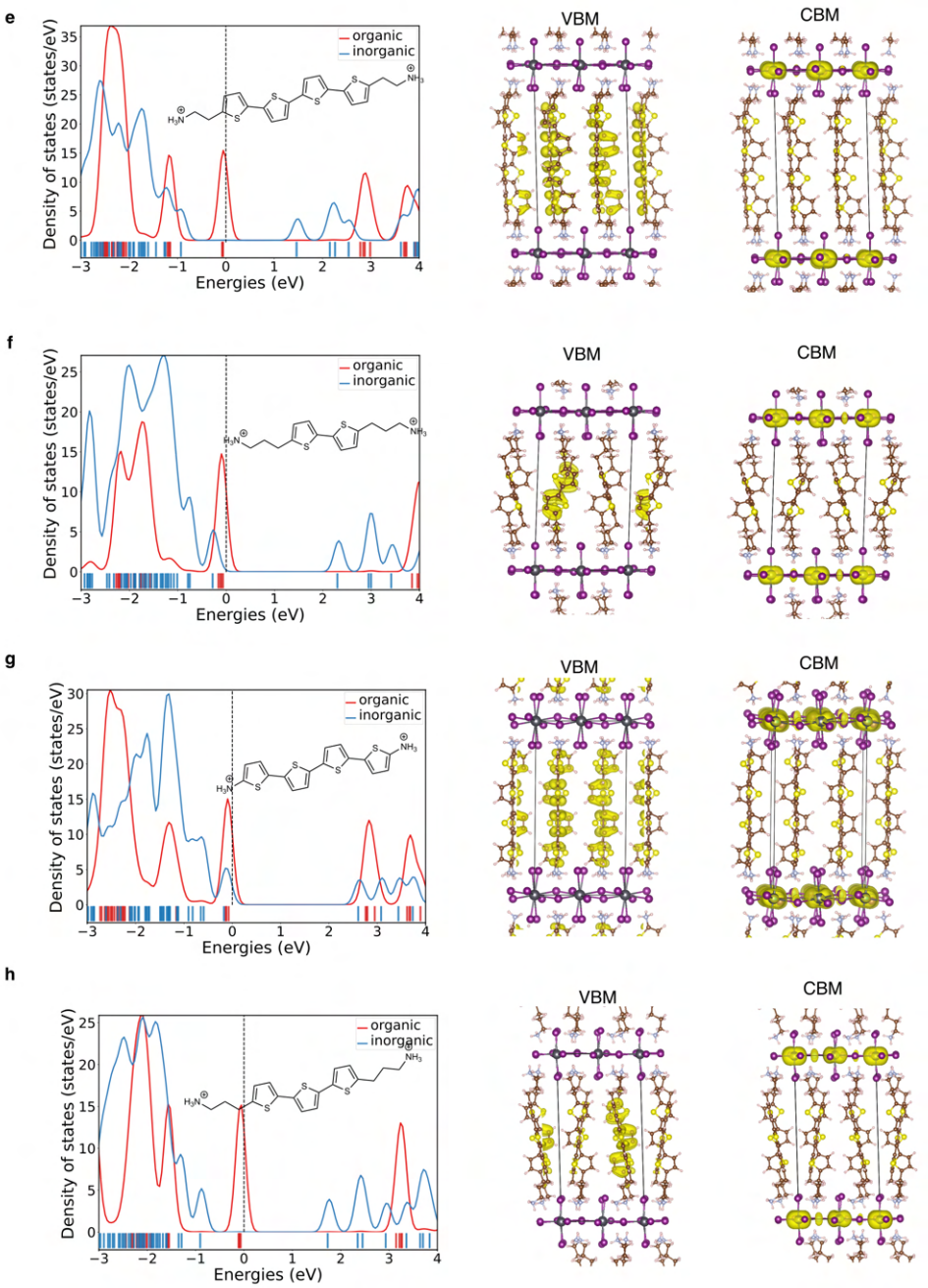


Figure 5.23: Electronic structure of inverse-designed type IIa DJ perovskites (Part 2, continued).

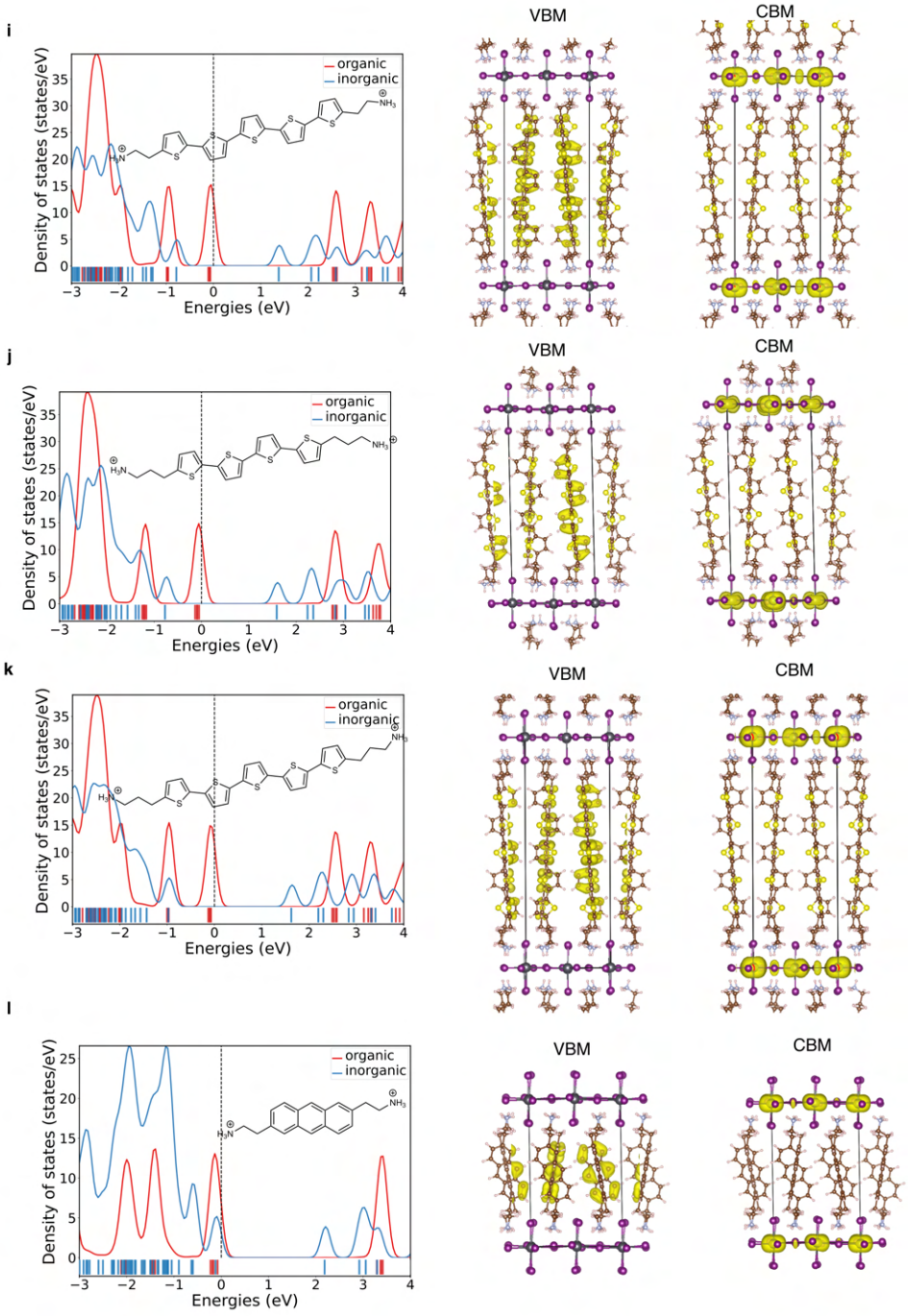


Figure 5.23: Electronic structure of inverse-designed type IIa DJ perovskites (Part 3, continued).

exhibit significantly simpler structures, suggesting improved synthetic accessibility and potential for experimental realization.

The electronic structure of the Type IIb DJ perovskites is shown in Figure 5.24. The bottom panel of the partial density of states (PDOS) displays the energy levels at the Z-point and Γ-point, respectively. Differences in energy levels between these two points indicate the presence of interlayer coupling between inorganic layers. Notably, most structures exhibit significant interlayer coupling, due to the small size of organic spacers within this alignment type.

