

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

An Autonomous Platform for Reticular Materials Discovery

Permalink

<https://escholarship.org/uc/item/2tk455m3>

ISBN

9798189270529

Author

Rong, Zichao

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

An Autonomous Platform for Reticular Materials Discovery

By

Zichao Rong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Omar M. Yaghi, Chair

Professor Teresa Head-Gordon

Professor Alexis T. Bell

Spring 2026

An Autonomous Platform for Reticular Materials Discovery

Copyright © 2026

By

Zichao Rong

Abstract

An Autonomous Platform for Reticular Materials Discovery

By

Zichao Rong

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Omar M. Yaghi, Chair

Reticular materials, including metal–organic frameworks (MOFs) and covalent organic frameworks (COFs), are porous crystalline structures constructed from molecular building blocks. Their structures can be rationally designed for a wide range of applications through control over building block composition, geometry, and functionality. However, the crystallization of reticular materials remains challenging. Strong chemical bonding between building units and the formation of competing phases result in a narrow crystallization window that requires precise control over reaction parameters. In the absence of a reliable theoretical framework for predicting crystallization outcomes, synthesis remains largely trial-and-error and dependent on human intuition. To address this limitation, this dissertation introduces an autonomous platform, termed Algorithmic Iterative Reticular Synthesis (AIRES), which integrates high-throughput experimentation and machine learning to identify correlations between successful crystallization outcomes, building block properties, and reaction parameters and therefore to accelerate the discovery of new reticular materials.

Chapter 1 introduces the conceptual framework of autonomous platforms for chemistry and materials science, with emphasis on the gradual transition from human-driven experimentation to data-driven, closed-loop discovery systems. The challenges associated with reticular materials discovery, including vast chemical design space and limited predictability in synthesis, are also highlighted. These characteristics make reticular materials discovery well-suited for autonomous approaches.

Chapter 2 describes the experimental platform of AIRES, including automated high-throughput synthesis, optical microscopy assisted by computer vision for rapid screening of crystallization outcomes, and single crystal X-ray diffraction (SCXRD) for definitive structural determination. Zeolitic imidazolate frameworks (ZIFs) are employed as a model system due to their structural complexity and

sensitivity to reaction conditions. The integration of scalable synthesis with rigorous structural verification enables systematic exploration of crystallization space.

Chapter 3 presents the construction of a crystallization dataset and the use of machine learning models for predicting crystallization outcomes. Chemical information is encoded through a combination of structural descriptors, physicochemical properties, and interaction-based features, enabling machine learning models to capture relationships between reaction input and experimental outcomes. Model performance is evaluated under both interpolation and extrapolation settings, revealing the challenges of generalization across chemical systems while also demonstrating the model's ability to learn meaningful relationships and provide predictive guidance in unexplored regions of chemical space.

Chapter 4 introduces an algorithmic decision-making framework tailored to discovery-oriented objectives, where the goal is to maximize the number of distinct discoveries. A probabilistic formulation of crystallization outcomes is employed, and candidate experiments are prioritized based on predicted success likelihood using greedy acquisition strategies. Batch selection methods are developed to accommodate high-throughput experimentation, and the framework is evaluated on 48 previously unexplored chemical systems, demonstrating approximately a twofold improvement in discovery efficiency compared to random search.

Chapter 5 presents the structure determination and topological analysis of newly discovered crystals. SCXRD-verified structures are analyzed in terms of underlying nets and natural tilings, providing insight into how variations in linker structure influence framework topology. Both ZIF and other crystalline products are examined, highlighting the diversity of structures accessible through autonomous exploration.

Chapter 6 serves as an outlook on the future of autonomous materials discovery, emphasizing the integration of generative modeling, advanced characterization, and scalable experimental platforms toward a fully autonomous discovery cycle.

Table of Contents

Table of Contents	i
Acknowledgments	iii
Chapter 1. Introduction to Autonomous Platforms and Reticular Materials Discovery	1
1.1 Evolution of Autonomous Platforms.....	1
1.1.1 Introduction.....	1
1.1.2 Combinatorial and High-Throughput Experimentation	2
1.1.3 Laboratory Automation	3
1.1.4 Computational Screening	3
1.1.5 Algorithmic Guidance	4
1.1.6 Fully Autonomous Platforms	5
1.2 Reticular Materials Discovery	6
1.2.1 Reticular Synthesis.....	6
1.2.2 Crystallization Challenges	7
1.3 References	8
Chapter 2 Experiment Platform for Reticular Materials Discovery	13
2.1 Introduction.....	13
2.2 Zeolitic Imidazolate Frameworks.....	15
2.3 Automated and High-throughput Synthesis	15
2.3 Automated Optical Microscopy and Computer Vision	17
2.4 Single Crystal X-ray Diffraction	22
2.5 References	24
Chapter 3. Machine Learning for Crystallization Outcome Classification	27
3.1 Introduction.....	27
3.2 ZIF Crystallization Dataset	28
3.3 Reaction Encoding	34
3.4 Predictive Performance of Random Forest Classifier	38
3.5 References	42
Chapter 4. Algorithmic Decision-Making for Accelerating Reticular Materials Discovery	45
4.1 Introduction.....	45
4.2 Acquisition Function and Batch Formalism	46

4.3	Discovery Performance	48
4.4	Analysis on Predictions and Feature Selections.....	54
4.5	References	68
Chapter 5. Structure Determination and Topological Analysis of ZIF Discovery		70
5.1	Introduction.....	70
5.2	Crystal Structures of ZIFs and LZIFs	71
5.3	Topologies of ZIFs and LZIFs	102
5.4	References	108
Chapter 6. Outlook.....		111

Acknowledgments

Time has flown since I first arrived at Berkeley in 2021, and I am now nearing the completion of my Ph.D. This has been a truly meaningful journey in my life, not only because of the people I have met, but also the work we have done together and the moments we have shared. In this section, I would like to express my sincere gratitude to the many individuals who have been part of this journey.

First, it is my greatest honor to have Prof. Omar M. Yaghi as my Ph.D. advisor. He is a pioneering scientist in the field of reticular chemistry, constantly asking important and forward-looking research questions. He holds exceptionally high standards for scientific rigor, writing, and the clear communication of results to a broad audience. These are the qualities I have learned from him and will carry with me throughout my career. In addition, Prof. Yaghi is an outstanding mentor. He continuously encourages me to strive for excellence, connects me with reliable collaborators, and fosters a friendly yet professional environment within the group. As a student who initiated a new interdisciplinary direction combining reticular chemistry and machine learning in the group, I am especially grateful for his strong support and the flexibility he has given me to pursue my research interests.

I am also deeply grateful to Prof. Joachim Sauer for mentoring me in computational chemistry. While the Yaghi group is primarily focused on experimental work, my interaction with Prof. Sauer provided me with a new perspective on rigorous computational modeling and its ability to reveal atomic-level insights and explain experimental phenomena. I deeply appreciate his guidance, as well as the time we spent investigating CO₂ chemisorption and developing new understanding in this area.

I would next like to thank my colleagues in the Yaghi group for their support, collaboration, and friendship throughout my Ph.D. journey. I am especially grateful to Dr. Xiaokun Pei for her mentorship in X-ray crystallography during my first semester, which had a lasting impact on my subsequent work, and to Dr. Zhiling (Zack) Zheng for his enthusiasm and involvement in multiple projects. I would also like to thank Dr. Haozhe Li, Dr. Kaiyu Wang, Dr. Zihui Zhou, and Dr. Ali H. Alawadhi, who joined the group in the same year as I did, for sharing this journey and providing companionship over the past five years. I am grateful to Kevin Greco and Abdullah A. Alghamdi, whom I had the opportunity to mentor, for their dedication and valuable contributions.

I further acknowledge the many members of the Yaghi group whose interactions have enriched my experience, including Dr. Zhe Ji, Dr. Xing Han, Dr. Tianqiong Ma, Dr. Ha Lac Nguyen, Dr. Woochul Song, Dr. Cornelius Gropp, Dr. Hao Lyu, Dr. Wentao Xu, Dr. Daria Kurandina, Dr. Chuanshuai Li, Dr. Magdalena Heiland, Dr. Nikita Hanikel, Dr. Haoze Wang, Dr. Fangying Jin, Dr. Lingxiang Hou, Dr. Saber Mirzaei, Dr. Meiyang Gao, Dr. Nakul Rampal, Dr. Saumil Chheda, Dr. S. Ephraim Neumann, Dr. Oscar Iu-Fan Chen, Dr. Théo Jaffrelot Inizan, Dr. Vivek Singh, Dr. Yuang Shi, Dr. Chengbin Zhao, Dr. Francesco Tavani, Dr. Xu Qiu, Dr. Matthew Burigana, Dr. Heyang Zhang, Kelvin Kam-Yun Li, Yen-hsu Lin, Zakir H. Talpur, Xinyu Chen, Zhiguo He, Jian Yin, Daehyun Daniel Ahn, Neda Sabeva, and Jackson Thomassian.

I am also grateful to the many collaborators outside the Yaghi group with whom I have had the privilege to work closely. I thank Dr. Zihao Chen, Dr. Felix Luong, and Prof. Mario Boley for their sustained efforts on our collaborative project and for the insightful discussions on machine learning during our weekly meetings. I also acknowledge our collaborators from VinUniversity, including Prof. Laurent El Ghaoui, Prof. Tung Nguyen-Dang, Prof. H. Hieu Pham, Prof. Dung D. Le, Prof. Viet Bac T. Phung, H. T. Nhan Luong, K. Huyen Bui, and Van Hao Vu, for their close collaboration on crystal recognition. Working across disciplines has been a particularly rewarding experience, allowing us to learn each other's "languages" and build new ideas together. I have also benefited from valuable support and insightful discussions from Prof. Stefan Wuttke, Prof. Davide M. Proserpio, Prof. Jennifer T. Chayes, Prof. Christian Borgs, Prof. Laura Gagliardi, Dr. Hilal Daglar, and Henning Windeck.

I would like to express my gratitude to the many faculty and staff members at University of California, Berkeley and Lawrence Berkeley National Lab. My qualifying exam committee members, Prof. Peidong Yang, Prof. Teresa Head-Gordon, Prof. Ke Xu, and Prof. Alexis T. Bell, provided valuable insights that helped shape the early direction of this dissertation. Dr. Kathleen Durkin and Dr. Dave Small from MGCF offered essential technical support in utilizing high-performance computing resources. Dr. Simon Teat provided invaluable assistance during our beamtime at the Advanced Light Source.

Beyond the academic environment, I am deeply grateful to the people in my life whose love, support, and encouragement have sustained me throughout this journey. I would first like to thank my parents, Dexiang Rong and Qin Zhu. Despite the distance, they have always been present in my life through our weekly video calls, constantly caring about how I am doing in Berkeley and encouraging me along the way. At the same time, they always reassured me that they were doing well in Shanghai and asked me not to worry about them, so that I could fully focus on my research. Their quiet strength, understanding, and selfless support have given me the confidence and peace of mind to pursue this path.

I would also like to remember my grandparents, Yushu Rong, Yulan Duan, Yongxiang Zhu, and Hongdi Zhu, who passed away before and during my Ph.D. When they were still with us, they always showed the deepest care for my life and well-being. I was especially touched that they would even watch television news to learn more about the United States, simply because their grandson was here. It was very difficult for me not to be able to attend their funerals due to travel restrictions, and this remains one of my greatest regrets. I want to tell them that I am doing well. I have grown, made progress in my work, published my research, and am now about to receive my Ph.D. I hope they would be proud of me.

I am deeply grateful to my wife, Qing (Runa) Pan, for her love, patience, and support throughout this journey. She has always given me confidence, showing understanding toward my weaknesses while celebrating my achievements with genuine joy. Her support has been present in both the defining moments and the everyday ones, guiding me through setbacks and uncertainty, while bringing warmth and stability to this journey. She has offered clarity when I felt lost, steadiness when I was overwhelmed, and perspective when I became too absorbed in my work. I have also learned so much from her: her dedication, her courage in facing uncertainty, and her warmth and kindness toward others. These qualities have shaped not only how I approach my research with greater perseverance and purpose, but also how I live my life and interact with the people around me. More than anything, her presence

has brought balance and meaning to my life, and none of this would have been possible without her. I would also like to thank my parents-in-law, Tianlin Pan and Suyu Ding, for their trust in us and for their continued support.

I have been especially grateful for the opportunity to reconnect with old friends while in Berkeley. My high school friends, Jiahao Huang, Zhenyi Zhang, Haoran Song, Zhengjie Xiong, Zhuoyang Ye, Yihan Lai, and Liyang Zhao, and my undergraduate friends, Fan Feng, Hanna Wang, Zongyang Peng, and Jianchun Wen, brought me great joy through their visits, during which we explored the Bay Area together and shared many memorable moments. I am also thankful for the new friendships I formed at UC Village, with university staff, Tim Cumagun, Analee Scott, and Victoria Bryant, and with student residents and their families, João Pedro Rangel Gomes da Silva, Camila Estela Cassis Augusto, Veronarindra Ramananjato, Finaritra Randimbarison, Rafael Meza Duriez, Gladys Prado Castillo, Kiyotaka Yageta, Ayano Yageta, Yuji, Mizuki, Yu-Wen Chen, Debbie Hsueh, Miryam, Nicholas, Wenyu Yang, Xiaoyi Du, Coco Papoi, Melissa, Behrooz Ebrazi, Marisol Rubio, Joel, Toti Nugmanova, Marat, Qingqing Zheng, Rocío Díaz, Jidong Chen, Sakina, Chenxia Zhang, and others, which added warmth and a sense of community to my life here.

Chapter 1

Introduction to Autonomous Platforms and Reticular Materials Discovery

1.1 Evolution of Autonomous Platforms

1.1.1 Introduction

Scientific research has historically followed a human-centered paradigm consisting of three essential stages: formulation of a research question, proposal of candidate solutions, and experimental or theoretical validation. In chemistry and materials science, these solutions typically take the form of new substances, improved strategies for utilizing known materials, or new theoretical models that explain observed phenomena. Traditionally, human researchers play a central role at every stage of this process. Research questions are formulated based on intellectual curiosity or societal needs, such as energy, health, and environment. Guided by prior knowledge and experience, researchers then define a feasible space of candidate solutions. Finally, experiments are designed and conducted to evaluate these candidates, often through iterative refinement informed by the outcomes. While this paradigm has driven centuries of scientific progress, it is inherently constrained. The process is labor-intensive, low-throughput, and often biased by human intuition and incomplete knowledge.

Recent advances in hardware, computational methods, and artificial intelligence (AI) are transforming this traditional research workflow¹. High-throughput experimental platforms enable rapid and parallelized exploration of candidates², while laboratory automation reduces human intervention³. In parallel, developments in data science and machine learning (ML) allow for the analysis of high-dimensional datasets, extraction of hidden patterns, and prediction of experimental outcomes⁴. More recently, generative models and large language models (LLMs) have further expanded the role of algorithms in scientific discovery, demonstrating the ability to generate synthesizable targets⁵, explore complex chemical spaces⁶, and assist in optimizing synthesis by integrating literature knowledge with experimental data⁷.

These developments have given rise to a new paradigm: autonomous or self-driving laboratories, in which experimentation, data analysis, and decision-making are integrated into a closed-loop system⁸. In such systems, algorithms can iteratively propose experiments, execute them through robotic platforms, analyze the results, and refine subsequent actions with minimal human intervention. This paradigm has been increasingly explored across chemistry and materials science. In this section, an overview of the evolution from traditional, human-driven research toward increasingly autonomous scientific workflows is presented. This evolution has experienced several stages: combinatorial and high-

throughput experimentation, computational screening, laboratory automation, algorithmic guidance, and fully autonomous platforms. For each stage, the enabling technologies, their impact on research efficiency and discovery, and representative achievements are discussed.

1.1.2 Combinatorial and High-Throughput Experimentation

A central challenge in chemistry and materials science is the vastness of the accessible chemical space. Even when restricting consideration to experimentally feasible systems, the number of possible combinations of compositions and spatial arrangements grows exponentially⁹. Moreover, the relevant chemical space extends far beyond structural variables. Experimental conditions, such as temperature, pressure, concentration, solvent media, and reaction time, also play a critical role in determining outcomes. These variables introduce additional, often continuous dimensions to the search space, significantly increasing its complexity. As a result, the effective discovery space is not only large but also high-dimensional and frequently nonlinear. This challenge is particularly pronounced in areas where predictive theoretical frameworks remain incomplete or unavailable. In such contexts, rational design alone is often insufficient, and systematic exploration through combinatorial experimentation becomes a necessary strategy. Beyond identifying target systems of interest, combinatorial approaches also generate comprehensive datasets that can be leveraged to extract underlying trends and mechanistic insights when coupled with modern data analysis tools¹⁰.

Combinatorial experimentation can be viewed as a form of brute-force exploration¹¹. By systematically sampling across multiple variables, through grid-based designs or advanced sampling techniques, it enables the identification of regions of interest that may not be anticipated by human intuition, thereby reducing bias introduced by prior assumptions. However, such approaches are inherently data-intensive and require efficient experimental and analytical methods to be practically feasible. High-throughput experimentation (HTE) provides the practical means to implement combinatorial strategies at scale¹². By leveraging parallelization, HTE enables the preparation and evaluation of large numbers of samples in a time-efficient manner. A typical HTE workflow relies on array-based formats, such as multi-well plates or substrate libraries, in which each position corresponds to a distinct combination of substances and experimental conditions. This design allows hundreds or even thousands of experiments to be conducted in parallel. Equally important is the development of rapid and scalable characterization techniques capable of providing high-content readouts across large sample sets¹³. Techniques such as optical microscopy, fluorescence spectroscopy, and UV-vis spectroscopy are commonly employed to rapidly assess sample properties and outcomes. These methods enable efficient identification of successful candidates within large combinatorial libraries. The impact of high-throughput experimentation has been widely demonstrated across multiple domains, including drug discovery¹⁴, thin-film materials^{15,16}, heterogeneous catalysis¹⁷, and polymer synthesis^{18–20}, where it has significantly accelerated the exploration of complex parameter spaces and the identification of desired targets.

1.1.3 Laboratory Automation

Although high-throughput experimentation is often associated with parallelized, array-based workflows, this format is not universally applicable. In many domains, the physical nature of the experiment or the constraints of instrumentation make strict parallelization difficult or impossible. For example, some characterization tools, such as liquid chromatography–mass spectrometry, nuclear magnetic resonance (NMR), gas chromatography, and many X-ray-based measurements, are serialized because each sample must be loaded, measured, and cleared before the next one can be analyzed. In such cases, automation increases throughput not by parallelizing every operation, but by removing the need for constant human supervision. Once instruments and handling routines are scripted, systems can execute repetitive tasks continuously, day and night, with high positional and procedural consistency.

Modern laboratory automation has evolved to encompass a wide range of operations. At the hardware level, automated systems can perform reagent handling, reaction execution, sample transfer, and characterization²¹. Liquid dispensing is commonly achieved through robotic pipetting systems, enabling precise and reproducible control²². In contrast, handling solid reagents remains more challenging due to the need for real-time mass measurement and variability in physical properties, often requiring specialized dispensing systems²³. Reaction environments, including temperature, pressure, and atmosphere, can be programmatically controlled using automated reactors and heating systems²⁴. Automation also extends to sample transfer and workflow integration across multiple instruments. Experimental pipelines may consist of modular stations for synthesis, processing, and analysis, connected through robotic arms, gantry systems, or mobile robotic platforms²⁵. These systems enable the physical movement of samples between workstations, effectively replicating the actions of a human operator while maintaining consistency and traceability. In parallel, many analytical instruments already support partial automation through autosamplers and scripted measurement routines, allowing for automated data acquisition once samples are loaded.

Beyond hardware, a critical component of modern laboratory automation is the software layer responsible for orchestrating experimental workflows^{26–28}. Early forms of automation primarily focused on individual instruments operating in isolation. In contrast, recent advances have enabled the coordination of multiple instruments into integrated, end-to-end experimental pipelines. Experimental protocols are translated into machine-executable workflows that define the sequence of operations, dependencies between tasks, and data flow across the system. More advanced implementations integrate scheduling algorithms that optimize instrument usage, minimize idle time, and manage dependencies between parallel and sequential tasks.

1.1.4 Computational Screening

The vastness of chemical space presents a fundamental challenge for experimental exploration alone. Computational screening provides a powerful strategy to reduce the size of the candidate space by evaluating large numbers of hypothetical systems prior to experimental validation. By filtering out unlikely or low-performing candidates, computational methods enable researchers to focus

experimental efforts on a much smaller and more promising subset of systems, thereby significantly improving the efficiency of discovery. The specific computational approaches used for screening vary across domains, depending on the underlying physical principles and target properties. In materials science, computational screening often relies on quantum mechanical calculations to evaluate structural stability and functional properties^{29–32}. In the context of drug discovery, computational screening typically involves virtual screening strategies, including molecular docking, in which large libraries of candidate molecules are evaluated for their binding affinity to biological targets, as well as ligand-based and data-driven approaches^{33,34}.

A key advantage of computational screening is that it is performed *in silico*, allowing for rapid and cost-effective evaluation without the need for immediate experimental resources. Large numbers of candidates can be assessed in parallel using computational infrastructure, enabling the exploration of chemical spaces that would be infeasible to probe experimentally. Despite its advantages, computational screening must be interpreted with caution and cannot replace experimental validation. All computational methods rely on approximations and assumptions, whether in the form of simplified physical models, limited basis sets, or imperfect training data in machine learning approaches. In many cases, structural models used for screening do not fully capture real experimental conditions, such as defects, disorder, solvent effects, or kinetic constraints. As a result, predictions of stability or performance may not directly translate to experimentally realizable systems. Consequently, computational screening is most effective when used as a complementary tool, guiding experiments rather than replacing them.

1.1.5 Algorithmic Guidance

Scientific research is often iterative: outcomes are analyzed, and subsequent experiments are designed based on accumulated knowledge. Traditionally, this process has been guided by human intuition, experience, and domain expertise. While effective, such approaches are often limited in their ability to systematically explore high-dimensional parameter spaces and fully leverage the growing volume of experimental data. Algorithmic guidance offers an alternative and complementary framework for directing this iterative process. At its core, this approach consists of two key components: models that extract patterns from data and decision-making strategies that use these models to perform subsequent actions. By learning relationships between inputs and outcomes, models can uncover complex patterns or correlations that may not be readily accessible through human reasoning alone. These learned representations can then be combined with explicitly defined objective functions, constraints, and selection strategies to enable a more transparent and reproducible approach to experiment planning.

A wide range of models has been developed, broadly falling under the umbrella of statistical learning and machine learning^{35,36}. Statistical learning emphasizes interpretability and uncertainty quantification, while machine learning more broadly focuses on flexible function approximation and predictive performance. Importantly, these categories are not mutually exclusive, and many methods, such as Gaussian processes, can be viewed within both frameworks depending on their application.

Common approaches include regression models, Gaussian processes, and ensemble methods such as random forests and gradient boosting, which can capture nonlinear relationships between input variables and outcomes. These models are often used as surrogate models to approximate complex systems, particularly in settings with limited or moderately sized datasets. Deep learning, a subset of machine learning based on multi-layer neural networks, has further expanded the capabilities of data-driven modeling³⁷. By learning hierarchical representations from large datasets, deep learning methods have demonstrated strong performance in tasks such as molecular property prediction, reaction outcome prediction, and generative design of chemical structures^{38–40}.

The choice of decision-making strategy depends on the objective of the study. In many cases, the goal is single-objective optimization, such as maximizing reaction yield, catalytic activity, or material performance. For example, Bayesian optimization has been widely used to optimize reaction conditions in chemical synthesis, where it iteratively proposes new experiments by balancing exploration of uncertain regions with exploitation of known high-performing conditions⁴¹. In more complex scenarios, multiple objectives must be considered simultaneously, leading to multi-objective optimization problems^{42,43}. For instance, in catalyst design, one may seek to maximize activity while maintaining stability and minimizing cost; in materials discovery, trade-offs may exist between conductivity, mechanical strength, and processability. In such cases, algorithms aim to identify a set of Pareto-optimal solutions that represent the best achievable trade-offs between competing objectives.

Algorithmic guidance enables a shift from passive data collection to active learning, in which each experiment is selected to maximize information gain or expected improvement toward a defined objective⁴⁴. By systematically incorporating feedback from previous experiments, these approaches allow for more efficient navigation of complex chemical spaces compared to unguided or brute-force exploration.

1.1.6 Fully Autonomous Platforms

The integration of computational screening, high-throughput experimentation, laboratory automation, and algorithmic guidance has led to the emergence of fully autonomous platforms for scientific discovery. In such systems, the research workflow is organized into a closed-loop process that begins with a high-level research objective defined by a human and proceeds through successive stages of hypothesis generation, experimental execution, and iterative refinement⁸.

Starting from a given research question or target property, the platform can first leverage existing knowledge by automatically accessing and analyzing the scientific literature and available databases. In combination with this prior knowledge, computational methods may be employed to screen large candidate spaces and identify a subset of promising systems for further investigation. This initial stage serves to reduce the experimental burden by focusing resources on candidates with a higher likelihood of success. Following this screening process, experimentally viable candidates are selected and evaluated using automated laboratory systems. Robotic platforms execute synthesis, processing, and characterization steps according to predefined or dynamically generated protocols. The results are then

automatically collected, structured, and fed back so that the platforms incorporate algorithmic guidance to iteratively refine the exploration of chemical space, expecting to improve performance, reduce uncertainty, or explore previously untested regions. This closed-loop process enables efficient and adaptive navigation of complex search spaces, significantly accelerating the discovery process compared to traditional trial-and-error approaches.

Despite these advances, fully autonomous platforms with minimal or no human intervention remain an ongoing goal. Current platforms typically rely on human input at various stages, such as constraining the search space, or validating experimental results. Nevertheless, several pioneering systems demonstrate key elements of autonomy^{45,46}. Looking forward, a central challenge is to increase the generality and robustness of these platforms. Most existing systems are tailored to specific classes of materials, reactions, or experimental setups, limiting their applicability across domains. Achieving more general autonomous platforms will require advances in modular hardware design, standardized data and software interfaces, improved perception and manipulation capabilities, and more flexible algorithmic frameworks capable of handling diverse scientific objectives.

1.2 Reticular Materials Discovery

1.2.1 Reticular Synthesis

Reticular materials, including metal–organic frameworks (MOFs) and covalent organic frameworks (COFs), are extended porous crystalline structures formed by assembling molecular building blocks through strong coordination or covalent bonds⁴⁷. These building blocks possess well-defined composition, connectivity, and geometry, which, in principle, enables the rational design of frameworks with tunable pore environments and functionalities^{48,49}.

In practice, however, the mapping from building block design to final structure is far from deterministic. In MOF synthesis, metal ions can adopt multiple coordination environments, and a variety of species, including solvents and counterions, may compete for coordination. Such competing interactions can lead to the formation of unintended phases or side products. Consequently, the synthetic outcome is governed not only by the intended connectivity of the building blocks, but also by complex chemical equilibria and reaction conditions⁵⁰. Moreover, even when building blocks share identical nominal connectivity, subtle differences in conformation, flexibility, or reactivity can result in distinct topologies⁵¹. Small variations in geometry or reaction environment may stabilize alternative network structures, obscuring the direct relationship between design and outcome. This inherent complexity limits the predictive power of purely rational design and necessitates extensive experimental verification.

These characteristics make reticular materials discovery particularly well suited to the methodologies discussed in the previous section. The large combinatorial space defined by building blocks and reaction conditions naturally motivates high-throughput experimentation, while the need to systematically explore and reproduce complex synthetic pathways benefits from laboratory automation.

Furthermore, computational screening and algorithmic guidance can help prioritize candidates and iteratively refine experimental strategies, enabling more efficient navigation of this complex design space.

1.2.2 Crystallization Challenges

In addition to design ambiguity, the formation of high-quality crystalline reticular materials presents intrinsic challenges. Reticular structures are constructed through relatively strong bonds, which are associated with high activation barriers for bond formation. While these strong bonds contribute to the thermodynamic stability of the final material, they also limit the system's ability to undergo error correction during crystallization. In contrast to more reversible systems, where incorrect connections can break and reform to yield ordered structures, the reduced reversibility in reticular systems can trap defects or result in poorly crystalline or amorphous products^{52,53}. As a result, achieving high crystallinity often requires careful control over both thermodynamic and kinetic factors during synthesis⁵⁴.

To address these challenges, a range of synthetic strategies has been developed to improve crystallinity. In MOF synthesis, the use of modulators, typically monodentate ligands such as carboxylic acids, can regulate nucleation and growth by competing with multidentate linkers for coordination sites, thereby slowing down the assembly of building blocks and promoting higher crystallinity⁵⁵⁻⁵⁷. Similarly, careful control of solvent composition, including the use of anhydrous conditions, can suppress competing coordination pathways and improve structural fidelity⁵⁸⁻⁶⁰. For COFs, achieving high crystallinity is even more challenging due to the formation of strong covalent bonds. Approaches, such as the use of modulators, templating agents, or optimized solvent systems, have enabled the growth of COFs of larger and more ordered crystalline domains and, in some cases, single crystals suitable for structural characterization⁶¹⁻⁶⁵. Despite these advances, crystallization in reticular materials remains highly sensitive to experimental conditions and often requires extensive empirical optimization. Small changes in conditions can sometimes significantly alter the outcome, leading to different phases, morphologies, or levels of crystallinity.

1.3 References

- (1) Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; Anandkumar, A.; Bergen, K.; Gomes, C. P.; Ho, S.; Kohli, P.; Lasenby, J.; Leskovec, J.; Liu, T.-Y.; Manrai, A.; Marks, D.; Ramsundar, B.; Song, L.; Sun, J.; Tang, J.; Veličković, P.; Welling, M.; Zhang, L.; Coley, C. W.; Bengio, Y.; Zitnik, M. Scientific Discovery in the Age of Artificial Intelligence. *Nature* **2023**, *620* (7972), 47–60.
- (2) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio De Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23* (6), 1213–1242.
- (3) Angelopoulos, A.; Cahoon, J. F.; Alterovitz, R. Transforming Science Labs into Automated Factories of Discovery. *Sci. Robot.* **2024**, *9* (95), eadm6991.
- (4) Haas, B. C.; Kalyani, D.; Sigman, M. S. Applying Statistical Modeling Strategies to Sparse Datasets in Synthetic Chemistry. *Sci. Adv.* **2025**, *11* (1), eadt3013.
- (5) Sun, K.; Bagni, D.; Cavanagh, J. M.; Wang, Y.; Sawyer, J. M.; Zhou, B.; Gritsevskiy, A.; Zhang, O.; Head-Gordon, T. SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models. *ACS Cent. Sci.* **2025**, *11* (11), 2108–2120.
- (6) Liu, Y.; Cavanagh, J.; Sun, K.; Toney, J.; Yuan, C.-Y.; Smith, A.; St Michel Li, R.; Graggs, P.; Toste, F. D.; Kulik, H.; Head-Gordon, T. Exploring Transition Metal Complexes with Large Language Models. November 17, 2025.
- (7) Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Cent. Sci.* **2023**, *9* (11), 2161–2170.
- (8) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124* (16), 9633–9732.
- (9) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 134–140.
- (10) Shahzad, K.; Mardare, A. I.; Hassel, A. W. Accelerating Materials Discovery: Combinatorial Synthesis, High-Throughput Characterization, and Computational Advances. *Sci. Technol. Adv. Mater. Methods* **2024**, *4* (1), 2292486.

- (11) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13* (6), 579–633.
- (12) Cawse, J. N. Experimental Strategies for Combinatorial and High-Throughput Materials Development. *Acc. Chem. Res.* **2001**, *34* (3), 213–221.
- (13) Vervoort, N.; Goossens, K.; Baeten, M.; Chen, Q. Recent Advances in Analytical Techniques for High Throughput Experimentation. *Anal. Sci. Adv.* **2021**, *2* (3–4), 109–127.
- (14) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *Int. J. Mol. Sci.* **2011**, *13* (1), 427–452.
- (15) Barber, Z. H.; Blamire, M. G. High Throughput Thin Film Materials Science. *Mater. Sci. Technol.* **2008**, *24* (7), 757–770.
- (16) Ludwig, A. Discovery of New Materials Using Combinatorial Synthesis and High-Throughput Characterization of Thin-Film Materials Libraries Combined with Computational Methods. *Npj Comput. Mater.* **2019**, *5* (1), 70.
- (17) Farrusseng, D. High-Throughput Heterogeneous Catalysis. *Surf. Sci. Rep.* **2008**, *63* (11), 487–513.
- (18) Hoogenboom, R.; Meier, M. A. R.; Schubert, U. S. Combinatorial Methods, Automated Synthesis and High-Throughput Screening in Polymer Research: Past and Present. *Macromol. Rapid Commun.* **2003**, *24* (1), 15–32.
- (19) Meier, M. A. R.; Hoogenboom, R.; Schubert, U. S. Combinatorial Methods, Automated Synthesis and High-Throughput Screening in Polymer Research: The Evolution Continues. *Macromol. Rapid Commun.* **2004**, *25* (1), 21–33.
- (20) Tweedie, C. A.; Anderson, D. G.; Langer, R.; Van Vliet, K. J. Combinatorial Material Mechanics: High-Throughput Polymer Synthesis and Nanomechanical Screening. *Adv. Mater.* **2005**, *17* (21), 2599–2604.
- (21) Olsen, K. The First 110 Years of Laboratory Automation: Technologies, Applications, and the Creative Scientist. *SLAS Technol.* **2012**, *17* (6), 469–480.
- (22) Kong, F.; Yuan, L.; Zheng, Y. F.; Chen, W. Automatic Liquid Handling for Life Science: A Critical Review of the Current State of the Art. *SLAS Technol.* **2012**, *17* (3), 169–185.
- (23) Jiang, Y.; Fakhruldeen, H.; Pizzuto, G.; Longley, L.; He, A.; Dai, T.; Clowes, R.; Rankin, N.; Cooper, A. I. Autonomous Biomimetic Solid Dispensing Using a Dual-Arm Robotic Manipulator. *Digit. Discov.* **2023**, *2* (6), 1733–1744.
- (24) McMullen, J. P.; Jensen, K. F. Integrated Microreactors for Reaction Automation: New Approaches to Reaction Development. *Annu. Rev. Anal. Chem.* **2010**, *3* (1), 19–42.

- (25) Vescovi, R.; Ginsburg, T.; Hippe, K.; Ozgulbas, D.; Stone, C.; Stroka, A.; Butler, R.; Blaiszik, B.; Brettin, T.; Chard, K.; Hereld, M.; Ramanathan, A.; Stevens, R.; Vriza, A.; Xu, J.; Zhang, Q.; Foster, I. Towards a Modular Architecture for Science Factories. *Digit. Discov.* **2023**, *2* (6), 1980–1998.
- (26) Munawar, A.; De Magistris, G.; Pham, T.-H.; Kimura, D.; Tatsubori, M.; Moriyama, T.; Tachibana, R.; Booch, G. MaestROB: A Robotics Framework for Integrated Orchestration of Low-Level Control and High-Level Reasoning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: Brisbane, QLD, 2018; pp 527–534.
- (27) Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories. *Matter* **2024**, *7* (9), 2959–2977.
- (28) Fei, Y.; Rendy, B.; Kumar, R.; Dartsi, O.; Sahasrabudde, H. P.; McDermott, M. J.; Wang, Z.; Szymanski, N. J.; Walters, L. N.; Milsted, D.; Zeng, Y.; Jain, A.; Ceder, G. AlabOS: A Python-Based Reconfigurable Workflow Management Framework for Autonomous Laboratories. *Digit. Discov.* **2024**, *3* (11), 2275–2288.
- (29) Colón, Y. J.; Snurr, R. Q. High-Throughput Computational Screening of Metal–Organic Frameworks. *Chem Soc Rev* **2014**, *43* (16), 5735–5749.
- (30) Singh, A. K.; Mathew, K.; Zhuang, H. L.; Hennig, R. G. Computational Screening of 2D Materials for Photocatalysis. *J. Phys. Chem. Lett.* **2015**, *6* (6), 1087–1098.
- (31) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational Screening of All Stoichiometric Inorganic Materials. *Chem* **2016**, *1* (4), 617–627.
- (32) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling Deep Learning for Materials Discovery. *Nature* **2023**, *624* (7990), 80–85.
- (33) Ou-Yang, S.; Lu, J.; Kong, X.; Liang, Z.; Luo, C.; Jiang, H. Computational Drug Discovery. *Acta Pharmacol. Sin.* **2012**, *33* (9), 1131–1140.
- (34) Sadybekov, A. V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nature* **2023**, *616* (7958), 673–685.
- (35) Vapnik, V. N. An Overview of Statistical Learning Theory. *IEEE Trans. Neural Netw.* **1999**, *10* (5), 988–999.
- (36) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, 2006.
- (37) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444.
- (38) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4* (4), 828–849.
- (39) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), 015016.

- (40) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54* (2), 263–270.
- (41) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96.
- (42) Yaochu Jin; Sendhoff, B. Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2008**, *38* (3), 397–415.
- (43) Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. Machine Learning Meets Continuous Flow Chemistry: Automated Optimization towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.* **2018**, *352*, 277–282.
- (44) Coley, C. W. Defining and Exploring Chemical Spaces. *Trends Chem.* **2021**, *3* (2), 133–145.
- (45) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. *Nature* **2023**, *624* (7990), 86–91.
- (46) Pilon, S.; Savino, E.; Bayley, O. M.; Vanzella, M.; Claros, M.; Siasiaridis, P.; Liu, J.; Lukas, F.; Damian, M.; Tseliou, V.; Intini, N.; Slattery, A.; SanJosé-Orduna, J.; Den Hartog, T.; Peters, R. A. H.; Gargano, A. F. G.; Mutti, F. G.; Noël, T. A Flexible and Affordable Self-Driving Laboratory for Automated Reaction Optimization. *Nat. Synth.* **2026**.
- (47) Yaghi, O. M.; Kalmutzki, M. J.; Diercks, C. S. *Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks*; Wiley-VCH Verlag GmbH & Co: Weinheim, Germany, 2019.
- (48) Yaghi, O. M.; O’Keeffe, M.; Ockwig, N. W.; Chae, H. K.; Eddaoudi, M.; Kim, J. Reticular Synthesis and the Design of New Materials. *Nature* **2003**, *423* (6941), 705–714.
- (49) Furukawa, H.; Cordova, K. E.; O’Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341* (6149), 1230444.
- (50) Stock, N.; Biswas, S. Synthesis of Metal-Organic Frameworks (MOFs): Routes to Various MOF Topologies, Morphologies, and Composites. *Chem. Rev.* **2012**, *112* (2), 933–969.
- (51) Kalmutzki, M. J.; Hanikel, N.; Yaghi, O. M. Secondary Building Units as the Turning Point in the Development of the Reticular Chemistry of MOFs. *Sci. Adv.* **2018**, *4* (10), eaat9180.
- (52) Bennett, T. D.; Cheetham, A. K. Amorphous Metal–Organic Frameworks. *Acc. Chem. Res.* **2014**, *47* (5), 1555–1562.
- (53) Smith, B. J.; Overholts, A. C.; Hwang, N.; Dichtel, W. R. Insight into the Crystallization of Amorphous Imine-Linked Polymer Networks to 2D Covalent Organic Frameworks. *Chem. Commun.* **2016**, *52* (18), 3690–3693.