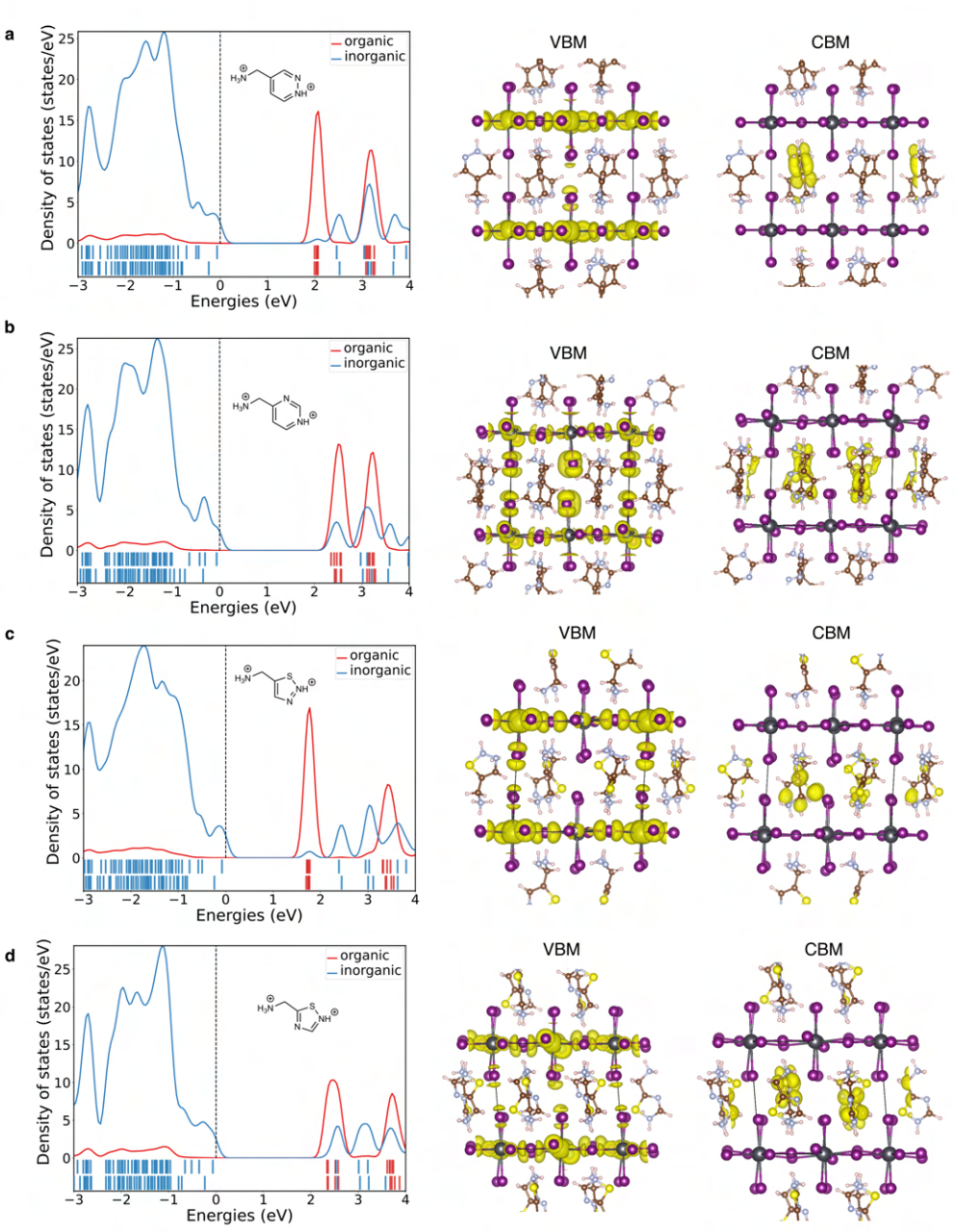


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 1).

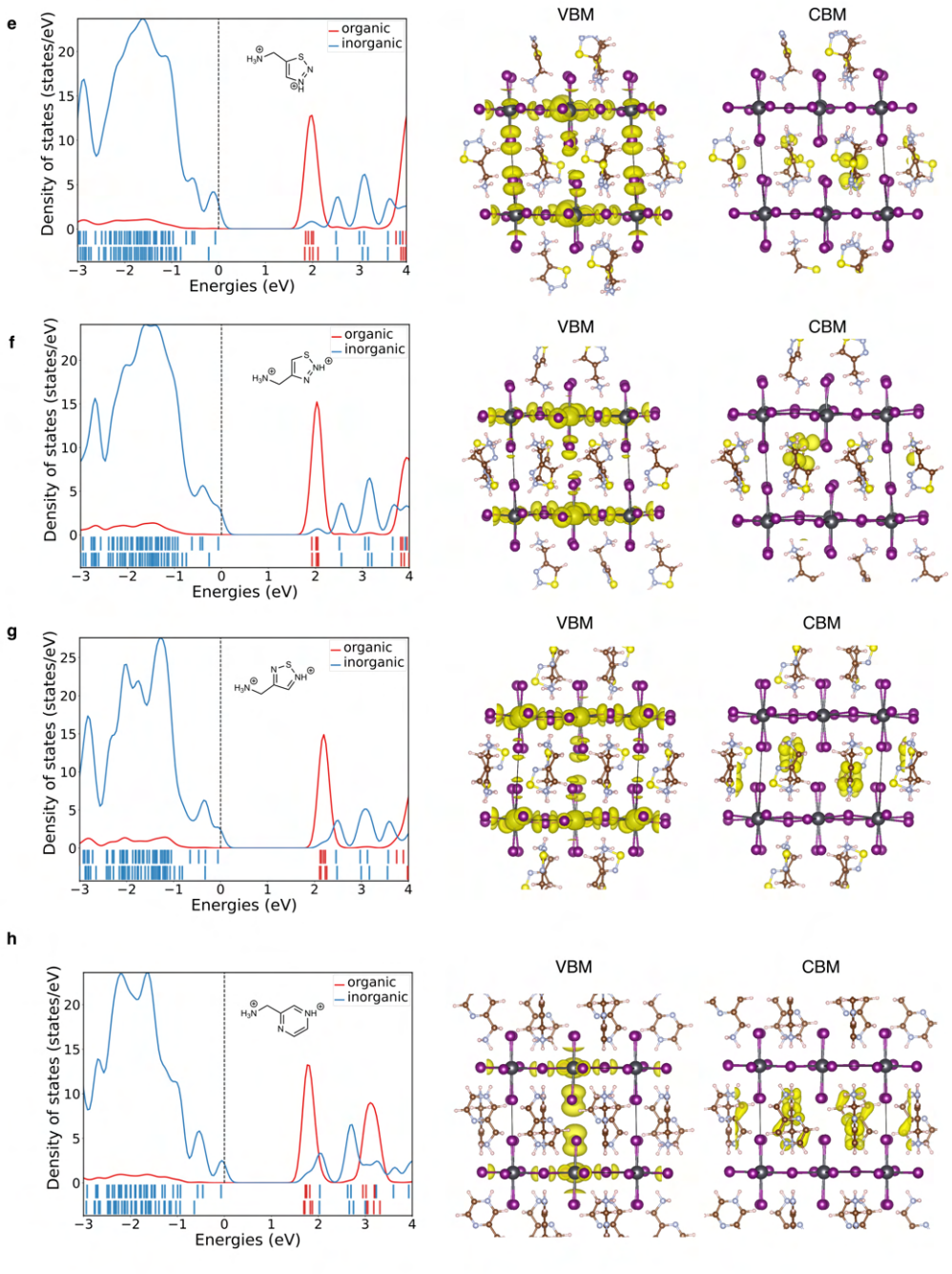


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 2, continued).

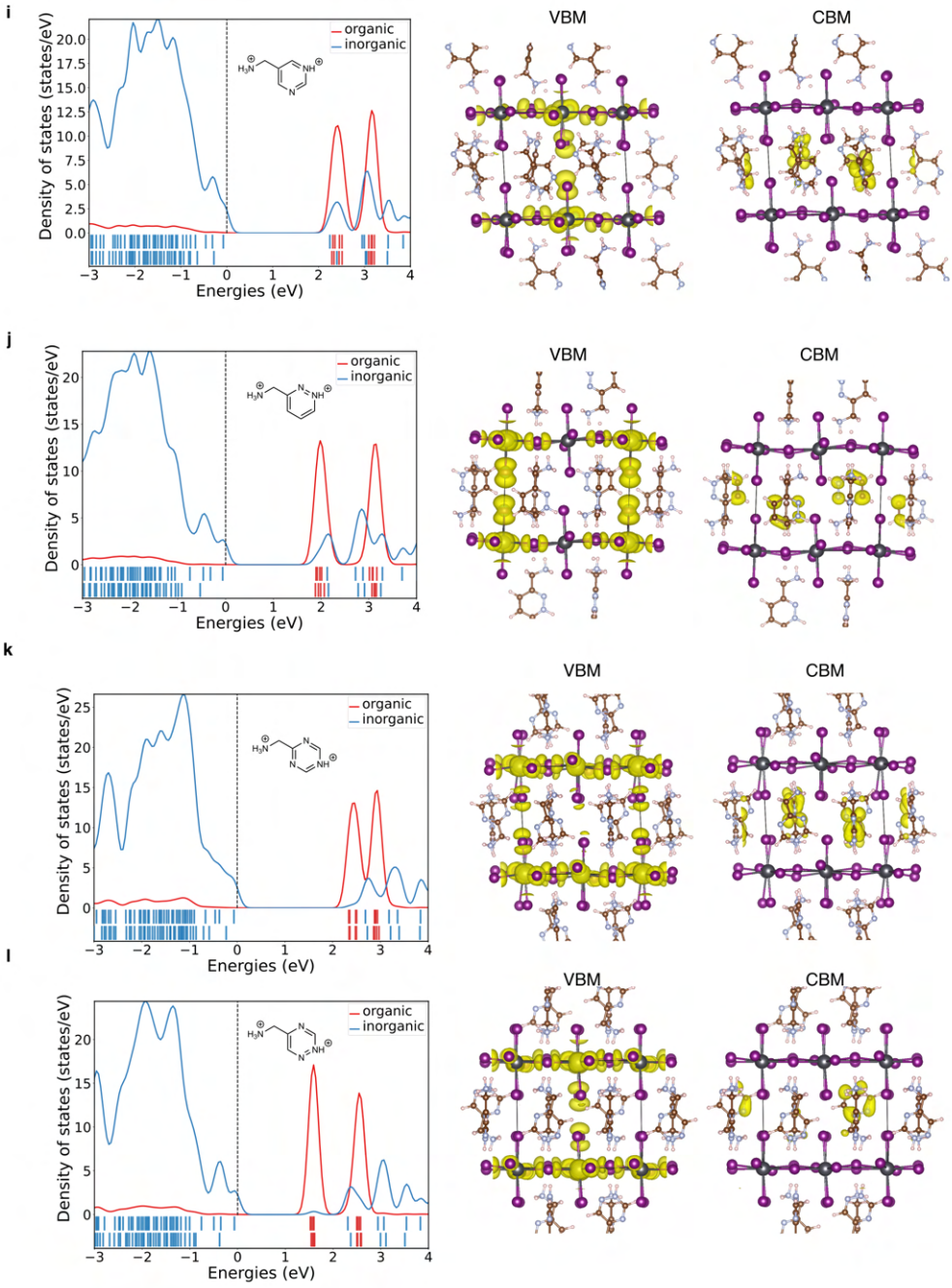


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 3, continued).

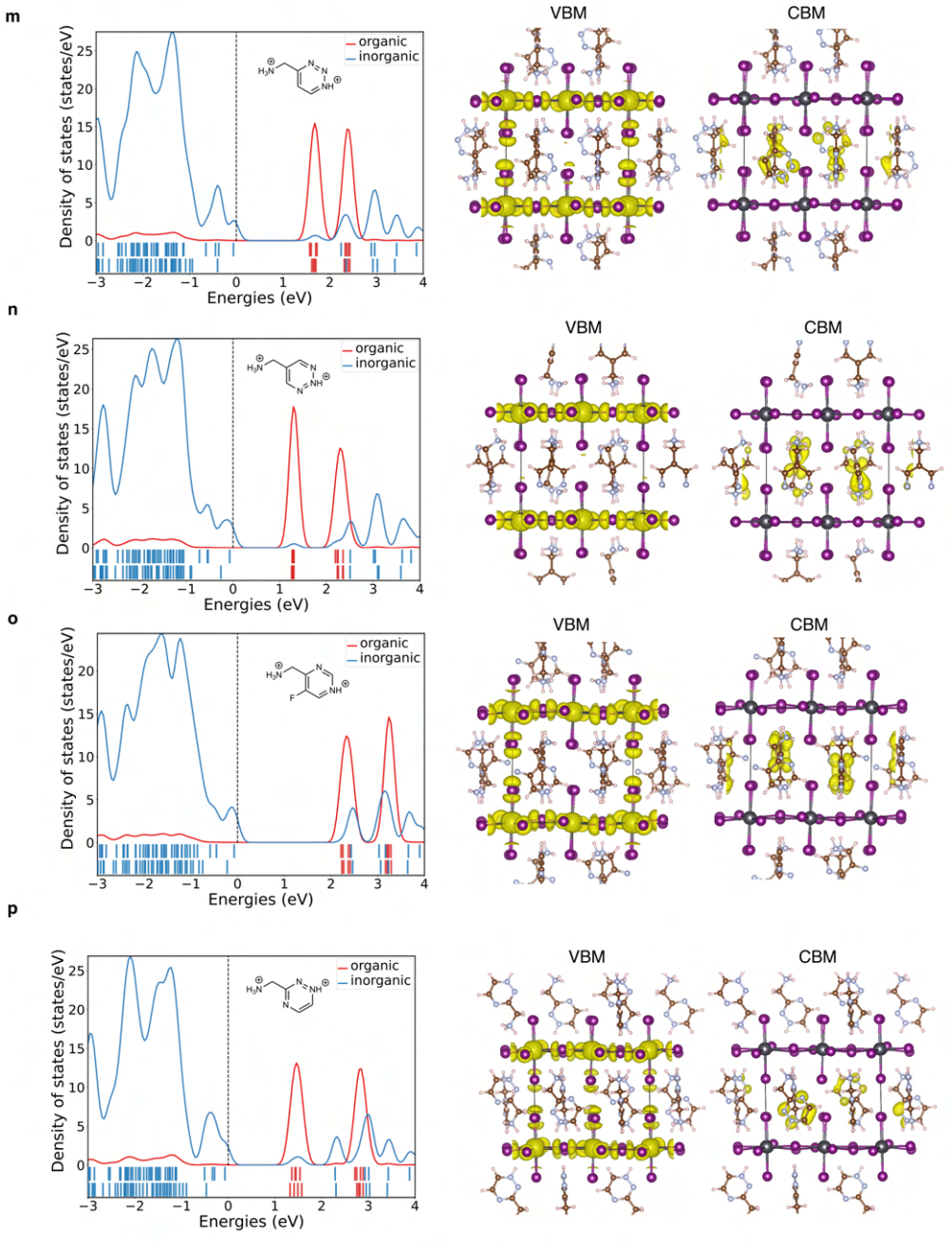


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 4, continued).

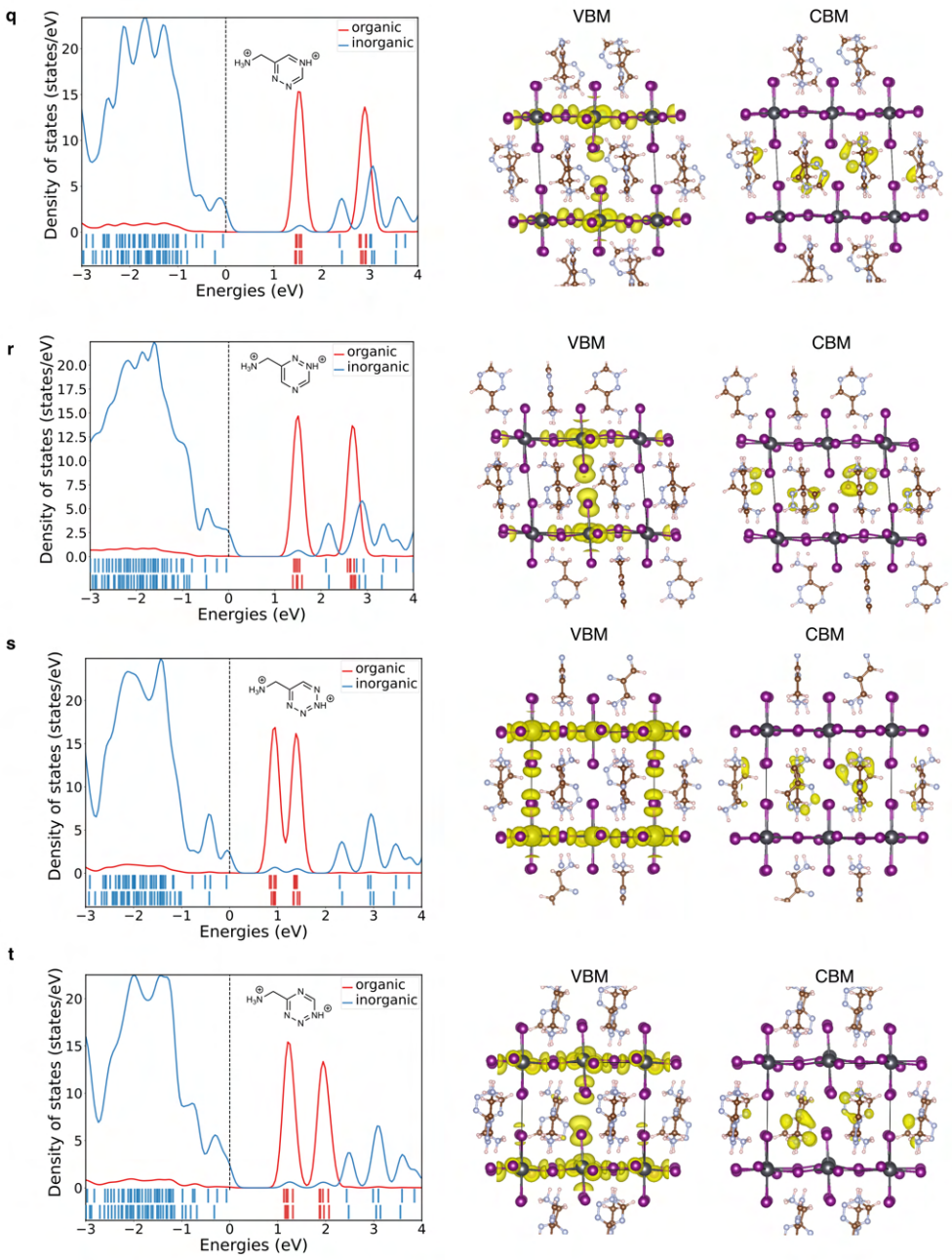


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 5, continued).

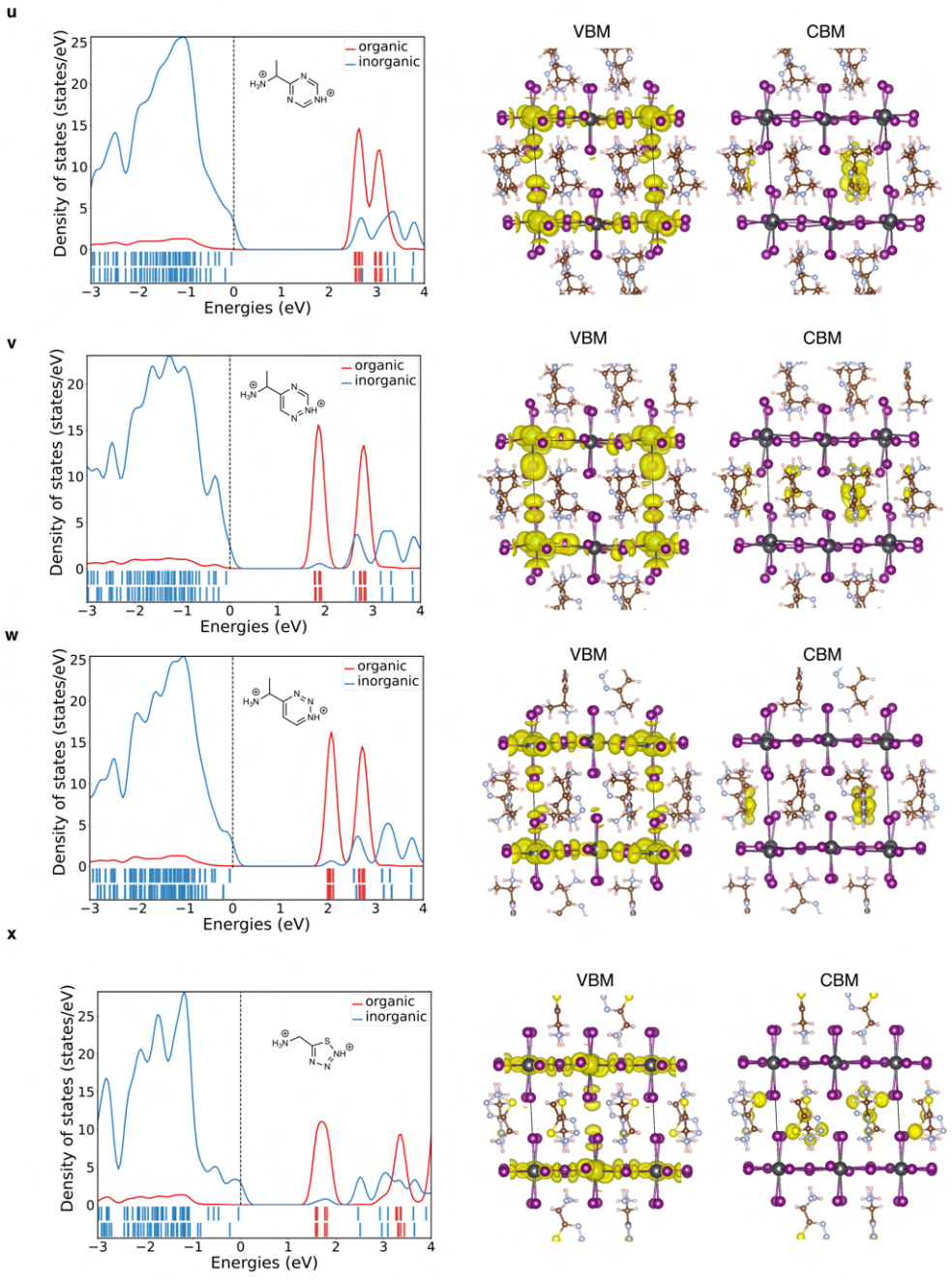


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 6, continued).