- (54) Cheetham, A. K.; Kieslich, G.; Yeung, H. H.-M. Thermodynamic and Kinetic Effects in the Crystallization of Metal–Organic Frameworks. *Acc. Chem. Res.* **2018**, *51* (3), 659–667.
- (55) Schaate, A.; Roy, P.; Godt, A.; Lippke, J.; Waltz, F.; Wiebcke, M.; Behrens, P. Modulated Synthesis of Zr-Based Metal–Organic Frameworks: From Nano to Single Crystals. *Chem. – Eur. J.* **2011**, *17* (24), 6643–6651.
- (56) Marshall, R. J.; Hobday, C. L.; Murphie, C. F.; Griffin, S. L.; Morrison, C. A.; Moggach, S. A.; Forgan, R. S. Amino Acids as Highly Efficient Modulators for Single Crystals of Zirconium and Hafnium Metal–Organic Frameworks. *J. Mater. Chem. A* **2016**, *4* (18), 6955–6963.
- (57) Hu, Z.; Castano, I.; Wang, S.; Wang, Y.; Peng, Y.; Qian, Y.; Chi, C.; Wang, X.; Zhao, D. Modulator Effects on the Water-Based Synthesis of Zr/Hf Metal–Organic Frameworks: Quantitative Relationship Studies between Modulator, Synthetic Condition, and Performance. *Cryst. Growth Des.* **2016**, *16* (4), 2295–2301.
- (58) Wang, B.; Côté, A. P.; Furukawa, H.; O’Keeffe, M.; Yaghi, O. M. Colossal Cages in Zeolitic Imidazolate Frameworks as Selective Carbon Dioxide Reservoirs. *Nature* **2008**, *453* (7192), 207–211.
- (59) Zhang, B.; Zhang, J.; Liu, C.; Sang, X.; Peng, L.; Ma, X.; Wu, T.; Han, B.; Yang, G. Solvent Determines the Formation and Properties of Metal–Organic Frameworks. *RSC Adv.* **2015**, *5* (47), 37691–37696.
- (60) Biserčić, M. S.; Marjanović, B.; Vasiljević, B. N.; Mentus, S.; Zasońska, B. A.; Ćirić-Marjanović, G. The Quest for Optimal Water Quantity in the Synthesis of Metal-Organic Framework MOF-5. *Microporous Mesoporous Mater.* **2019**, *278*, 23–29.
- (61) Zhang, Y.-B.; Su, J.; Furukawa, H.; Yun, Y.; Gándara, F.; Duong, A.; Zou, X.; Yaghi, O. M. Single-Crystal Structure of a Covalent Organic Framework. *J. Am. Chem. Soc.* **2013**, *135* (44), 16336–16339.
- (62) Beaudoin, D.; Maris, T.; Wuest, J. D. Constructing Monocrystalline Covalent Organic Networks by Polymerization. *Nat. Chem.* **2013**, *5* (10), 830–834.
- (63) Ma, T.; Kapustin, E. A.; Yin, S. X.; Liang, L.; Zhou, Z.; Niu, J.; Li, L.-H.; Wang, Y.; Su, J.; Li, J.; Wang, X.; Wang, W. D.; Wang, W.; Sun, J.; Yaghi, O. M. Single-Crystal x-Ray Diffraction Structures of Covalent Organic Frameworks. *Science* **2018**, *361* (6397), 48–52.
- (64) Han, J.; Feng, J.; Kang, J.; Chen, J.-M.; Du, X.-Y.; Ding, S.-Y.; Liang, L.; Wang, W. Fast Growth of Single-Crystal Covalent Organic Frameworks for Laboratory x-Ray Diffraction. *Science* **2024**, *383* (6686), 1014–1019.
- (65) Yi, L.; Gao, Y.; Luo, S.; Wang, T.; Deng, H. Structure Evolution of 2D Covalent Organic Frameworks Unveiled by Single-Crystal X-Ray Diffraction. *J. Am. Chem. Soc.* **2024**, *146* (29), 19643–19648.

Chapter 2

Experiment Platform for Reticular Materials Discovery

2.1 Introduction

The discovery of reticular materials has traditionally relied on trial-and-error experimentation, in which building blocks and reaction conditions are selected based on prior knowledge and iteratively adjusted to obtain crystalline products. While this approach has enabled the synthesis of a wide range of frameworks, it remains inherently limited in its ability to systematically explore the vast combinatorial space defined by building block chemistry and conditions. Moreover, the formation of reticular materials is highly sensitive to subtle variations in reaction parameters, often leading to complex and unpredictable outcomes that are difficult to rationalize a priori¹.

Recent advances in automation and high-throughput experimentation have begun to address these limitations by enabling more systematic and reproducible exploration of synthesis conditions^{2–4}. In parallel, substantial progress has been made in the automation of characterization techniques. Automated and high-throughput platforms have been developed for methods such as powder X-ray diffraction (PXRD)^{5–7}, vibrational spectroscopy^{8,9}, and electron microscopy^{10,11}, enabling rapid measurement across large reaction libraries. However, in many cases, these characterization techniques provide indirect or ensemble-averaged information, and their interpretation often relies on comparison with reference databases or the construction of structural models followed by simulation. For example, PXRD patterns are typically analyzed through pattern matching or refinement against candidate structures^{12,13}, while spectroscopic signatures must be assigned based on prior knowledge or computational predictions^{14–16}. As a result, although data collection can be automated, the determination of atomic-level structure for previously unknown materials frequently still requires human intervention or iterative model fitting.

In the context of reticular materials discovery, two experimental challenges remain central. First, efficient exploration of reaction space requires the ability to perform large numbers of synthesis experiments in parallel with precise control over composition and conditions. Second, and more critically, the reliable identification of newly formed structures requires definitive structural characterization. Because reticular materials can exhibit substantial structural complexity, reaction products often include multiple crystalline phases or unexpected structures that cannot be

Parts of this chapter have been adapted from:

Rong, Z.; Chen, Z.; Luong, F.; Chheda, S.; Luong, H. T. N.; Zheng, Z.; Greco, K.; Alghamdi, A. A.; Bui, K. H.; Jaffrelot Inizan, T.; Nguyen-Dang, T.; Pham, H. H.; Le, D. D.; Sauer, J.; Phung, V. B. T.; Chayes, J. T.; Borgs, C.; Boley, M.; El Ghaoui, L.; Yaghi, O. M. Algorithmic Iterative Reticular Synthesis of Zeolitic Imidazolate Framework Crystals. *Nat. Synth.* **2025**, 5 (3), 409–419.

unambiguously resolved using indirect characterization techniques¹⁷. In this context, single-crystal X-ray diffraction (SCXRD) provides the most definitive determination of atomic structure and is therefore essential for establishing the identity of newly discovered reticular materials¹⁸. However, its integration into high-throughput workflows presents practical challenges, including the identification of suitable crystals from large numbers of reaction outcomes. This motivates the need for experimental platforms that combine scalable synthesis, rapid screening of crystallization outcomes, and rigorous structural verification within a unified framework.

This dissertation introduces an autonomous platform termed AIRES (Algorithmic Iterative Reticular Synthesis) for accelerating the discovery of reticular materials. The platform integrates high-throughput synthesis, automated optical microscopy, single-crystal X-ray diffraction, and machine learning-guided decision-making into a closed-loop workflow. Throughout this dissertation, zeolitic imidazolate frameworks (ZIFs) are employed as a representative system to demonstrate the capabilities of the AIRES platform. In this chapter, the experimental components of the platform (Figure 2.1) are presented.

High-throughput experimentation

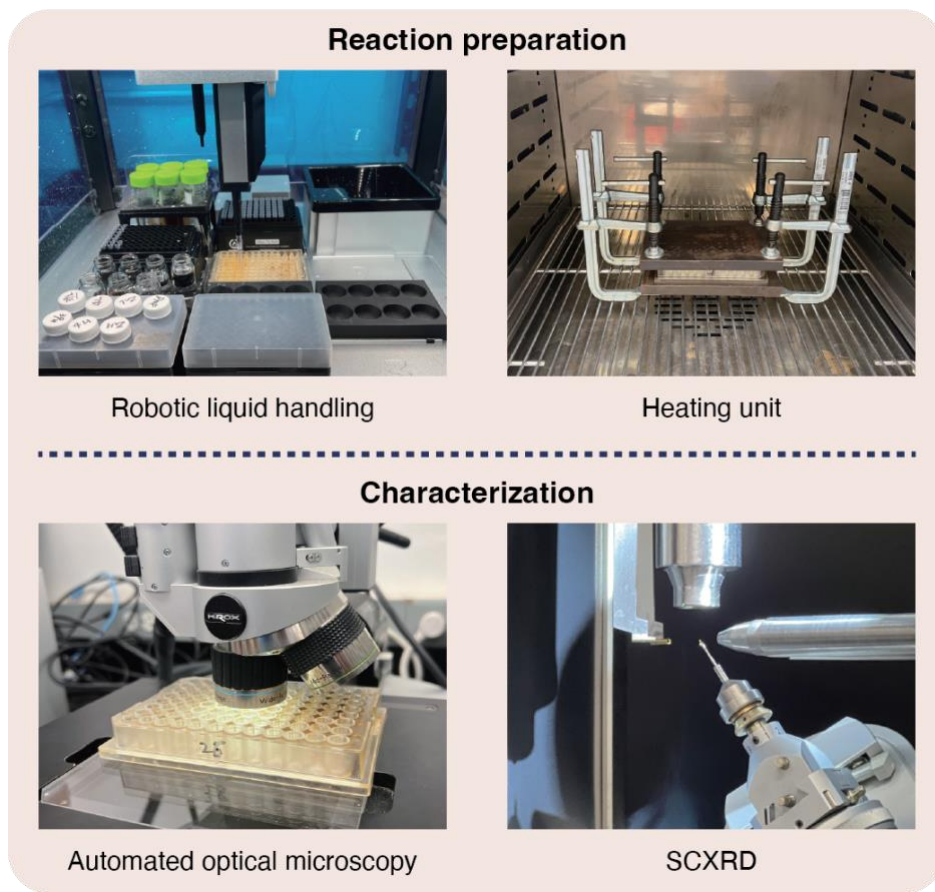


Figure 2.1. Components of the experimental platform for reticular materials discovery.

2.2 Zeolitic Imidazolate Frameworks

ZIFs constitute a subclass of MOFs characterized by tetrahedrally coordinated metal centers bridged by imidazolate linkers, forming structures analogous to zeolites¹⁹. ZIFs present unique challenges that make them particularly suitable for evaluating autonomous discovery platforms.

First, although ZIFs share a common tetrahedral building block geometry, linker–linker interactions play a critical role in determining the conformation of adjacent building blocks. As a result, multiple framework arrangements can be thermodynamically accessible, making it difficult to design isorecticular structures or predict which topology will ultimately form²⁰. Subtle variations in linker functionality or reaction conditions can therefore lead to distinct structures, even when the connectivity remains unchanged²¹. Second, the single-point metal–ligand coordination in ZIFs introduces competing synthesis pathways. Metal centers can adopt both four- and six-coordinate environments, while solvents, counterions, and other species may compete for coordination sites. These competing interactions often lead to the formation of undesired phases or side products, further complicating synthesis and reducing predictability.

This combination of structural diversity arising from subtle chemical variations and the presence of multiple competing pathways makes ZIF synthesis highly sensitive to experimental conditions. In contrast, other classes of reticular materials, such as carboxylate-based MOFs or COFs, more frequently exhibit isorecticular behavior, where changes in functionality preserve the underlying topology^{22–25}.

2.3 Automated and High-throughput Synthesis

The synthesis of reticular materials in this work follows a direct approach, in which precursor species corresponding to building blocks are combined under controlled conditions to form extended frameworks. To enable efficient exploration, an automated and high-throughput experiment platform was developed to perform synthesis, reaction control, and post-processing in a reproducible and scalable manner. The platform consists of three stages: automated reaction preparation, controlled thermal treatment, and post-reaction processing.

Reaction preparation is carried out using a robotic liquid-handling system, which enables precise and reproducible dispensing of reagents into a multi-well format. The system is based on an Opentrons OT-2 platform equipped with dual pipetting modules, allowing accurate handling of both small and moderate liquid volumes. The experimental setup utilizes a 96-well plate format, where each well corresponds to a distinct chemical system and reaction condition. Stock solutions of metal precursors and organic linkers are prepared at predefined concentrations and combined with solvent to achieve the desired reaction composition. The use of stock solutions allows flexible control over concentration and stoichiometry while maintaining consistency across experiments. All dispensing operations are controlled through custom Python scripts executed through a Jupyter-based interface. Reaction conditions are defined in structured input files, which are translated into pipetting

instructions for the robotic system. To prevent cross-contamination, pipette tips are replaced between transfers involving different stock solutions. The preparation of a full 96-well plate can be completed within approximately 30 minutes, enabling efficient generation of large reaction libraries.

Following automated dispensing, the reaction plate is sealed and subjected to controlled thermal treatment. Each plate is sealed using PTFE/silicone caps and mechanically secured between steel plates to ensure uniform pressure and minimize solvent loss during heating. The sealed assembly is then placed in an isothermal oven at the desired reaction temperature. Given the sensitivity of reticular synthesis to temperature variations, thermal uniformity across the reaction plate is critical. To assess this, thermal imaging experiments were conducted using an infrared camera to monitor temperature distribution during heating. Measurements were performed at multiple positions across the plate and at different time points during both the heating ramp and steady-state conditions (Figure 2.2). The results indicate that the well plate maintains a relatively uniform temperature distribution under operating conditions, both across its lateral dimensions and over time. Additional measurements from a top-view perspective further confirm spatial uniformity (Figure 2.3). These observations validate that the thermal environment is consistent across wells, ensuring that variations in reaction outcomes primarily arise from controlled experimental parameters rather than unintended temperature gradients.

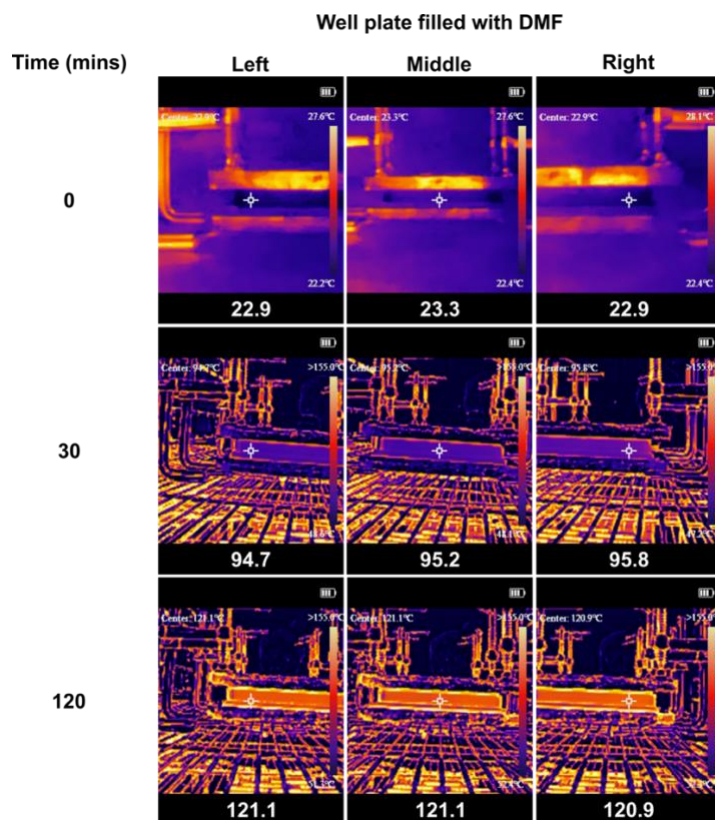


Figure 2.2. Thermal imaging of a DMF-filled well plate during heating. Thermal images of the side view of a well plate filled with DMF, sealed with PTFE/silicone cap, and secured between steel bars at 0, 30, and 120 minutes after placement in a preheated 120 °C oven. Images were captured at left, middle, and right positions along the well plate's side, with corresponding temperatures (°C) displayed below each image.

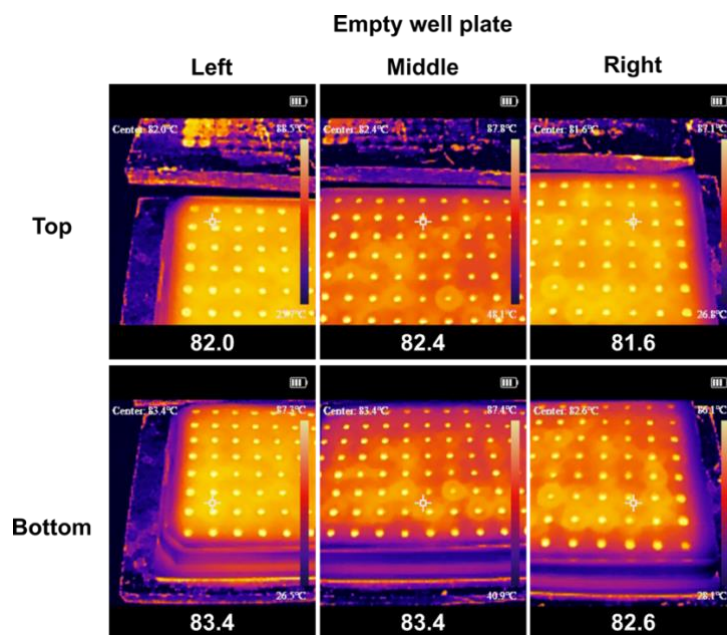


Figure 2.3. Thermal imaging of an empty well plate during cooling. Top-view thermal images of an empty well plate after heating to thermal equilibrium and cooling for 5 minutes. Images were captured from the upper and lower rows at left, middle, and right positions, with corresponding temperatures (°C) displayed below each image.

After thermal treatment, the reaction plate is allowed to cool to room temperature before further processing. The plate is then unsealed and returned to the robotic system for automated handling of reaction products. A key step in post-processing is the removal of supernatant solution while preserving solid products for subsequent optical imaging. This is achieved through a scripted aspiration procedure, in which the pipette tip is positioned slightly above the bottom of each well to avoid disturbing the solid phase. This automated post-processing step ensures consistent sample preparation across all wells.

2.3 Automated Optical Microscopy and Computer Vision

Optical microscopy plays a critical role in the AIRES platform by providing a rapid and scalable method for assessing reaction outcomes. Given the large number of reactions generated through high-throughput synthesis, manual inspection of samples becomes a major bottleneck. To address this challenge, an integrated workflow combining automated optical imaging and computer vision-based analysis was developed to enable efficient identification of promising single crystals for further SCXRD characterization.

Optical image acquisition is performed using an HRX-01 digital microscope, equipped with a multi-illumination high-resolution auto-turret zoom lens and a motorized 100 mm × 100 mm automated XY stage. Imaging is conducted in transmission mode, where light passes through the well plate and sample from below the stage, enabling clear visualization of crystalline features. For each

reaction well, five images are acquired. One image is captured at 50 $\times$ magnification, centered on the well to provide an overview of the sample morphology. In addition, four images are collected at 400 $\times$ magnification, each focusing on a different quadrant of the well (top left, top right, bottom left, and bottom right). The microscope automatically determines the optimal focal plane by scanning a predefined range along the z-axis and maximizing a contrast-based metric. This autofocus procedure ensures consistent image quality across wells with varying sample distributions. To enable reproducible imaging across experiments, well coordinates are manually calibrated and stored in a CSV file, which is used to guide the automated positioning of the microscope stage during each imaging session.

To enable automated identification of single crystals, a labeled image dataset was constructed. A total of 20,365 images were collected from high-throughput experiments on ZIF synthesis, including 7,184 images containing crystals and 13,181 images classified as non-crystals, which encompass both clear solutions and amorphous or particulate precipitates. All images were manually labeled by an expert crystallographer and subsequently verified by a second expert to ensure labeling consistency. Prior to model training, the images were resized from their original resolution of 2040 $\times$ 1530 $\times$ 3 to 512 $\times$ 512 $\times$ 3, balancing computational efficiency and retention of relevant features. The dataset was randomly divided into 18,327 training images and 2,038 test images. Representative examples of labeled images are shown in Figures 2.4 & 2.5, highlighting the visual variability present within both crystal and non-crystal categories.

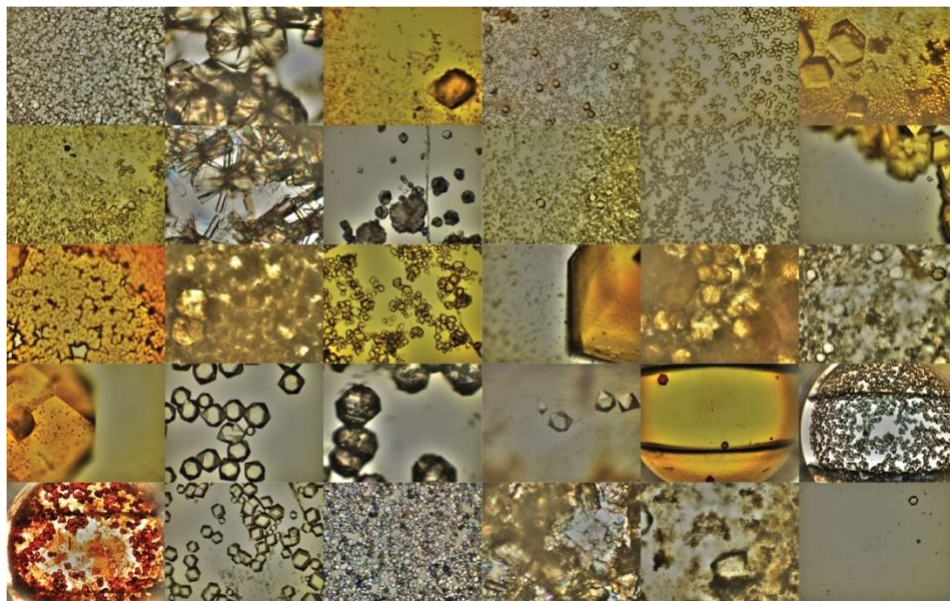


Figure 2.4. Representative images with positive (crystal) labels in the image dataset. A random collection of 30 different images arranged in a 5 $\times$ 6 array.

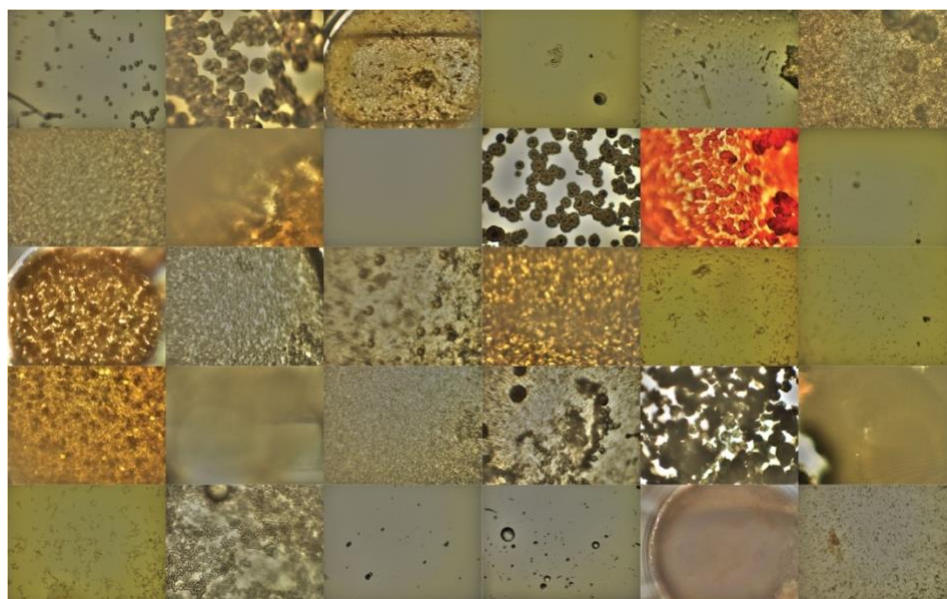


Figure 2.5. Representative images with negative (non-crystals) labels in the image dataset. A random collection of 30 different images arranged in a 5×6 array.

A convolutional neural network was developed to classify optical images as either crystal or non-crystal, enabling automated screening of reaction outcomes. The model is based on the EfficientNetV2-S architecture²⁶ implemented in PyTorch²⁷, which contains approximately 20.18 million trainable parameters and requires 14.87 billion multiply–add operations (Table 2.1). Transfer learning was employed by initializing the model with ImageNet-1K pretrained weights, allowing the network to leverage previously learned feature representations. The output layer was modified for binary classification. The model was trained using the ADAM optimizer²⁸ with a batch size of 64, a learning rate of 0.001, and a weight decay of 0.0001 over 300 epochs. All layers of the network were fine-tuned to adapt to the specific visual features associated with crystalline materials. To address class imbalance, oversampling of crystal images was performed through data augmentation. These augmentations included image flipping, contrast enhancement using CLAHE²⁹, the introduction of defocus and motion blur, as well as random adjustments to brightness, contrast, and RGB channels. Training and validation loss curves are shown in Figure 2.6, while model performance on the test set is summarized in the confusion matrix and the precision–recall curve presented in Figures 2.7 & 2.8, respectively. The best model that maximized the validation accuracy was evaluated on the test set.

The human identification of suitable crystals relies on a combination of visual criteria, including well-defined geometric features, uniform coloration, and sufficient size, typically on the order of approximately $20\ \mu\text{m}$ in the shortest dimension. Based on these labels, the trained model achieved a recall of 0.88 on the test set, indicating that most crystal-containing images were correctly identified. Although the model substantially accelerates the screening process, human validation remains an important step. Images predicted to contain crystals are reviewed by expert researchers to confirm their suitability for single-crystal X-ray diffraction analysis.

Table 2.1. Detailed layer structure of the EfficientNetV2-S architecture configured for binary crystal recognition. Abbreviations: Conv2d (2D Convolutional layer), BN (Batch Normalization), SiLU (Sigmoid Linear Unit activation function), FusedMBConv (Fused Mobile Inverted Bottleneck Convolution block), MBConv (Mobile Inverted Bottleneck Convolution block), AdaptiveAvgPool2d (2D Adaptive Average Pooling layer).

Block	Layer Type	Output Shape	Parameters
Input	-	[1, 3, 256, 256]	-
Entry Block	Conv2d + BN + SiLU	[1, 24, 256, 256]	696
Block 1	2× FusedMBConv	[1, 24, 256, 256]	10,464
Block 2	4× FusedMBConv	[1, 48, 128, 128]	303,552
Block 3	4× FusedMBConv	[1, 64, 64, 64]	589,184
Block 4	6× MBConv	[1, 128, 32, 32]	917,584
Block 5	9× MBConv	[1, 160, 32, 32]	3,064,840
Block 6	15× MBConv	[1, 256, 16, 16]	15,076,800
Exit Block	Conv2d + BN + SiLU	[1, 1280, 16, 16]	330,240
Head	AdaptiveAvgPool2d	[1, 1280, 1, 1]	0
Classifier	Dropout + Linear	[1, 2]	2,562

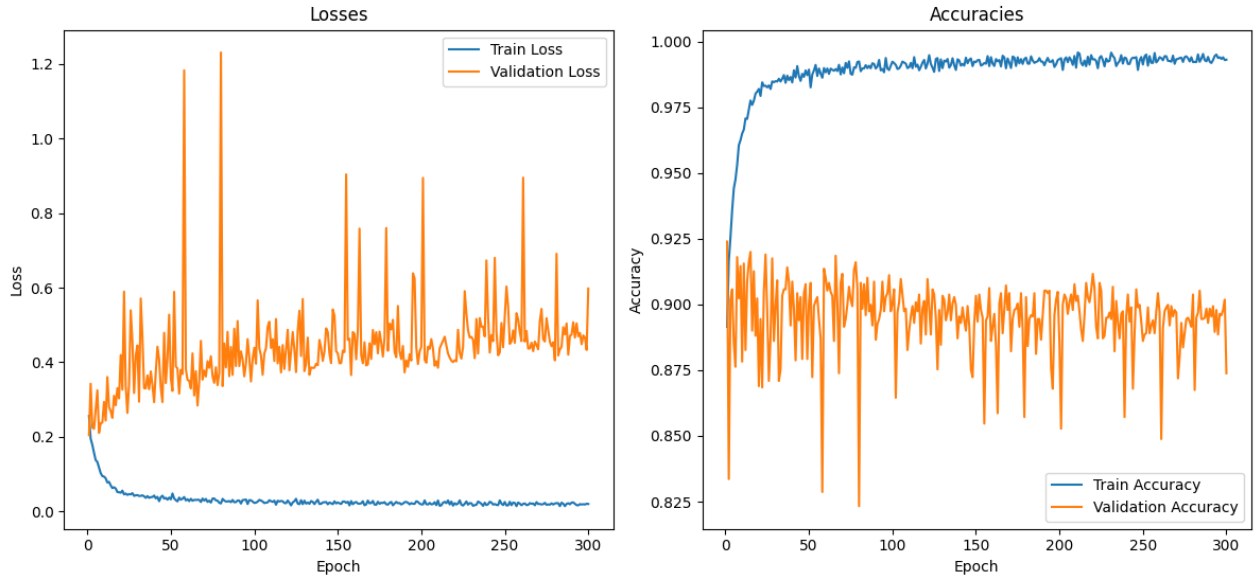


Figure 2.6. Training and validation curves for the EfficientV2-S model showing loss (left) and accuracy (right) over 300 epochs.

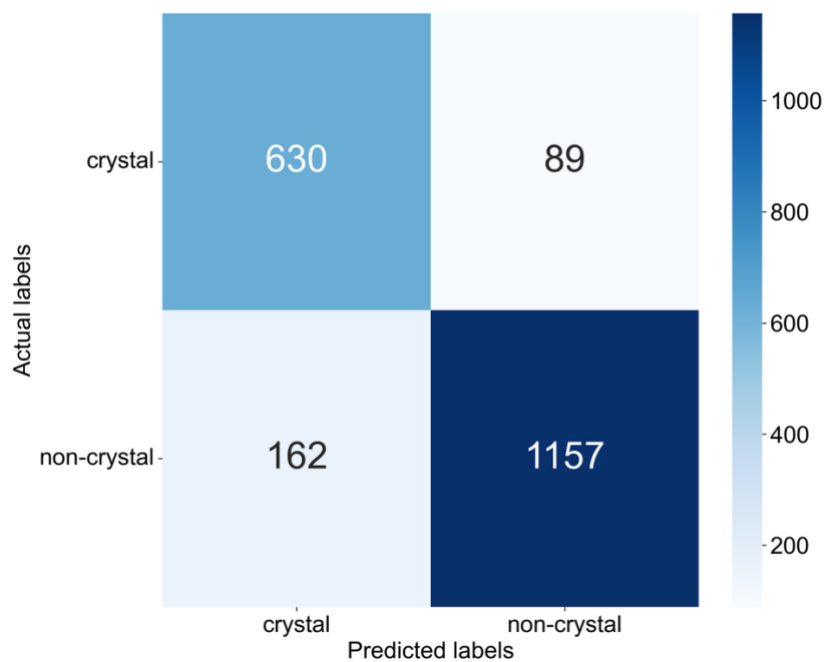


Figure 2.7. Confusion matrix depicting the best EfficientV2-S model’s classification results across 2,038 test images. The threshold value of 0.5 was used to compute the matrix values.

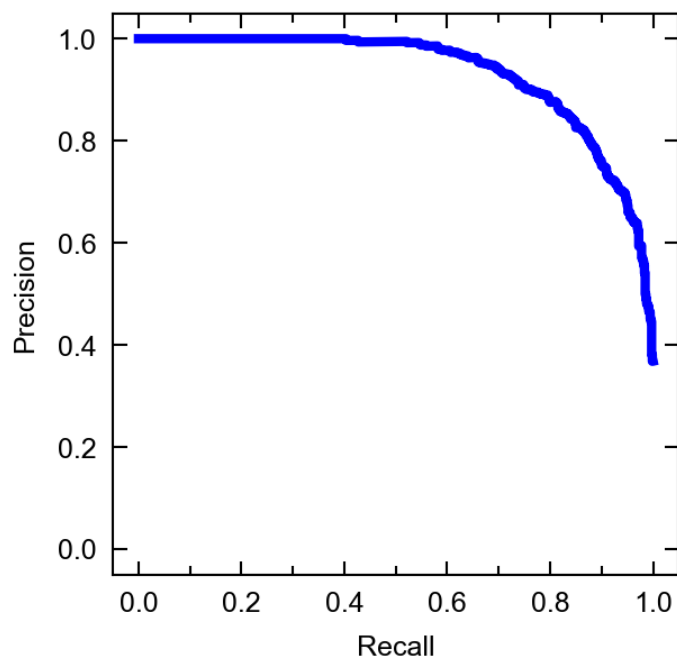


Figure 2.8. Precision-recall curve of the best EfficientV2-S model on 2,038 test images.

2.4 Single Crystal X-ray Diffraction

Single-crystal X-ray diffraction (SCXRD) plays a central role in the AIRES platform by providing definitive structural determination of reaction products. In the context of reticular materials discovery, structural verification is not merely a characterization step but a fundamental requirement for determining whether a desired framework has been successfully formed. This necessity arises from the intrinsic complexity of reticular systems, where multiple crystalline phases with similar compositions or morphologies may coexist under comparable synthesis conditions. Conventional characterization techniques that rely on bulk properties or averaged structural information are often insufficient to unambiguously distinguish between desired and undesired products. For example, PXRD is commonly used to assess crystallinity and phase identity. While PXRD can provide information on lattice parameters and allow indexing of diffraction peaks, it does not directly resolve atomic connectivity, which defines the identity of reticular structures.

The importance of SCXRD in resolving such ambiguity is illustrated in several ZIF synthesis systems using N,N-dimethylformamide (DMF) as the solvent (Figure 2.9). For instance, the reaction between zinc nitrate hexahydrate and benzimidazole (bIm) yields not only the desired ZIF-7 structure, $\text{Zn}(\text{bIm})_2$, but also a secondary crystalline product with composition $\text{Zn}_5(\text{bIm})_6(\text{HCOO})_4(\text{DMF})_2$. Similarly, reactions involving 2-nitroimidazole (2nIm) or 9H-purine (9pur) can produce both the targeted ZIF frameworks, such as $\text{Zn}(\text{2nIm})_2$ and $\text{Zn}(\text{9pur})_2$, and competing phases including $\text{Zn}(\text{2nIm})(\text{HCOO})$ and $\text{Zn}_2(\text{9pur})_2(\text{HCOO})_2(\text{DMF})$. Although these products may exhibit differences in crystal morphology, such distinctions are often insufficient to reliably determine their structural identity. These examples demonstrate that solvent molecules and other species such as formate derived from DMF, can participate in coordination and give rise to structurally distinct phases.

To enable efficient integration of SCXRD into a high-throughput workflow, practical strategies were adopted to reduce experimental burden while maintaining reliability. Crystals exhibiting similar morphology within reactions involving the same linker are assumed to share identical structures, based on the observation that macroscopic crystal symmetry often reflects microscopic symmetry. In addition, crystals with unit cell parameters matching those of known ZIF structures are classified as known frameworks, allowing rapid identification without requiring full structure refinement. These approximations enable efficient screening while preserving the accuracy necessary for reliable structural assignment.

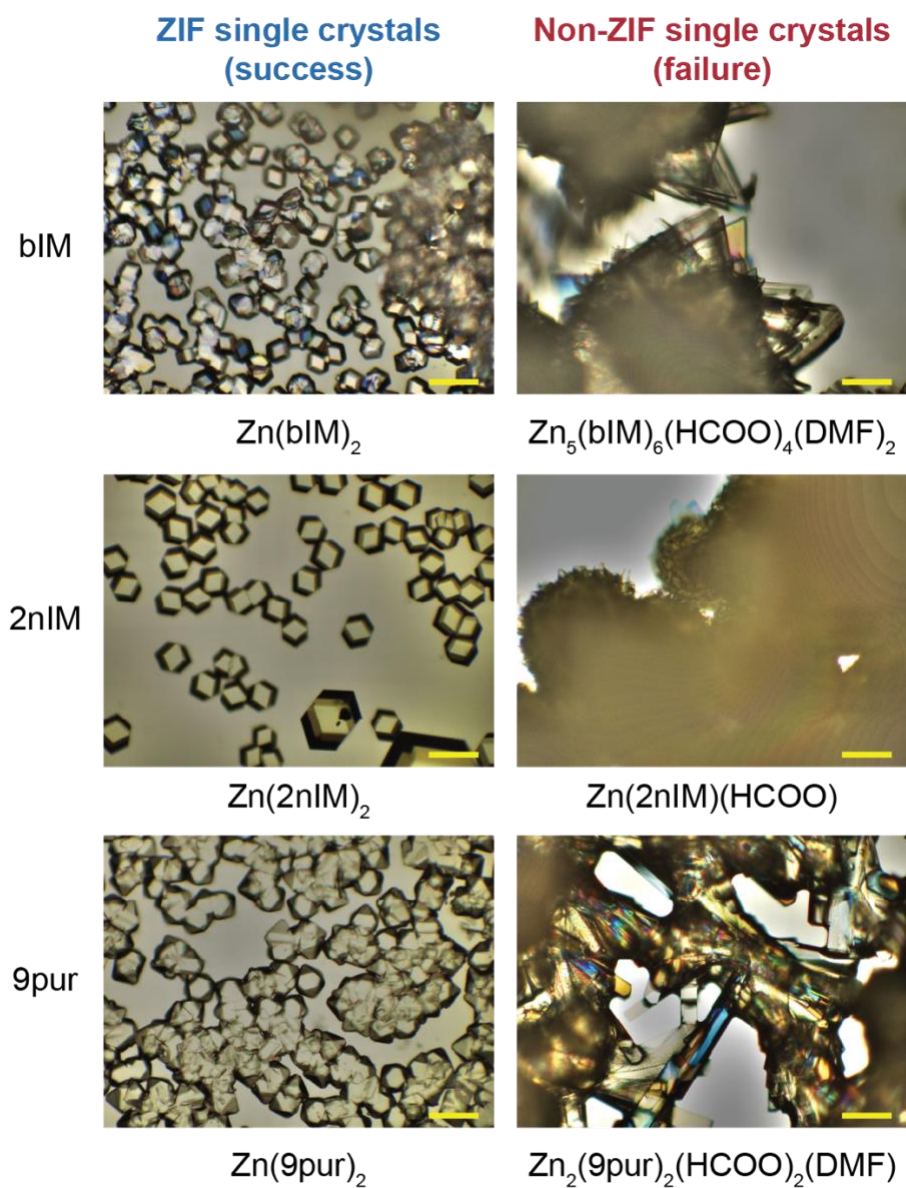


Figure 2.9. Optical microscopy images of crystalline products from ZIF synthesis, with crystal identities confirmed by SCXRD. Yellow scale bar 100 μm .

2.5 References

- (1) Stock, N.; Biswas, S. Synthesis of Metal-Organic Frameworks (MOFs): Routes to Various MOF Topologies, Morphologies, and Composites. *Chem. Rev.* **2012**, *112* (2), 933–969.
- (2) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282–291.
- (3) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.* **2020**, *6* (20), eaaz8867.
- (4) Zhao, H.; Chen, W.; Huang, H.; Sun, Z.; Chen, Z.; Wu, L.; Zhang, B.; Lai, F.; Wang, Z.; Adam, M. L.; Pang, C. H.; Chu, P. K.; Lu, Y.; Wu, T.; Jiang, J.; Yin, Z.; Yu, X.-F. A Robotic Platform for the Synthesis of Colloidal Nanocrystals. *Nat. Synth.* **2023**, *2* (6), 505–514.
- (5) Gregoire, J. M.; Van Campen, D. G.; Miller, C. E.; Jones, R. J. R.; Suram, S. K.; Mehta, A. High-Throughput Synchrotron X-Ray Diffraction for Combinatorial Phase Mapping. *J. Synchrotron Radiat.* **2014**, *21* (6), 1262–1268.
- (6) Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Diallo, M.; Kim, H.; Ceder, G. Adaptively Driven X-Ray Diffraction Guided by Machine Learning for Autonomous Phase Identification. *Npj Comput. Mater.* **2023**, *9* (1), 31.
- (7) Yotsumoto, Y.; Nakajima, Y.; Takamoto, R.; Takeichi, Y.; Ono, K. Autonomous Robotic Experimentation System for Powder X-Ray Diffraction. *Digit. Discov.* **2024**, *3* (12), 2523–2532.
- (8) Horgan, C. C.; Jensen, M.; Nagelkerke, A.; St-Pierre, J.-P.; Vercauteren, T.; Stevens, M. M.; Bergholt, M. S. High-Throughput Molecular Imaging via Deep-Learning-Enabled Raman Spectroscopy. *Anal. Chem.* **2021**, *93* (48), 15850–15860.
- (9) Atia, K.; Hunter, R.; Asare-Werehene, M.; Tsang, B. K.; Anis, H. RamanBot: Versatile High Throughput Raman System. *PLOS One* **2026**, *21* (1), e0334679.
- (10) Kornfeld, J.; Denk, W. Progress and Remaining Challenges in High-Throughput Volume Electron Microscopy. *Curr. Opin. Neurobiol.* **2018**, *50*, 261–267.
- (11) Kim, J.; Rhee, J.; Kang, S.; Jung, M.; Kim, J.; Jeon, M.; Park, J.; Ham, J.; Kim, B. H.; Lee, W. C.; Roh, S.-H.; Park, J. Self-Supervised Machine Learning Framework for High-Throughput Electron Microscopy. *Sci. Adv.* **2025**, *11* (14), eads5552.
- (12) Rietveld, H. M. A Profile Refinement Method for Nuclear and Magnetic Structures. *J. Appl. Crystallogr.* **1969**, *2* (2), 65–71.
- (13) *The Rietveld Method*, 1. publ. in paperb., reprint.; Young, R. A., Ed.; Monographs on crystallography; Oxford Univ. Press: Oxford, 2002.

- (14) Scott, A. P.; Radom, L. Harmonic Vibrational Frequencies: An Evaluation of Hartree–Fock, Møller–Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.* **1996**, *100* (41), 16502–16513.
- (15) Bagno, A.; Rastrelli, F.; Saielli, G. Predicting¹³ C NMR Spectra by DFT Calculations. *J. Phys. Chem. A* **2003**, *107* (46), 9964–9973.
- (16) Han, R.; Ketkaew, R.; Lubner, S. A Concise Review on Recent Developments of Machine Learning for the Prediction of Vibrational Spectra. *J. Phys. Chem. A* **2022**, *126* (6), 801–812.
- (17) Van Gele, S.; Bette, S.; Lotsch, B. V. The Devil Is in the Details: Pitfalls and Ambiguities in the Analysis of X-Ray Powder Diffraction Data of 2D Covalent Organic Frameworks. *JACS Au* **2025**, *5* (1), 388–398.
- (18) Girolami, G. S. *X-Ray Crystallography*; University science books: Mill Valley (Calif.), 2016.
- (19) Park, K. S.; Ni, Z.; Côté, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O’Keeffe, M.; Yaghi, O. M. Exceptional Chemical and Thermal Stability of Zeolitic Imidazolate Frameworks. *Proc. Natl. Acad. Sci.* **2006**, *103* (27), 10186–10191.
- (20) Baburin, I. A.; Leoni, S.; Seifert, G. Enumeration of Not-Yet-Synthesized Zeolitic Zinc Imidazolate MOF Networks: A Topological and DFT Approach. *J. Phys. Chem. B* **2008**, *112* (31), 9437–9443.
- (21) Hayashi, H.; Côté, A. P.; Furukawa, H.; O’Keeffe, M.; Yaghi, O. M. Zeolite A Imidazolate Frameworks. *Nat. Mater.* **2007**, *6* (7), 501–506.
- (22) Eddaoudi, M.; Kim, J.; Rosi, N.; Vodak, D.; Wachter, J.; O’Keeffe, M.; Yaghi, O. M. Systematic Design of Pore Size and Functionality in Isoreticular MOFs and Their Application in Methane Storage. *Science* **2002**, *295* (5554), 469–472.
- (23) Deng, H.; Grunder, S.; Cordova, K. E.; Valente, C.; Furukawa, H.; Hmadeh, M.; Gándara, F.; Whalley, A. C.; Liu, Z.; Asahina, S.; Kazumori, H.; O’Keeffe, M.; Terasaki, O.; Stoddart, J. F.; Yaghi, O. M. Large-Pore Apertures in a Series of Metal-Organic Frameworks. *Science* **2012**, *336* (6084), 1018–1023.
- (24) Dong, J.; Wang, Y.; Liu, G.; Cheng, Y.; Zhao, D. Isoreticular Covalent Organic Frameworks for Hydrocarbon Uptake and Separation: The Important Role of Monomer Planarity. *CrystEngComm* **2017**, *19* (33), 4899–4904.
- (25) Li, L.; Yun, Q.; Zhu, C.; Sheng, G.; Guo, J.; Chen, B.; Zhao, M.; Zhang, Z.; Lai, Z.; Zhang, X.; Peng, Y.; Zhu, Y.; Zhang, H. Isoreticular Series of Two-Dimensional Covalent Organic Frameworks with the Kgd Topology and Controllable Micropores. *J. Am. Chem. Soc.* **2022**, *144* (14), 6475–6482.
- (26) Tan, M.; Le, Q. Efficientnetv2: Smaller Models and Faster Training. In *International conference on machine learning*; PMLR, 2021; pp 10096–10106.
- (27) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others. Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

- (28) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*; 2015.
- (29) Zuiderveld, K. Contrast Limited Adaptive Histogram Equalization. In *Graphics gems IV*; 1994; pp 474–485.

Chapter 3

Machine Learning for Crystallization Outcome Classification

3.1 Introduction

In addition to the automated high-throughput experimental platform, a data-driven framework for analyzing and predicting crystallization outcomes is needed to drive reticular materials discovery. Central to this approach is the construction of a crystallization dataset that captures both successful and unsuccessful reactions, providing a structured representation of reaction inputs and outcomes for subsequent modeling¹.

Machine learning offers a systematic approach to analyzing such datasets. Reaction outcome prediction can be formulated as a supervised classification problem, where the goal is to learn a mapping from experimental inputs to binary outcomes indicating successful or unsuccessful crystallization. While conditions are already expressed as numerical variables, chemical identity cannot be directly incorporated without appropriate representation. This necessitates the development of encoding strategies that translate chemical information into quantitative descriptors while preserving relevant physical and chemical characteristics.

In general, encoding approaches can be broadly categorized into several classes. One common strategy is the use of handcrafted descriptors, which quantifies molecular properties such as topological indices, steric parameters, and physicochemical properties^{2,3}. These descriptors are widely used in cheminformatics due to their interpretability and direct connection to chemical intuition. Another approach is fingerprint-based representations, which encode the presence of local substructures in a binary or count-based format and are effective for capturing structural similarity^{4,5}. More recently, graph-based representations have gained prominence, in which molecules and crystals are treated as graphs with atoms as nodes and bonds as edges. Graph neural networks can then learn task-specific representations directly from structures themselves^{6–12}. In addition, string-based representations, such as SMILES¹³, can be used in conjunction with sequence models to learn latent embeddings of molecules^{14,15}.

While these approaches are effective for representing individual molecules, reaction outcome

Parts of this chapter have been adapted from:

Rong, Z.; Chen, Z.; Luong, F.; Chheda, S.; Luong, H. T. N.; Zheng, Z.; Greco, K.; Alghamdi, A. A.; Bui, K. H.; Jaffrelot Inizan, T.; Nguyen-Dang, T.; Pham, H. H.; Le, D. D.; Sauer, J.; Phung, V. B. T.; Chayes, J. T.; Borgs, C.; Boley, M.; El Ghaoui, L.; Yaghi, O. M. Algorithmic Iterative Reticular Synthesis of Zeolitic Imidazolate Framework Crystals. *Nat. Synth.* **2025**, 5 (3), 409–419.

prediction often depends not only on the identity of each component but also on the interactions between multiple precursors under specific conditions^{16,17}. Therefore, an appropriate encoding strategy must capture not only the intrinsic properties of each precursor but also how these properties influence their interactions. To more explicitly account for interactions, descriptors can be constructed to reflect pairwise or higher-order relationships between components.

In this chapter, both the crystallization dataset and the corresponding encoding strategy are presented. The encoding focuses on the linker as the primary source of variation, using descriptors derived from molecular structure and physicochemical properties. Crucially, the effects of the metal center and solvent are incorporated implicitly through their interactions with the linker. Using these representations, machine learning models are trained to predict crystallization outcomes from combined chemical descriptors and reaction parameters, with labels defined by SCXRD-verified results. Model performance is evaluated using appropriate classification metrics to assess predictive capability and generalization behavior.

3.2 ZIF Crystallization Dataset

The initial design of the ZIF crystallization dataset was guided by prior knowledge extracted from the literature. Reported synthesis conditions and successful linkers provide valuable starting points for identifying regions of chemical space that are known to yield crystalline frameworks. However, literature data exhibit an inherent limitation: published reports typically document only successful synthesis conditions, while unsuccessful or suboptimal conditions are rarely disclosed. In reality, successful crystallization occurs within a broader parameter space that includes many failed attempts. This positive bias in reported data presents a significant challenge for data-driven modeling, as it prevents the construction of balanced datasets necessary for training predictive models¹⁸. Without information on unsuccessful conditions, it is difficult to learn the boundaries that separate crystallization from non-crystallization regimes. To address this limitation, an initial ZIF synthesis dataset was constructed through high-throughput experimentation, using Zn(II) and linkers from previously reported ZIF structures. In this approach, literature information is treated as a set of seeds, around which a broader and more representative exploration of synthesis conditions is performed.

The reaction space was systematically defined and explored through a grid-based sampling strategy. This dissertation focuses on Zn-based ZIFs constructed from a single type of linker, a class of materials that, despite containing over 30 known structures derived from 16 linkers, has seen limited expansion in recent years¹⁹. From this pool, eight representative linkers were selected (Figure 3.1), including imidazole (IM), 2-methylimidazole (2mIM), 2-nitroimidazole (2nIM), 4,5-dichloroimidazole (45dcIM), benzimidazole (bIM), 9H-purine (9pur), 5-chlorobenzimidazole (5cbIM), and 5-methyl-4-imidazolecarboxaldehyde (4a5mIM). These linkers span a range of functional groups and substitution patterns on the imidazole ring, providing diversity in both steric and electronic properties.

Each linker was reacted with zinc nitrate hexahydrate in N,N-dimethylformamide (DMF) under systematically varied conditions. Three primary reaction parameters were identified as key variables

governing crystallization behavior: total concentration, linker-to-metal ratio (expressed as the logarithm of the ratio), and reaction temperature. The boundaries of this parameter space were determined based on literature knowledge and preliminary experiments (Table 3.1 & Figure 3.2). Within this defined space, grid sampling was performed by selecting discrete values for each parameter, using six values of total concentration (0.05, 0.10, 0.15, 0.25, 0.35 and 0.50 mol L⁻¹), eight values of log linker-to-metal ratio (−0.30, 0.00, 0.30, 0.48, 0.70, 0.90, 1.00 and 1.18), and seven values of temperature (65, 85, 100, 110, 120, 130 and 140 °C). This strategy generated 336 distinct reactions for each starting linker, leading to a total of 2,688 reactions for the initial dataset. Stock solutions of metal salts and linkers were prepared at concentrations of 0.15 and 0.50 mol L⁻¹, with DMF used as the solvent and diluent.

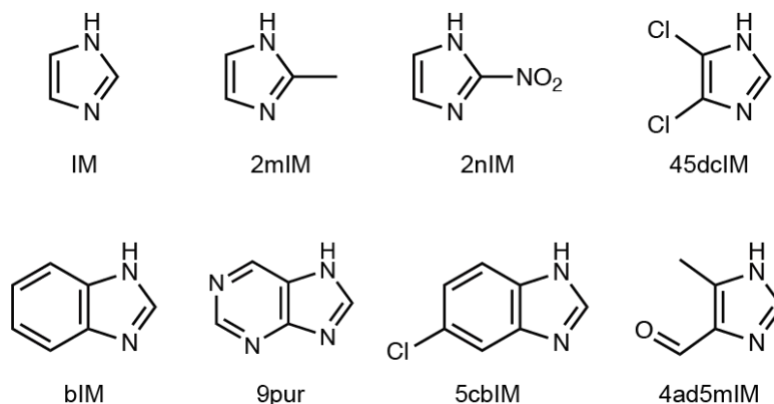


Figure 3.1. Chemical structures and abbreviations of eight starting linkers. Abbreviations denote neutral linkers when referring to reactants but represent their deprotonated forms in ZIF chemical formulae unless specified otherwise.

Table 3.1. Synthesis conditions of reported Zn-ZIFs used for determining the reaction condition space.

ZIF	DOI	Total concentration (mol/L)	Log linker-to- metal ratio	Temperature (°C)
ZIF-1	10.1073/pnas.0602439103	0.141	0.81	85
ZIF-2	10.1073/pnas.0602439103	0.141	1.18	85
ZIF-3	10.1073/pnas.0602439103	0.160	1.06	85
ZIF-4	10.1073/pnas.0602439103	0.198	0.46	130
ZIF-6	10.1073/pnas.0602439103	0.139	1.08	85
ZIF-7	10.1073/pnas.0602439103	0.066	-0.13	130
ZIF-8	10.1073/pnas.0602439103	0.085	-0.04	140
ZIF-10	10.1073/pnas.0602439103	0.160	1.06	85
ZIF-11	10.1073/pnas.0602439103	0.104	1.18	100
ZIF-11	10.1126/science.1152516	0.200	0.43	100

ZIF-14	10.1126/science.1152516	0.200	0.43	85
ZIF-20	10.1126/science.1152516	0.200	0.40	65
ZIF-20	10.1038/nmat1927	0.300	0.70	65
ZIF-60	10.1126/science.1152516	0.120	0.48	85
ZIF-61	10.1126/science.1152516	0.090	0.30	100
ZIF-62	10.1126/science.1152516	0.150	0.30	100
ZIF-64	10.1126/science.1152516	0.200	0.95	100
ZIF-68	10.1126/science.1152516	0.160	0.60	100
ZIF-68	10.1021/ja809459e	0.200	0.30	100
ZIF-69	10.1126/science.1152516	0.200	0.60	100
ZIF-69	10.1021/ja809459e	0.200	0.30	85
ZIF-70	10.1126/science.1152516	0.200	0.60	100
ZIF-70	10.1021/ja809459e	0.200	0.40	110
ZIF-71	10.1126/science.1152516	0.075	0.60	85
ZIF-71	10.1021/ja104035j	0.200	0.48	85
ZIF-72	10.1126/science.1152516	0.200	0.60	65
ZIF-73	10.1126/science.1152516	0.110	0.43	85
ZIF-74	10.1126/science.1152516	0.190	0.73	85
ZIF-76	10.1126/science.1152516	0.150	0.48	65
ZIF-77	10.1126/science.1152516	0.200	0.64	100
ZIF-78	10.1021/ja809459e	0.120	0.60	100
ZIF-79	10.1021/ja809459e	0.200	0.60	85
ZIF-80	10.1021/ja809459e	0.200	0.60	85
ZIF-81	10.1021/ja809459e	0.200	0.60	85
ZIF-82	10.1021/ja809459e	0.200	0.60	85

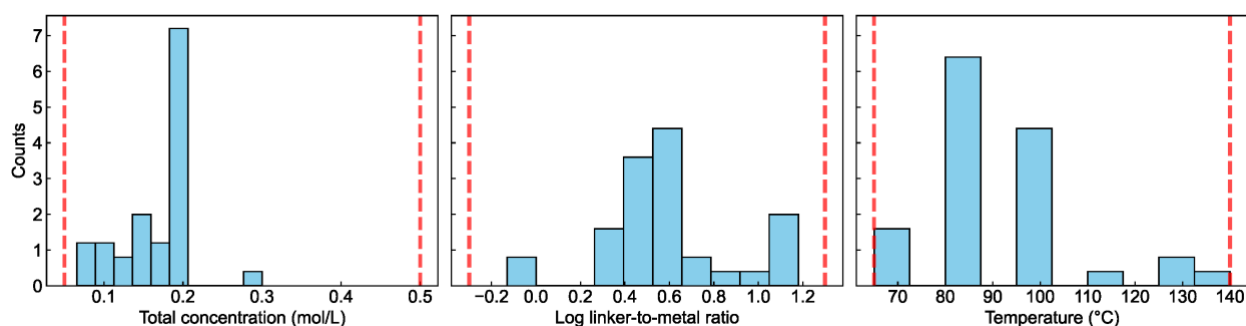


Figure 3.2. Distribution of primary synthesis parameters. Dashed vertical lines indicate the boundaries of selected ranges.

Following synthesis and characterization, reaction outcomes were assigned binary labels based on structural verification. Only reactions that produced single crystals confirmed by SCXRD corresponding to ZIF structures were labeled as successful (assigned a value of “1”). All other outcomes, including amorphous products, non-crystalline precipitates, and crystalline phases that do not correspond to ZIF frameworks, were labeled as unsuccessful (assigned a value of “0”).

To analyze the relationship between reaction conditions and crystallization outcomes, the binary results were projected onto two-dimensional parameter spaces (Figures 3.3–3.5) and one-dimensional spaces (Figures 3.6–3.8). These visualizations reveal complex, non-monotonic relationships between synthesis parameters and success rates. From a physical perspective, these patterns can be understood in terms of both thermodynamic and kinetic considerations. Classical nucleation theory suggests that optimal crystal growth requires a balance between parameters such as temperature, concentration, and supersaturation, which collectively govern nucleation and growth rates²⁰. At the same time, coordination chemistry plays a critical role, as intermediate species formed during synthesis, such as partially coordinated zinc complexes, can evolve into extended frameworks depending on their stability and reactivity²¹. The crystallization behavior varies significantly across linkers. For example, IM exhibits moderate success rates at high linker-to-metal ratios, corresponding to overlapping formation regions of multiple ZIF structures. In contrast, 2mIM shows a more localized region of successful crystallization at lower linker-to-metal ratios and elevated temperatures. Among the linkers studied, 2nIM demonstrates the greatest versatility, achieving high success rates across a wide range of conditions, particularly at intermediate linker-to-metal ratios and lower temperatures. Other linkers exhibit more constrained or distinct behaviors. For instance, 45dcIM yields successful crystallization only within a narrow region at lower temperatures. The crystallization behavior of bIM and 9pur reveals an inverse relationship, with bIM favoring low linker-to-metal ratios and 9pur favoring higher ratios. Additionally, while both 5cbIM and 4a5mIM can form multiple ZIF structures, they display different preferences in temperature and composition space, indicating that even subtle changes in functionalization can significantly influence crystallization behavior.

These observations highlight two important insights. First, the overall success rate across all linkers is approximately $25\% \pm 16\%$, indicating that random exploration of the reaction space is highly inefficient, particularly for more challenging linkers. Second, each linker exhibits a distinct and complex crystallization landscape that does not follow simple or intuitive trends. The absence of clear, generalizable patterns underscores the difficulty of relying solely on human intuition for experimental design.

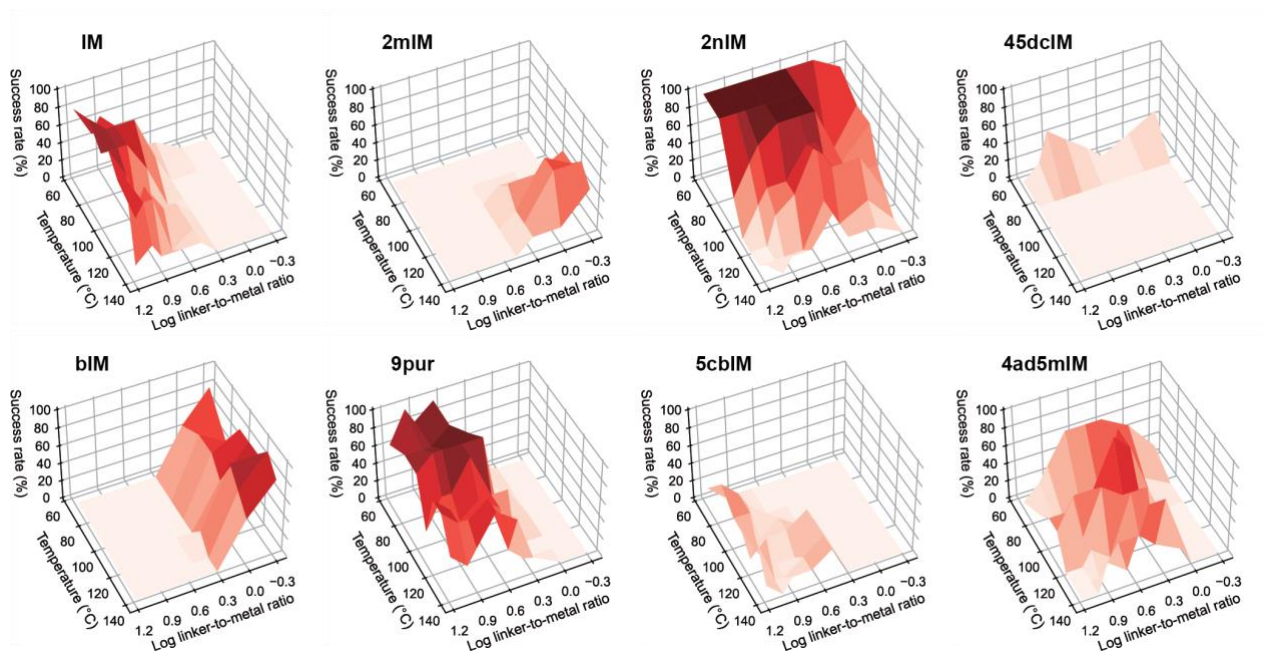


Figure 3.3. Success rate maps of individual starting linkers as functions of total concentration and log linker-to-metal ratio. The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.

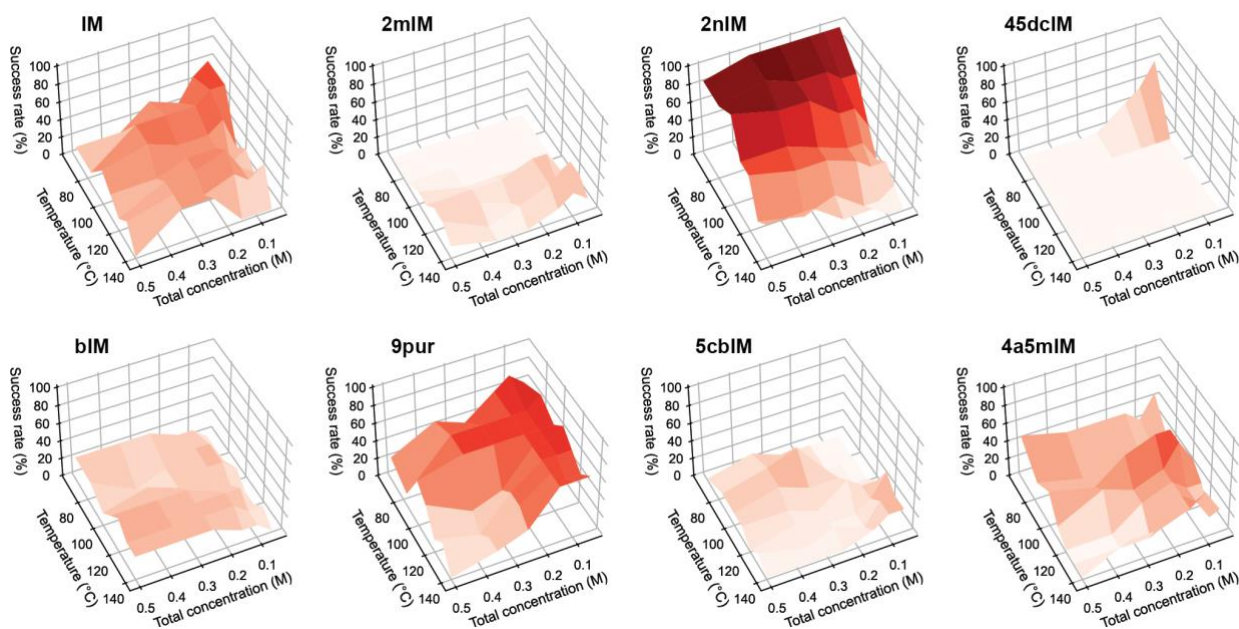


Figure 3.4. Success rate maps of individual starting linkers as functions of total concentration and temperature. The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.

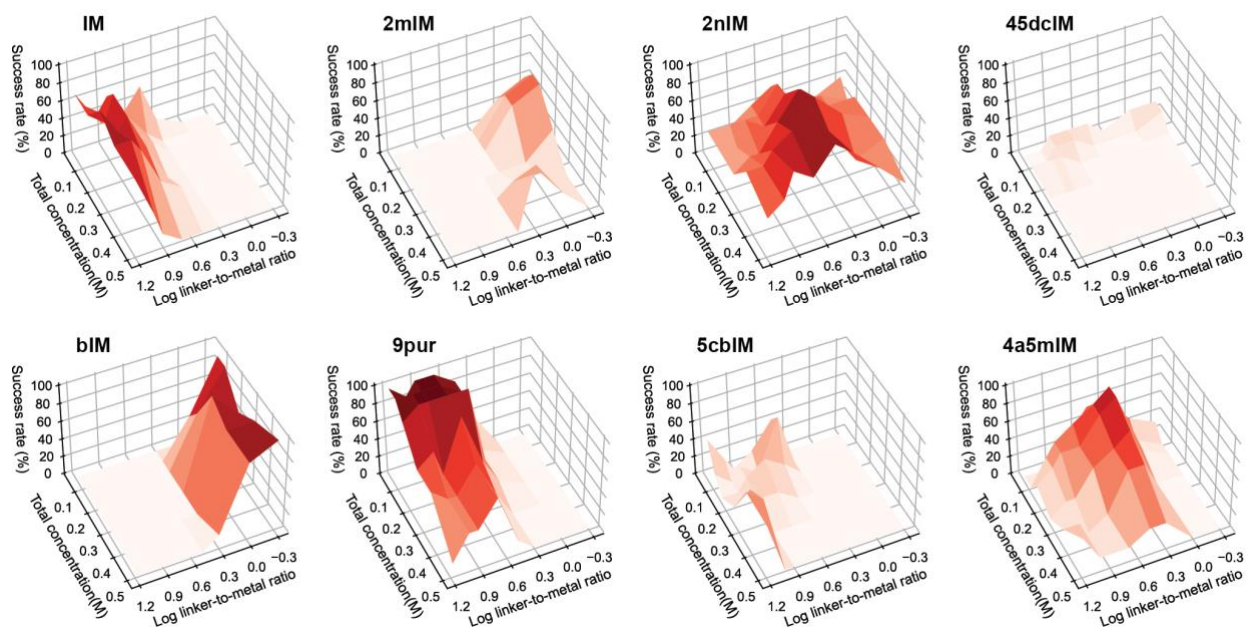


Figure 3.5. Success rate maps of individual starting linkers as functions of total concentration and log linker-to-metal ratio. The success rate is shown as surface plots for each starting linker. The color intensity corresponds to the success rate, with darker red indicating higher success rates.

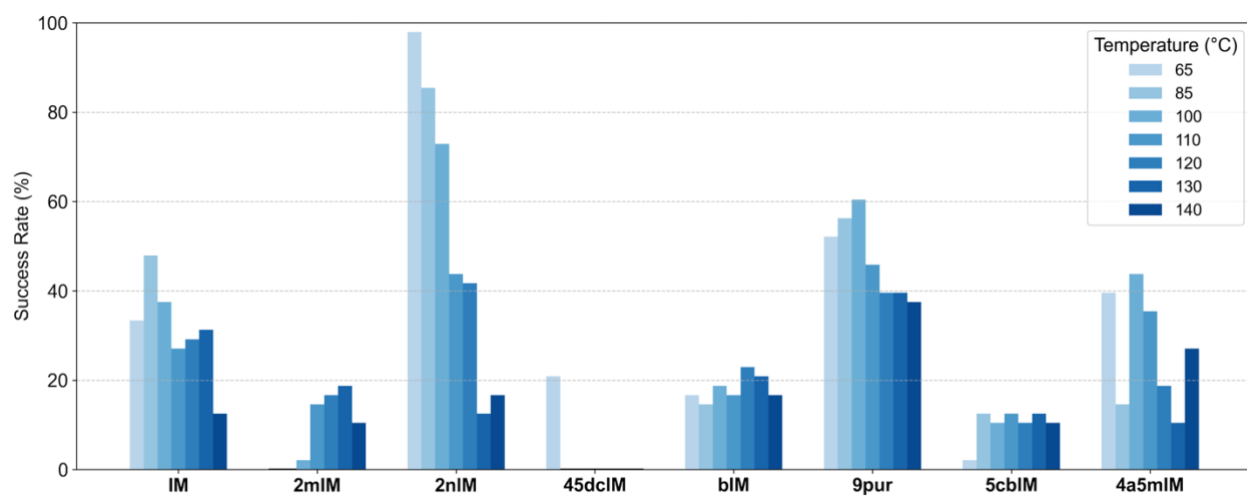


Figure 3.6. Success rates of individual starting linkers as functions of temperature. Bar plots show the success rates for each linker, with darker blue indicating higher temperature.

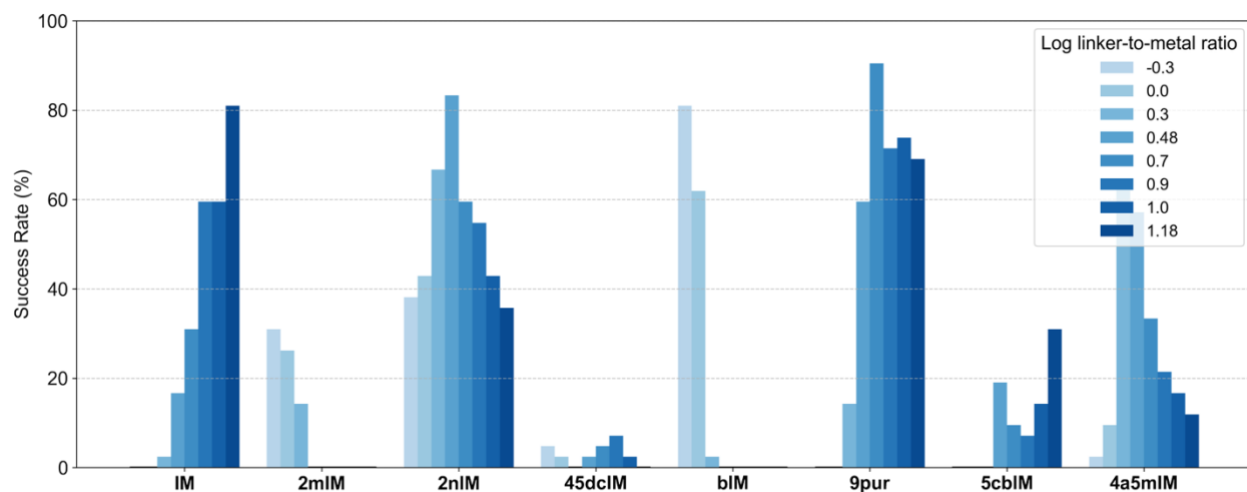


Figure 3.7. Success rates of individual starting linkers as functions of log linker-to-metal ratio. Bar plots show the success rates for each linker, with darker blue indicating higher ratio.

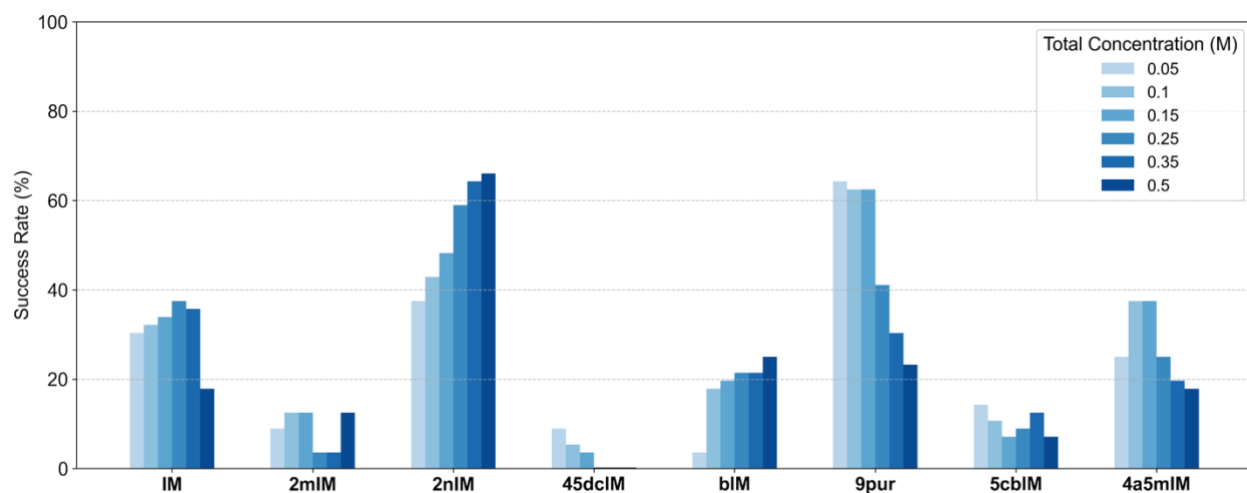


Figure 3.8. Success rates of individual starting linkers as functions of total concentration. Bar plots show the success rates for each linker, with darker blue indicating higher concentration.

3.3 Reaction Encoding

Reaction encoding is achieved by constructing feature vectors that combine linker-specific descriptors with reaction parameters. The encoding is developed from four complementary perspectives, namely functional group composition, molecular geometry, quantum mechanical properties of the linker, and quantum mechanical properties of a representative prenucleation species. These descriptors are subsequently concatenated with experimental parameters for model training. A schematic overview of the encoding strategy and a complete list of features are provided in Figure 3.9 and Table 3.2, respectively.

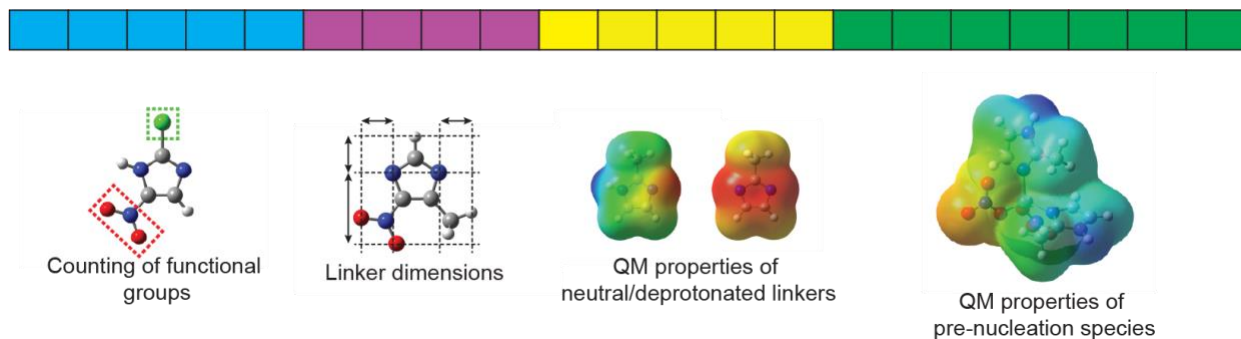


Figure 3.9. Schematic illustration of reaction encoding used for machine learning model.

Table 3.2. The names and meanings of features used in the reaction encoding.