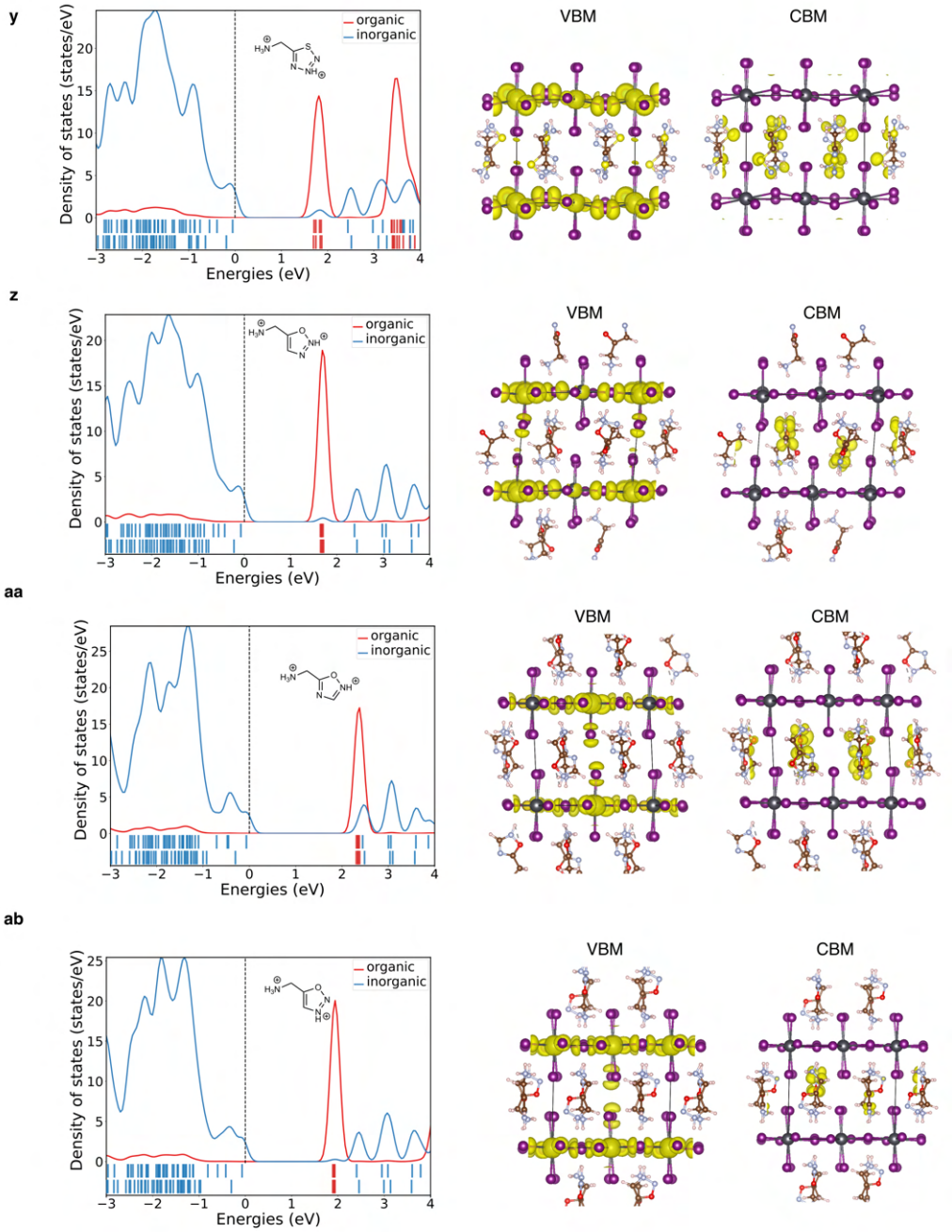


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 7, continued).

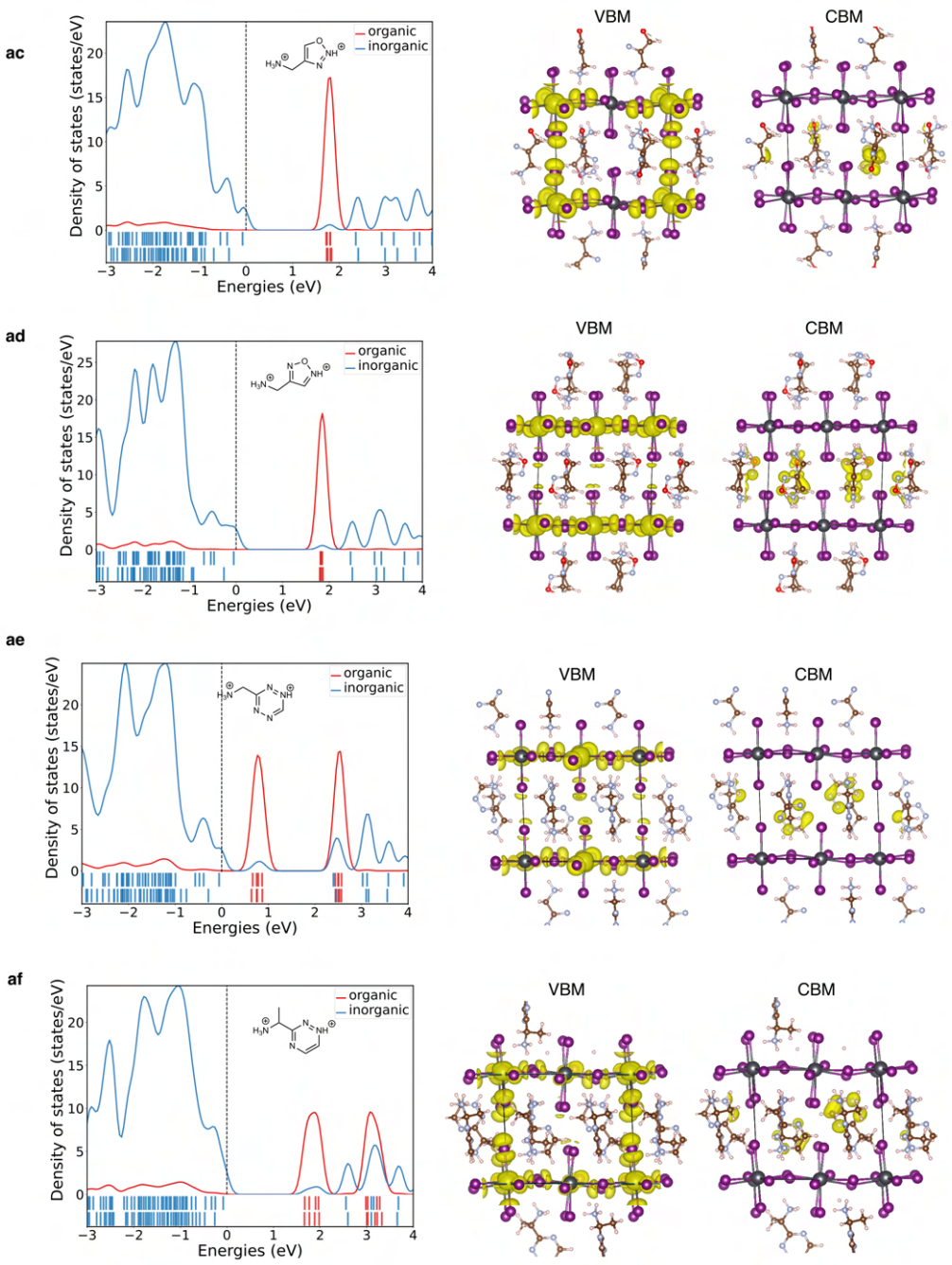


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 8, continued).

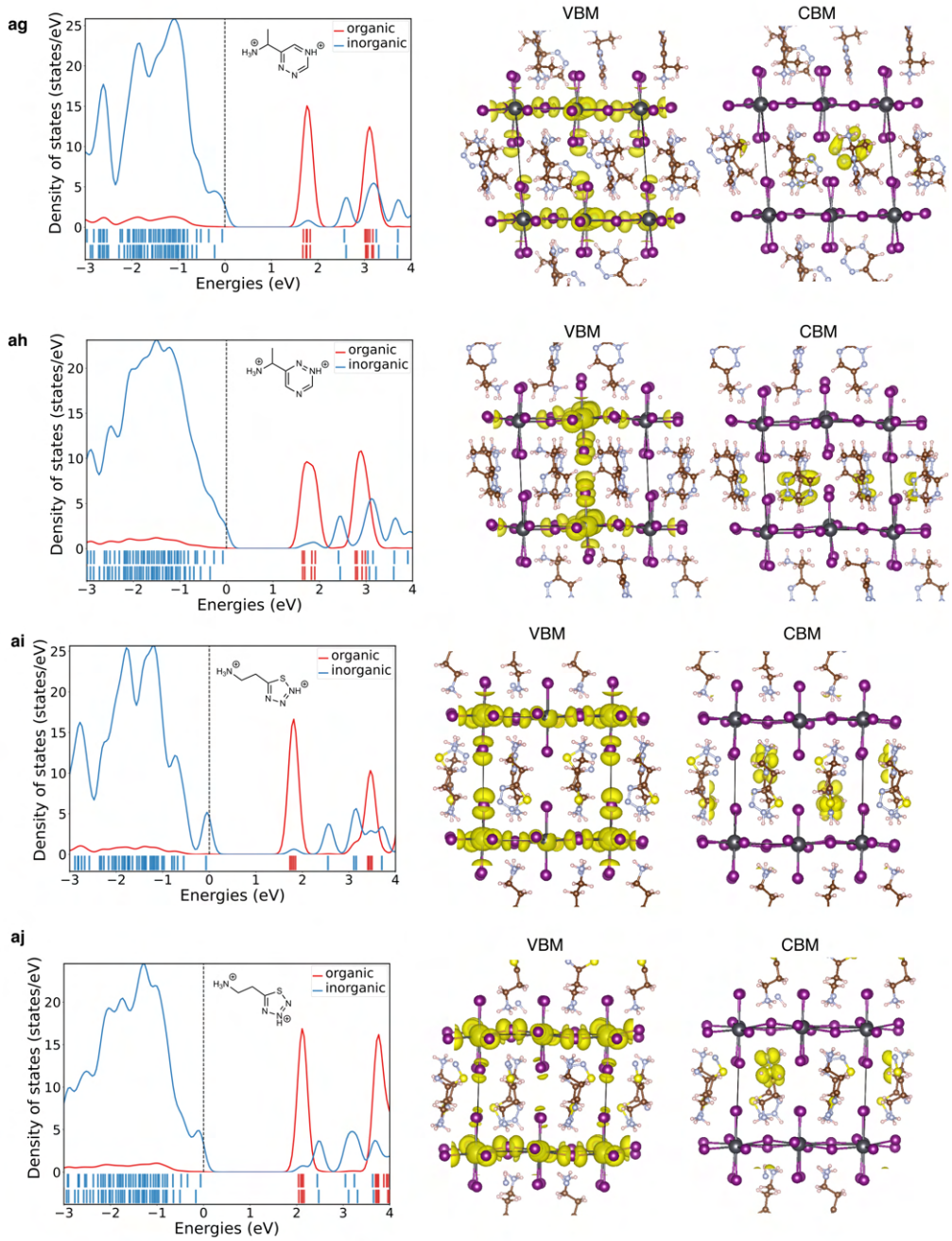


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 9, continued).