Feature	Description
feat_HbondAcceptorCount	Total number of hydrogen bond acceptors present on the neutral linker.
feat_HalogenOnC2	Binary indicator (0 or 1) for the presence of a halogen atom at the C2 position of the linker.
feat_HalogenOnC4orC5_ IM-based	Count of halogen substituents at C4/C5 positions, applicable only to imidazole-based linkers (set to 0 for benzimidazole-based linkers).
feat_NitroCount	Total number of nitro (-NO ₂) groups present on the linker.
feat_CyanoCount	Total number of cyano (-CN) groups present on the linker.
feat_TopDistance (Å)	Distance from the furthest C2 substituent atom to the line segment connecting the two imidazole nitrogen atoms, measured using the optimized geometry of the deprotonated linker.
feat_BottomDistance (Å)	Distance from the furthest C4/C5 substituent atom to the line segment connecting the two imidazole nitrogen atoms, measured using the optimized geometry of the deprotonated linker.
feat_LeftRightDistanceMax (Å)	The larger of the horizontal distances between vertical planes containing outermost atoms at C2 or C4/C5 positions and their nearest imidazole nitrogen planes in the optimized deprotonated structure.
feat_LeftRightDistanceMin (Å)	The smaller of the horizontal distances between vertical planes containing outermost atoms at C2 or C4/C5

	positions and their nearest imidazole nitrogen planes in the optimized deprotonated structure.
feat_N1ESPChargeNeutralLowEnergy	Electrostatic potential charge on N1 from the lower-energy neutral tautomer.
feat_H1ESPChargeNeutralLowEnergy	Electrostatic potential charge on H1 (attached to nitrogen) from the lower-energy neutral tautomer.
feat_ElectronicEnergyDifference-NeutralLowEnergyCharged (kJ/mol)	Electronic energy difference between the lower-energy neutral tautomer and the deprotonated (charged) form of the linker.
feat_N1N3ESPChargeChargedMax	The more positive electrostatic potential charge between the two N atoms in the deprotonated (charged) form.
feat_N1N3ESPChargeChargedMin	The more negative electrostatic potential charge between the two N atoms in the deprotonated (charged) form.
feat_dH_in_DMF_298K	Enthalpy (ΔH) of $\text{Zn}(\text{H}_2\text{O})_4(\text{NO}_3)_2 + 2\text{L} \rightarrow \text{ZnL}_2(\text{NO}_3)_2 + 4\text{H}_2\text{O}$ at 298 K.
feat_dG_in_DMF_298K	Gibbs free energy (ΔG) of $\text{Zn}(\text{H}_2\text{O})_4(\text{NO}_3)_2 + 2\text{L} \rightarrow \text{ZnL}_2(\text{NO}_3)_2 + 4\text{H}_2\text{O}$ at 298 K.
feat_d(Zn-N)Avg	Average Zn–N bond length for all coordinating N atoms in the prenucleation cluster.
feat_A(N-Zn-N)	N–Zn–N angle formed by coordinating N atoms and the central Zn atom in the prenucleation cluster.
feat_qZn(APT)	APT charge of the central Zn atom in the prenucleation cluster.
feat_q[N]-Zn(APT)Avg	Average APT charge of coordinating N atoms bonded to Zn in the prenucleation cluster.
feat_q[N]-H(APT)Avg	Average APT charge of terminal N atoms bonded to H (not Zn) in the prenucleation cluster.
rxn_TotalConcentration (mol/L)	Combined concentration of metal ions and linkers in the reaction mixture.
rxn_Log10LinkerMetalRatio	Log10 of the linker-to-metal concentration ratio.
rxn_Temperature (°C)	Reaction temperature during heating.
rxn_MetalConcentration (mol/L)	Concentration of metal ions in the reaction mixture.
rxn_LinkerConcentration (mol/L)	Concentration of linkers in the reaction mixture.

The simplest level of chemical representation involves encoding the presence of functional groups on the linker. Such features are particularly important because specific functional groups can exhibit systematic dependencies on reaction conditions. For example, halogen substituents at the C2, C4, or C5 positions of the imidazole backbone are more susceptible to substitution reactions under synthesis conditions, whereas halogen atoms at the C4–C7 positions of the benzimidazole backbone are comparatively less reactive due to differences in electronic structure and steric environment. Similarly, cyano groups may undergo hydrolysis at elevated temperatures to form amide species, thereby diverting the system away from the intended ZIF crystallization process. By incorporating such functional group descriptors, the model can capture these chemically meaningful trends and their interactions with reaction parameters.

Beyond functional group identity, both steric and electronic effects of substituents play critical roles in determining framework topology and crystallization behavior. To capture these factors, quantum mechanical calculations are performed on both neutral and deprotonated forms of each linker. Geometry optimization is carried out using density functional theory (DFT) at the B3LYP/6-31G* level²², with solvent effects incorporated through a polarizable continuum model representing DMF²³. From the optimized deprotonated structures, geometric descriptors are extracted to quantify steric effects. Distances are measured relative to a reference axis defined by the two imidazolate nitrogen atoms, which represent the coordination sites. These descriptors include distances from substituents at the C2 and C4/C5 positions to the coordination axis, as well as lateral distances that characterize the overall width of the linker. Both maximum and minimum lateral distances are included to account for asymmetry in substitution patterns. In parallel, electronic descriptors are derived from the same DFT calculations to capture properties governing coordination and reactivity. These include electrostatic potential (ESP) charges²⁴, particularly the nitrogen atoms involved in coordination and the hydrogen atom in the neutral form. The relative stability of neutral tautomers is evaluated, and the lowest-energy configuration is used for descriptor extraction. Additionally, the electronic energy difference between the neutral and deprotonated forms is computed, providing a quantitative measure of the ease of deprotonation and, consequently, the coordination propensity of the linker.

In addition to isolated linker properties, the early stages of ZIF formation involve coordination between metal ions and linkers, leading to the formation of intermediate species. To explicitly account for these interactions, descriptors are also derived from a representative prenucleation cluster. The model system consists of a zinc center coordinated by two neutral linkers and two nitrate ions, representing a plausible intermediate in ZIF formation. For each linker type, multiple configurations with different linker orientations are generated and optimized using the M06-L functional²⁴ and the def2-SVP basis set²⁵ with the SMD implicit continuum solvent model²⁶. The most stable configuration is identified, from which the descriptors are extracted. These include the formation enthalpy and Gibbs free energy of the coordination process, average Zn–N bond lengths, coordination angles, and atomic polar tensor (APT)²⁷ charges on zinc and nitrogen atoms. These features provide a direct representation of metal–linker interactions, which are central to framework nucleation and growth. All calculations were performed in Gaussian 16²⁸.

3.4 Predictive Performance of Random Forest Classifier

To assess the ability of machine learning models to learn patterns within the crystallization dataset, first an interpolation setting was employed. In this setting, all reactions, regardless of linker identity, were pooled together and randomly shuffled to remove any ordering bias. The dataset was then split into training and validation subsets using five-fold cross-validation²⁹. In each fold, the model was trained on 80% of the data and evaluated on the remaining 20%.

Random forest is selected for making predictions. Random forest is an ensemble learning method that constructs a collection of decision trees and aggregates their predictions to improve robustness and predictive performance. Each tree is trained on a bootstrap sample of the dataset, where training data are randomly sampled with replacement. This process, known as bagging (bootstrap aggregating), reduces variance by averaging predictions across multiple independently trained models³⁰. In addition to bootstrapping, random forest introduces randomness during tree construction by selecting a subset of features at each split. This further decorrelates the trees and improves generalization performance. For classification tasks, each decision tree produces a class prediction, and the final output of the random forest is obtained by aggregating these predictions across all trees. In this case, the model outputs a probability of successful crystallization, computed as the fraction of trees that predict the positive class.

The performance of the random forest model depends on several hyperparameters that control model complexity and learning behavior. Key hyperparameters include the number of trees (`n_estimators`), the maximum depth of each tree (`max_depth`), the minimum number of samples required to split a node (`min_samples_split`), and the minimum number of samples required at a leaf node (`min_samples_leaf`). To quantify their influence, hyperparameter tuning was performed using five-fold cross-validation, with log loss as the primary evaluation metric (Figure 3.10). The implementation of random forest in scikit-learn library was used³¹. Among these parameters, `max_depth` was found to have the most significant impact on model performance. Specifically, log loss exhibits a non-monotonic dependence on `max_depth`, with an optimal value at intermediate depths. Shallow trees are insufficient to capture the complexity of the dataset, leading to underfitting, while overly deep trees tend to overfit, resulting in less well-calibrated probability estimates and increased log loss.

The interpolation performance was further evaluated across multiple metrics, including, accuracy, ROC AUC³², and recall. Model performance was compared against a baseline dummy classifier, which predicts outcomes based solely on class frequency without utilizing input features. Across all evaluated metrics except recall, the random forest model significantly outperforms the dummy classifier. Improvements in log loss indicate that the model produces better-calibrated probability estimates, while higher accuracy and ROC AUC reflect improved classification performance and discriminative ability. For recall, the dummy classifier does not provide a meaningful comparison, as its predictions are determined entirely by class imbalance and do not systematically identify positive cases.

Notably, the dependence of performance on `max_depth` differs across evaluation metrics

(Figure 3.11). While log loss reaches an optimum at intermediate depths, both accuracy and ROC AUC tend to improve monotonically with increasing `max_depth`. This difference arises from the nature of these metrics. Accuracy and recall rely on a fixed classification threshold of 0.5, and therefore primarily reflect the correctness of binary predictions. In contrast, log loss evaluates the full predicted probability distribution and penalizes overconfident or poorly calibrated predictions, making it more sensitive to overfitting. As a result, log loss provides a more rigorous assessment of model performance in probabilistic classification tasks, particularly when prediction confidence is important. Overall, these results demonstrate that, under interpolation conditions, the random forest model can capture meaningful relationships between encoded chemical features, reaction parameters, and crystallization outcomes.

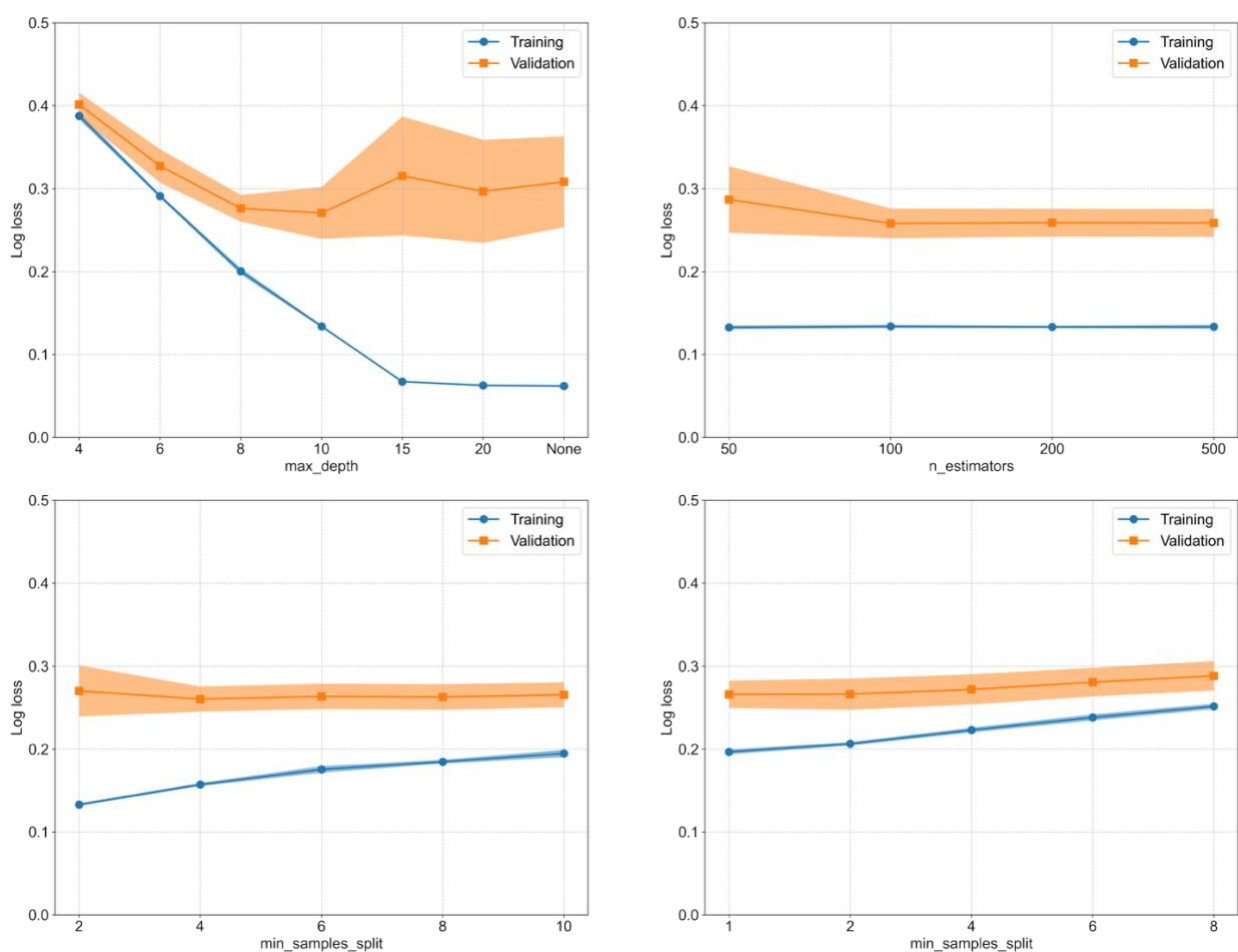


Figure 3.10. Hyperparameter dependence of random forest log loss on ZIF crystallization dataset. Training and validation log loss as a function of random forest hyperparameters, including `max_depth`, `n_estimators`, `min_samples_split`, and `min_samples_leaf`, are evaluated using five-fold cross-validation. Shaded regions indicate standard deviation across folds. For each plot, only the specified hyperparameter is varied while all others are fixed at their default values in the scikit-learn implementation.

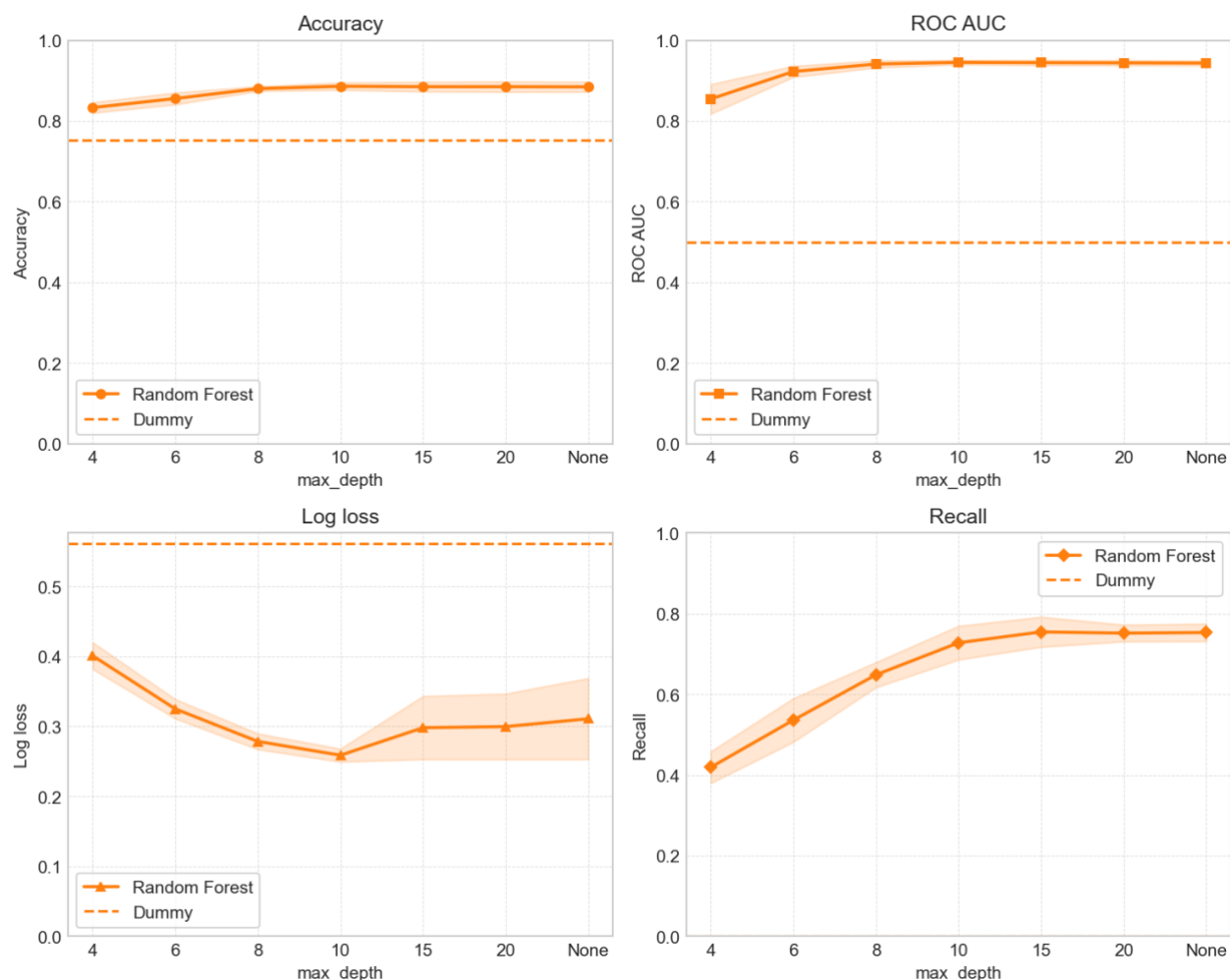


Figure 3.11. Interpolation performance of the random forest model on the ZIF crystallization dataset as a function of `max_depth`, evaluated by accuracy, ROC AUC, log loss, and recall. Shaded regions indicate standard deviation across folds. Other hyperparameters are fixed at their default values in the scikit-learn implementation.

In practical materials discovery scenarios, the chemical system of interest often has no prior experimental data. This setting differs from interpolation, where predictions are made within a previously sampled space. Instead, the task becomes one of extrapolation, where the model must make predictions for new chemical systems that lie outside the distribution of the training data. In the context of this dissertation, extrapolation arises when predicting crystallization outcomes for previously unseen linkers.

To evaluate extrapolation performance, the dataset was partitioned at the level of linkers rather than individual reactions. Given the eight linkers in the initial ZIF crystallization dataset, several leave- k -out strategies were employed, including leave-four-out, leave-three-out, leave-two-out, and leave-one-out schemes. In each case, all data associated with selected linkers were excluded from the training set and used exclusively for validation. The model was trained on the remaining linkers and evaluated on the held-out linkers, ensuring that no information about the target chemical system was present during

training. The random forest model was used for all experiments, with log loss as the primary evaluation metric due to its sensitivity to probabilistic predictions (Figure 3.12).

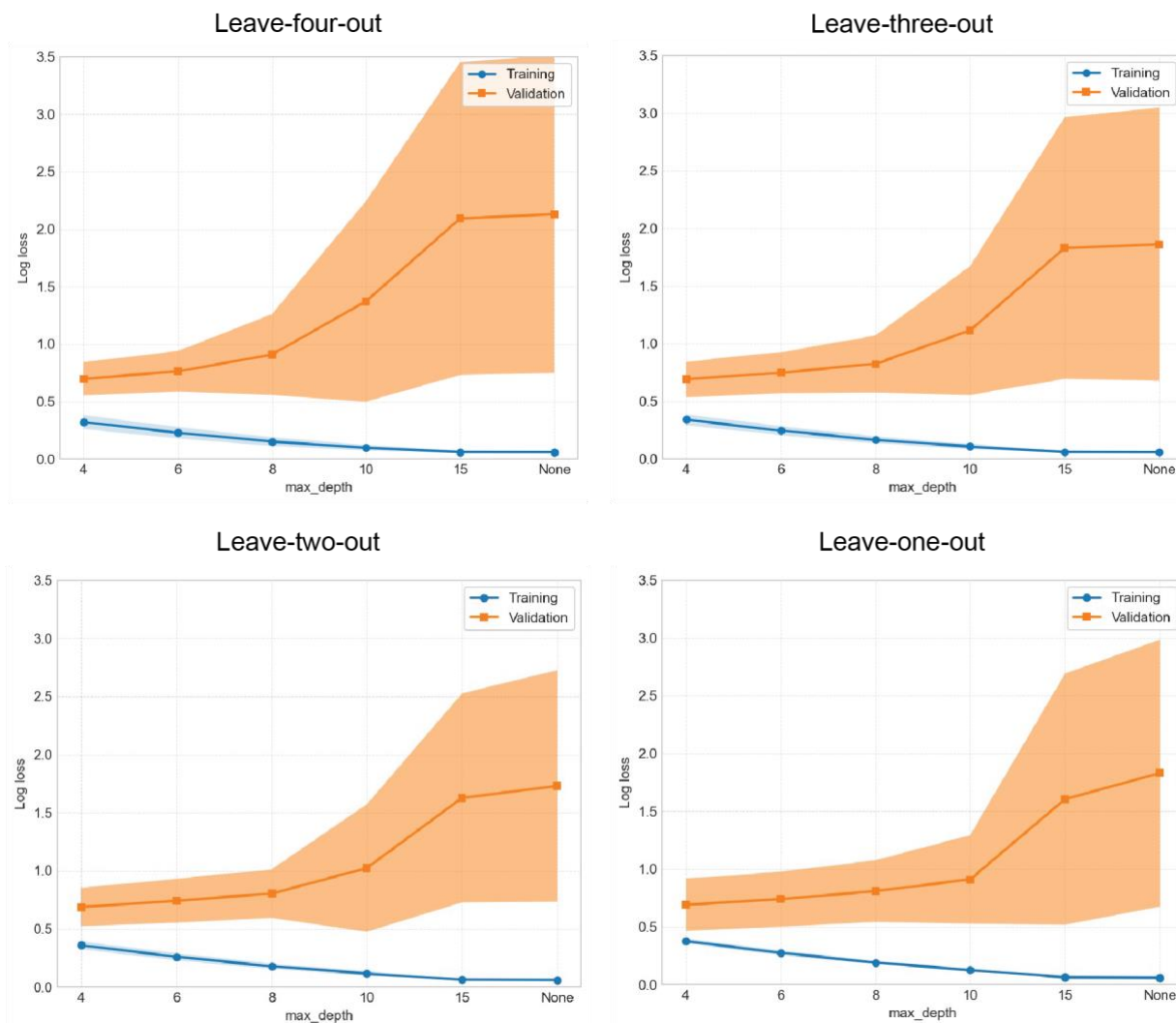


Figure 3.12. Extrapolation performance of the random forest model on the ZIF crystallization dataset as a function of max_depth using the leave- k -out scheme. Shaded regions indicate standard deviation across folds. Other hyperparameters are fixed at their default values in the scikit-learn implementation.

A consistent trend is observed across all four scenarios: increasing the maximum tree depth leads to higher validation log loss. This behavior contrasts with the interpolation setting, where intermediate tree depths provided optimal performance. The degradation in performance at larger max_depth indicates that deeper trees tend to overfit the training linkers, capturing patterns that do not generalize to unseen chemical systems. Comparing different leave- k -out schemes reveals that increasing the number of linkers in the training set provides some improvement in extrapolation performance. However, this improvement is modest.

Although reticular materials discovery often begins as an extrapolation problem, it is inherently iterative. As new experimental data are generated for previously unexplored linkers, these data can be incorporated into the training set, gradually reducing the distributional gap between training and target systems. In this process, the prediction task transitions from purely extrapolative toward interpolation within an expanded chemical space. Given that the initial dataset contains only eight linkers, the observed modest improvements with increasing training diversity suggest that incorporating additional linkers is expected to further enhance predictive performance. This highlights the importance of coupling machine learning models with experimental platforms in a closed-loop framework, where iterative data acquisition enables progressively more accurate and generalizable predictions.

3.5 References

- (1) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76.
- (2) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 1st ed.; Methods and Principles in Medicinal Chemistry; Wiley, 2009.
- (3) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.
- (4) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
- (5) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.
- (6) Gori, M.; Monfardini, G.; Scarselli, F. A New Model for Learning in Graph Domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*; IEEE: Montreal, Que., Canada, 2005; Vol. 2, pp 729–734.
- (7) Scarselli, F.; Gori, M.; Ah Chung Tsoi; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20* (1), 61–80.
- (8) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, 28.
- (9) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

- (10) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminformatics* **2021**, *13* (1), 12.
- (11) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4* (3), 279–287.
- (12) Corso, G.; Stark, H.; Jegelka, S.; Jaakkola, T.; Barzilay, R. Graph Neural Networks. *Nat. Rev. Methods Primer* **2024**, *4* (1), 17.
- (13) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
- (14) Bjerrum, E. J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). arXiv 2017.
- (15) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, *10* (6), 1692–1701.
- (16) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443.
- (17) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190.
- (18) Huo, H.; Bartel, C. J.; He, T.; Trewartha, A.; Dunn, A.; Ouyang, B.; Jain, A.; Ceder, G. Machine-Learning Rationalization and Prediction of Solid-State Synthesis Conditions. *Chem. Mater.* **2022**, *34* (16), 7323–7336.
- (19) Zheng, Z.; Rong, Z.; Nguyen, H. L.; Yaghi, O. M. Structural Chemistry of Zeolitic Imidazolate Frameworks. *Inorg. Chem.* **2023**, *62* (51), 20861–20873.
- (20) Kashchiev, D. *Nucleation: Basic Theory with Applications*; Butterworth Heinemann: Oxford ; Boston, 2000.
- (21) Lim, I. H.; Schrader, W.; Schüth, F. Insights into the Molecular Assembly of Zeolitic Imidazolate Frameworks by ESI-MS. *Chem. Mater.* **2015**, *27* (8), 3088–3095.
- (22) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627.
- (23) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3094.
- (24) Zhao, Y.; Truhlar, D. G. A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Phys.* **2006**, *125* (19), 194101.

- (25) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*(18), 3297.
- (26) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396.
- (27) Cioslowski, J. A New Population Analysis Based on Atomic Polar Tensors. *J. Am. Chem. Soc.* **1989**, *111* (22), 8333–8336.
- (28) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Revision A.03, 2016.
- (29) Kohavi, R.; others. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Ijcai*; Montreal, Canada, 1995; Vol. 14, pp 1137–1145.
- (30) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140.
- (31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. **2012**.
- (32) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27* (8), 861–874.

Chapter 4

Algorithmic Decision-Making for Accelerating Reticular Materials Discovery

4.1 Introduction

Discovery differs fundamentally from conventional optimization tasks. Rather than maximizing a continuous property or refining an existing system, the objective is to identify at least one successful instance of a desired structure. Once such a structure is discovered, further optimization does not contribute additional value in its existence. This distinction necessitates a different approach to decision-making. For crystallization, outcomes are naturally represented as binary labels indicating whether a reaction yields the desired single crystals. As shown in Chapter 3, probabilistic machine learning models can be used to estimate the likelihood of crystallization success. More generally, in data-driven decision-making, classification models provide probabilities of discrete outcomes, while regression models can quantify predictive uncertainty, both of which enable principled strategies for balancing exploration of uncertain regions and exploitation of high-confidence candidates. Such probabilistic or uncertainty-aware predictions have been widely applied across chemistry and materials science, including organic molecules¹⁻³, metal complexes^{4,5}, perovskites⁶⁻⁸, and nanoparticles⁹⁻¹¹. Under the present discovery-oriented formulation, these probabilities are directly used to prioritize experiments, focusing on selecting conditions that maximize the likelihood of obtaining at least one successful outcome. Accordingly, a greedy policy is adopted, in which candidate experiments are ranked based on their predicted probability of success and selected in descending order to guide experimental execution.

Decision-making is integrated with the experimental platform in a closed-loop framework. Suggested experiments are executed through automated synthesis and characterization, and the resulting outcomes are incorporated into the dataset to update the model. Such self-driving lab paradigms have been increasingly explored to accelerate scientific discovery¹²⁻¹⁵. This iterative process enables progressive refinement of predictive capability as the model adapts to newly explored regions of chemical space. Moreover, the high-throughput nature of the platform requires experiments to be conducted in batches, making batch design an essential component for maximizing discovery efficiency¹⁶.

Parts of this chapter have been adapted from:

Rong, Z.; Chen, Z.; Luong, F.; Chheda, S.; Luong, H. T. N.; Zheng, Z.; Greco, K.; Alghamdi, A. A.; Bui, K. H.; Jaffrelot Inizan, T.; Nguyen-Dang, T.; Pham, H. H.; Le, D. D.; Sauer, J.; Phung, V. B. T.; Chayes, J. T.; Borgs, C.; Boley, M.; El Ghaoui, L.; Yaghi, O. M. Algorithmic Iterative Reticular Synthesis of Zeolitic Imidazolate Framework Crystals. *Nat. Synth.* **2025**, 5 (3), 409–419.

This chapter presents the algorithmic decision-making framework for discovery objective, including the acquisition function and batch selection formalism. To evaluate this framework, 48 new candidate linkers, spanning diverse chemical structures, were selected and encoded using the feature representation described in Chapter 3. Discovery performance on the new linker set is evaluated and compared with baseline strategies, followed by analysis of model suggestions and ablation studies on feature representations. Together, these results demonstrate how algorithmic decision-making, coupled with automated experimentation, can accelerate reticular materials discovery.

4.2 Acquisition Function and Batch Formalism

The objective of the decision-making process for the discovery objective is to maximize the number of unique discoveries within a limited experimental budget. This objective differs from traditional optimization frameworks, such as Bayesian optimization¹⁷ and multi-armed bandits¹⁸, which are designed to maximize success probability or cumulative (potentially non-unique) rewards. The discovery objective naturally leads to a greedy decision-making strategy, in which experiments are prioritized based on their predicted probability of success.

Given a probabilistic machine learning model, each candidate reaction is associated with a predicted probability of yielding a successful crystallization outcome. The acquisition function is therefore defined directly in terms of this probability, ranking candidates according to their likelihood of success. In the sequential setting, the greedy algorithm operates as follows:

1. Initialization phase
 - Construct initial model M using training data D_{train}
 - Define experiment space X containing all possible reaction candidates
2. Discovery cycle (repeated until budget exhaustion)
 - Selection: choose experiment x_t maximizing predicted success probability from M
 - Execution: conduct the experiment and record outcome y_t
 - Update:
 - Incorporate new datapoint (x_t, y_t) into D_{train}
 - For successful outcomes ($y_t = 1$), remove all remaining experiments with the same linker from X
 - Retrain model M using updated dataset

In practice, experiments are conducted in batches due to the high-throughput nature of the experimental platform. This introduces an additional challenge: selecting a set of experiments that collectively maximize discovery performance. A naive extension of the greedy strategy would independently select the top-ranked candidates. However, this approach neglects dependencies between candidates, particularly when multiple experiments involve the same linker. Therefore, batch selection must account for redundancy within the batch. To address this, a batch construction strategy is developed as follows:

1. Initialization phase
 - Construct initial model M using training data D_{train}
 - Define experiment space X containing all possible reaction candidates
2. Batch construction cycle
 - Initialize: create empty batch $B_t = \{\}$
 - Batch assembly: for $b = 1$ to N_B :
 - Calculate expected reward $R(x)$ for all remaining candidates $x \in X$
 - Select candidate x_t^b maximizing $R(x)$
 - Add selected candidate to batch: $B_t \leftarrow B_t \cup \{x_t^b\}$
 - Update reward function $R(x)$
 - Remove selected candidate from consideration: $X \leftarrow X \setminus \{x_t^b\}$
 - Execution: conduct all experiment in B_t and record outcome $\{y_t^b\}$
 - Update:
 - Incorporate new datapoint $(\{x_t^b\}, \{y_t^b\})$ into D_{train}
 - For successful outcomes ($y_t^b = 1$), remove all remaining experiments with the same linker from X
 - Retrain model $\mathcal{M}$ using updated dataset

The reward function $R(x)$ is defined as follows:

$$\begin{aligned}
 R(x_{\text{new}}) &= \mathbb{P} [\text{success with } x_{\text{new}}] - \mathbb{P} [\text{success without } x_{\text{new}}] \\
 &= \mathbb{P} [\exists x \in B_t^l \cup \{x_{\text{new}}\} : Y(x) = 1] - \mathbb{P} [\exists x \in B_t^l : Y(x) = 1] \\
 &= \mathbb{P} [Y(x_{\text{new}}) = 1 \mid \forall x \in B_t^l : Y(x) = 0] \times \mathbb{P} [\forall x \in B_t^l : Y(x) = 0]
 \end{aligned}$$

where B_t^l denotes the set of previously selected candidates in the current batch that share the same linker l , and $Y(x)$ represents the random outcome of candidate x . Accordingly, estimating the expected reward reduces to evaluating two quantities: a conditional probability and a joint probability. The joint probability can be computed using the chain rule, decomposing it into a product of conditional probabilities that can be estimated individually:

$$\mathbb{P} [\forall x \in B_t^l : Y(x) = 0] = \prod_{k=1}^l \mathbb{P} [Y(x_t^{b_k}) = 0 \mid Y(x_t^{b_{k-1}}) = \dots = Y(x_t^{b_1}) = 0]$$

Estimating this marginal contribution requires evaluating conditional and joint probabilities. In practice, the conditional probabilities are approximated by temporarily augmenting the training dataset with virtual failure observations for previously selected candidates associated with the same linker, followed by retraining the model. While this procedure is exact for fully probabilistic models such as Gaussian process classifiers, it serves as an effective heuristic for models such as random forests, enabling practical estimation of conditional probabilities.

Despite the need for repeated model updates within each batch, the computational cost remains comparable to the sequential setting. For each candidate addition, the expected reward contributions of

remaining candidates sharing the same linker are updated by refitting the model with augmented data. By storing intermediate conditional probabilities, joint probabilities can be computed efficiently.

The batch formalism also accommodates practical experimental constraints. For example, in the present system, reactions within a batch must share the same temperature due to equipment limitations. This constraint is incorporated by restricting the candidate search space during batch construction. Specifically, candidate batches are evaluated across a discrete set of allowable temperatures, and the batch configuration that maximizes the expected reward is selected. These constraints are enforced without altering the underlying greedy principle.

4.3 Discovery Performance

To evaluate the effectiveness of the machine learning-guided discovery framework, a set of 48 new linker candidates was selected to define an unexplored chemical space. The selection process was designed to be largely unbiased, subject to two practical constraints. First, linkers were required to exhibit sufficient solubility in the solvent, which is DMF in this case, at ambient temperature to ensure compatibility with the robotic liquid handling system. Second, linkers containing strongly coordinating functional groups, such as hydroxyl or amino groups, were excluded to avoid competition with imidazole nitrogen atoms for metal coordination, which could lead to undesired side reactions. The 48 new linkers are divided between imidazole- and benzimidazole-based structures (Figure 4.1). Among these, 2-bromoimidazole (2BrIM), 2-cyanoimidazole (2cyIM), and 2-propylimidazole (2pIM) have previously been reported in Zn-ZIF synthesis under alcohol-based conditions^{19,20}. The remaining 45 linkers have not been independently synthesized into ZIF single crystals.

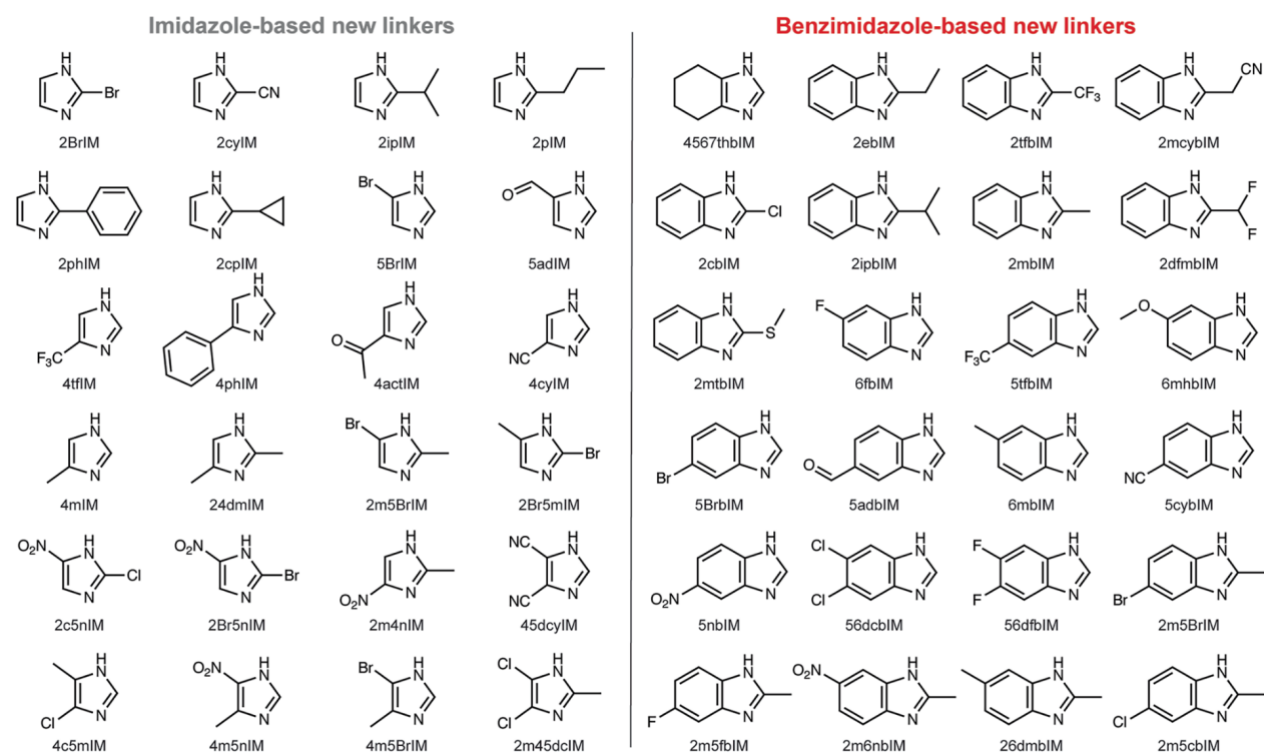


Figure 4.1. Chemical structures and abbreviations of 48 new linkers.

The newly selected linkers were encoded using the same feature representation described in Chapter 3, ensuring consistency between training and prediction. Principal component analysis (PCA) of the linker features reveals clustering of imidazole- and benzimidazole-based linkers (Figure 4.2). However, the relatively low explained variance (50.8% captured by the first two principal components) underscores the intrinsic complexity of the chemical space, which cannot be adequately described by simple linear combinations of features (Figure 4.3). Given the higher cost of these linkers and insights gained from the initial crystallization dataset, a reduced sampling was employed while maintaining coverage of the same reaction condition space. Specifically, reactions were conducted across two values of total concentration (0.15 and 0.40 mol L⁻¹), six values of the logarithmic linker-to-metal ratio (0.00, 0.25, 0.50, 0.75, 1.00, and 1.25), and three temperature settings (70, 100, and 130 °C). This design resulted in 36 candidate reactions per new linker, compared to 336 reactions per linker in the initial dataset.

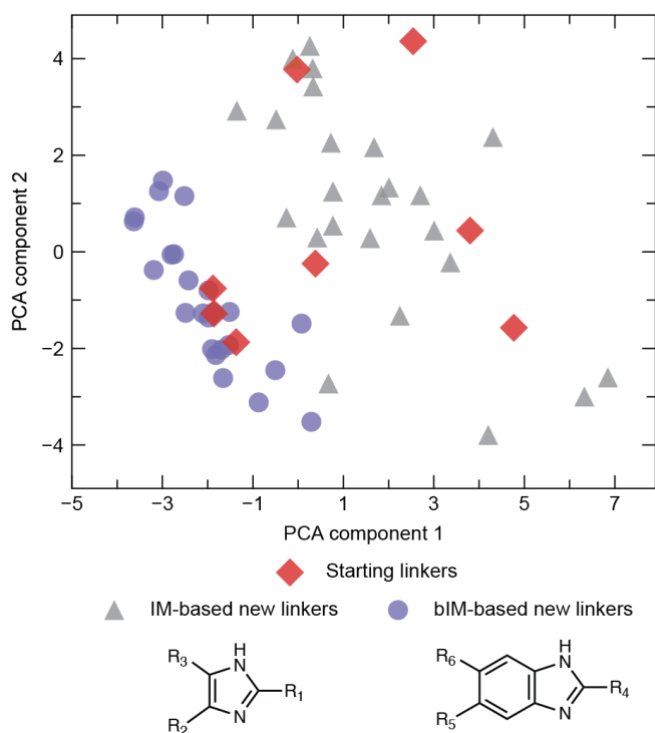


Figure 4.2. Two-dimensional PCA projection of the chemical space, showing the distribution of 8 starting linkers (red diamonds), 24 new IM-based (grey triangles) and 24 new bIM-based linkers (purple circles).

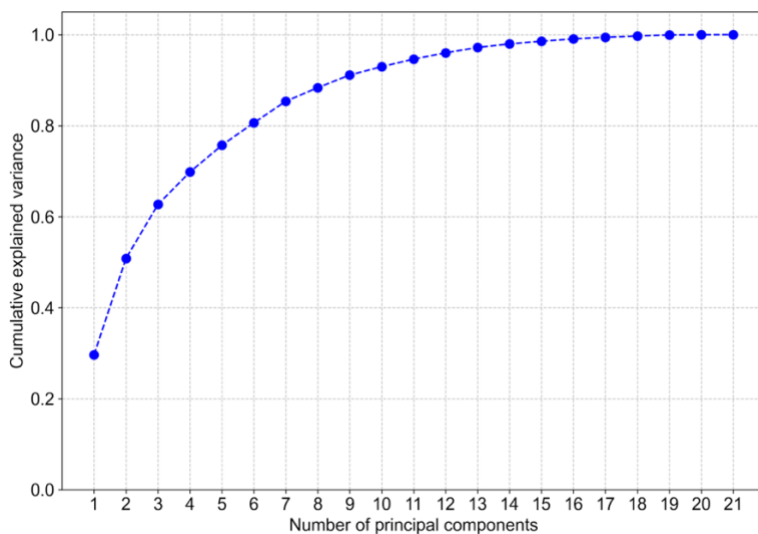


Figure 4.3. Cumulative explained variance ratio as a function of principal components for linker features.

Two probabilistic machine learning models were employed for guiding discovery: random forest classifier (RFC)²¹ and Gaussian process classifier (GPC)^{22,23}. The RFC model captures complex, nonlinear relationships through an ensemble of decision trees, while the GPC model provides naturally

calibrated probability estimates through a kernel-based probabilistic framework. A key challenge in this setting arises from data imbalance. Each initial training linker is associated with hundreds of data points, whereas new linkers are explored sequentially with only a small number of experiments. This imbalance, combined with the inherent distributional shift between known and new linkers, can lead to diminished sensitivity of the model to newly acquired data.

To address this issue, two strategies were implemented. First, data augmentation techniques were used to increase the influence of newly obtained experimental results during training. The degree of augmentation reflects the similarity between new and training linkers. When a new linker resembles known systems, minimal augmentation is required; when it differs substantially, stronger augmentation helps the model prioritize learning from new data. An augmentation factor of 10 was selected, approximately matching the ratio between the number of experiments per initial (336) and new linker (36), providing a balance between leveraging prior knowledge and adapting to new chemical space. Second, dynamic hyperparameter optimization was applied to adjust model complexity as new data were incorporated, effectively increasing the weight of recent observations. Every 100 iterations, five-fold cross-validation was performed on newly collected data to determine the optimal maximum tree depth, using log loss as the objective. Other hyperparameters were kept at their default values. The GPC model provides an alternative probabilistic framework based on a latent Gaussian process, offering naturally calibrated probability estimates and uncertainty quantification. Analysis of the initial dataset revealed distinct behaviors: outcomes vary smoothly with reaction conditions for a fixed linker but exhibit less continuity across linker identities. To capture this structure, a composite kernel was employed:

$$k(x, x') = k_{\text{RBF}}(x_e, x_e') \times k_{\text{Mat52}}(x_l, x_l')$$

where k_{RBF} denotes a radial basis function kernel applied to the experimental features x_e , capturing smooth variations across reaction conditions, and k_{Mat52} represents a Matérn-5/2 kernel applied to the linker features x_l , allowing for less smooth variations across molecular structures. Posterior probabilities and hyperparameters were computed using variational inference implemented in the GPyTorch library²⁴. While for RFC, probability estimate was achieved through the proportion of trees predicting success and the implementation in the scikit-learn library was used²⁵.

Although more complex models such as neural networks could in principle be applied, the limited number of unique linkers in the training dataset makes such approaches less suitable. RFC and GPC were therefore selected for their robustness, interpretability, and ability to provide uncertainty-aware predictions in low-data regimes.

To rigorously compare machine learning-guided discovery with random exploration, multiple trials are required to account for stochastic variation. Because repeated experimental campaigns are resource intensive, an equivalent validation strategy was adopted. A comprehensive grid of reaction conditions was first executed experimentally for all 48 new linkers, generating a dataset of 1,728 reactions. This dataset enables simulation of different discovery strategies by iteratively selecting candidate experiments and retrieving their outcomes, eliminating the need for repeated physical experiments. Using this framework, both machine learning-guided and random search (RS) strategies

were evaluated under identical conditions.

Distinct discovery behaviors were observed across different strategies (Figure 4.4). Random search required approximately 1,400 experiments to identify all ZIF-forming systems among the 10 successful linkers, exhibiting a gradually diminishing discovery rate characteristic of exhaustive exploration. The GPC-based approach demonstrated rapid early discoveries but plateaued at a performance level comparable to random search. This behavior likely reflects the difficulty of extrapolating from the initial dataset to chemically distinct linkers. In contrast, RFC achieved substantially improved performance, identifying all successful systems in approximately 700 experiments, representing a twofold reduction in experimental effort compared to random search. This result demonstrates the effectiveness of RFC in capturing relevant patterns and prioritizing promising candidates within the chemical space.

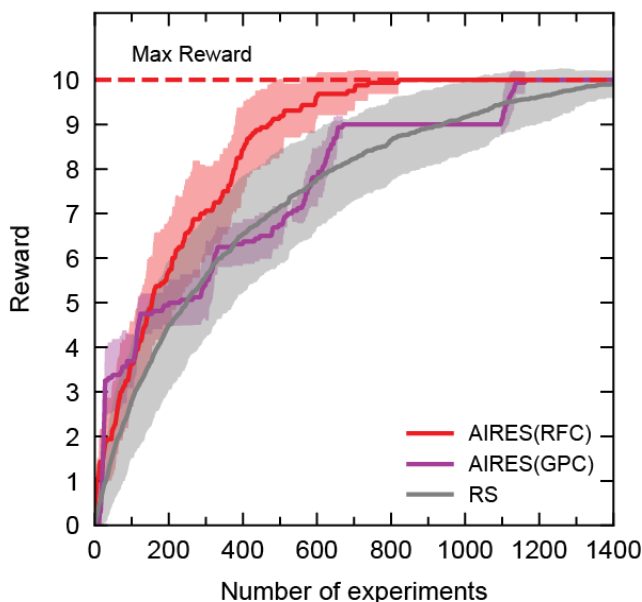


Figure 4.4. Reward (number of distinct discoveries) of AIRES equipped with RFC (red) and GPC (magenta), and RS under the sequential setting. Solid lines represent the mean, and shaded areas represent the range of ± 1 standard deviation (16 repetitions for RFC/GPC; 100 repetitions for RS). The red dashed line indicates the maximum achievable reward.

The batch implementation preserves the efficiency of the RFC-based approach while enabling high-throughput experimentation. Across a range of batch sizes, the batch algorithm consistently required approximately half the number of experiments compared to random search to achieve the same number of discoveries (Figure 4.5). This result is non-trivial. A naive strategy that selects the top-k candidates based solely on predicted probability introduces significant redundancy, as multiple experiments may target the same linker. Such redundancy leads to substantially worse performance than even batch random search (Figure 4.6). In contrast, the proposed batch selection strategy explicitly

accounts for dependencies between candidates through conditional probability estimation, ensuring that experiments within a batch provide complementary information. This enables efficient extension of machine learning-guided discovery to parallel experimental settings.

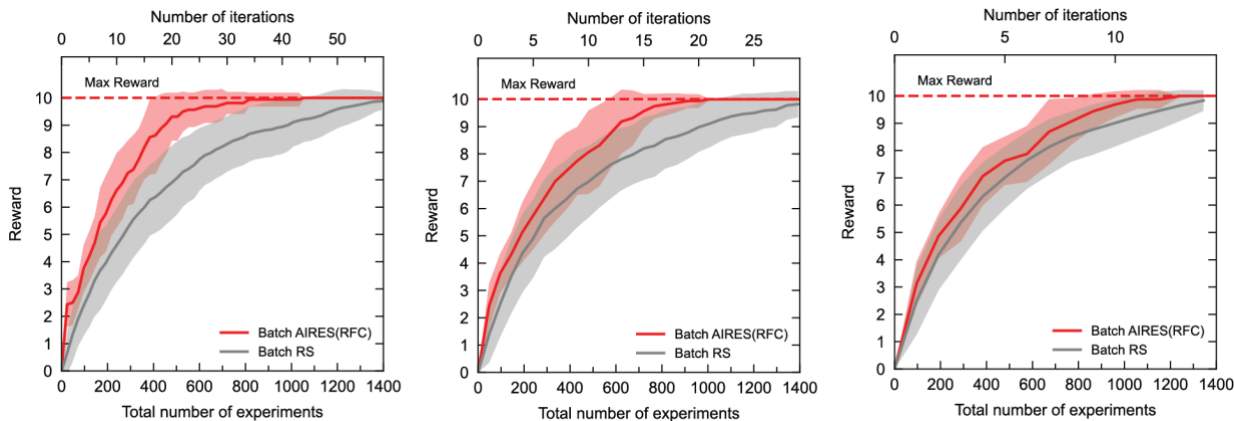


Figure 4.5. Reward (number of distinct discoveries) of Batch AIRES(RFC) and Batch RS under different batch sizes. Solid lines represent the mean, and shaded areas represent the range of ± 1 standard deviation (16 repetitions for Batch AIRES(RFC); 100 repetitions for Batch RS). Batch sizes: 24 reactions per iteration (left), 48 reactions per iteration (middle), and 96 reactions per iteration (right).

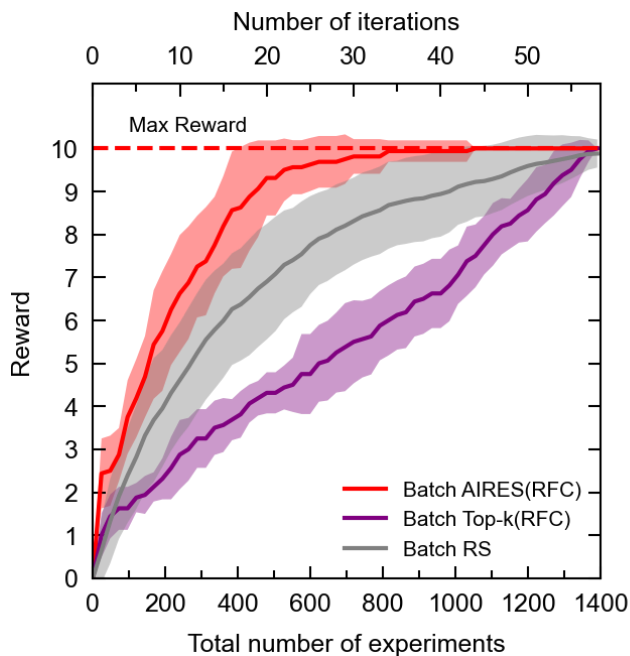


Figure 4.6. Reward (number of distinct discoveries) of Batch AIRES(RFC), Batch Top-k(RFC) and Batch RS under batch size of 24. Solid lines represent the mean, and shaded areas represent the range of ± 1 standard deviation (16 repetitions for Batch AIRES(RFC) and Batch Top-k(RFC); 100 repetitions for Batch RS). The procedures of data augmentation and hyperparameter tuning are the same for Batch AIRES(RFC) and Batch Top-k(RFC).

Further insight is obtained by examining the distribution of experiments required to discover each successful linker based on the results of the sequential setting (Figure 4.7). For some linkers, such as 2cyIM, both random search and RFC achieved early discovery, indicating that crystallization occurs under relatively broad conditions. In such cases, the advantage of model guidance is modest but still present. For more challenging systems, such as 6mbIM, the difference is more pronounced. Random search discovers these systems at a roughly uniform rate through exhaustive exploration, whereas RFC significantly accelerates discovery by identifying patterns in the reaction space. An interesting exception is 2pIM, where random search outperformed RFC. Although this linker exhibits moderate overall success rates, its structural features and resulting ZIF topology differ substantially from those in the training data. As a result, the model assigns lower predicted probabilities and deprioritizes these candidates, while random search benefits from uniform sampling. Despite such exceptions, the overall performance demonstrates that the AIRES framework, particularly when using RFC, consistently outperforms random exploration and enables substantial acceleration of reticular materials discovery.

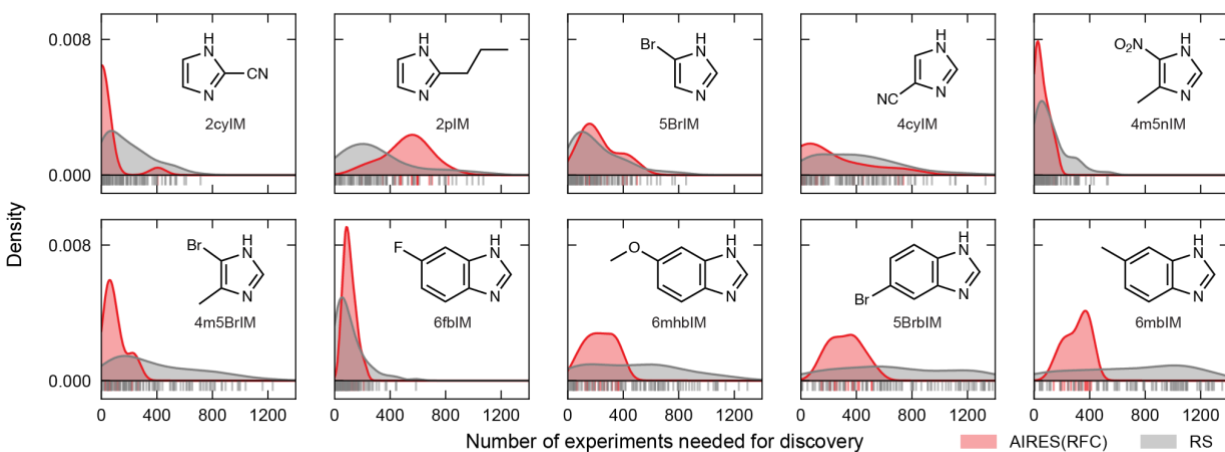


Figure 4.7. Kernel density estimation showing the distribution of the number of experiments required to discover new ZIF crystals for the ten successful linkers using AIRES (RFC) versus RS. Horizontal bars beneath each distribution indicate individual discovery events across multiple repetitions.

4.4 Analysis on Predictions and Feature Selections

Further analysis of the AIRES (RFC) discovery trajectory (Figure 4.8) reveals that the model effectively captures crystallization tendencies and prioritizes promising candidates during exploration. This behavior is reflected in the non-uniform frequency of suggested linkers, which contrasts sharply with the uniform sampling pattern of random search. Certain linkers are sampled more frequently, indicating that the model identifies them as having higher potential for ZIF formation. This prioritization arises from the model's ability to learn chemically meaningful patterns from the feature space. For example, linkers with bulky substituents at the 2 and 4/5 positions are suggested less frequently, consistent with steric hindrance around the metal center that disfavors framework

formation. Such trends demonstrate that the model can capture physically relevant constraints governing crystallization.

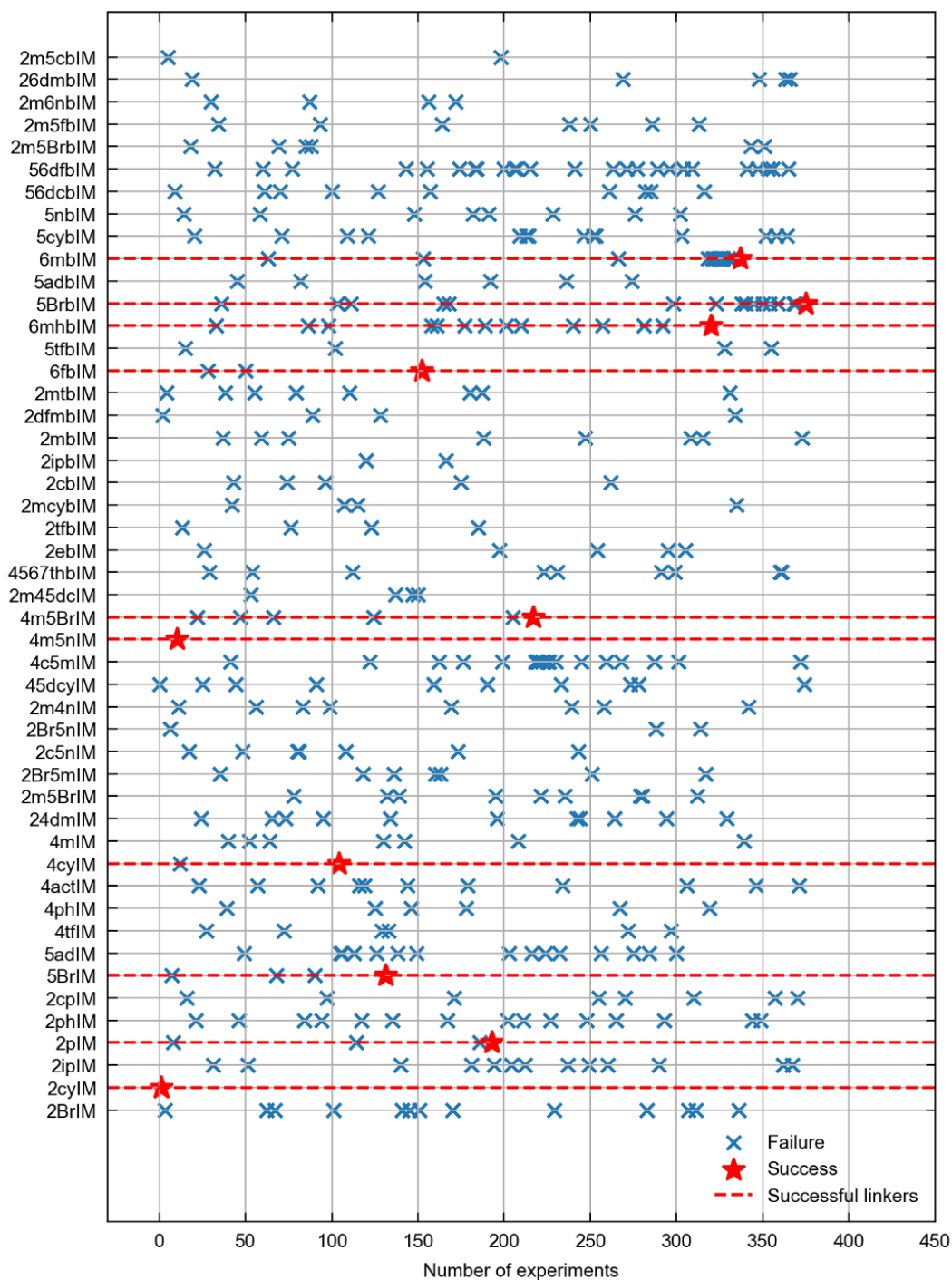


Figure 4.8. Linker suggestions from a single sequential trajectory of RFC. Each linker suggestion is paired with specific reaction conditions (not shown in this plot).

A deeper understanding of model behavior is obtained by examining predicted crystallization landscapes across the reaction condition space. For each successful linker, probability heatmaps are constructed at two key timepoints: iteration 0, where the model is trained only on the initial dataset, and the final iteration prior to discovery, where the model incorporates all previously acquired data for that linker (Figures 4.9–4.18). In these visualizations, white crosses indicate previously tested conditions that resulted in negative outcomes, while the red star marks the experiment that ultimately yielded a successful crystal. For consistency, the heatmaps are generated using an RFC model retrained on the trajectory data, with the maximum tree depth matched to that used during the discovery process. Although this retrospective analysis introduces minor numerical differences, it preserves the overall trends in the predicted landscapes. Two distinct model behaviors are observed. For some linkers, such as 2cyIM and 4m5nIM, the model rapidly forms a well-defined probability landscape early in the process, enabling accurate identification of favorable regions and leading to early discovery. In contrast, for more challenging linkers such as 6mhbIM and 6mbIM, the predicted landscapes remain relatively flat throughout much of the exploration. Despite this uncertainty, the model still identifies these linkers as promising and allocates additional experimental resources to them, ultimately accelerating discovery compared to random search.

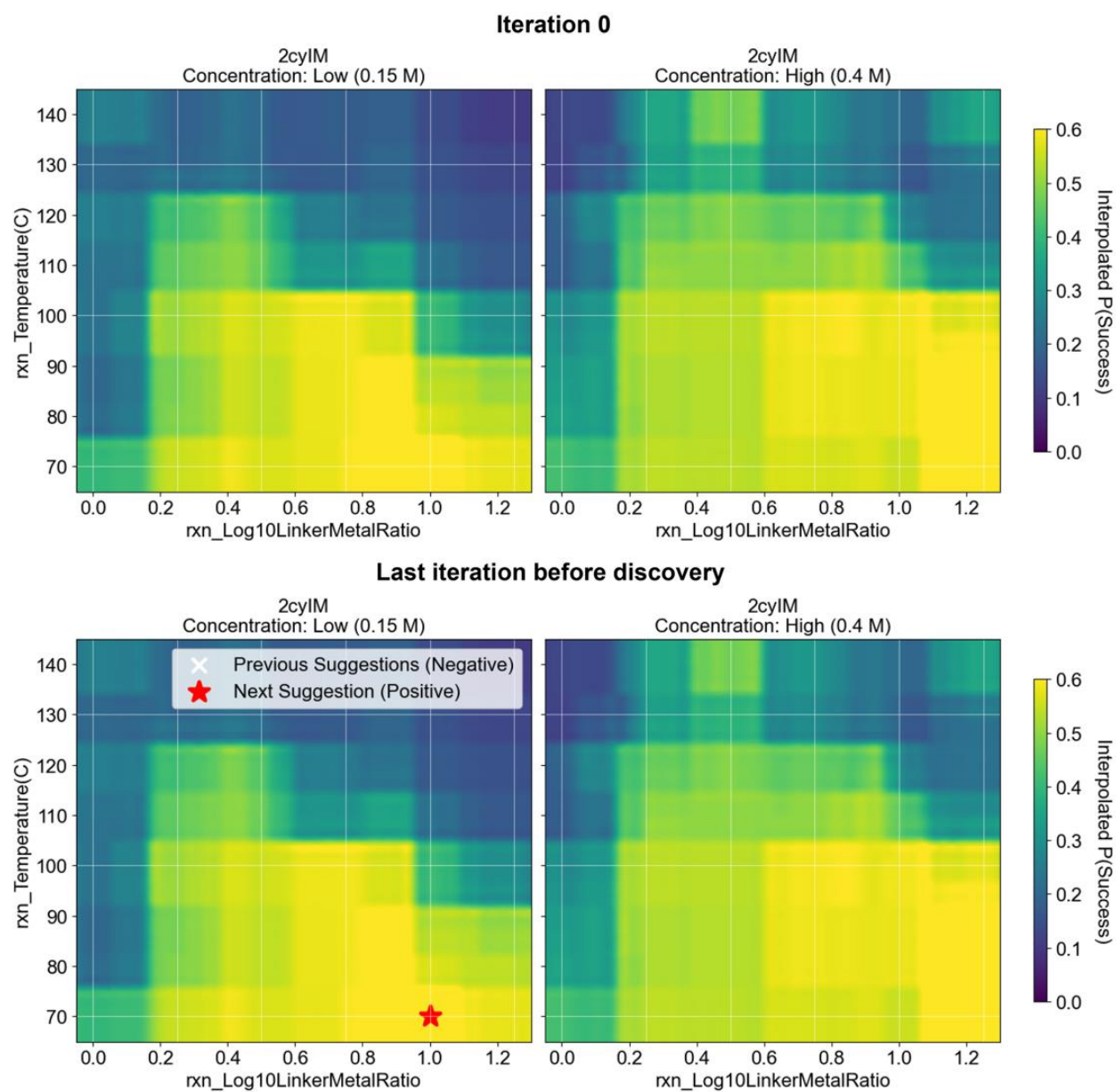


Figure 4.9. Heatmaps of predicted crystallization landscapes for 2cyIM in a single sequential trajectory from AIRES(RFC).

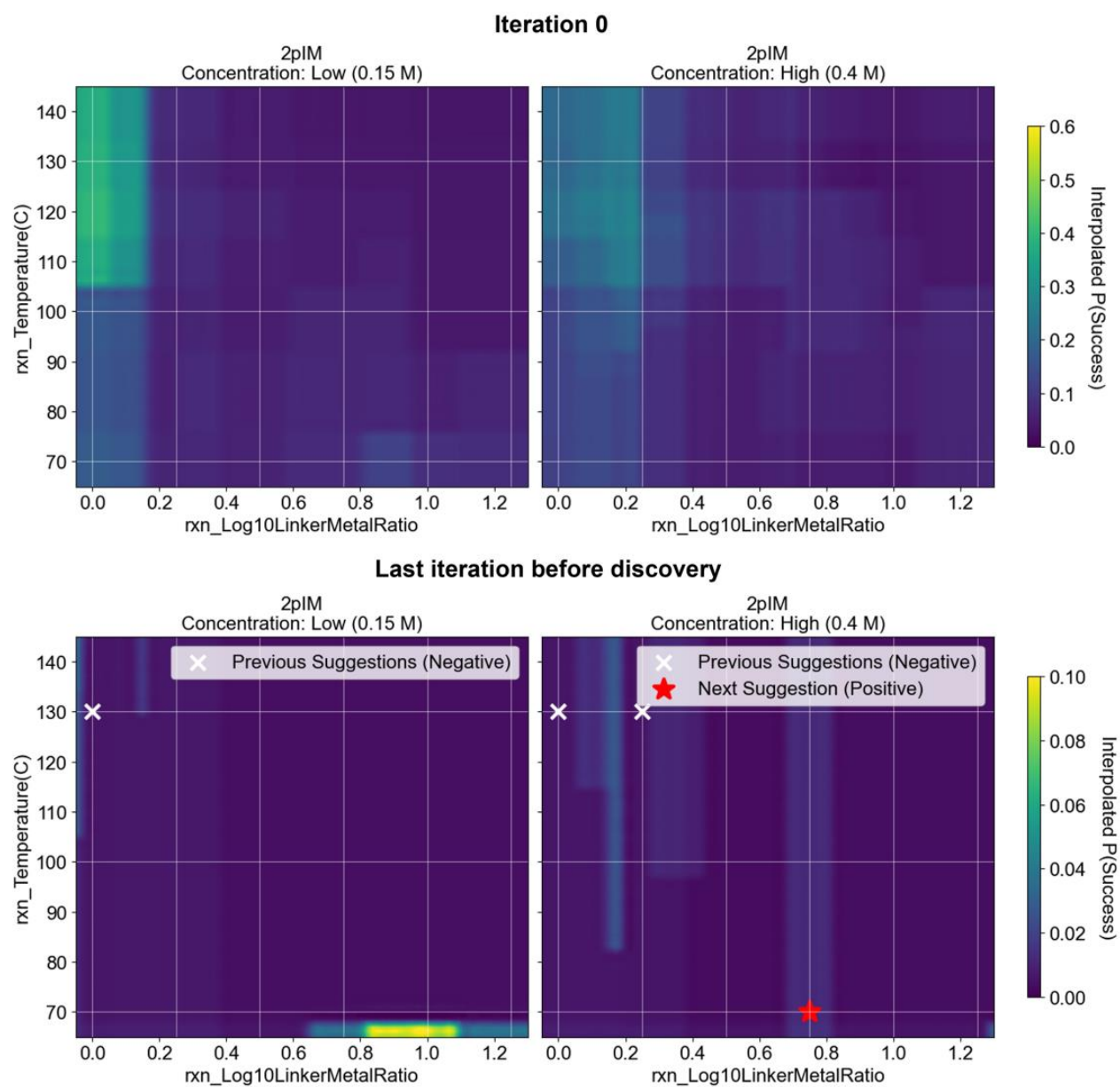


Figure 4.10. Heatmaps of predicted crystallization landscapes for 2pIM in a single sequential trajectory from AIRES(RFC).

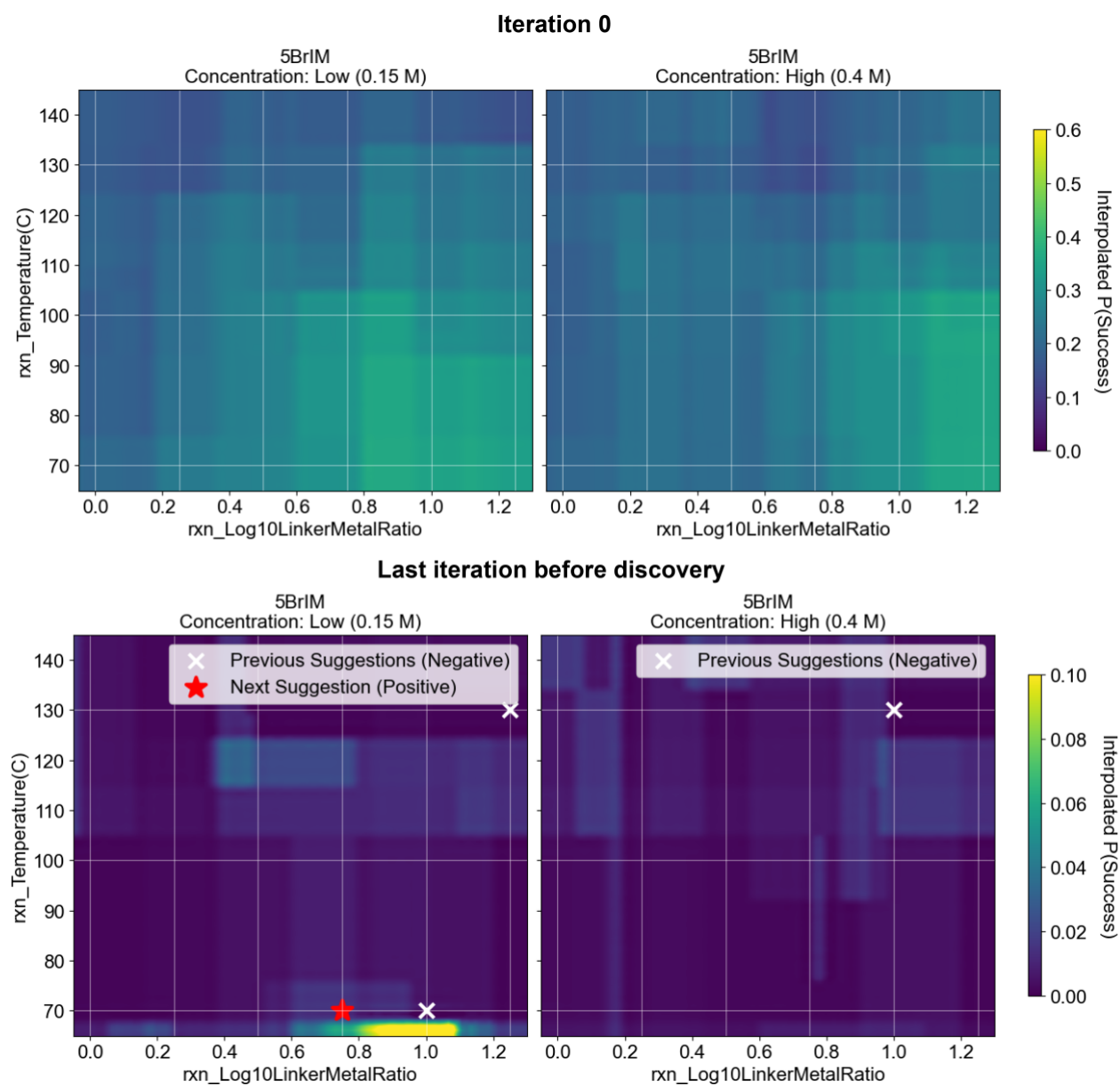


Figure 4.11. Heatmaps of predicted crystallization landscapes for 5BrIM in a single sequential trajectory from AIRES(RFC).

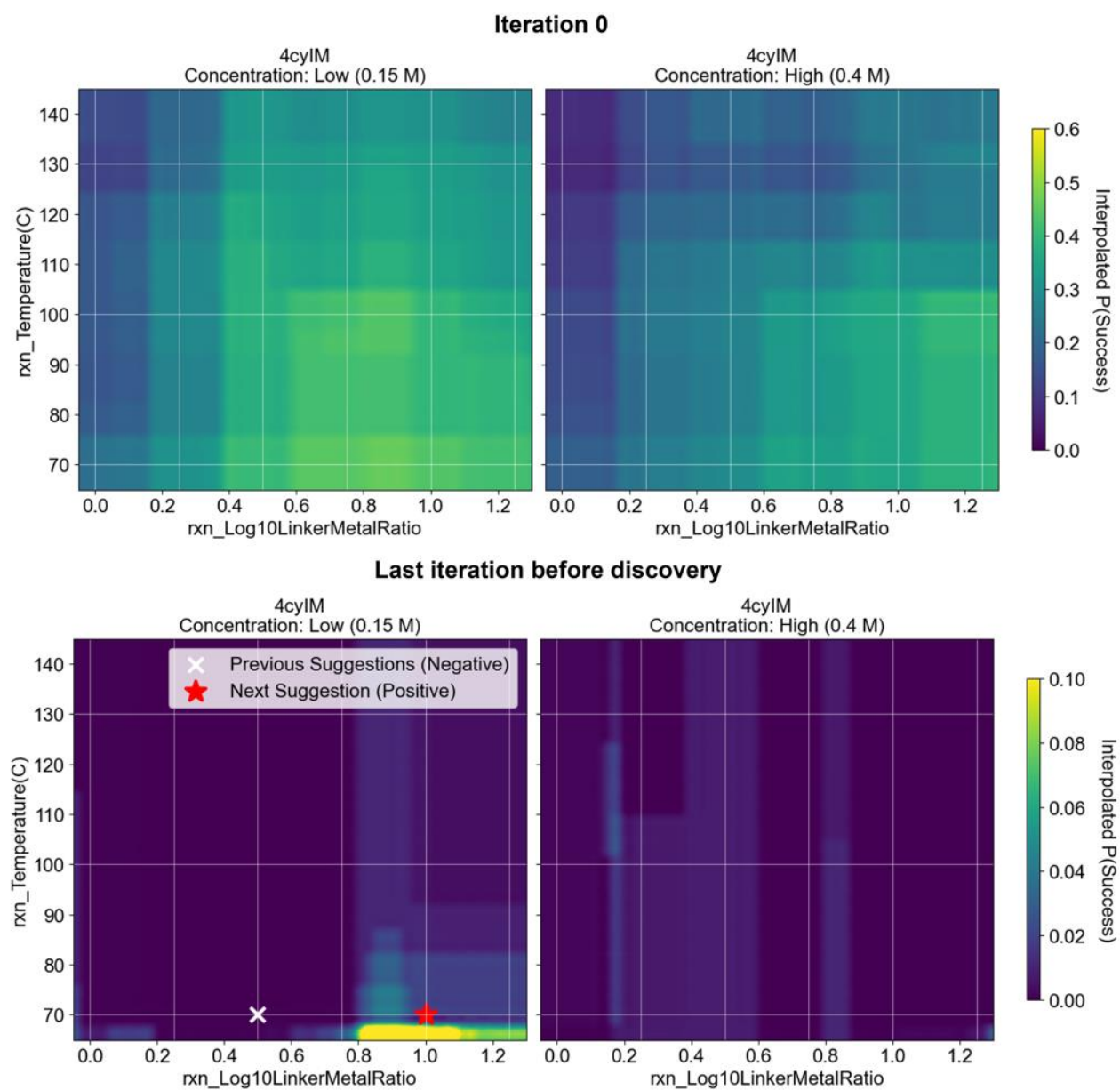


Figure 4.12. Heatmaps of predicted crystallization landscapes for 4cyIM in a single sequential trajectory from AIRES(RFC).

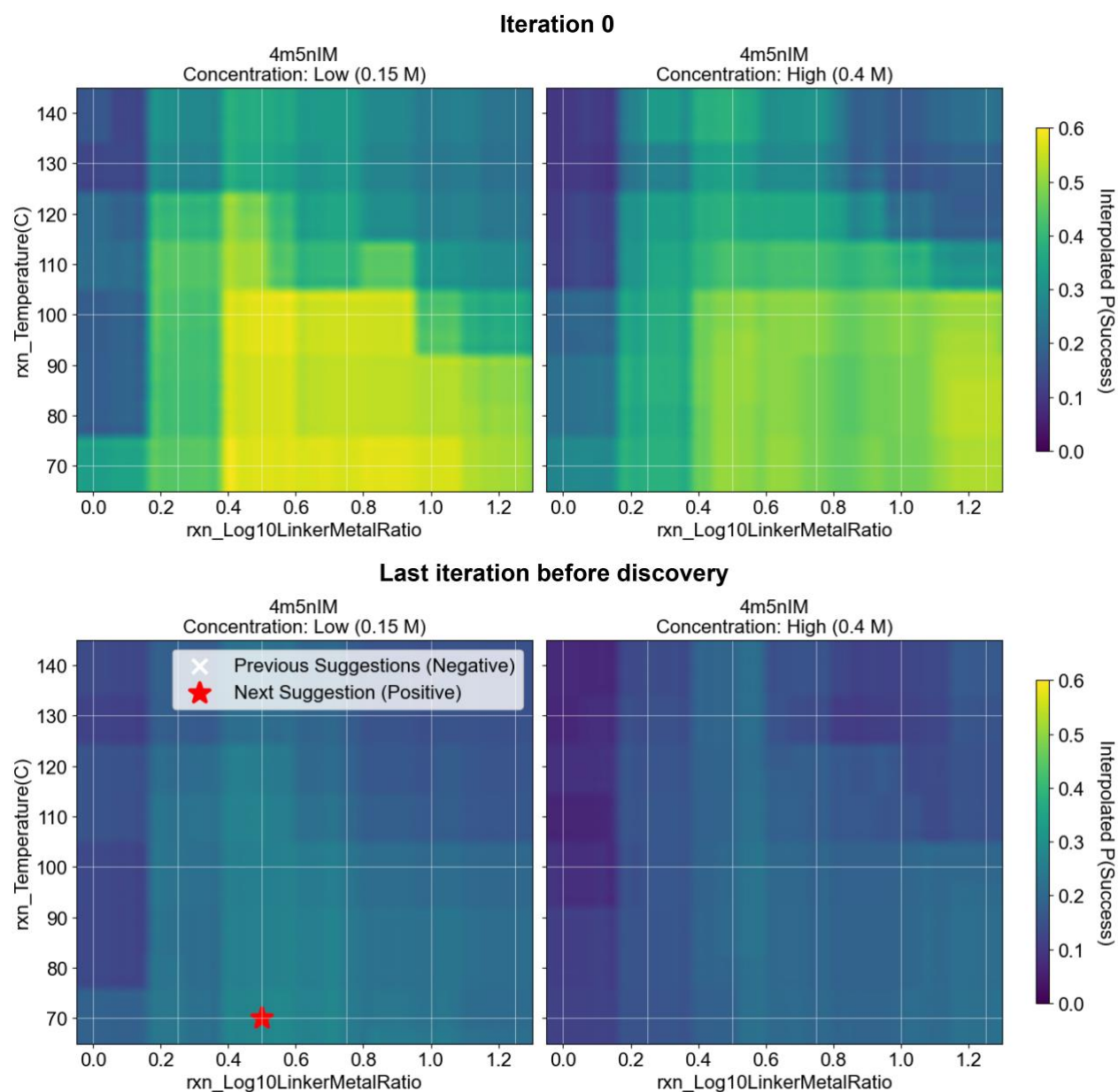


Figure 4.13. Heatmaps of predicted crystallization landscapes for 4m5nIM in a single sequential trajectory from AIRES(RFC).