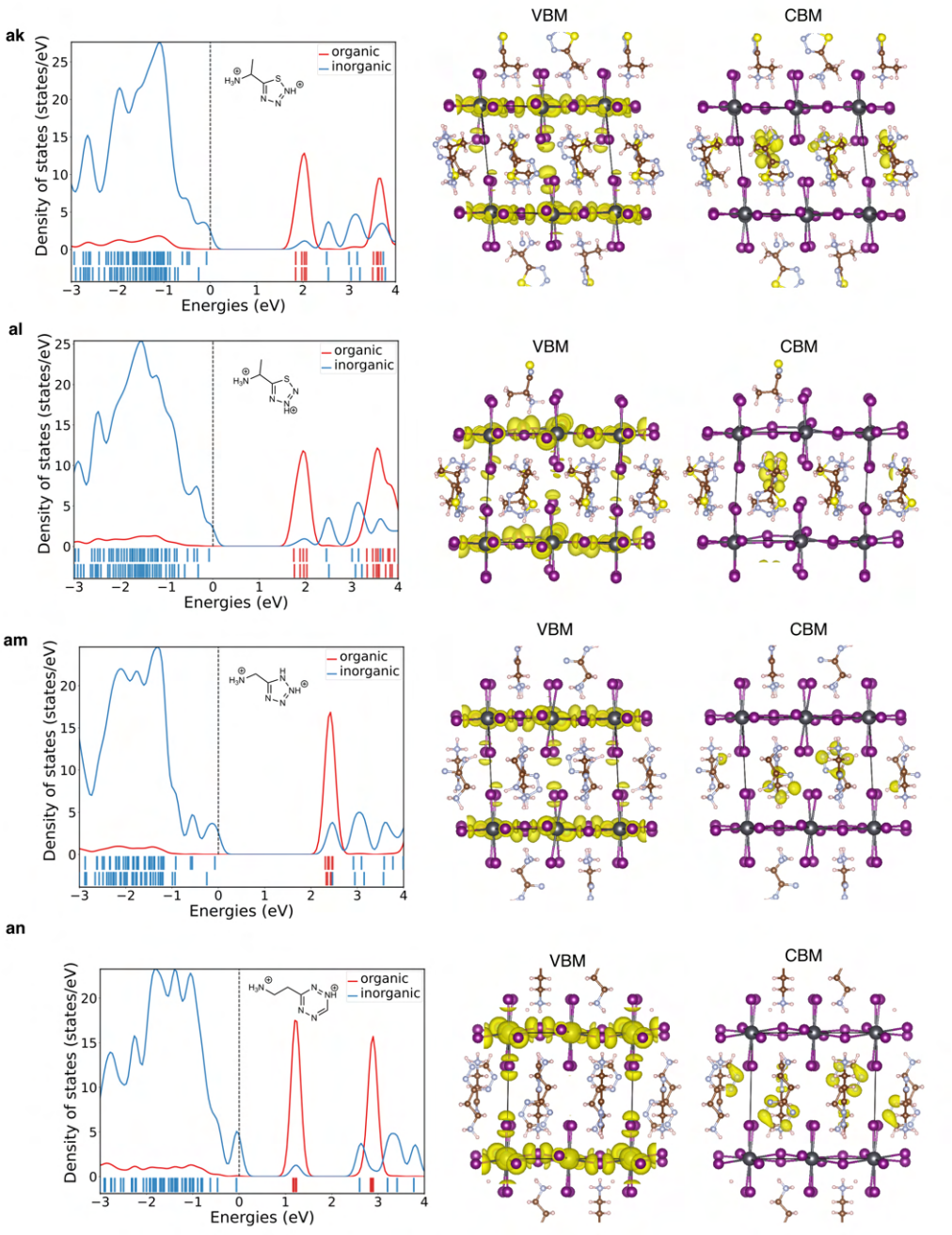


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 10, continued).

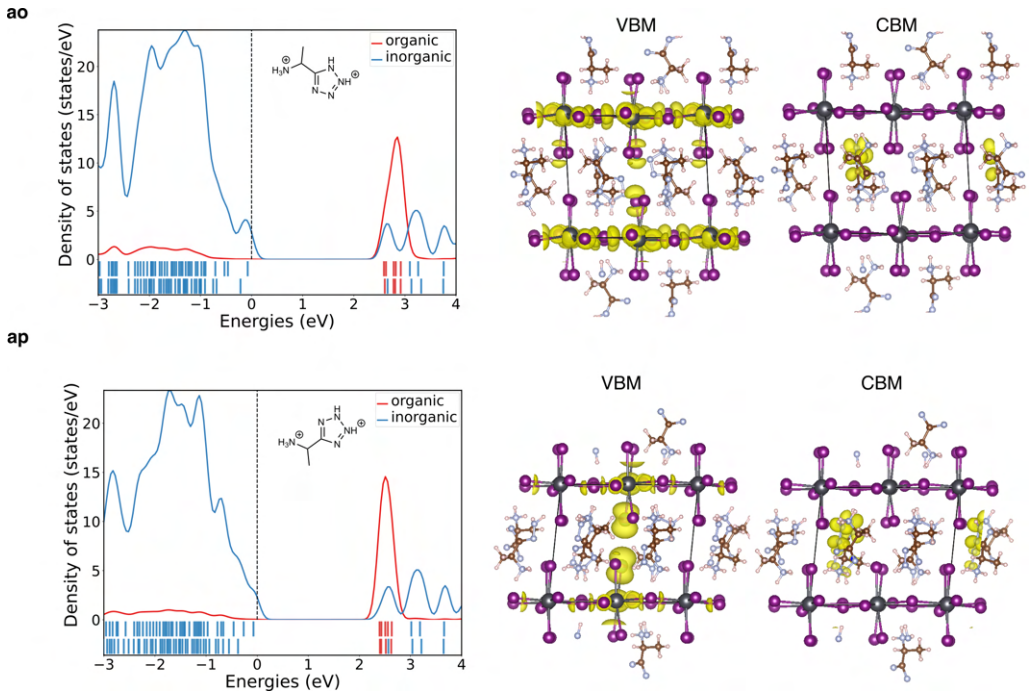


Figure 5.24: Electronic structure of inverse-designed type IIb DJ perovskites (Part 11, continued).

5.2.5 Chemical space visualization of final candidates

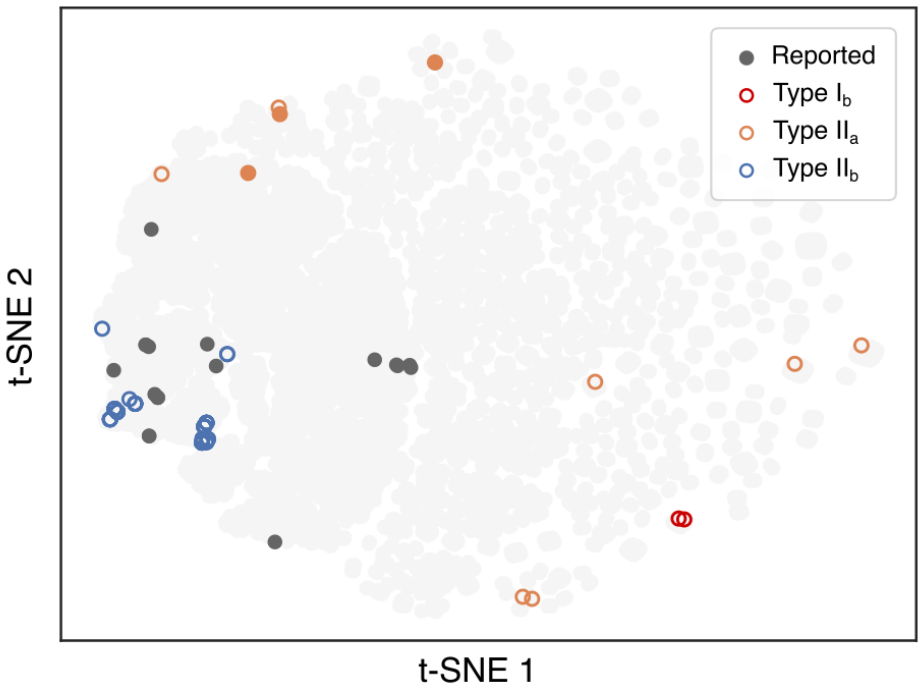


Figure 5.25: Chemical space visualization of inverse-designed DJ perovskites.

Figure 5.25 presents a t-SNE projection of the chemical space defined by the 12-digit fingerprint, highlighting both experimentally reported DJ-phase organic spacers (solid circles) and newly discovered candidates (open circles) with favorable energy level alignment types: Type Ib, Type IIa, and Type IIb. The reported spacers cluster primarily in the left-central region of the map, indicating a relatively narrow exploration of the broader chemical space. This cluster is dominated by molecules with a single aromatic ring and short linker lengths. In contrast, the newly discovered candidates—identified through inverse design and machine learning—are more broadly distributed, often occupying regions sparsely populated or entirely unoccupied by known compounds.

Type Ib candidates are located predominantly in the bottom-right region, far from the known spacers. These molecules tend to have five-membered rings and relatively short linkers. Type IIa candidates are more widely spread, including both known structures

near the top of the map and newly identified candidates in remote regions, indicating chemical diversity within this alignment category. Type IIb spacers are mostly found in the center-left region, within a moderately dense area of the chemical space, yet still occupy a subspace not covered by any reported spacers.

These observations highlight the ability of the proposed workflow to identify chemically distinct and previously unexplored candidates. The clear spatial separation between known and newly designed molecules reinforces the novelty of the final candidates and supports the claim that this data-driven approach effectively expands the chemically relevant design space for DJ-phase perovskites.

5.3 Chapter summary

This chapter addressed the critical challenge of bridging computational predictions with experimental feasibility in 2D perovskite design. A two-step screening strategy was introduced to assess the synthetic accessibility of candidate organic spacers, incorporating cheminformatics-based filtering and literature cross-referencing. This was followed by the inverse design of final DJ-phase candidates, which balanced desired energy level alignment with practical constraints. The resulting shortlist of spacers represents a set of experimentally viable materials with targeted electronic properties, laying the groundwork for future experimental validation.

Chapter 6

Summary and Outlook

6.1 Summary of key contributions

This thesis presents a comprehensive framework for the AI-assisted inverse design of DJ-phase 2D perovskites with targeted electronic properties. By integrating an invertible molecular fingerprint, high-throughput DFT calculations, and a synthesis feasibility screening strategy, the proposed workflow enables property-driven generation of organic spacers that are both theoretically promising and experimentally viable. This framework supports efficient exploration of vast chemical spaces and offers a systematic approach to designing lab-synthesizable DJ-phase perovskite materials. Beyond DJ systems, the underlying principles—particularly the use of interpretable and invertible molecular representations—provide a transferable strategy applicable to a broad range of hybrid materials.

The key contributions of this work are as follows:

First, we developed a 12-digit fingerprint representation tailored to encode key structural features of organic spacers in a machine-readable and human-interpretable format. This invertible fingerprint simplifies the inverse design process by enabling direct generation of candidate structures without relying on complex AI architectures such as deep neural

networks. The modular design of the fingerprint also allows for adaptation to other families of small organic molecules with user-defined structural features.

Second, we significantly expanded the design space of DJ-phase spacers—from approximately $\tilde{2}0$ known experimental molecules to over $\tilde{1}0^6$ hypothetical candidates—through systematic morphing operations. From this expanded space, $\tilde{7}0$ final organic spacers were identified with favorable energy level alignments (Types Ib, IIa, and IIb). Many of these candidates exhibit structural features distinct from reported molecules, thereby offering new insights and opportunities for experimental realization.

Third, we demonstrated that interpretable, physics-informed machine learning models, specifically linear regression using domain-relevant descriptors, can achieve high predictive performance. This supports the notion that incorporating expert knowledge into model design can offer a more transparent and effective alternative to complex, black-box approaches, especially when training data is limited.

Finally, we introduced a synthesis feasibility screening filter based on synthetic accessibility of organic spacers (from PubChem) and assessment of 2D formability based on hydrogen bonding between organic and inorganic components. To our knowledge, this is the first virtual synthesis feasibility filter tailored specifically for DJ-phase perovskites. This screening step enabled the prioritization of a shortlist of realistic, lab-accessible candidates with desirable electronic properties, bridging the gap between computational design and experimental synthesis.

Addressing research gaps

This work directly addresses several research gaps outlined in Chapter 2:

- Data scarcity is addressed through the generation of a large, systematically constructed dataset based on molecular morphing and high-throughput DFT calculations, significantly expanding the known chemical space of DJ-phase organic spacers.
- Structure–property relationships are revealed through interpretable machine learning

models trained on physics-informed descriptors. These descriptors are derived from insights gained through high-throughput DFT simulations and are designed to reflect relevant molecular and electronic features.

- Constraints of hybrid materials are incorporated through the fingerprinting scheme, which defines a chemically meaningful and computationally tractable scope of organic spacers compatible with 2D perovskite structures.

- The interaction between the organic and inorganic components are reflected in the synthesis feasibility filter, which includes a 2D formability assessment based on potential hydrogen bonding interactions, providing an initial proxy for evaluating the compatibility between organic cations and the inorganic lattice.

6.2 Limitations and challenges

While the ML-assisted workflow has proven effective in the inverse design of organic spacers with targeted energy level alignments, several challenges remain. The first limitation arises from the inability to identify certain organic spacers previously designed by organic chemistry experts. This shortfall stem from the trade-off inherent in the fingerprint representation, which, while compact and interpretable, confines the scope of explored chemical space. As discussed in Chapter 3, this fragment-based fingerprint vector restricts exploration to a specific subset of organic spacers, for example, excluding organic spacers with additional ring on vertical direction, or inclusion of triple bonds as those designed by chemistry experts. As a result, compounds that fall outside this scope—such as triple bonds and additional rings on vertical direction—remain unaddressed[72], [106]. This limitation underscores a broader challenge in AI-driven materials discovery: bridging the gap between machine exploration and the expert intuition cultivated through decades of experimental research.

The second limitation pertains to reduced prediction accuracy, particularly as the molecular structure becomes more complex. This reflects another fundamental limitation of

machine learning models: their performance is intrinsically tied to the quality and diversity of the training data provided. We identified two key aspects of chemistry insights that were not included in the machine learning model as a trade-off to reduce computational cost:

(1) Input feature limitations: the input features (fingerprint) do not adequately capture the increased structural complexity of the final candidates, which can significantly influence their energy levels. This includes distinctions in energy levels among certain isomers with identical fingerprints and variations due to the conformations of organic spacers in hybrid perovskite structures.

(2) Training data scope: the training data, drawn primarily from $G_0 - G_3$ spacers, encompasses a narrow feature space, predominantly featuring spacers with 1-4 rings (61% containing 1-2 rings). This limited dataset leads the model to learn structure-property relationships within this range. However, the model struggles with higher-generation spacers outside this feature space, where chemistry deviates significantly, because it lacks prior exposure to such chemistries. Addressing these limitations, through more comprehensive fingerprints or including higher generation spacers in the training data would require a substantial increase in computational resources.

6.3 Outlook

Despite these challenges, the workflow represents a versatile tool for materials discovery, with several opportunities for refinement and broader application:

1. **Expanding target properties:** the workflow can be adapted to optimize other properties in DJ perovskites, such as chirality, charge mobility, etc.

2. Applicability to other systems: While this workflow is directly applicable to 2D perovskite organic spacers, it can be extended to other materials systems, especially those involving small organic molecules by customizing the molecular fingerprint.

3. **Flexible data sources:** The data generation need not rely solely on high-throughput DFT calculations, Alternative sources, such as high-throughput experiments or other simulation techniques, can be integrated into the pipeline.

4. **Advanced machine learning models:** The pipeline could be enhanced with more sophisticated machine learning models to capture nonlinear and intricate structure-property relationships. Strategies such as active learning or Bayesian optimization could further refine the selection of final candidates.

5. **Multi-objective optimization:** Future work should incorporate additional performance-relevant properties such as exciton binding energy, defect-formation energy, and charge carrier mobility. These properties are critical for practical device deployment, and their inclusion would support a more comprehensive understanding of structure–function relationships in 2D perovskites.

6. **Application-specific material selection guidelines:** To enhance the practical relevance of this work, future studies could establish device-oriented selection criteria based on predicted material properties. For instance, mapping band alignment and mobility into charts tailored for specific applications—such as LEDs (requiring type-I alignment and high radiative efficiency), photovoltaics (type-II with optimal offsets), or transistors (requiring low effective mass and high carrier mobility)—would significantly enhance the utility of AI-assisted material discovery.

References

[1] H. Wang *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, 2023.

[2] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, 2018.

[3] Z. Yao *et al.*, "Inverse design of nanoporous crystalline reticular materials with deep generative models," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 76–86, 2021.

[4] Z. Ren *et al.*, "An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties," *Matter*, vol. 5, no. 1, pp. 314–335, 2022.

[5] J. Wu *et al.*, "Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells," *Science*, vol. 386, no. 6727, pp. 1256–1264, 2024.

[6] X. Li, J. M. Hoffman, and M. G. Kanatzidis, "The 2d halide perovskite rulebook: How the spacer influences everything from the structure to optoelectronic device efficiency," *Chem Rev*, vol. 121, no. 4, pp. 2230–2291, 2021.

[7] Y. Wu *et al.*, "Universal machine learning aided synthesis approach of two-dimensional perovskites in a typical laboratory," *Nat Commun*, vol. 15, no. 1, p. 138, 2024.

[8] Z. Y. Lin *et al.*, "Design rules for two-dimensional organic semiconductor-incorporated perovskites (osip) gleaned from thousands of simulated structures," *Angew Chem Int Ed Engl*, vol. 62, no. 33, e202305298, 2023.

[9] C. Liu *et al.*, "Tunable semiconductors: Control over carrier states and excitations in layered hybrid organic-inorganic perovskites," *Phys Rev Lett*, vol. 121, no. 14, p. 146 401, 2018.

[10] E. O. Pyzer-Knapp *et al.*, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, no. 1, 2022.

[11] K. T. Butler *et al.*, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018.

[12] K. M. Jablonka *et al.*, "Using collective knowledge to assign oxidation states of metal cations in metal-organic frameworks," *Nat Chem*, vol. 13, no. 8, pp. 771–777, 2021.

[13] S. M. Moosavi *et al.*, "A data-science approach to predict the heat capacity of nanoporous materials," *Nat Mater*, vol. 21, no. 12, pp. 1419–1425, 2022.

[14] A. S. Larsen, T. Rekis, and A. Ø. Madsen, "Phai: A deep-learning approach to solve the crystallographic phase problem," *Science*, vol. 385, no. 6708, pp. 522–528, 2024.

[15] R. Gomez-Bombarelli *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," *Nat Mater*, vol. 15, no. 10, pp. 1120–7, 2016.

[16] Z. Rao *et al.*, "Machine learning–enabled high-entropy alloy discovery," *Science*, vol. 378, no. 6615, pp. 78–85, 2022.

[17] X. Jia *et al.*, "Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis," *Nature*, vol. 573, no. 7773, pp. 251–255, 2019.

[18] N. H. Angello *et al.*, "Closed-loop optimization of general reaction conditions for heteroaryl suzuki-miyaura coupling," *Science*, vol. 378, no. 6618, pp. 399–405, 2022.