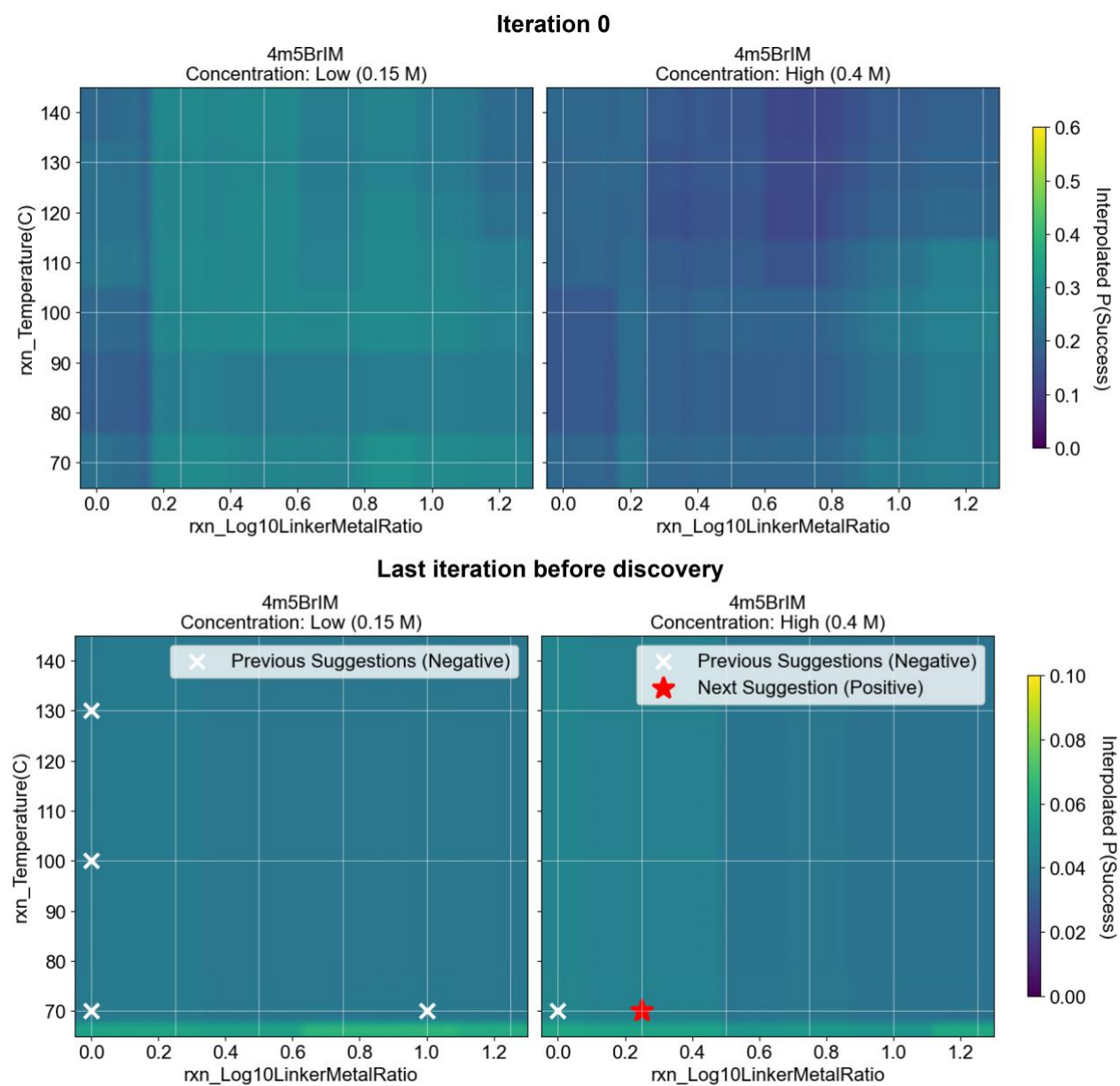


Figure 4.14. Heatmaps of predicted crystallization landscapes for 4m5BrIM in a single sequential trajectory from AIRES(RFC).

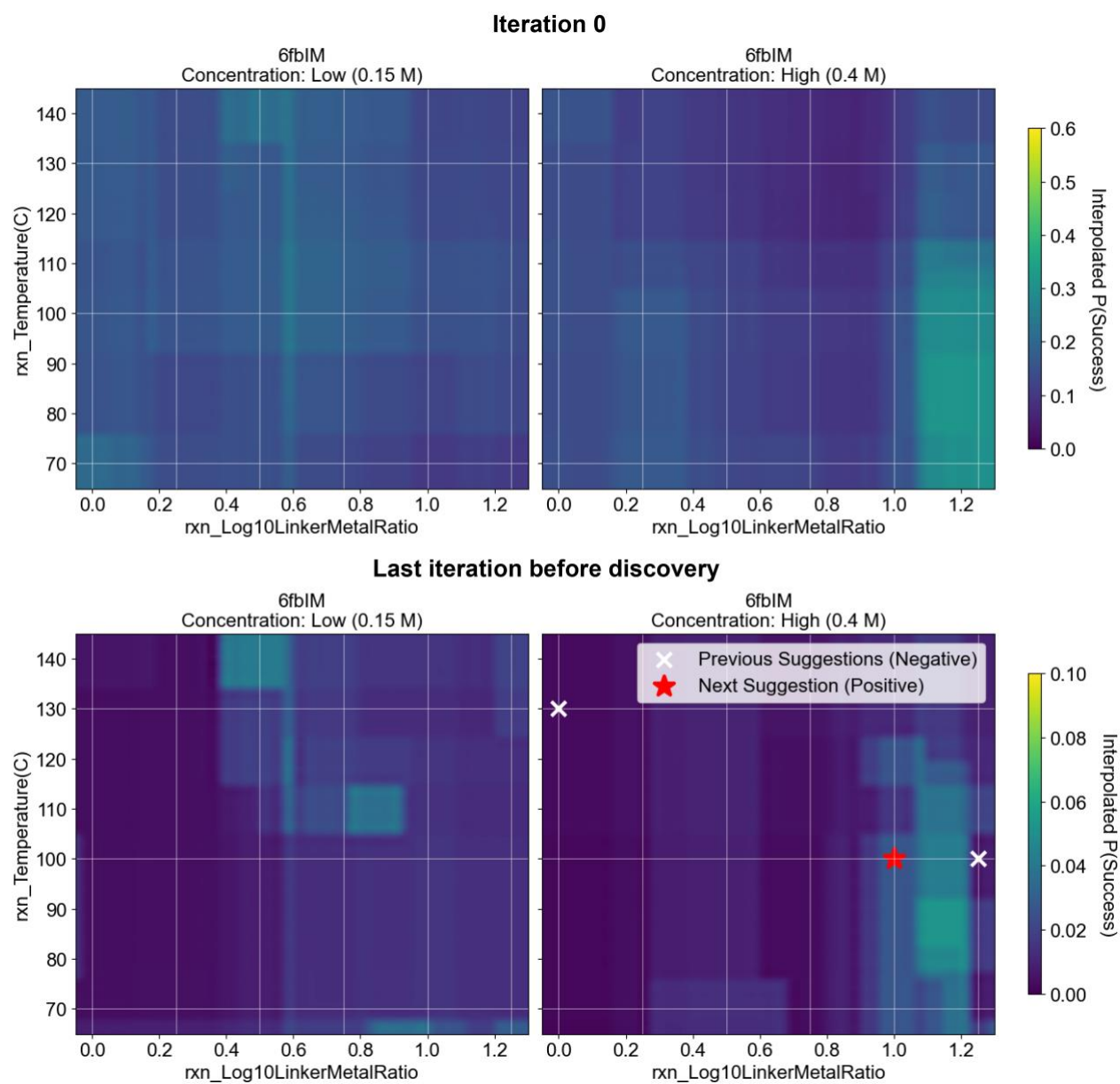


Figure 4.15. Heatmaps of predicted crystallization landscapes for 6fbIM in a single sequential trajectory from AIRES(RFC).

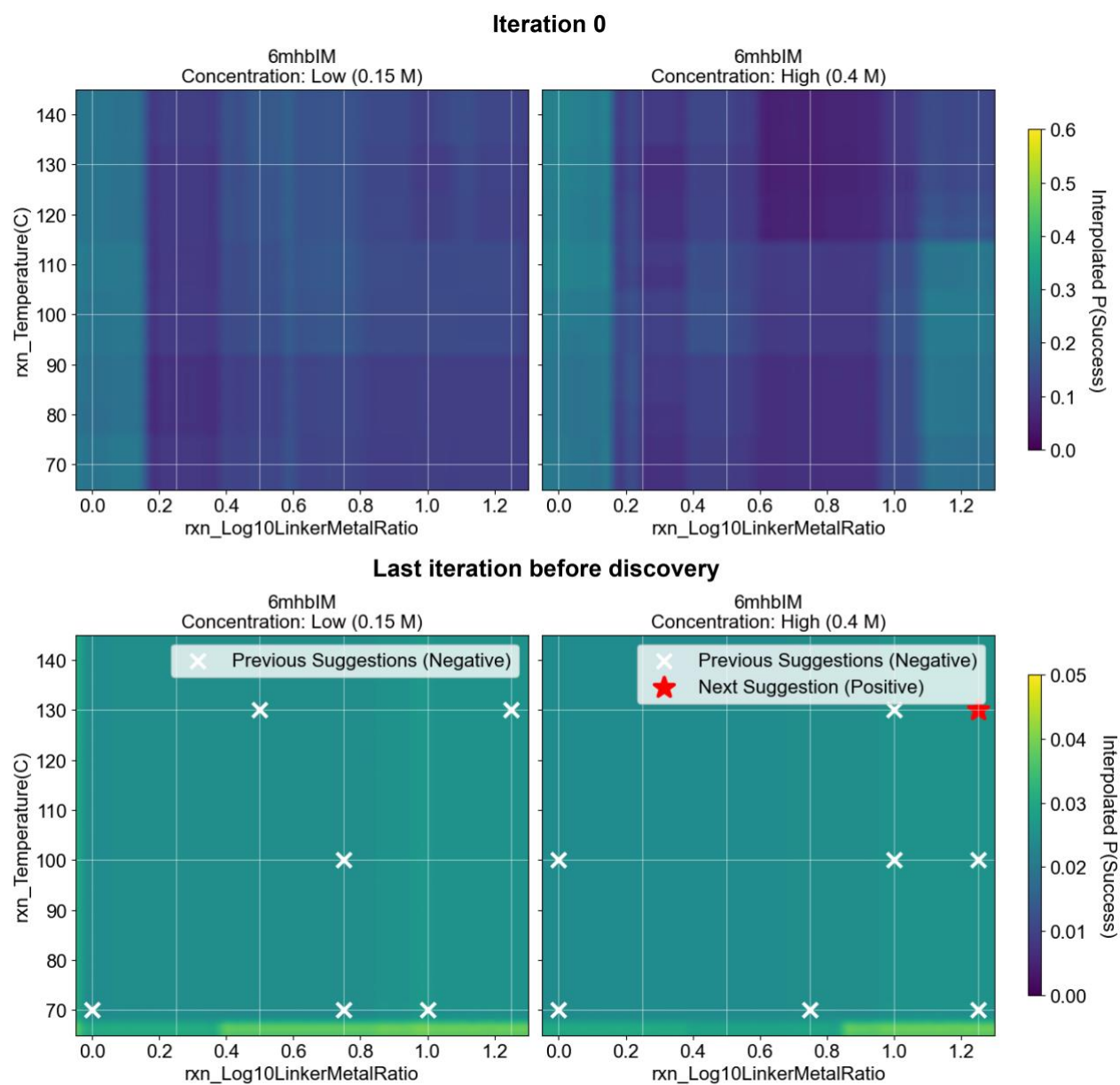


Figure 4.16. Heatmaps of predicted crystallization landscapes for 6mhbIM in a single sequential trajectory from AIRES(RFC).

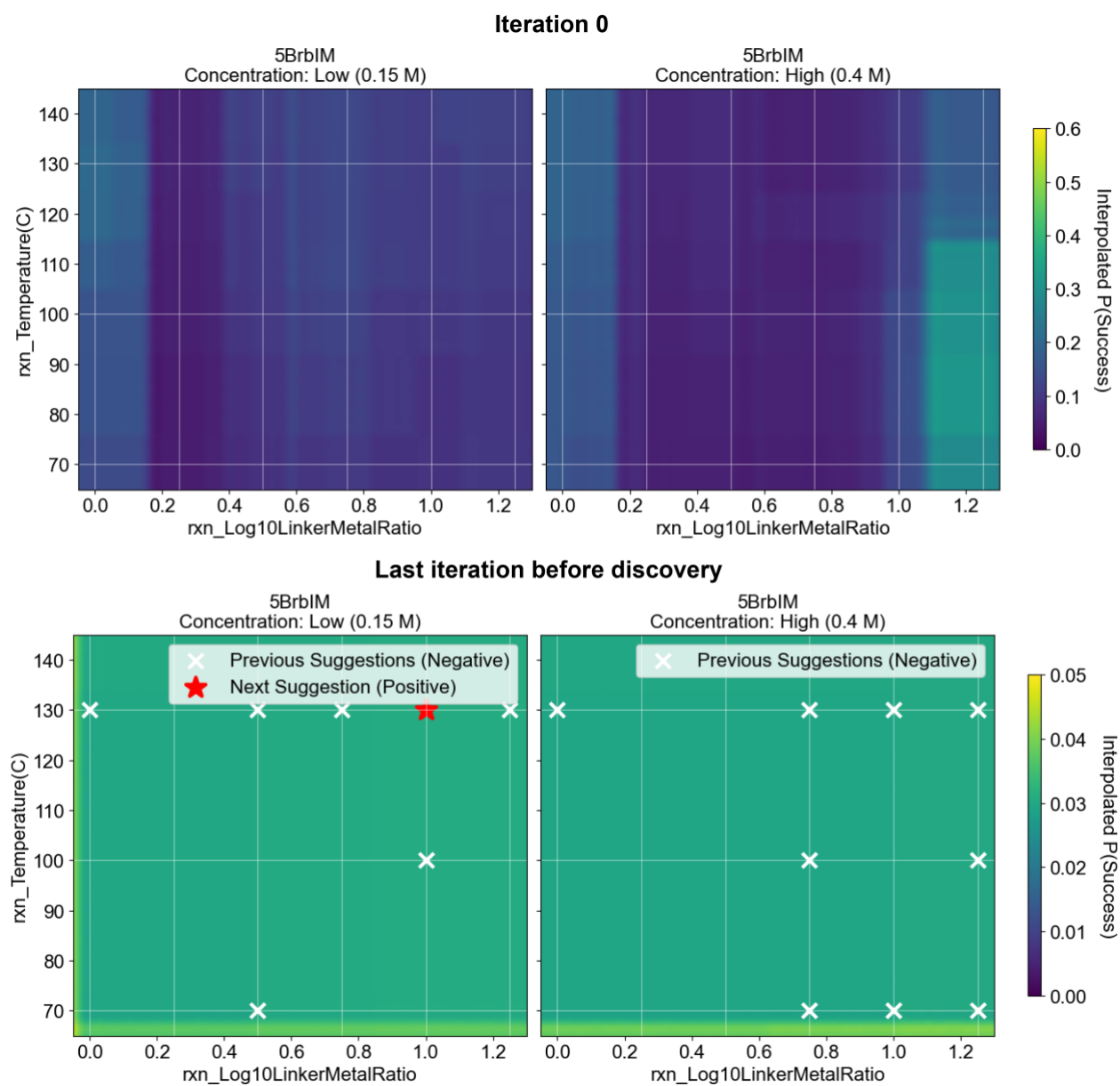


Figure 4.17. Heatmaps of predicted crystallization landscapes for 5BrbIM in a single sequential trajectory from AIRES(RFC).

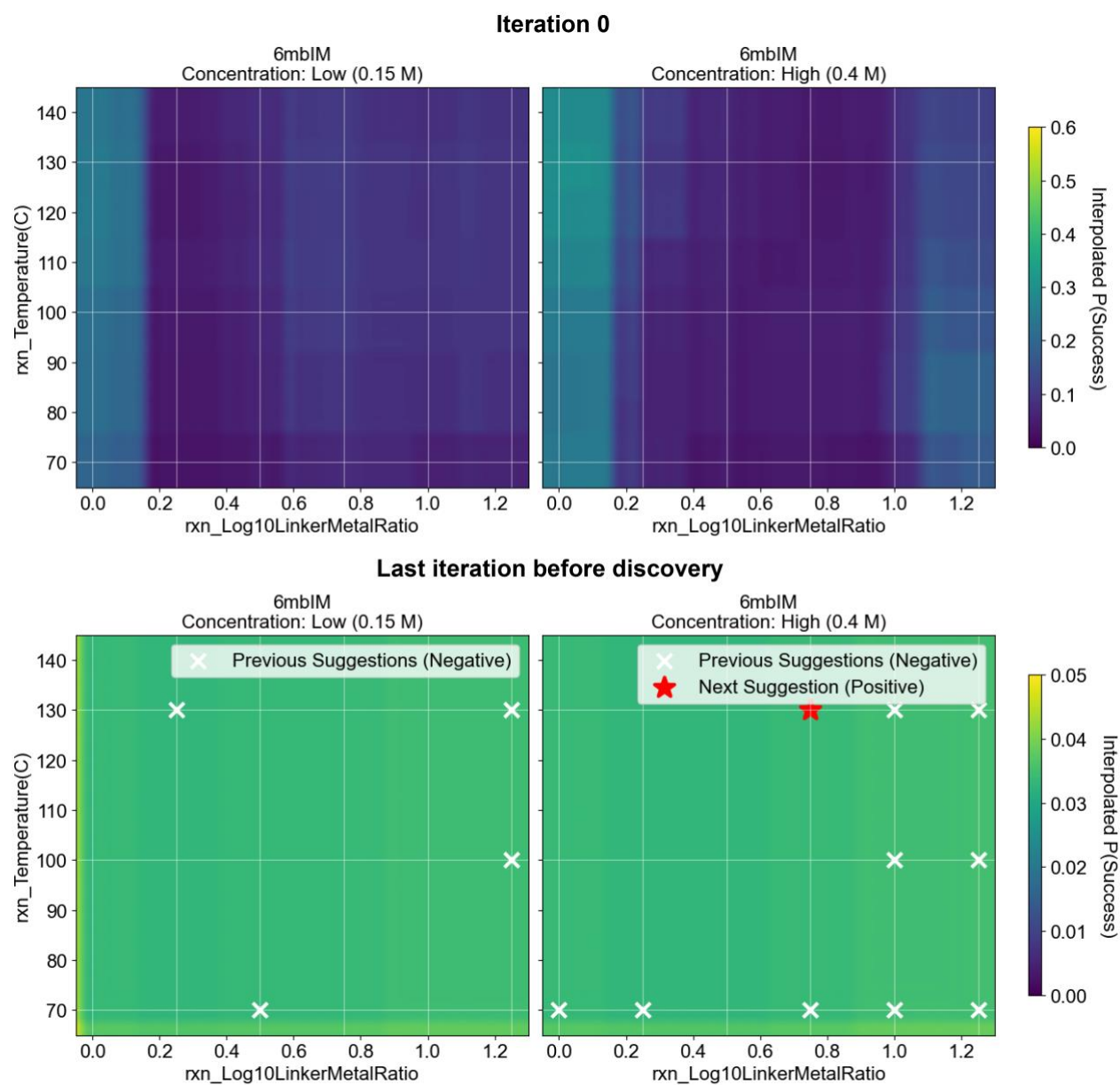


Figure 4.18. Heatmaps of predicted crystallization landscapes for 6mbIM in a single sequential trajectory from AIRES(RFC).

Ablation studies are often used in machine learning to evaluate the influence of different features on model performance by selectively removing or modifying them^{26,27}. Here two ablation studies were conducted. In the first study, all quantum mechanical-derived features, ranging from electrostatic potential charges to prenucleation cluster properties, were removed, while retaining functional group descriptors, geometric features, and experimental parameters. This resulted in a reduced input vector of length 14 (9 chemical features and 5 experimental features). In the second study, all chemically derived features were replaced with a simple one-hot encoding of linker identity derived from SMILES strings, while retaining the experimental parameters. This representation yields an input vector of length 61 (56 linker identities and 5 experimental features). The results clearly demonstrate the importance of incorporating physics-informed descriptors (Figure 4.19). The one-hot encoding approach produces discovery curves that are statistically indistinguishable from random search, indicating that such representations fail to capture transferable chemical information. Removing only the quantum mechanical-derived features leads to modest improvement over random search but still performs significantly worse than the full feature set.

These findings highlight that quantum mechanical-derived features provide critical information for guiding discovery, particularly in capturing electronic and coordination properties that govern ZIF formation. More broadly, they emphasize the importance of incorporating physically meaningful descriptors in machine learning models for materials discovery, rather than relying solely on abstract or identity-based representations.

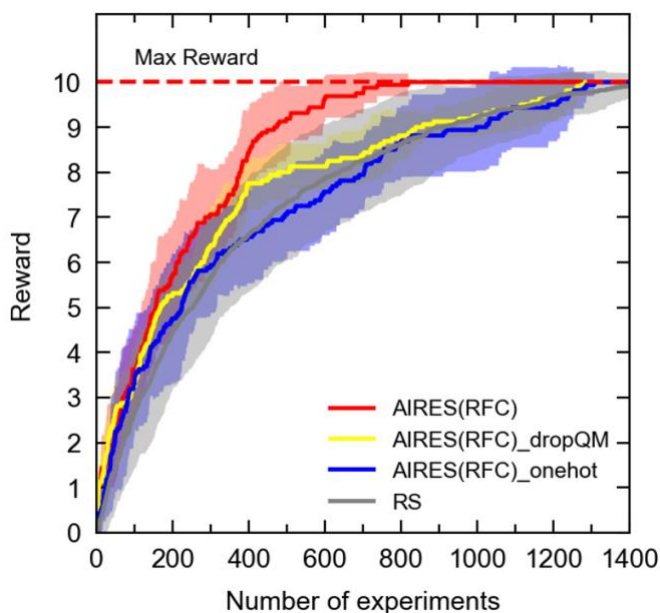


Figure 4.19. Impact of encoding strategies on AIRES discovery performance. The plot compares discovery rates between RS and AIRES(RFC) using different chemical features. Solid lines represent the mean, and shaded areas represent the range of ± 1 standard deviation (16 repetitions for RFC; 100 repetitions for RS). The reward (y-axis) indicates the number of successful distinct discoveries, with a maximum possible reward of 10 (dashed red line).

4.5 References

- (1) Eyke, N. S.; Green, W. H.; Jensen, K. F. Iterative Experimental Design Based on Active Machine Learning Reduces the Experimental Burden Associated with Reaction Screening. *React. Chem. Eng.* **2020**, *5* (10), 1963–1972.
- (2) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96.
- (3) Desimpel, S.; Dorbec, M.; Van Geem, K. M.; Stevens, C. V. Bayesian Optimization for Chemical Reactions. *Chem. Soc. Rev.* **2026**, *55* (5), 2731–2775.
- (4) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121* (16), 9927–10000.
- (5) Duan, C.; Nandy, A.; Terrones, G. G.; Kastner, D. W.; Kulik, H. J. Active Learning Exploration of Transition-Metal Complexes to Discover Method-Insensitive and Synthetically Accessible Chromophores. *JACS Au* **2023**, *3* (2), 391–401.
- (6) Balachandran, P. V.; Kowalski, B.; Sehirlioglu, A.; Lookman, T. Experimental Search for High-Temperature Ferroelectric Perovskites Guided by Two-Step Machine Learning. *Nat. Commun.* **2018**, *9* (1), 1668.
- (7) Li, Z.; Achenie, L. E. K.; Xin, H. An Adaptive Machine Learning Strategy for Accelerating Discovery of Perovskite Electrocatalysts. *ACS Catal.* **2020**, *10* (7), 4377–4384.
- (8) Moon, J.; Beker, W.; Siek, M.; Kim, J.; Lee, H. S.; Hyeon, T.; Grzybowski, B. A. Active Learning Guides Discovery of a Champion Four-Metal Perovskite Oxide for Oxygen Evolution Electrocatalysis. *Nat. Mater.* **2024**, *23* (1), 108–115.
- (9) Voznyy, O.; Levina, L.; Fan, J. Z.; Askerka, M.; Jain, A.; Choi, M.-J.; Ouellette, O.; Todorović, P.; Sagar, L. K.; Sargent, E. H. Machine Learning Accelerates Discovery of Optimal Colloidal Quantum Dot Synthesis. *ACS Nano* **2019**, *13* (10), 11122–11128.
- (10) Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; Hippalgaonkar, K.; Khan, S.; Buonassisi, T.; Li, Q.; Wang, X. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. *Npj Comput. Mater.* **2021**, *7* (1), 55.
- (11) Guo, H.; Lu, Y.; Lei, Z.; Bao, H.; Zhang, M.; Wang, Z.; Guan, C.; Tang, B.; Liu, Z.; Wang, L. Machine Learning-Guided Realization of Full-Color High-Quantum-Yield Carbon Quantum Dots. *Nat. Commun.* **2024**, *15* (1), 4843.
- (12) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241.

- (13) Xie, Y.; Zhang, C.; Deng, H.; Zheng, B.; Su, J.-W.; Shutt, K.; Lin, J. Accelerate Synthesis of Metal–Organic Frameworks by a Robotic Platform and Bayesian Optimization. *ACS Appl. Mater. Interfaces* **2021**, *13* (45), 53485–53491.
- (14) Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. *ACS Cent. Sci.* **2022**, *8* (6), 825–836.
- (15) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Inorganic Materials. *Nature* **2023**, *624* (7990), 86–91.
- (16) Desautels, T.; Krause, A.; Burdick, J. W. Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization. *J. Mach. Learn. Res.* **2014**, *15* (119), 4053–4103.
- (17) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148–175.
- (18) Lattimore, T.; Szepesvári, C. *Bandit Algorithms*; Cambridge University Press: Cambridge New York Port Melbourne New Delhi Singapore, 2020.
- (19) Li, K.; Olson, D. H.; Seidel, J.; Emge, T. J.; Gong, H.; Zeng, H.; Li, J. Zeolitic Imidazolate Frameworks for Kinetic Separation of Propane and Propene. *J. Am. Chem. Soc.* **2009**, *131* (30), 10368–10369.
- (20) Tian, Y.; Yao, S.; Gu, D.; Cui, K.; Guo, D.; Zhang, G.; Chen, Z.; Zhao, D. Cadmium Imidazolate Frameworks with Polymorphism, High Thermal Stability, and a Large Surface Area. *Chem. – Eur. J.* **2010**, *16* (4), 1137–1141.
- (21) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
- (22) Williams, C. K. I.; Barber, D. Bayesian Classification with Gaussian Processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20* (12), 1342–1351.
- (23) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*, 3. print.; Adaptive computation and machine learning; MIT Press: Cambridge, Mass., 2008.
- (24) Gardner, J.; Pleiss, G.; Weinberger, K. Q.; Bindel, D.; Wilson, A. G. Gpytorch: Blackbox Matrix-Matrix Gaussian Process Inference with Gpu Acceleration. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
- (25) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (26) Merrick, L. Randomized Ablation Feature Importance. arXiv 2019.
- (27) Meyes, R.; Lu, M.; de Puiseau, C. W.; Meisen, T. Ablation Studies in Artificial Neural Networks. arXiv 2019.

Chapter 5

Structure Determination and Topological Analysis of ZIF Discovery

5.1 Introduction

The autonomous platform described in previous chapters enables the systematic discovery of new reticular materials by integrating high-throughput experimentation, machine learning, and rigorous structural characterization. Single crystal X-ray diffraction (SCXRD) provides definitive determination of atomic structure, allowing newly synthesized materials to be unambiguously identified. Based on these SCXRD-verified structures, crystals of zeolitic imidazolate frameworks (ZIFs), characterized by tetrahedral metal–imidazolate connectivity, can be identified¹. Despite this shared structural motif, substantial diversity arises in how these tetrahedral building units are assembled^{2–5}. Therefore, beyond confirming chemical identity, it is necessary to further analyze the commonalities and differences among these structures in a way that transcends compositional variation and captures their underlying architectural relationships.

This motivates the need for topological analysis as a complementary perspective for understanding reticular materials, and it has become a standard practice for reporting new reticular structures^{6–8}. Beyond simple classification, it provides a powerful abstraction that enables comparison across chemically diverse systems and offering insights into structure–property relationships. For example, prior studies have demonstrated correlations between topology and mechanical properties^{9–11}, porosity^{12,13}, and other functional properties¹⁴, highlighting topology as an important descriptor in materials design.

Topological analysis involves simplifying a crystal structure into a graph representation that captures only its connectivity^{15–17}. In this abstraction for reticular structures, building blocks of higher connectivity are treated as vertices, and linkers are treated as edges, yielding an underlying net, which is mathematically defined as a periodic graph^{18,19}. This representation removes geometric and chemical details while preserving the essential connectivity of the framework. Such simplification enables mapping of experimentally determined structures onto known network topologies, typically referenced using standardized identifiers from databases such as the Reticular Chemistry Structure Resource (RCSR)²⁰. Importantly, different structures with distinct geometries may share the same underlying net,

Parts of this chapter have been adapted from:

Rong, Z.; Chen, Z.; Luong, F.; Chheda, S.; Luong, H. T. N.; Zheng, Z.; Greco, K.; Alghamdi, A. A.; Bui, K. H.; Jaffrelot Inizan, T.; Nguyen-Dang, T.; Pham, H. H.; Le, D. D.; Sauer, J.; Phung, V. B. T.; Chayes, J. T.; Borgs, C.; Boley, M.; El Ghaoui, L.; Yaghi, O. M. Algorithmic Iterative Reticular Synthesis of Zeolitic Imidazolate Framework Crystals. *Nat. Synth.* **2025**, 5 (3), 409–419.

while conversely, identical building blocks may assemble into different nets, reflecting the intrinsic complexity of reticular synthesis²¹.

In addition to underlying nets, natural tiling provides a complementary description of framework topology. Natural tiling partitions the structure into space-filling polyhedral units that reflect the local environment around nodes and the organization of pores^{22,23}. While the underlying net captures connectivity, natural tiling offers a geometric perspective that is closely related to pore architecture and spatial arrangement. Together, these two representations provide a comprehensive description of reticular structures from both connectivity and geometric viewpoints.

In this chapter, the single crystal structures obtained from the autonomous platform applied onto ZIF discovery are presented, followed by topological analysis. These include both ZIFs and other extended crystalline structures, which are termed as layered-ZIFs (LZIFs) due to their two-dimensional structures, identified through SCXRD. The underlying nets and tiles of these structures are determined and discussed, providing insight into how different connectivity patterns emerge from similar or competing building units. This analysis not only highlights the structural diversity uncovered by the autonomous discovery process but also situates both ZIF and non-ZIF frameworks within the broader landscape of reticular chemistry.

5.2 Crystal Structures of ZIFs and LZIFs

X-ray diffraction data were collected either using a Bruker D8 Venture diffractometer or at beamline 12.2.1 of the Advanced Light Source. The Bruker diffractometer was equipped with dual radiation sources, Cu ($\lambda = 1.54178 \text{ \AA}$) and Mo ($\lambda = 0.71073 \text{ \AA}$), while the ALS beamline, utilizing a Si (111) monochromator, provided radiation at $\lambda = 0.7288 \text{ \AA}$. In both cases, data were recorded using PHOTON-II CMOS detectors. Crystals were mounted on Kapton® MiTeGen MicroMount™ loops with a minimal amount of Paratone®N oil and maintained under a cold nitrogen gas stream generated by an Oxford Cryosystems 800 Series during data collection. Diffraction data were processed using the Bruker APEX4 software suite²⁴, with SAINT²⁵ used for data integration and SADABS²⁶ applied for absorption correction. Structures were solved by either intrinsic phasing using SHELXT²⁷ or direct methods using SHELXS²⁸, followed by full-matrix least-squares refinement on F^2 using SHELXL²⁹. Structure solution, refinement, and visualization were carried out within the Olex2 software³⁰. For structures exhibiting significant disorder from solvent molecules residing in framework pores, solvent masking was applied using the Bypass method implemented in Olex2 to account for residual electron density³¹.

Refinement details of ZIF-A1. The entire dataset was cut to a resolution of 0.80 Å, resulting in an I/σ value of 69.7, an R_{int} values of 3.98%, and a completeness of 99.8%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. The weight parameters a and b converged to 0.183 and 0, respectively.

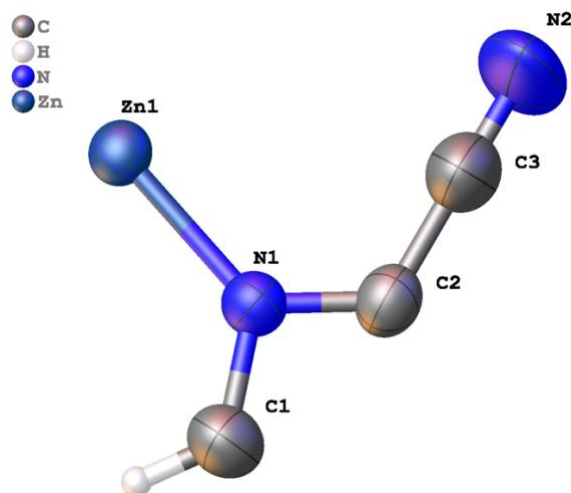


Figure 5.1. Asymmetric unit of the crystal structure of ZIF-A1. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue. CCDC reference number: 2411199.

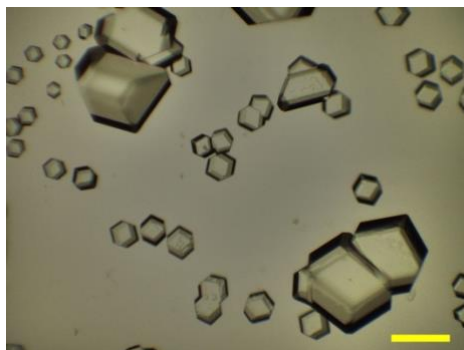


Figure 5.2. Representative optical image of ZIF-A1 crystallized from 2cyIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 70 °C. Yellow scale bar 100 μm .

Table 5.1. Crystal data and structure refinement for ZIF-A1.

Empirical formula	C ₈ H ₄ N ₆ Zn
Formula weight	249.54
Temperature/K	100
Crystal system	cubic
Space group	I-43m
$a/\text{\AA}$	16.96440(10)
$b/\text{\AA}$	16.96440(10)
$c/\text{\AA}$	16.96440(10)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	4882.20(9)
Z	12
ρ_{calc} (g/cm ³)	1.018
μ (mm ⁻¹)	1.979
$F(000)$	1488.0
Crystal size/mm ³	0.1 $\times$ 0.1 $\times$ 0.1
Radiation	Cu K $_{\alpha}$ (λ = 1.54178)
2θ range for data collection/ $^\circ$	7.37 to 149.142
Index ranges	$-20 \leq h \leq 16$, $-18 \leq k \leq 19$, $-19 \leq l \leq 18$
Reflections collected	13548
Independent reflections	946 [R_{int} = 0.0398, R_{sigma} = 0.0143]
Data/restraints/parameters	946/0/39
Goodness-of-fit on F^2	1.176
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0821$, $wR_2 = 0.2221$
Final R indexes [all data]	$R_1 = 0.0837$, $wR_2 = 0.2254$
Largest diff. peak/hole / e $\text{\AA}^3$	1.44 / -0.54
Flack parameter	0.035(19)

Refinement details of ZIF-A2. The entire dataset was cut to a resolution of 0.86 Å, resulting in an I/σ value of 74.8, an R_{int} values of 2.17%, and a completeness of 99.7%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. A pore volume of 1928 Å³ containing 81 electrons per unit cell was revealed after applying solvent masks. The positional disorder between the methyl and nitro groups was resolved by setting each group's occupancy to 0.5. The weight parameters a and b converged to 0.065 and 13.2, respectively.

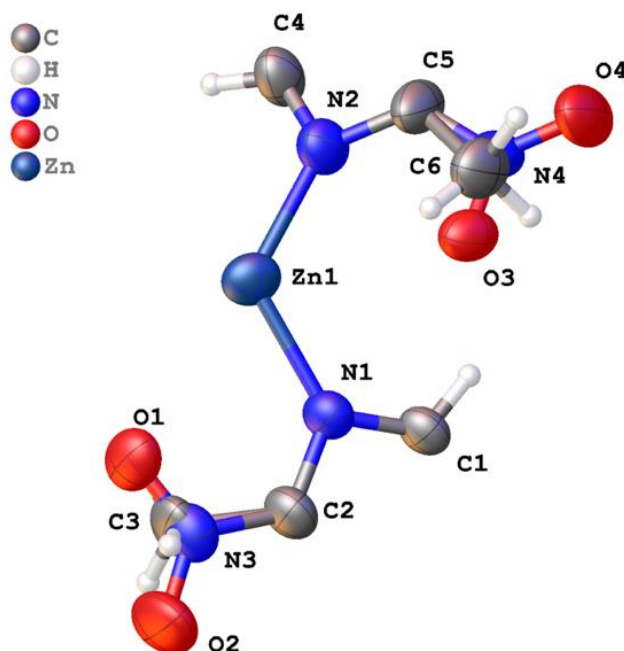


Figure 5.3. Asymmetric unit of the crystal structure of ZIF-A2. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue. CCDC reference number: 2411197.