[19] J. Bures and I. Larrosa, "Organic reaction mechanism classification using machine learning," *Nature*, vol. 613, no. 7945, pp. 689–695, 2023.

[20] N. I. Rinehart *et al.*, "A machine-learning tool to predict substrate-adaptive conditions for pd-catalyzed c–n couplings," *Science*, vol. 381, no. 6661, pp. 965–972, 2023.

[21] P. Raccuglia *et al.*, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, pp. 73–6, 2016.

[22] B. Huang and O. A. von Lilienfeld, "Quantum machine learning using atom-in-molecule-based fragments selected on the fly," *Nat Chem*, vol. 12, no. 10, pp. 945–951, 2020.

[23] G. Hautier *et al.*, "Finding nature's missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.

[24] R. Batra *et al.*, "Machine learning overcomes human bias in the discovery of self-assembling peptides," *Nat Chem*, vol. 14, no. 12, pp. 1427–1435, 2022.

[25] H. Zhao *et al.*, "A robotic platform for the synthesis of colloidal nanocrystals," *Nature Synthesis*, vol. 2, no. 6, pp. 505–514, 2023.

[26] S. Lu *et al.*, "Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning," *Nat Commun*, vol. 9, no. 1, p. 3405, 2018.

[27] N. T. P. Hartono *et al.*, "How machine learning can help select capping layers to suppress perovskite degradation," *Nat Commun*, vol. 11, no. 1, p. 4172, 2020.

[28] M. Zhong *et al.*, "Accelerated discovery of co(2) electrocatalysts using active machine learning," *Nature*, vol. 581, no. 7807, pp. 178–183, 2020.

[29] J. Xu *et al.*, "Anion optimization for bifunctional surface passivation in perovskite solar cells," *Nat Mater*, vol. 22, no. 12, pp. 1507–1514, 2023.

[30] J. A. Hueffel *et al.*, "Accelerated dinuclear palladium catalyst identification through unsupervised machine learning," *Science*, vol. 374, no. 6571, pp. 1134–1140, 2021.

[31] W. Sun *et al.*, "A map of the inorganic ternary metal nitrides," *Nat Mater*, vol. 18, no. 7, pp. 732–739, 2019.

[32] H. Li *et al.*, "Machine learning-accelerated discovery of heat-resistant polysulfates for electrostatic energy storage," *Nature Energy*, 2024.

[33] H. Lu *et al.*, "Machine learning-aided engineering of hydrolases for pet depolymerization," *Nature*, vol. 604, no. 7907, pp. 662–667, 2022.

[34] N. J. Szymanski *et al.*, "An autonomous laboratory for the accelerated synthesis of novel materials," *Nature*, vol. 624, no. 7990, pp. 86–91, 2023.

[35] C. K. Schissel *et al.*, "Deep learning to design nuclear-targeting abiotic miniproteins," *Nat Chem*, vol. 13, no. 10, pp. 992–1000, 2021.

[36] F. Wong *et al.*, "Discovery of a structural class of antibiotics with explainable deep learning," *Nature*, vol. 626, no. 7997, pp. 177–185, 2024.

[37] D. F. Nippa *et al.*, "Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning," *Nat Chem*, vol. 16, no. 2, pp. 239–248, 2024.

[38] F. Strieth-Kalthoff *et al.*, "Delocalized, asynchronous, closed-loop discovery of organic laser emitters," *Science*, vol. 384, no. 6697, eadk9227, 2024.

[39] A. Merchant *et al.*, "Scaling deep learning for materials discovery," *Nature*, vol. 624, no. 7990, pp. 80–85, 2023.

[40] B. A. Koscher *et al.*, "Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back," *Science*, vol. 382, no. 6677, eadi1407, 2023.

[41] C. Zeni *et al.*, "A generative model for inorganic materials design," *Nature*, 2025.

[42] T. Mou *et al.*, "Bridging the complexity gap in computational heterogeneous catalysis with machine learning," *Nature Catalysis*, vol. 6, no. 2, pp. 122–136, 2023.

[43] X. Li *et al.*, "Sequential closed-loop bayesian optimization as a guide for organic molecular metallophotocatalyst formulation discovery," *Nat Chem*, vol. 16, no. 8, pp. 1286–1294, 2024.

[44] P. Shetty *et al.*, "A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing," *NPJ Comput Mater*, vol. 9, no. 1, p. 52, 2023.

[45] Z. Pei *et al.*, "Toward the design of ultrahigh-entropy alloys via mining six million texts," *Nat Commun*, vol. 14, no. 1, p. 54, 2023.

[46] Z. Pei *et al.*, "Towards the holistic design of alloys with large language models," *Nature Reviews Materials*, vol. 9, no. 12, pp. 840–841, 2024.

[47] A. Zunger, "Inverse design in search of materials with target functionalities," *Nature Reviews Chemistry*, vol. 2, no. 4, p. 0121, 2018.

[48] H. Choubisa *et al.*, "Crystal site feature embedding enables exploration of large chemical spaces," *Matter*, vol. 3, no. 2, pp. 433–448, 2020.

[49] C. Bilodeau *et al.*, "Generative models for molecular discovery: Recent advances and challenges," *WIREs Computational Molecular Science*, vol. 12, no. 5, 2022.

[50] J. Westermayr *et al.*, "High-throughput property-driven generative design of functional organic molecules," *Nature Computational Science*, 2023.

[51] R. Gomez-Bombarelli *et al.*, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Cent Sci*, vol. 4, no. 2, pp. 268–276, 2018.

[52] J. Noh *et al.*, "Inverse design of solid-state materials via a continuous representation," *Matter*, vol. 1, no. 5, pp. 1370–1384, 2019.

[53] K. Kim *et al.*, "Deep-learning-based inverse design model for intelligent discovery of organic molecules," *npj Computational Materials*, vol. 4, no. 1, 2018.

[54] H. Xiao *et al.*, "An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning," *Nat Commun*, vol. 14, no. 1, p. 7027, 2023.

[55] H. Tsai *et al.*, "High-efficiency two-dimensional ruddlesden-popper perovskite solar cells," *Nature*, vol. 536, no. 7616, pp. 312–6, 2016.

[56] B. Saparov and D. B. Mitzi, "Organic-inorganic perovskites: Structural versatility for functional materials design," *Chem Rev*, vol. 116, no. 7, pp. 4558–96, 2016.

[57] C. C. Stoumpos *et al.*, "Ruddlesden–popper hybrid lead iodide perovskite 2d homologous semiconductors," *Chemistry of Materials*, vol. 28, no. 8, pp. 2852–2867, 2016.

[58] L. Mao *et al.*, "Tunable white-light emission in single-cation-templated three-layered 2d perovskites (ch(3)ch(2)nh(3))(4)pb(3)br(10-x)cl(x)," *J Am Chem Soc*, vol. 139, no. 34, pp. 11 956–11 963, 2017.

[59] Z. Chu *et al.*, "Blue light-emitting diodes based on quasi-two-dimensional perovskite with efficient charge injection and optimized phase distribution via an alkali metal salt," *Nature Electronics*, 2023.

[60] L. Mao *et al.*, "Hybrid dion-jacobson 2d lead iodide perovskites," *J Am Chem Soc*, vol. 140, no. 10, pp. 3775–3783, 2018.

[61] S. Ahmad *et al.*, "Dion-jacobson phase 2d layered perovskites for solar cells with ultrahigh stability," *Joule*, vol. 3, no. 3, pp. 794–806, 2019.

[62] C. M. M. Soe *et al.*, "New type of 2d perovskites with alternating cations in the interlayer space, (c(nh(2))(3))(ch(3)nh(3))(n)pb(n)i(3n+1): Structure, properties, and photovoltaic performance," *J Am Chem Soc*, vol. 139, no. 45, pp. 16 297–16 309, 2017.

[63] L. Mao *et al.*, "Layered hybrid lead iodide perovskites with short interlayer distances," *ACS Energy Letters*, vol. 7, no. 8, pp. 2801–2806, 2022.

[64] T. Luo *et al.*, "Compositional control in 2d perovskites with alternating cations in the interlayer space for photovoltaics with efficiency over 18," *Adv Mater*, vol. 31, no. 44, e1903848, 2019.

[65] Y. Zhang *et al.*, "Dynamical transformation of two-dimensional perovskites with alternating cations in the interlayer space for high-performance photovoltaics," *J Am Chem Soc*, vol. 141, no. 6, pp. 2684–2694, 2019.

[66] J. C. Blancon *et al.*, "Semiconductor physics of organic-inorganic 2d halide perovskites," *Nat Nanotechnol*, vol. 15, no. 12, pp. 969–985, 2020.

[67] T. He *et al.*, "Reduced-dimensional perovskite photovoltaics with homogeneous energy landscape," *Nat Commun*, vol. 11, no. 1, p. 1672, 2020.

[68] M. Yuan *et al.*, "Perovskite energy funnels for efficient light-emitting diodes," *Nat Nanotechnol*, vol. 11, no. 10, pp. 872–877, 2016.

[69] B. Traore *et al.*, "Composite nature of layered hybrid perovskites: Assessment on quantum and dielectric confinements and band alignment," *ACS Nano*, vol. 12, no. 4, pp. 3321–3332, 2018.

[70] L. Pedesseau *et al.*, "Advances and promises of layered halide hybrid perovskite semiconductors," *ACS Nano*, vol. 10, no. 11, pp. 9776–9786, 2016.

[71] J. Sun *et al.*, "Emerging two-dimensional organic semiconductor-incorporated perovskites horizontal line a fascinating family of hybrid electronic materials," *J Am Chem Soc*, vol. 145, no. 38, pp. 20 694–20 715, 2023.

[72] Y. Gao *et al.*, "Molecular engineering of organic-inorganic hybrid perovskites quantum wells," *Nat Chem*, vol. 11, no. 12, pp. 1151–1157, 2019.

[73] E. Shi *et al.*, "Two-dimensional halide perovskite nanomaterials and heterostructures," *Chem Soc Rev*, vol. 47, no. 16, pp. 6046–6072, 2018.

[74] J. Y. Park *et al.*, "Thickness control of organic semiconductor-incorporated perovskites," *Nat Chem*, 2023.

[75] D. H. Cao *et al.*, "2d homologous perovskites as light-absorbing materials for solar cell applications," *J Am Chem Soc*, vol. 137, no. 24, pp. 7843–50, 2015.