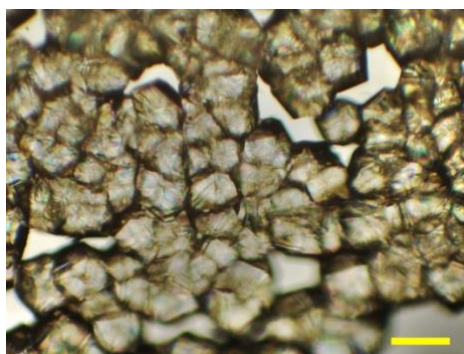


Figure 5.4. Representative optical image of ZIF-A1 crystallized from 2cyIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 70 °C. Yellow scale bar 100 μm .

Table 5.2. Crystal data and structure refinement for ZIF-A2

Empirical formula	$\text{C}_8\text{H}_8\text{N}_6\text{O}_4\text{Zn}$
Formula weight	317.57
Temperature/K	100.00
Crystal system	trigonal
Space group	R-3m
$a/\text{\AA}$	22.9854(3)
$b/\text{\AA}$	22.9854(3)
$c/\text{\AA}$	15.4842(3)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	120
Volume/ $\text{\AA}^3$	7084.7(2)
Z	18
ρ_{calc} (g/cm^3)	1.340
μ (mm^{-1})	2.355
$F(000)$	2880.0
Crystal size/ mm^3	$0.05 \times 0.05 \times 0.05$
Radiation	Cu K_α ($\lambda = 1.54178$)
2θ range ($^\circ$)	7.232 to 128.028
Index ranges	$-25 \leq h \leq 25, -26 \leq k \leq 22, -17 \leq l \leq 16$
Reflections collected	9480
Independent reflections	1428 [$R_{\text{int}} = 0.0217, R_{\text{sigma}} = 0.0134$]
Data/restraints/parameters	1428/47/127
Goodness-of-fit on F^2	1.170
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0387, wR_2 = 0.1204$
Final R indexes [all data]	$R_1 = 0.0408, wR_2 = 0.1226$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.32 / -0.26

Refinement details of ZIF-A3. The entire dataset was cut to a resolution of 0.86 Å, resulting in an I/σ value of 52.3, an R_{int} values of 4.20%, and a completeness of 99.9%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. Two distinct pore types were revealed after applying solvent masks: one pore with a volume of 177.4 Å³ containing 80.7 electrons, and another with a volume of 107.8 Å³ containing 33 electrons. Two types of disorder were identified in the structure. The first involves the benzene ring position, where the occupancies of both configurations were set at 0.5. The second relates to the flipping of fluorine atom positions. F1's occupancy was fixed at 0.5 due to its mirror plane symmetry relationship. For F3, its position is shared between two 6fbIM linkers when the asymmetric unit is grown, leading to the constraint: $\text{Occupancy}(\text{F2}) + 0.5 \times \text{Occupancy}(\text{F3}) = 0.5$. Refinement yielded occupancies of 0.476(6) and 0.047(12) for F2 and F3, respectively. The weight parameters a and b converged to 0.052 and 70.5, respectively.

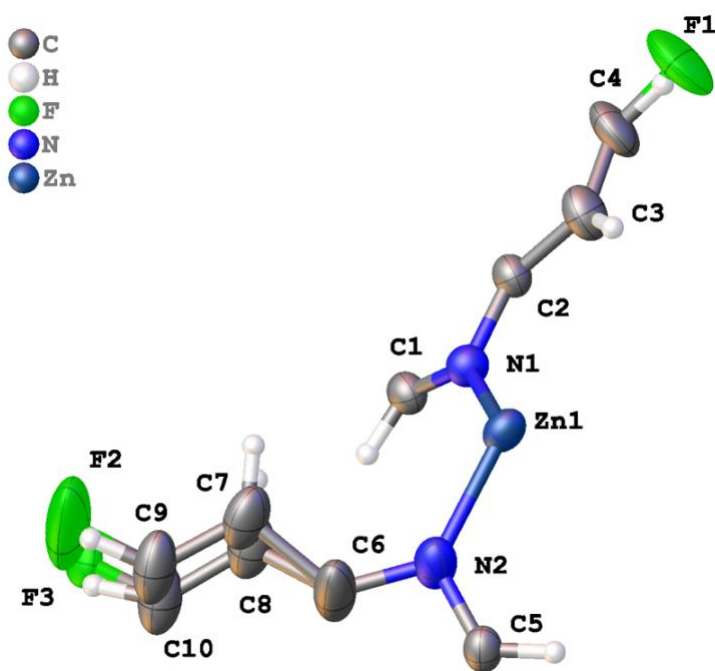


Figure 5.5. Asymmetric unit of the crystal structure of ZIF-A3. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; F, lime green. CCDC reference number: 2411200.

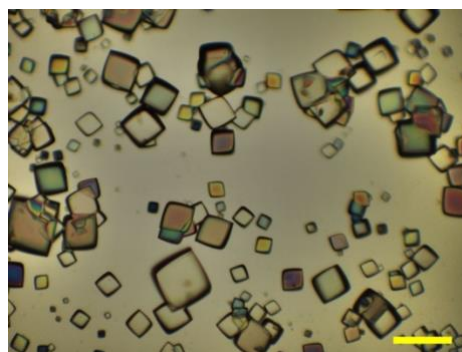


Figure 5.6. Representative optical image of ZIF-A3 crystallized from 6fbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 0.00, and a temperature of 70 °C. Yellow scale bar 100 μm .

Table 5.3. Crystal data and structure refinement for ZIF-A3.

Empirical formula	$\text{C}_{14}\text{H}_8\text{F}_2\text{N}_4\text{Zn}$
Formula weight	335.61
Temperature/K	100.00
Crystal system	trigonal
Space group	R-3m
$a/\text{\AA}$	22.8041(2)
$b/\text{\AA}$	22.8041(2)
$c/\text{\AA}$	16.0117(3)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	120
Volume/ $\text{\AA}^3$	7210.97(19)
Z	18
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.391
$\mu (\text{mm}^{-1})$	1.549
$F(000)$	3024.0
Crystal size/ mm^3	$0.04 \times 0.04 \times 0.03$
Radiation	Mo K_α ($\lambda = 0.71073$)
2θ range ($^\circ$)	4.846 to 48.616
Index ranges	$-26 \leq h \leq 26, -26 \leq k \leq 26, -18 \leq l \leq 18$
Reflections collected	16432
Independent reflections	1426 [$R_{\text{int}} = 0.0420, R_{\text{sigma}} = 0.0191$]
Data/restraints/parameters	1426/38/132
Goodness-of-fit on F^2	1.149
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0506, wR_2 = 0.1335$
Final R indexes [all data]	$R_1 = 0.0540, wR_2 = 0.1356$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.58 / -0.65

Refinement details of ZIF-A4. The entire dataset was cut to a resolution of 0.86 Å, resulting in an I/σ value of 99.0, an R_{int} values of 4.74%, and a completeness of 99.8%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. The weight parameters a and b were refined to 0.076 and 37.2, respectively.

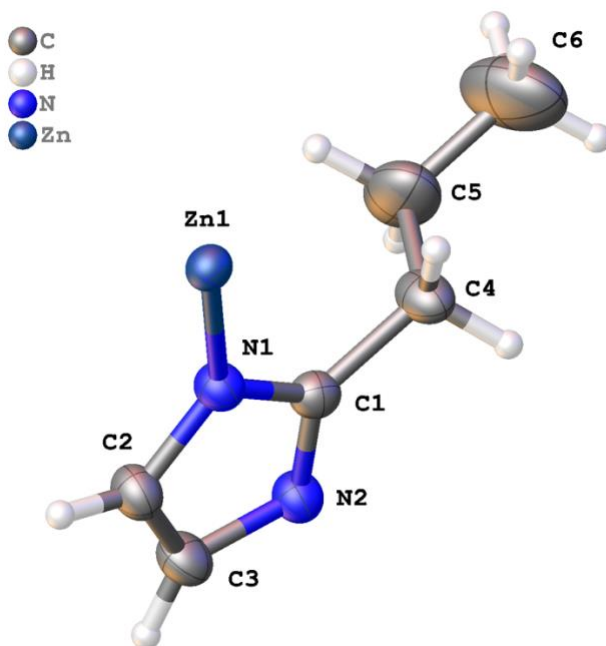


Figure 5.7. Asymmetric unit of the crystal structure of ZIF-A4. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue. CCDC reference number: 2411198.

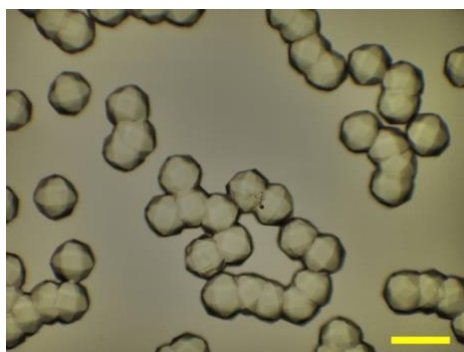


Figure 5.8. Representative optical image of ZIF-A4 crystallized from 2pIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 0.50, and a temperature of 100 °C. Yellow scale bar 100 μm .

Table 5.4. Crystal data and structure refinement for ZIF-A4.

Empirical formula	$\text{C}_{12}\text{H}_{18}\text{N}_4\text{Zn}$
Formula weight	283.67
Temperature/K	100.00
Crystal system	cubic
Space group	Ia-3d
$a/\text{\AA}$	26.4649(2)
$b/\text{\AA}$	26.4649(2)
$c/\text{\AA}$	26.4649(2)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	18535.8(4)
Z	48
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.220
$\mu (\text{mm}^{-1})$	1.576
$F(000)$	7104.0
Crystal size/ mm^3	$0.1 \times 0.1 \times 0.1$
Radiation	Mo K_α ($\lambda = 0.71073$)
2θ range ($^\circ$)	4.354 to 49.006
Index ranges	$-30 \leq h \leq 30, -30 \leq k \leq 30, -30 \leq l \leq 30$
Reflections collected	59631
Independent reflections	1301 [$R_{\text{int}} = 0.0474, R_{\text{sigma}} = 0.0101$]
Data/restraints/parameters	1301/0/79
Goodness-of-fit on F^2	1.120
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0322, wR_2 = 0.1150$
Final R indexes [all data]	$R_1 = 0.0371, wR_2 = 0.1204$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.76 / -0.35

Refinement details of ZIF-A5. The entire dataset was cut to a resolution of 0.90 Å, resulting in an I/σ value of 24.1, an R_{int} values of 9.55%, and a completeness of 99.7%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. Two identical pores per unit cell, each with a volume of 248 Å³ and containing 56 electrons, were revealed after applying solvent masks. The positional disorder of the bromine atom was resolved by setting the occupancies of Br1 and Br2 to 0.5. Interestingly, Br3 doesn't show positional disorder. The weight parameters a and b converged to 0.079 and 91.5, respectively.

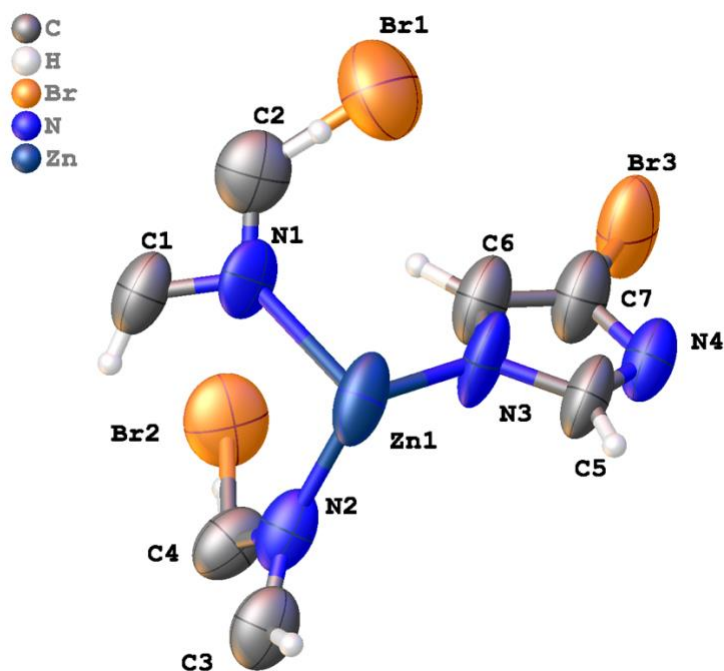


Figure 5.9. Asymmetric unit of the crystal structure of ZIF-A5. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; Br, orange. CCDC reference number: 2411067.

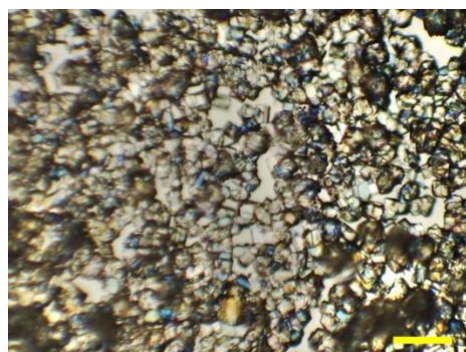


Figure 5.10. Representative optical image of ZIF-A5 crystallized from 5BrIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 0.50, and a temperature of 130 °C. Yellow scale bar 100 μm .

Table 5.5. Crystal data and structure refinement for ZIF-A5.

Empirical formula	$\text{C}_6\text{H}_4\text{Br}_2\text{N}_4\text{Zn}$
Formula weight	357.32
Temperature/K	100
Crystal system	orthorhombic
Space group	Pnnm
$a/\text{\AA}$	9.6668(9)
$b/\text{\AA}$	15.7456(13)
$c/\text{\AA}$	15.7888(15)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	2403.2(4)
Z	8
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.975
$\mu (\text{mm}^{-1})$	9.210
$F(000)$	1344.0
Crystal size/ mm^3	$0.03 \times 0.015 \times 0.015$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	3.746 to 48.028
Index ranges	$-10 \leq h \leq 10, -17 \leq k \leq 17, -17 \leq l \leq 17$
Reflections collected	22010
Independent reflections	1824 [$R_{\text{int}} = 0.0955, R_{\text{sigma}} = 0.0414$]
Data/restraints/parameters	1824/41/132
Goodness-of-fit on F^2	1.091
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.1268, wR_2 = 0.2929$
Final R indexes [all data]	$R_1 = 0.1461, wR_2 = 0.3035$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.85 / -0.94

Refinement details of ZIF-A6. The entire dataset was cut to a resolution of 0.90 Å, resulting in an I/σ value of 16.3, an R_{int} values of 14.35%, and a completeness of 99.7%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. The structure exhibits two types of disorder: orientational disorder of the imidazolate rings and positional disorder of the bromine atoms. The three 5BrIM ligands coordinated to Zn show positional disorder, with each configuration modeled at 0.5 occupancy. The bromine atoms display positional disorder through ring flipping. Br3 was modeled with 0.5 occupancy due to this flipping motion. While Br1 and Br2 also undergo flipping, they are each shared between two disordered imidazolate rings. As a result, despite the ring flipping, their effective occupancies remain at 0.5 (calculated as $0.5 \times 2 \text{ rings} \div 2 \text{ positions}$). Two types of pores were revealed after applying solvent masks. The first one has a volume of 969.7 Å³ containing 306 electrons, while the second one has a volume of 114.2 Å³ containing 44 electrons. The weight parameters a and b converged to 0.163 and 3.29, respectively.

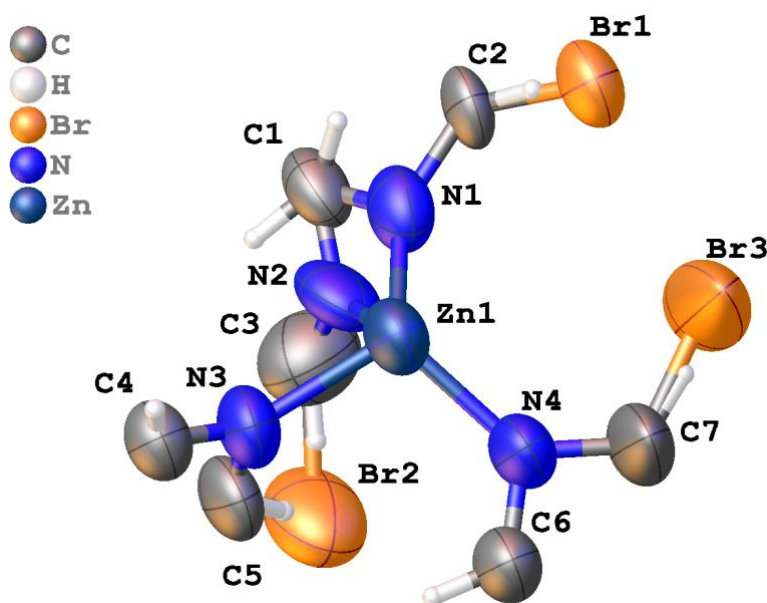


Figure 5.11. Asymmetric unit of the crystal structure of ZIF-A6. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; Br, orange. CCDC reference number: 2411068.

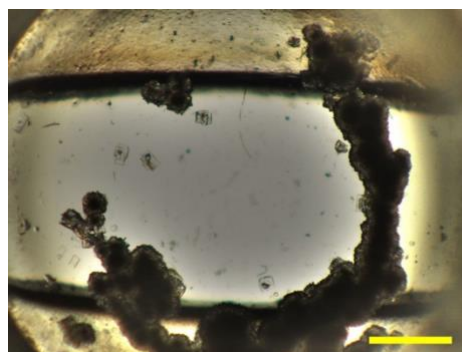


Figure 5.12. Representative optical image of ZIF-A6 crystallized from 5BrIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 0.00, and a temperature of 100 °C. Yellow scale bar 1 mm

Table 5.6. Crystal data and structure refinement for ZIF-A6.

Empirical formula	C ₆ H ₄ Br ₂ N ₄ Zn
Formula weight	357.32
Temperature/K	100
Crystal system	tetragonal
Space group	P4/mmm
<i>a</i> /Å	14.095(2)
<i>b</i> /Å	14.095(2)
<i>c</i> /Å	14.829(3)
α /°	90
β /°	90
γ /°	90
Volume/Å ³	2946.0(10)
<i>Z</i>	8
ρ_{calc} (g/cm ³)	1.611
μ (mm ⁻¹)	7.513
<i>F</i> (000)	1344.0
Crystal size/mm ³	0.08 × 0.02 × 0.01
Radiation	Synchrotron (λ = 0.7288)
2 θ range (°)	4.088 to 47.976
Index ranges	-15 ≤ <i>h</i> ≤ 15, -15 ≤ <i>k</i> ≤ 15, -16 ≤ <i>l</i> ≤ 16
Reflections collected	40465
Independent reflections	1301 [<i>R</i> _{int} = 0.1435, <i>R</i> _{sigma} = 0.0612]
Data/restraints/parameters	1301/17/116
Goodness-of-fit on <i>F</i> ²	1.077
Final <i>R</i> indexes [<i>I</i> ≥ 2 σ (<i>I</i>)]	<i>R</i> ₁ = 0.0804, <i>wR</i> ₂ = 0.2373
Final <i>R</i> indexes [all data]	<i>R</i> ₁ = 0.0990, <i>wR</i> ₂ = 0.2601
Largest diff. peak/hole (e Å ⁻³)	1.28 / -0.94

Refinement details of ZIF-A7. The entire dataset was cut to a resolution of 0.82 Å, resulting in an I/σ value of 62.5, an R_{int} values of 5.99%, and a completeness of 99.8%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. A pore volume of 2082.5 Å³ containing 437 electrons per unit cell was revealed after applying solvent masks. The cyano group disorder was modeled as follows: C8/N7 were assigned occupancies of 0.5 due to mirror symmetry. The C4/N5 and C5/N6 occupancies were freely refined to 0.586(9) and 0.414(9), respectively. Since N6 is shared between two disordered cyano group configurations, its occupancy is $0.414(9) \times 2 = 0.827(18)$. The weight parameters a and b converged to 0.149 and 8.93, respectively.

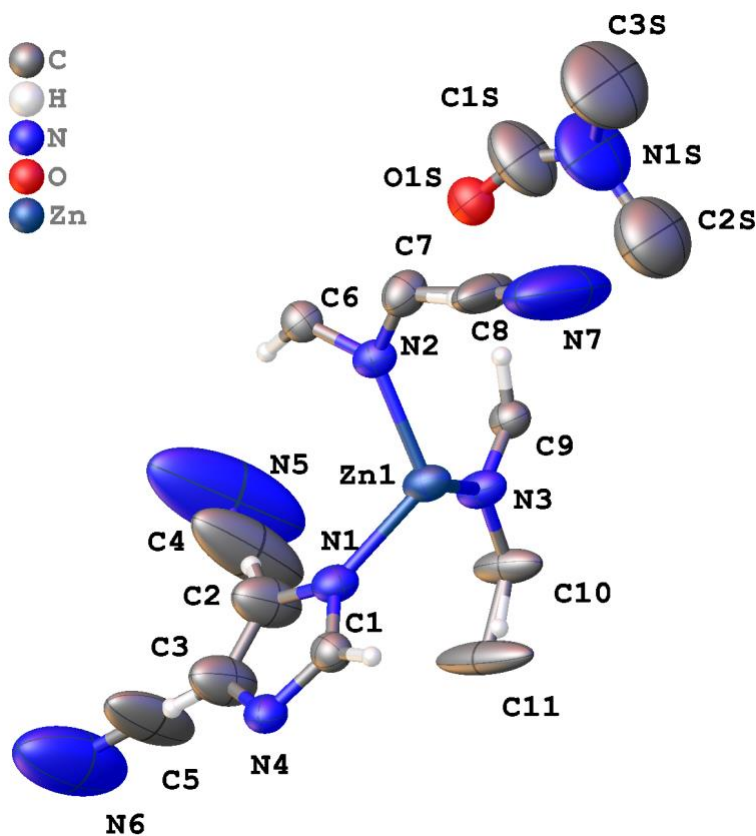


Figure 5.13. Asymmetric unit of the crystal structure of ZIF-A7. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue. CCDC reference number: 2411070.

number: 2411070.

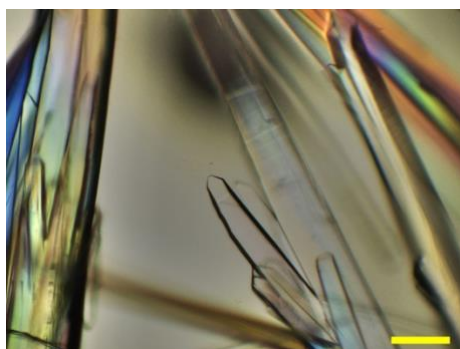


Figure 5.14. Representative optical image of ZIF-A7 crystallized from 4cyIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 70 °C. Yellow scale bar 100 μm .

Table 5.7. Crystal data and structure refinement for ZIF-A7.

Empirical formula	$\text{C}_8\text{H}_4\text{N}_6\text{Zn}\cdot 0.25(\text{C}_3\text{NO})$
Formula weight	266.05
Temperature/K	100
Crystal system	tetragonal
Space group	$\text{P4}_2/\text{mm}$
$a/\text{\AA}$	19.2055(16)
$b/\text{\AA}$	19.2055(16)
$c/\text{\AA}$	16.9062(19)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	6235.9(13)
Z	16
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.134
$\mu (\text{mm}^{-1})$	1.662
$F(000)$	2116.0
Crystal size/ mm^3	$0.17 \times 0.08 \times 0.08$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	3.076 to 52.488
Index ranges	$-23 \leq h \leq 23, -23 \leq k \leq 23, -19 \leq l \leq 20$
Reflections collected	107971
Independent reflections	3148 [$R_{\text{int}} = 0.0599, R_{\text{sigma}} = 0.0160$]
Data/restraints/parameters	3148/60/190
Goodness-of-fit on F^2	1.141
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0673, wR_2 = 0.2178$
Final R indexes [all data]	$R_1 = 0.0777, wR_2 = 0.2323$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.98 / -0.96

Refinement details of ZIF-A8. The entire dataset was cut to a resolution of 1.30 Å, resulting in an I/σ value of 31.9, an R_{int} values of 7.87%, and a completeness of 99.6%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. A pore volume of 10050.3 Å³ containing 2029 electrons per unit cell was revealed after applying solvent masks. The positional disorder between the methyl group and the bromine atom was resolved by setting each group's occupancy to 0.5. The weight parameters a and b converged to 0.131 and 16.0, respectively.

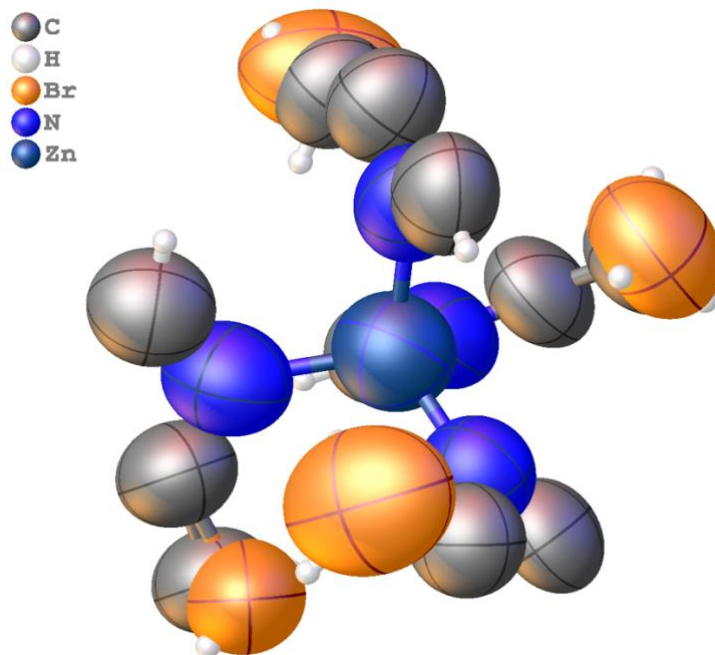


Figure 5.15. Asymmetric unit of the crystal structure of ZIF-A8. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; Br, orange. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411069.

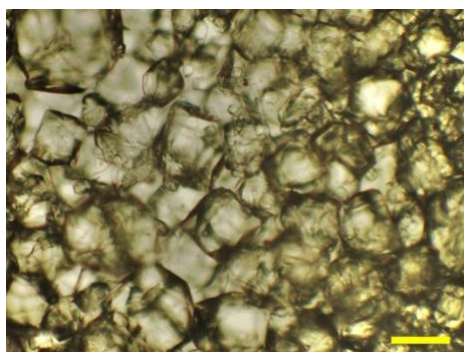


Figure 5.16. Representative optical image of ZIF-A8 crystallized from 4m5BrIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 0.00, and a temperature of 100 °C. Yellow scale bar 100 μm .

Table 5.8. Crystal data and structure refinement for ZIF-A8.

Empirical formula	$\text{C}_8\text{H}_8\text{Br}_2\text{N}_4\text{Zn}$
Formula weight	385.37
Temperature/K	100.00
Crystal system	cubic
Space group	Pm-3m
$a/\text{\AA}$	28.3885(12)
$b/\text{\AA}$	28.3885(12)
$c/\text{\AA}$	28.3885(12)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	22878(3)
Z	48
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.343
$\mu (\text{mm}^{-1})$	5.810
$F(000)$	8832.0
Crystal size/ mm^3	$0.04 \times 0.04 \times 0.04$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	2.942 to 32.596
Index ranges	$-21 \leq h \leq 21, -21 \leq k \leq 21, -21 \leq l \leq 21$
Reflections collected	86762
Independent reflections	1146 [$R_{\text{int}} = 0.0787, R_{\text{sigma}} = 0.0313$]
Data/restraints/parameters	1146/33/152
Goodness-of-fit on F^2	1.058
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0506, wR_2 = 0.1748$
Final R indexes [all data]	$R_1 = 0.0650, wR_2 = 0.1968$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.22 / -0.14

Refinement details of ZIF-A9. The entire dataset was cut to a resolution of 1.49 Å, resulting in an I/σ value of 18.3, an R_{int} values of 17.87%, and a completeness of 98.8%. Initial solution of SHELXS Direct Methods gave the positions of zinc atoms and nitrogen atoms, as they are less disordered in the structure. The benzimidazole cores and methoxy groups were then constructed by adding geometric constraints. Due to the poor data quality, heavy atoms excluding zinc were refined isotropically, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. SIMU was applied on to the atoms on the linkers to stabilize the refinement. While extensive disorder in the benzimidazole cores and methoxy groups prevented complete refinement convergence, the overall framework structure is well-defined. Solvent masks revealed a pore volume of 22374.0 Å³ containing 8768 electrons per unit cell. The structure is an inversion twin. Therefore, the twin law $[-1, 0, 0, 0, -1, 0, 0, 0, -1]$ was applied, resulting in a refined BASF value of 0.51(5).

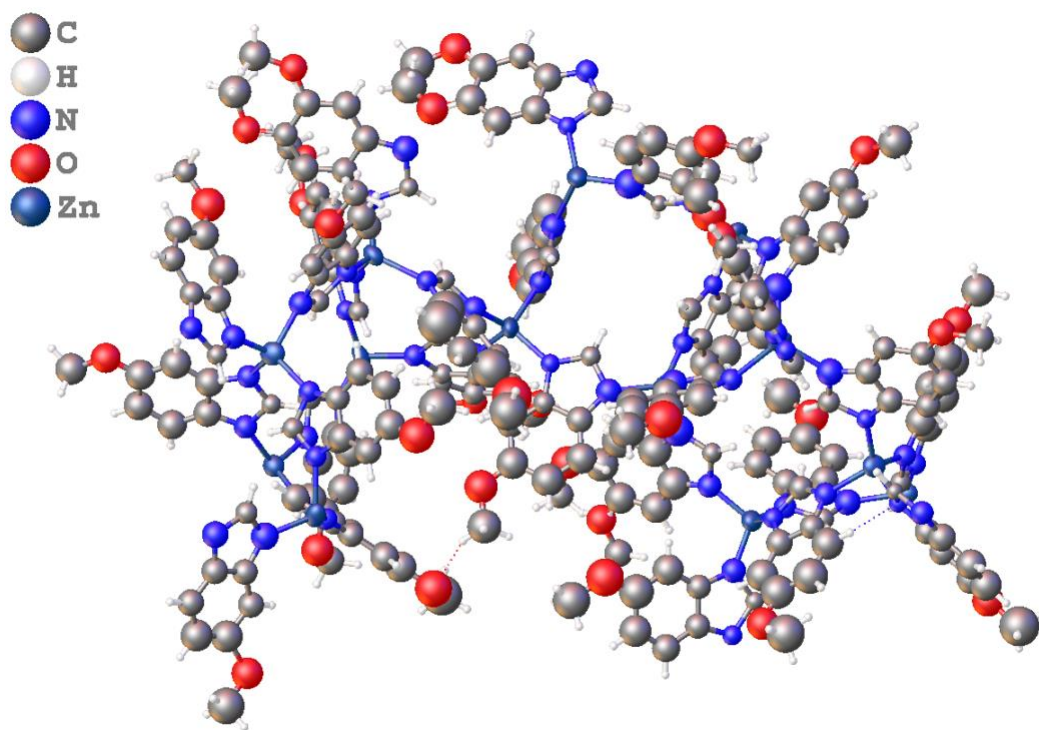


Figure 5.17. Asymmetric unit of the crystal structure of ZIF-A9. Ellipsoids are plotted with 15% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411079.

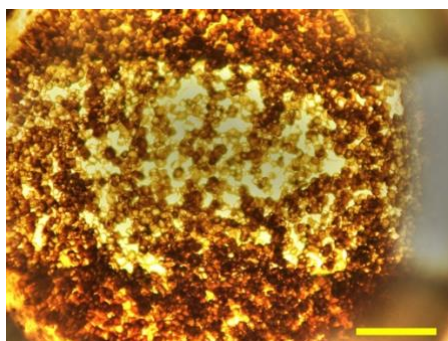


Figure 5.18. Representative optical image of ZIF-A9 crystallized from 6mhbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 0.75, and a temperature of 130 °C. Yellow scale bar 1 mm.

Table 5.9. Crystal data and structure refinement for ZIF-A9.

Empirical formula	C ₁₆ H ₁₄ N ₄ O ₂ Zn
Formula weight	359.68
Temperature/K	100
Crystal system	tetragonal
Space group	P4 ₃ 2 ₁ 2
<i>a</i> /Å	29.235(5)
<i>b</i> /Å	29.235(5)
<i>c</i> /Å	68.891(16)
α /°	90
β /°	90
γ /°	90
Volume/Å ³	58879(23)
<i>Z</i>	104
ρ_{calc} (g/cm ³)	1.055
μ (mm ⁻¹)	1.163
<i>F</i> (000)	19136.0
Crystal size/mm ³	0.05 × 0.05 × 0.03
Radiation	Synchrotron (λ = 0.7288)
2 θ range (°)	1.552 to 28.24
Index ranges	-19 ≤ <i>h</i> ≤ 19, -19 ≤ <i>k</i> ≤ 19, -46 ≤ <i>l</i> ≤ 46
Reflections collected	172886
Independent reflections	9156 [<i>R</i> _{int} = 0.1787, <i>R</i> _{sigma} = 0.0547]
Data/restraints/parameters	9156/694/1019
Goodness-of-fit on <i>F</i> ²	1.440
Final <i>R</i> indexes [<i>I</i> ≥ 2 σ (<i>I</i>)]	<i>R</i> ₁ = 0.1098, <i>wR</i> ₂ = 0.2997
Final <i>R</i> indexes [all data]	<i>R</i> ₁ = 0.1324, <i>wR</i> ₂ = 0.3285
Largest diff. peak/hole (e Å ⁻³)	0.43 / -0.41
Flack parameter	0.51(5)

Refinement details of ZIF-A10. The entire dataset was cut to a resolution of 1.50 Å, resulting in an I/σ value of 16.5, an R_{int} values of 13.42%, and a completeness of 99.7%. Initial solution of SHELXS Direct Methods gave the positions of zinc atoms and nitrogen atoms, as they are less disordered in the structure. The benzimidazole cores were then constructed by adding geometric constraints. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. SIMU was applied on to the atoms on the linkers to stabilize the refinement. Disordered bromine atoms were resolved based on difference map. A pore volume of 12585.8 Å³ containing 3750 electrons per unit cell was revealed after applying solvent masks. The weight parameters a and b converged to 0.200 and 0, respectively.

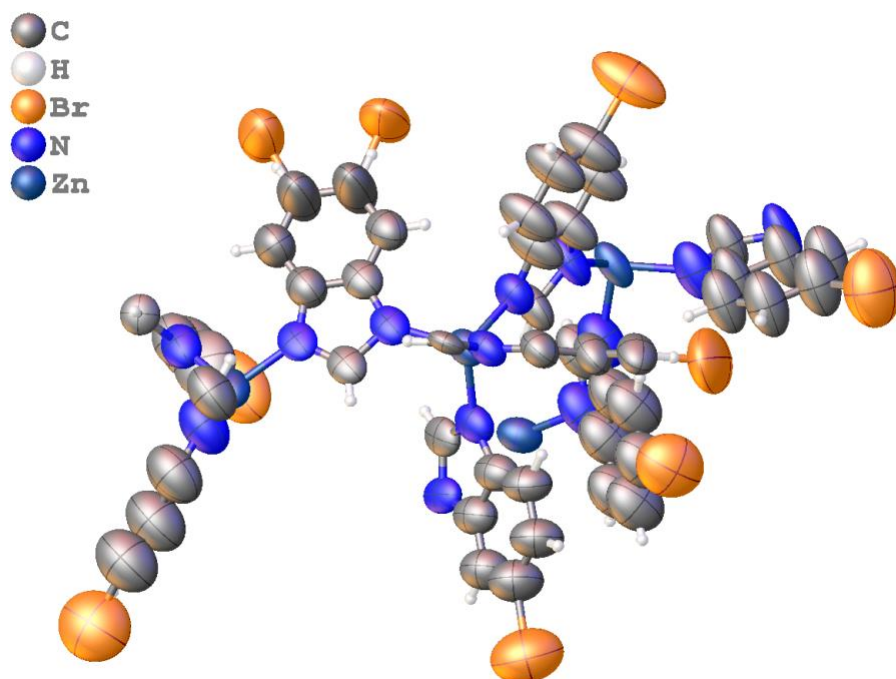


Figure 5.19. Asymmetric unit of the crystal structure of ZIF-A10. Ellipsoids are plotted with 15% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; Br, orange. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411078.

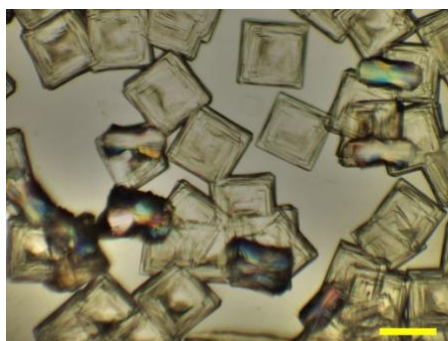


Figure 5.20. Representative optical image of ZIF-A10 crystallized from 5BrbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 130 °C. Yellow scale bar 100 μm .

Table 5.10. Crystal data and structure refinement for ZIF-A10.

Empirical formula	$\text{C}_{14}\text{H}_8\text{Br}_2\text{N}_4\text{Zn}$
Formula weight	457.43
Temperature/K	100
Crystal system	tetragonal
Space group	$P4_2/\text{ncm}$
$a/\text{\AA}$	30.099(5)
$b/\text{\AA}$	30.099(5)
$c/\text{\AA}$	34.042(8)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	30841(13)
Z	52
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.281
$\mu (\text{mm}^{-1})$	4.682
$F(000)$	11440.0
Crystal size/ mm^3	$0.05 \times 0.05 \times 0.02$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	2.314 to 28.144
Index ranges	$-20 \leq h \leq 20, -20 \leq k \leq 20, -22 \leq l \leq 22$
Reflections collected	79220
Independent reflections	2532 [$R_{\text{int}} = 0.1342, R_{\text{sigma}} = 0.0605$]
Data/restraints/parameters	2532/680/558
Goodness-of-fit on F^2	3.188
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.2646, wR_2 = 0.5945$
Final R indexes [all data]	$R_1 = 0.2897, wR_2 = 0.6278$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	1.62 / -0.84

Refinement details of ZIF-A11. The entire dataset was cut to a resolution of 1.27 Å, resulting in an I/σ value of 30.6, an R_{int} values of 5.11%, and a completeness of 99.9%. Initial solution of SHELXS Direct Methods gave the positions of zinc atoms and nitrogen atoms, as they are less disordered in the structure. The benzimidazole cores and methyl groups were then constructed by adding geometric constraints. Due to the poor data quality, heavy atoms excluding zinc were refined isotropically, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. SIMU was applied on to the atoms on the linkers to stabilize the refinement. While extensive disorder in the benzimidazole cores and methyl groups prevented complete refinement convergence, the overall framework structure is well-defined. Solvent masks revealed a pore volume of 27664.0 Å³ containing 5474 electrons per unit cell. The structure is an inversion twin. Therefore, the twin law [-1, 0, 0, 0, -1, 0, 0, 0, -1] was applied, resulting in a refined BASF value of 0.45(8).