[76] E. R. Dohner, E. T. Hoke, and H. I. Karunadasa, "Self-assembly of broadband white-light emitters," *J Am Chem Soc*, vol. 136, no. 5, pp. 1718–21, 2014.

[77] P. Cheng *et al.*, "Highly efficient ruddlesden–popper halide perovskite pa2ma4pb5i16 solar cells," *ACS Energy Letters*, vol. 3, no. 8, pp. 1975–1982, 2018.

[78] I. C. Smith *et al.*, "A layered hybrid perovskite solar-cell absorber with enhanced moisture stability," *Angew Chem Int Ed Engl*, vol. 53, no. 42, pp. 11 232–5, 2014.

[79] X. Li *et al.*, "Two-dimensional dion-jacobson hybrid lead iodide perovskites with aromatic diammonium cations," *J Am Chem Soc*, vol. 141, no. 32, pp. 12 880–12 890, 2019.

[80] J. V. Passarelli *et al.*, "Enhanced out-of-plane conductivity and photovoltaic performance in n = 1 layered perovskites through organic cation design," *J Am Chem Soc*, vol. 140, no. 23, pp. 7313–7323, 2018.

[81] J. W. Lee *et al.*, "Rethinking the a cation in halide perovskites," *Science*, vol. 375, no. 6583, eabj1186, 2022.

[82] D. G. Billing and A. Lemmerer, "Synthesis, characterization and phase transitions of the inorganic–organic layered perovskite-type hybrids [(cnh2n+1nh3)2pbi4] (n = 12, 14, 16 and 18)," *New Journal of Chemistry*, vol. 32, no. 10, 2008.

[83] D. G. Billing and A. Lemmerer, "Synthesis, characterization and phase transitions in the inorganic-organic layered perovskite-type hybrids [(cnh2n+1nh3)2pbi4], n = 4, 5 and 6," *Acta Crystallogr B*, vol. 63, no. Pt 5, pp. 735–47, 2007.

[84] A. Lemmerer and D. G. Billing, "Synthesis, characterization and phase transitions of the inorganic-organic layered perovskite-type hybrids [(c(n)h(2n+1)nh3)2pbi4], n = 7, 8, 9 and 10," *Dalton Trans*, vol. 41, no. 4, pp. 1146–57, 2012.

[85] D. B. Mitzi, "Templating and structural engineering in organic–inorganic perovskites," *Journal of the Chemical Society, Dalton Transactions*, no. 1, pp. 1–12, 2001.

[86] D. G. Billing and A. Lemmerer, "Inorganic–organic hybrid materials incorporating primary cyclic ammonium cations: The lead iodide series," *CrystEngComm*, vol. 9, no. 3, pp. 236–244, 2007.

[87] ——, "Inorganic–organic hybrid materials incorporating primary cyclic ammonium cations: The lead bromide and chloride series," *CrystEngComm*, vol. 11, no. 8, 2009.

[88] L. Mao *et al.*, "Structural diversity in white-light-emitting hybrid lead bromide perovskites," *J Am Chem Soc*, vol. 140, no. 40, pp. 13 078–13 088, 2018.

[89] M. E. Kamminga *et al.*, "Confinement effects in low-dimensional lead iodide perovskite hybrids," *Chemistry of Materials*, vol. 28, no. 13, pp. 4554–4562, 2016.

[90] P. Fu *et al.*, "Short aromatic diammonium ions modulate distortions in 2d lead bromide perovskites for tunable white-light emission," *Chemistry of Materials*, vol. 34, no. 21, pp. 9685–9698, 2022.

[91] A. Ducinskas *et al.*, "The role of alkyl chain length and halide counter ion in layered dion-jacobson perovskites with aromatic spacers," *J Phys Chem Lett*, vol. 12, no. 42, pp. 10 325–10 332, 2021.

[92] J. Shi *et al.*, "Fluorinated low-dimensional ruddlesden-popper perovskite solar cells with over 17% power conversion efficiency and improved stability," *Adv Mater*, vol. 31, no. 37, e1901673, 2019.

[93] C. Ma *et al.*, "2d perovskites with short interlayer distance for high-performance solar cell application," *Adv Mater*, vol. 30, no. 22, e1800710, 2018.

[94] Q. Li *et al.*, "Fluorinated aromatic formamidinium spacers boost efficiency of layered ruddlesden–popper perovskite solar cells," *ACS Energy Letters*, vol. 6, no. 6, pp. 2072–2080, 2021.

[95] Z. Xu *et al.*, "Highly efficient and stable dion-jacobson perovskite solar cells enabled by extended pi-conjugation of organic spacer," *Adv Mater*, vol. 33, no. 51, e2105083, 2021.

[96] J. Xue *et al.*, "Reconfiguring the band-edge states of photovoltaic perovskites by conjugated organic cations," *Science*, vol. 371, no. 6529, pp. 636–640, 2021.

[97] W. A. Dunlap-Shohl *et al.*, "Tunable internal quantum well alignment in rationally designed oligomer-based perovskite films deposited by resonant infrared matrix-assisted pulsed laser evaporation," *Materials Horizons*, vol. 6, no. 8, pp. 1707–1716, 2019.

[98] Y. Wu *et al.*, "Accelerated design of promising mixed lead-free double halide organic-inorganic perovskites for photovoltaics using machine learning," *Nanoscale*, vol. 13, no. 28, pp. 12 250–12 259, 2021.

[99] G. H. Gu *et al.*, "Perovskite synthesizability using graph neural networks," *npj Computational Materials*, vol. 8, no. 1, p. 71, 2022.

[100] R. Lyu *et al.*, "Predictive design model for low-dimensional organic-inorganic halide perovskites assisted by machine learning," *J Am Chem Soc*, vol. 143, no. 32, pp. 12 766–12 776, 2021.

[101] C. Zhi *et al.*, "Machine-learning-assisted screening of interface passivation materials for perovskite solar cells," *ACS Energy Letters*, vol. 8, no. 3, pp. 1424–1433, 2023.

[102] Q. Zhang *et al.*, "Machine-learning-assisted design of buried-interface engineering materials for high-efficiency and stable perovskite solar cells," *ACS Energy Letters*, vol. 9, no. 12, pp. 5924–5934, 2024.

[103] Y. Wu *et al.*, "Two-dimensional perovskites with tunable room-temperature phosphorescence," *Advanced Functional Materials*, vol. 32, no. 39, p. 2 204 579, 2022.

[104] M. Bhatt, P. K. Nayak, and D. Ghosh, "Data-driven design of electroactive spacer molecules to tune charge carrier dynamics in layered halide perovskite heterostructures," *ACS Nano*, vol. 18, no. 35, pp. 24 484–24 494, 2024.

[105] E. Mahal *et al.*, "Machine learning-driven prediction of band-alignment types in 2d hybrid perovskites," *Journal of Materials Chemistry A*, vol. 11, no. 43, pp. 23 547–23 555, 2023.

[106] K. Wang *et al.*, "Two-dimensional-lattice-confined single-molecule-like aggregates," *Nature*, vol. 633, no. 8030, pp. 567–574, 2024.

[107] J. Moon *et al.*, "Active learning guides discovery of a champion four-metal perovskite oxide for oxygen evolution electrocatalysis," *Nat Mater*, vol. 23, no. 1, pp. 108–115, 2024.

[108] A. M. Virshup *et al.*, "Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds," *J Am Chem Soc*, vol. 135, no. 19, pp. 7296–303, 2013.

[109] Q. Dai *et al.*, "Evolution of the nature of excitons and electronic couplings in hybrid 2d perovskites as a function of organic cation [U+2010]conjugation," *Advanced Functional Materials*, vol. 32, no. 10, 2021.

[110] M. K. Jana *et al.*, "Resolving rotational stacking disorder and electronic level alignment in a 2d oligothiophene-based lead iodide perovskite," *Chemistry of Materials*, vol. 31, no. 20, pp. 8523–8532, 2019.

[111] L. Gao *et al.*, "M-phenylenediammonium as a new spacer for dion-jacobson two-dimensional perovskites," *J Am Chem Soc*, vol. 143, no. 31, pp. 12 063–12 073, 2021.

[112] Y. Li *et al.*, "Bifunctional organic spacers for formamidinium-based hybrid dion-jacobson two-dimensional perovskite solar cells," *Nano Lett*, vol. 19, no. 1, pp. 150–157, 2019.

[113] M. Almalki *et al.*, "Nanosegregation in arene-perfluoroarene pi-systems for hybrid layered dion-jacobson perovskites," *Nanoscale*, vol. 14, no. 18, pp. 6771–6776, 2022.

[114] R. Zhao *et al.*, "Rigid conjugated diamine templates for stable dion-jacobson-type two-dimensional perovskites," *J Am Chem Soc*, vol. 143, no. 47, pp. 19 901–19 908, 2021.

[115] C. Kunkel *et al.*, "Active discovery of organic semiconductors," *Nat Commun*, vol. 12, no. 1, p. 2422, 2021.

[116] H. Bronstein *et al.*, "The role of chemical design in the performance of organic semiconductors," *Nature Review Chemistry*, vol. 4, no. 2, pp. 66–77, 2020.

[117] H. Zhang *et al.*, "Modulating the dipole moment of secondary ammonium spacers for efficient 2d ruddlesden-popper perovskite solar cells," *Angew Chem Int Ed Engl*, vol. 63, no. 7, e202318206, 2024.

[118] A. Zunger, "Beware of plausible predictions of fantasy materials," *Nature*, 2019.

[119] A. K. Cheetham and R. Seshadri, "Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery," *Chem Mater*, vol. 36, no. 8, pp. 3490–3495, 2024.

[120] J. Jang *et al.*, "Structure-based synthesizability prediction of crystals using partially supervised learning," *J Am Chem Soc*, vol. 142, no. 44, pp. 18 836–18 843, 2020.

[121] Z. Luo *et al.*, "Side-chain engineering of nonfullerene small-molecule acceptors for organic solar cells," *Energy Environmental Science*, vol. 16, no. 7, pp. 2732–2758, 2023.

[122] C. Wang *et al.*, "Semiconducting pi-conjugated systems in field-effect transistors: A material odyssey of organic electronics," *Chem Rev*, vol. 112, no. 4, pp. 2208–67, 2012.

[123] H. Jiang and W. Hu, "The emergence of organic single-crystal electronics," *Angew Chem Int Ed Engl*, vol. 59, no. 4, pp. 1408–1428, 2020.