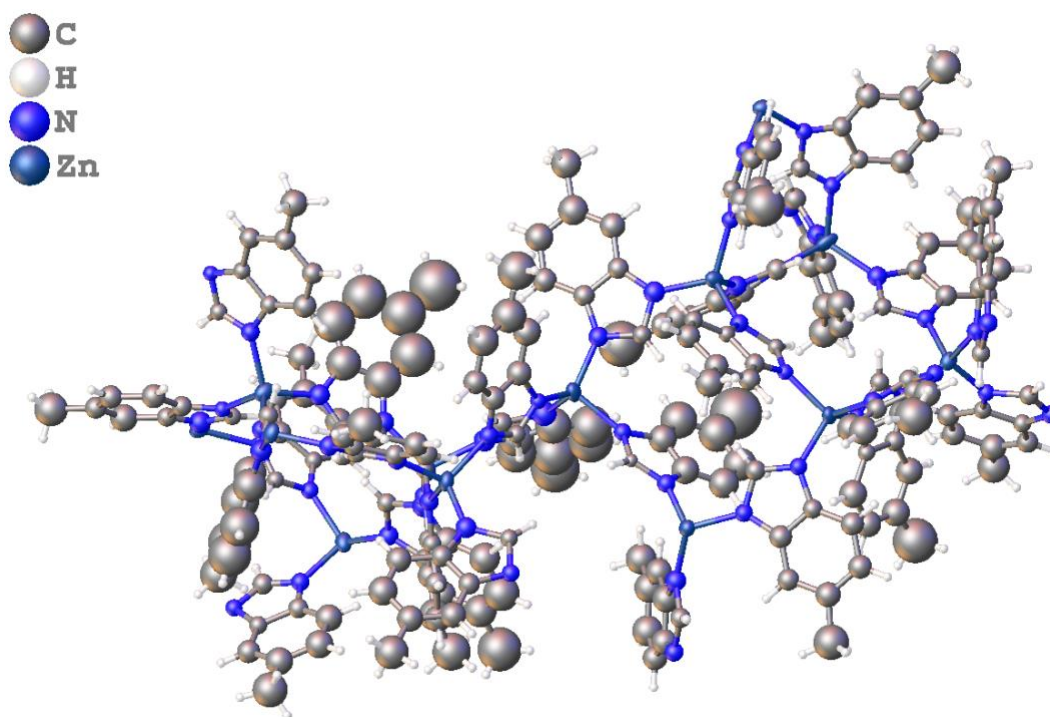


Figure 5.21. Asymmetric unit of the crystal structure of ZIF-A11. Ellipsoids are plotted with 15% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411080.

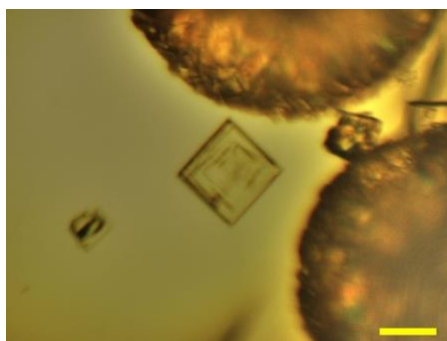


Figure 5.22. Representative optical image of ZIF-A11 crystallized from 6mbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 0.75, and a temperature of 130 °C. Yellow scale bar 100 μm .

Table 5.11. Crystal data and structure refinement for ZIF-A11.

Empirical formula	$\text{C}_{16}\text{H}_{14}\text{N}_4\text{Zn}$
Formula weight	327.68
Temperature/K	200.00
Crystal system	tetragonal
Space group	$P4_32_12$
$a/\text{\AA}$	29.5539(3)
$b/\text{\AA}$	29.5539(3)
$c/\text{\AA}$	70.0610(12)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	61193.6(16)
Z	104
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	0.925
$\mu (\text{mm}^{-1})$	1.433
$F(000)$	17472.0
Crystal size/ mm^3	$0.06 \times 0.06 \times 0.02$
Radiation	$\text{CuK}\alpha (\lambda = 1.54178)$
2θ range ($^\circ$)	3.912 to 74.742
Index ranges	$-23 \leq h \leq 23, -23 \leq k \leq 22, -55 \leq l \leq 50$
Reflections collected	100215
Independent reflections	15629 [$R_{\text{int}} = 0.0511, R_{\text{sigma}} = 0.0327$]
Data/restraints/parameters	15629/530/855
Goodness-of-fit on F^2	1.028
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0669, wR_2 = 0.1832$
Final R indexes [all data]	$R_1 = 0.0740, wR_2 = 0.1902$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.30 / -0.33
Flack parameter	0.48(5)

Refinement details of LZIF-A1. The entire dataset was cut to a resolution of 0.77 Å, resulting in an I/σ value of 45.1, an R_{int} values of 3.55%, and a completeness of 99.6%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. The weight parameters a and b converged to 0.33 and 19.0, respectively.

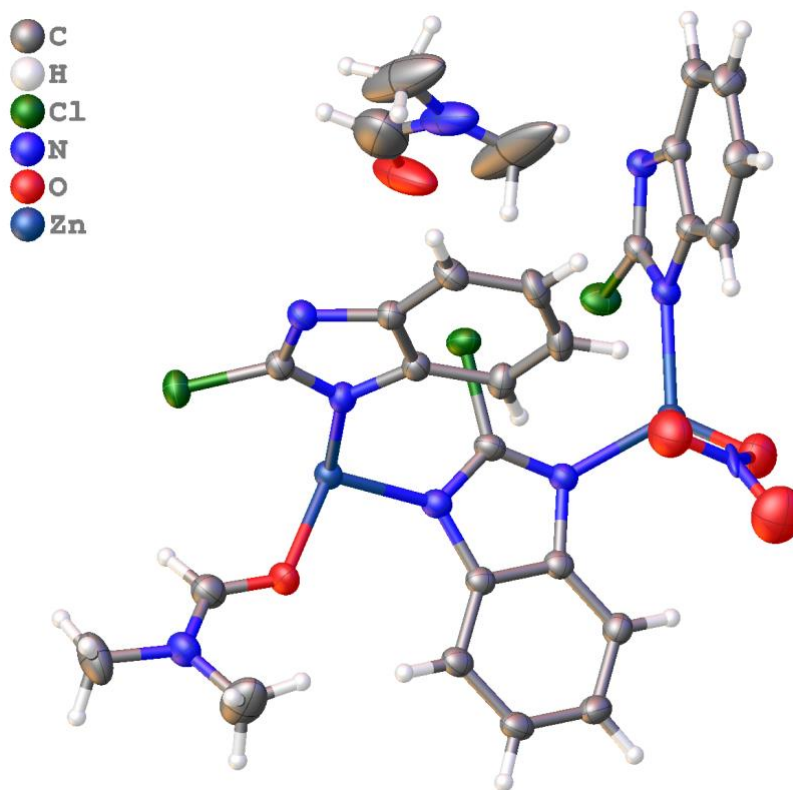


Figure 5.23. Asymmetric unit of the crystal structure of LZIF-A1. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue; Cl, green. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411239.

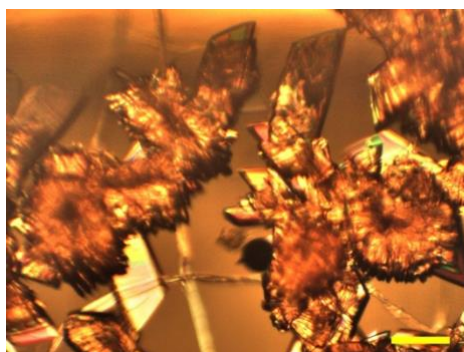


Figure 5.24. Representative optical image of LZIF-A1 crystallized from 2cbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 0.25, and a temperature of 70 °C. Yellow scale bar 100 μm .

Table 5.12. Crystal data and structure refinement for LZIF-A1.

Empirical formula	$\text{C}_{24}\text{H}_{19}\text{Cl}_3\text{N}_8\text{O}_4\text{Zn}_2 \cdot (\text{C}_3\text{H}_7\text{NO})$
Formula weight	793.66
Temperature/K	100
Crystal system	monoclinic
Space group	$\text{P2}_1/\text{n}$
$a/\text{\AA}$	10.2127(9)
$b/\text{\AA}$	23.446(2)
$c/\text{\AA}$	12.8222(11)
$\alpha/^\circ$	90
$\beta/^\circ$	92.915(3)
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	3066.2(5)
Z	4
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.719
$\mu (\text{mm}^{-1})$	2.000
$F(000)$	1608.0
Crystal size/ mm^3	$0.07 \times 0.05 \times 0.02$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	4.466 to 56.432
Index ranges	$-13 \leq h \leq 12, -30 \leq k \leq 29, -16 \leq l \leq 16$
Reflections collected	49352
Independent reflections	6315 [$R_{\text{int}} = 0.0355, R_{\text{sigma}} = 0.0222$]
Data/restraints/parameters	6315/18/419
Goodness-of-fit on F^2	1.123
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0565, wR_2 = 0.1317$
Final R indexes [all data]	$R_1 = 0.0622, wR_2 = 0.1344$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	1.82 / -0.82

Refinement details of LZIF-A2. The entire dataset was cut to a resolution of 0.85 Å, resulting in an I/σ value of 8.6, an R_{int} values of 10.67%, and a completeness of 99.8%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. The weight parameters a and b converged to 0.087 and 0, respectively.

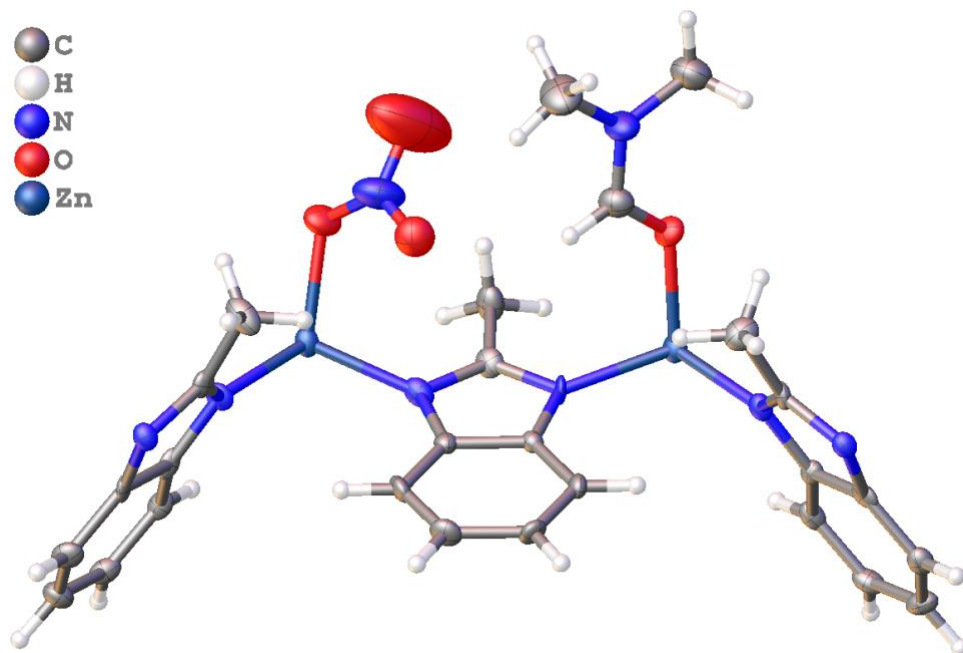


Figure 5.25. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue. Atomic labels are omitted due to the crowded atomic arrangement. CCDC reference number: 2411238.

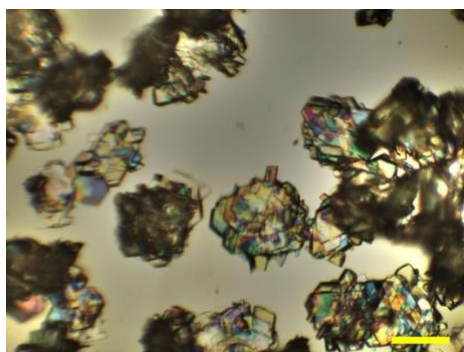


Figure 5.26. Representative optical image of LZIF-A2 crystallized from 2mbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.40 mol/L, a log linker-to-metal ratio of 1.00, and a temperature of 100 °C. Yellow scale bar 100 μm .

Table 5.13. Crystal data and structure refinement for LZIF-A2.

Empirical formula	$\text{C}_{27}\text{H}_{28}\text{N}_8\text{O}_4\text{Zn}_2$
Formula weight	659.31
Temperature/K	100
Crystal system	triclinic
Space group	P-1
$a/\text{\AA}$	10.0427(9)
$b/\text{\AA}$	10.4078(10)
$c/\text{\AA}$	14.8995(14)
$\alpha/^\circ$	82.247(4)
$\beta/^\circ$	82.335(4)
$\gamma/^\circ$	62.190(3)
Volume/ $\text{\AA}^3$	1360.5(2)
Z	2
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.609
$\mu (\text{mm}^{-1})$	1.927
$F(000)$	676.0
Crystal size/ mm^3	$0.03 \times 0.01 \times 0.01$
Radiation	Synchrotron ($\lambda = 0.7288$)
2θ range ($^\circ$)	2.838 to 50.77
Index ranges	$-11 \leq h \leq 11, -12 \leq k \leq 12, -17 \leq l \leq 17$
Reflections collected	15172
Independent reflections	4647 [$R_{\text{int}} = 0.1067, R_{\text{sigma}} = 0.1164$]
Data/restraints/parameters	4647/3/375
Goodness-of-fit on F^2	0.979
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0549, wR_2 = 0.1435$
Final R indexes [all data]	$R_1 = 0.0840, wR_2 = 0.1575$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.85 / -0.94

Refinement details of LZIF-A3. The entire dataset was cut to a resolution of 0.80 Å, resulting in an I/σ value of 34.2, an R_{int} values of 4.12%, and a completeness of 96.4%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. Four pores per unit cell, each with a volume of 278.7 Å³ and containing 79 electrons, were revealed after applying solvent masks. The positional disorder of the fluorine atoms was resolved by setting the occupancies of F1 and F2 to 0.5 due to the mirror plane symmetry relationship, while refining the occupancies of F3 and F4 to 0.628(6) and 0.372(6), respectively. The weight parameters a and b converged to 0.066 and 11.1, respectively.

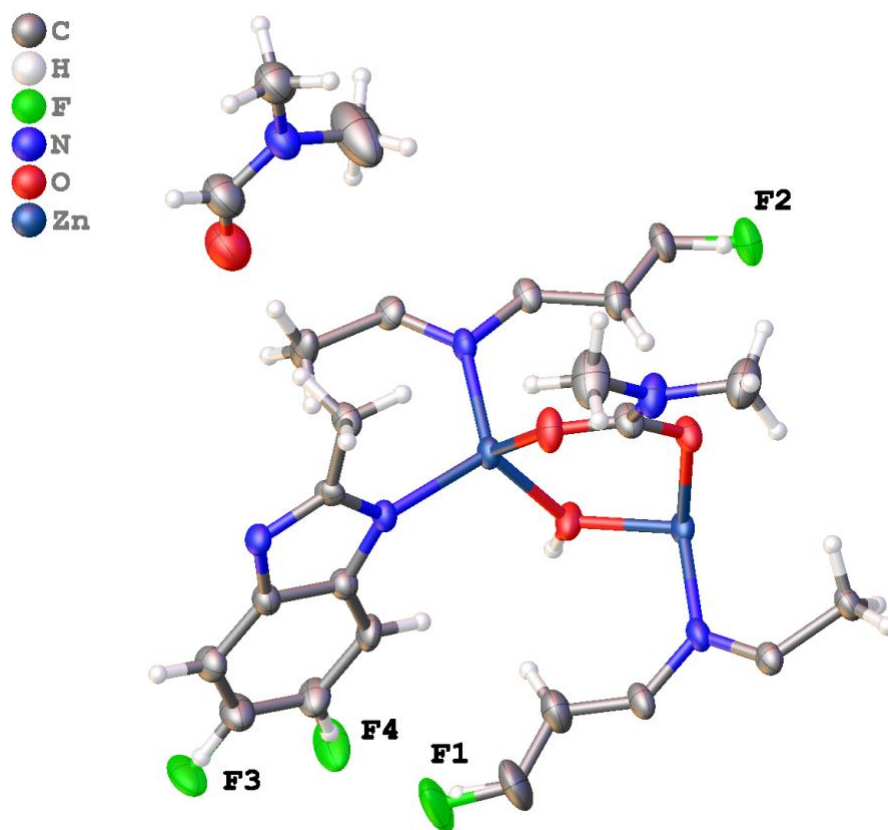


Figure 5.27. Asymmetric unit of the crystal structure of LZIF-A3. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; H, white; Zn, deep blue; F, lime green. Atomic labels except fluorine are omitted due to the crowded atomic arrangement. CCDC reference number: 2411236.

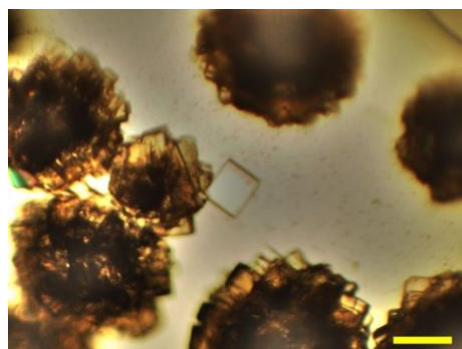


Figure 5.28. Representative optical image of LZIF-A3 crystallized from 2m5fbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 0.00, and a temperature of 130 °C. Yellow scale bar 100 μm .

Table 5.14. Crystal data and structure refinement for LZIF-A3.

Empirical formula	$\text{C}_{19}\text{H}_{19}\text{F}_2\text{N}_5\text{O}_3\text{Zn}_2 \cdot 0.5(\text{C}_3\text{H}_7\text{NO})$
Formula weight	570.68
Temperature/K	100
Crystal system	orthorhombic
Space group	Pnma
$a/\text{\AA}$	20.1801(4)
$b/\text{\AA}$	16.0267(4)
$c/\text{\AA}$	16.7307(3)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	5411.1(2)
Z	8
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.401
$\mu (\text{mm}^{-1})$	2.580
$F(000)$	2320.0
Crystal size/ mm^3	$0.1 \times 0.1 \times 0.03$
Radiation	Cu K_α ($\lambda = 1.54178$)
2θ range ($^\circ$)	6.862 to 149.002
Index ranges	$-23 \leq h \leq 21, -17 \leq k \leq 13, -20 \leq l \leq 16$
Reflections collected	26886
Independent reflections	5292 [$R_{\text{int}} = 0.0412, R_{\text{sigma}} = 0.0292$]
Data/restraints/parameters	5292/1/342
Goodness-of-fit on F^2	1.044
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0462, wR_2 = 0.1235$
Final R indexes [all data]	$R_1 = 0.0613, wR_2 = 0.1359$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	1.08 / -0.62

Refinement details of LZIF-A4. The entire dataset was cut to a resolution of 1.00 Å, resulting in an I/σ value of 26.3, an R_{int} values of 7.37%, and a completeness of 99.7%. Anisotropic refinement was performed on all non-hydrogen atoms, while hydrogen atoms were placed on the calculated positions. Bad reflections with $(I_{\text{obs}} - I_{\text{calc}})/\sigma > 10$ were excluded. Four pores per unit cell, each with a volume of 290.7 Å³ and containing 78 electrons, were revealed after applying solvent masks. The positional disorder of the chlorine atoms was resolved by setting the occupancies of Cl1 and Cl2 to 0.5 due to the mirror plane symmetry relationship, while refining the occupancies of Cl3 and Cl4 to 0.737(5) and 0.263(5), respectively. The weight parameters a and b converged to 0.078 and 14.4, respectively.

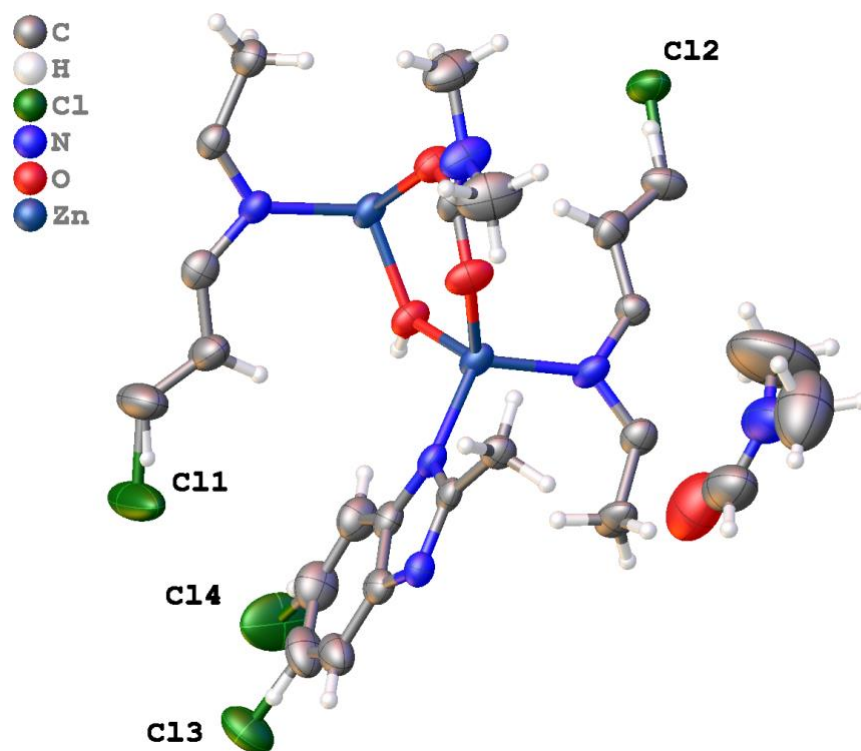


Figure 5.29. Asymmetric unit of the crystal structure of LZIF-A4. Ellipsoids are plotted with 50% probability. Color code: C, grey; N, blue; O, red; H, white; Zn, deep blue; Cl, green. Atomic labels except chlorine are omitted due to the crowded atomic arrangement. CCDC reference number: 2411237.

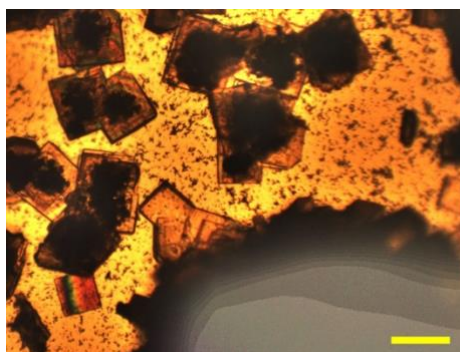


Figure 5.30. Representative optical image of LZIF-A4 crystallized from 2m5cbIM. This image was obtained from a reaction whose conditions were: a total concentration of 0.15 mol/L, a log linker-to-metal ratio of 0.00, and a temperature of 130 °C. Yellow scale bar 100 μm .

Table 5.15. Crystal data and structure refinement for LZIF-A4.

Empirical formula	$\text{C}_{19}\text{H}_{19}\text{Cl}_2\text{N}_5\text{O}_3\text{Zn}_2 \cdot 0.5(\text{C}_3\text{H}_7\text{NO})$
Formula weight	603.58
Temperature/K	150
Crystal system	orthorhombic
Space group	Pnma
$a/\text{\AA}$	21.4794(5)
$b/\text{\AA}$	16.1386(4)
$c/\text{\AA}$	16.6108(4)
$\alpha/^\circ$	90
$\beta/^\circ$	90
$\gamma/^\circ$	90
Volume/ $\text{\AA}^3$	5758.1(2)
Z	8
$\rho_{\text{calc}} (\text{g}/\text{cm}^3)$	1.392
$\mu (\text{mm}^{-1})$	4.020
$F(000)$	2448.0
Crystal size/ mm^3	$0.04 \times 0.02 \times 0.01$
Radiation	Cu K_α ($\lambda = 1.54178$)
2θ range ($^\circ$)	6.726 to 100.988
Index ranges	$-21 \leq h \leq 21, -16 \leq k \leq 16, -16 \leq l \leq 15$
Reflections collected	29065
Independent reflections	3152 [$R_{\text{int}} = 0.0737, R_{\text{sigma}} = 0.0380$]
Data/restraints/parameters	3152/2/342
Goodness-of-fit on F^2	1.057
Final R indexes [$I \geq 2 \sigma(I)$]	$R_1 = 0.0483, wR_2 = 0.1338$
Final R indexes [all data]	$R_1 = 0.0568, wR_2 = 0.1412$
Largest diff. peak/hole ($\text{e \AA}^{-3}$)	0.48 / -0.43

5.3 Topologies of ZIFs and LZIFs

Topological analysis of the SCXRD-determined crystal structures was carried out using the ToposPro software package³². The crystallographic information files (CIFs) obtained from SCXRD refinement were first imported into ToposPro. From these structural models, the connectivity of the framework was determined using the AutoCN module. This step constructs the adjacency matrix of the structure by automatically identifying bonding interactions based on geometric and Voronoi-based criteria. The resulting adjacency matrix defines the network of connections between atoms. The crystal structures were then simplified into node–edge representations using the ADS module.

Topological descriptors of the resulting nets were then computed using the ADS routines. These include coordination sequences, vertex symbols, and point symbols, which characterize the local and global connectivity of the network. Among these, point symbols provide a concise description of the ring environments surrounding each node, serving as a key identifier for distinguishing different topologies. To assign the identified nets to known topological types, the computed descriptors were compared against entries in the TopCryst database³³, which contains a large collection of known network topologies, including those from RCSR²⁰, EPINET³⁴, and other crystallographic repositories. Matching the computed descriptors to database entries enables automated identification of known nets, while unmatched cases indicate new topologies. In addition to the underlying net analysis, natural tiling of the structures was determined using ToposPro. Natural tiling partitions the periodic network into space-filling polyhedral units (tiles) that correspond to cages or cavities within the framework. The visualization of natural tiling is achieved through the 3dt program³⁵.

Seven topologies were identified among the eleven ZIF structures, and two additional topologies were identified for the four LZIF structures. Table 5.16 summarizes the topological descriptors of all nine topologies, including point symbols, tiles, and transitivity. The corresponding tile representations are also presented (Figures 5.31–5.38). Taken together, these results reveal clear correlations between linker structure and framework topology.

The SOD and ANA topologies, represented by ZIF-A1 and ZIF-A4 and obtained from 2cyIM and 2pIM, respectively, correspond to structures that were previously reported only under alkyl-alcohol solvent systems^{36,37}, in contrast to the DMF-based synthesis used here. Additional SOD structures, ZIF-A2 and ZIF-A3, were obtained from 4m5nIM and 6fbIM. Comparison with linkers yielding other topologies suggests that SOD favors imidazole-based linkers bearing small substituents at the 2-position, as well as benzimidazole-based linkers bearing relatively small substituents at the 5 or 6 positions. When the substituent at the 2-position becomes larger, the topology shifts from SOD to ANA, as illustrated by 2pIM. Likewise, when the substituents at the 5 or 6-positions of benzimidazole-based linkers become too extended, the resulting structures adopt the ykh topology, which is the only non-uninodal topology observed in this set, suggesting that stronger steric hindrance lowers the structural symmetry and likely introduces more complex assembly pathways involving a greater diversity of intermediate species. Linkers with double substitution at the 4 and 5 positions of the imidazole ring can lead to either SOD or RHO topologies, whereas linkers with a single substituent at the 4 or 5 position give rise to several particularly interesting structures. Specifically, ZIF-A5, obtained from 5BrIM and adopting the crb

topology, and ZIF-A7, obtained from 4cyIM and adopting the DFT topology, demonstrate that such singly substituted linkers can access topologies that were previously realized only with unsubstituted imidazole linkers, such as ZIF-1, ZIF-2, and ZIF-3¹.

Table 5.16. Topological classification and descriptors of ZIF and LZIF structures.

Topology	Compound	Point symbol	Tile	Transitivity ^a
SOD	ZIF-A1, ZIF-A2, ZIF-A3,	[4 ⁴ 2.6 ⁴]	[4 ⁶ 6.6 ⁸]	[1 1 2 1]
ANA	ZIF-4	[4 ⁴ 2.6 ² 8 ²]	3[4 ² 8 ²] + 2[6 ² 8 ³]	[1 1 3 2]
crb	ZIF-A5	[4.6 ⁵]	2[6 ² 8 ²] + [4 ² 6 ⁴]	[1 3 4 3]
fes (double)*	ZIF-A6	[4 ⁴ 4.8 ²]	[4 ⁶] + [4 ⁸ 8 ²]	[1 3 4 2]
DFT	ZIF-A7	[4 ⁴ 2.6 ² 8 ²]	[4 ² 6 ²] + [6 ² 8 ²] + [4 ² 8 ⁴]	[1 3 6 4]
RHO	ZIF-A8	[4 ⁴ 3.6 ³]	3[4 ⁸ 8 ²] + [4 ¹² 6 ⁸ 8 ⁶]	[1 4 6 4]
ykh	ZIF-A9, ZIF-A10, ZIF-A11	8[3.4 ² 5.6.7] + 4[3.4 ² 5 ² 7] + [4 ⁴ 6 ²]	2[3 ² 4.12 ²] + [4 ⁴ 12 ² 14 ²] + [3 ⁴ 4 ⁸ 10 ² 14 ²]	[13 26 18 5]
hcb	LZIF-A1, LZIF-A2	[6 ³]	/	[1 1 1]
fes	LZIF-A3, LZIF-A4	[4.8 ²]	/	[1 2 2]

All the identified ZIF topologies are three-dimensional except for the fes (double) topology observed for ZIF-A6. While there are previously reported structures exhibiting a single atomic layer built from tetrahedral units³⁸, ZIF-A6 shows a new double-thick-layer arrangement in which two layers of metal nodes are covalently bridged by 5BrIM, representing a previously unreported structural motif in ZIF chemistry. In this sense, it occupies an intermediate position between a conventional 2D and 3D framework. Because the standard definition of natural tiling is not directly applicable to this double-layer arrangement, only a single layer was considered for the tiling representation. The stacking of these layers is expected to involve halogen–halogen interactions associated with the bromo substituents³⁹.

In contrast, the LZIF structures are all two-dimensional. In these structures, the zinc centers are coordinated not only by the bidentate imidazolate-derived linker but also by solvent molecules or monodentate anions. In LZIF-A1, derived from 2cbIM, and LZIF-A2, derived from 2mbIM, DMF and

nitrate compete for coordination sites. In LZIF-A3, derived from 2m5fbIM, and LZIF-A4, derived from 2m5cbIM, the competing ligand is dimethylcarbamate, which forms through oxidation of DMF. Their point symbols therefore involve only three circuits.

A deeper understanding of how linker structure and reaction conditions determine topology will require more advanced modeling of both thermodynamics and kinetics during nucleation. It will be necessary to determine which species dominate under given conditions, what conformations these species preferentially adopt, and how those conformations evolve into the final network topology.

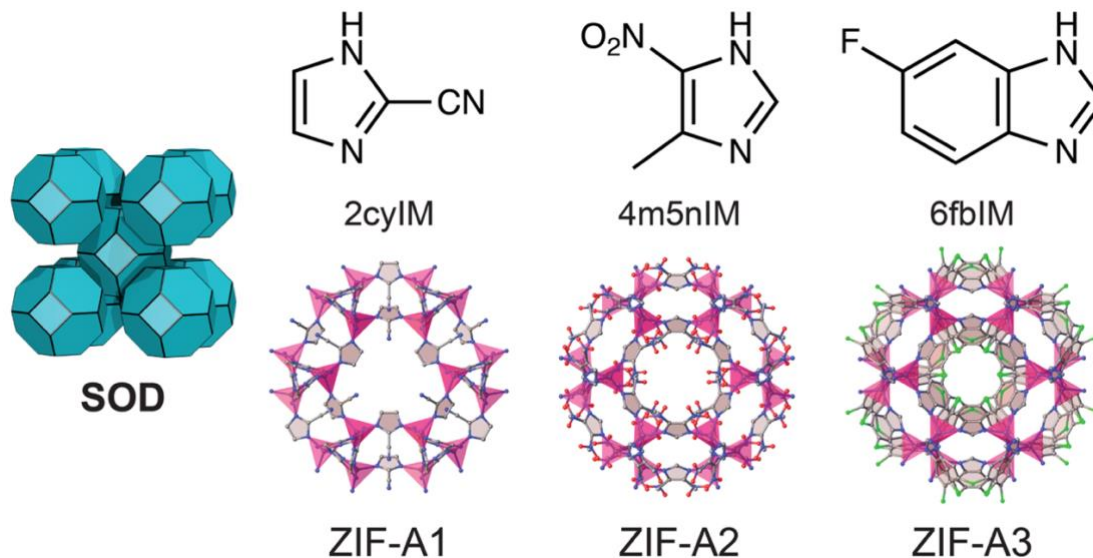


Figure 5.31. Tile representation of SOD topology and the corresponding ZIF structures (ZIF-A1, ZIF-A2, and ZIF-A3).

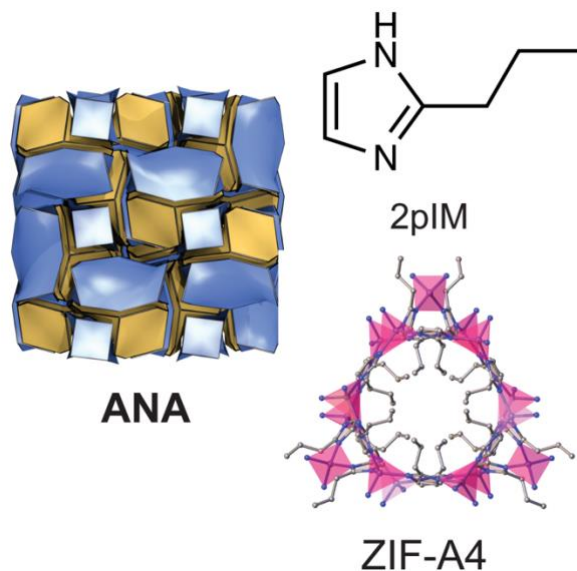


Figure 5.32. Tile representation of ANA topology and the corresponding ZIF structure (ZIF-A4).

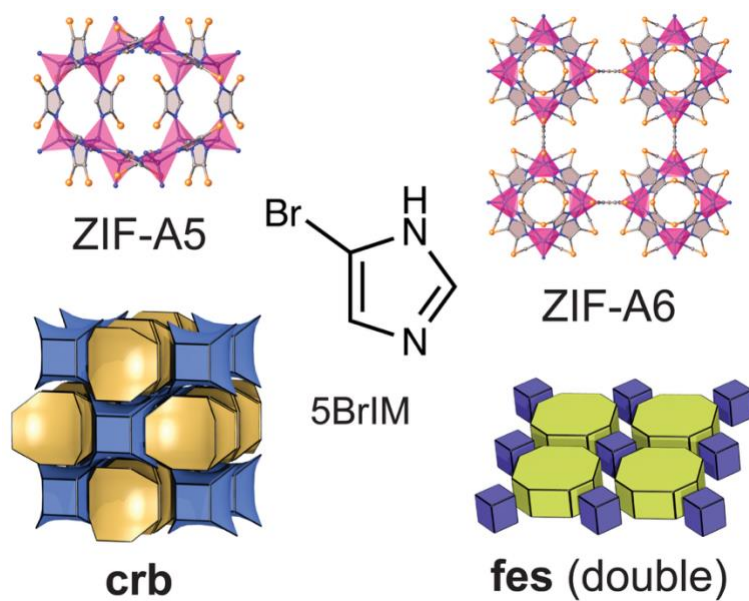


Figure 5.33. Tile representation of **crb** and **fes(double)** topology and their corresponding ZIF structures (ZIF-A5 and ZIF-A6, respectively).

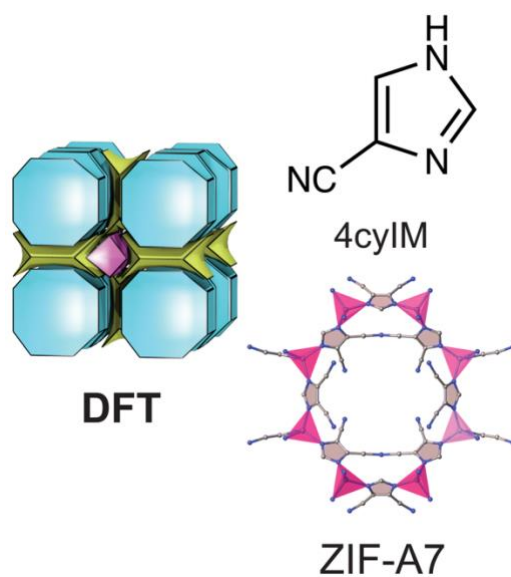


Figure 5.34. Tile representation of **DFT** topology and the corresponding ZIF structure (ZIF-A7).

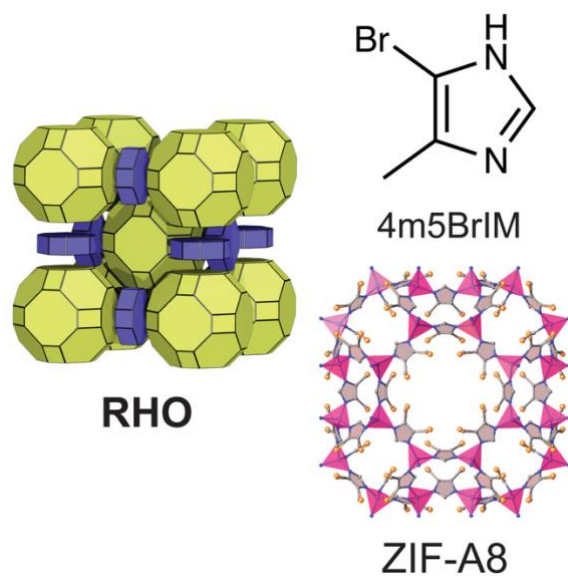


Figure 5.35. Tile representation of RHO topology and the corresponding ZIF structure (ZIF-A8).

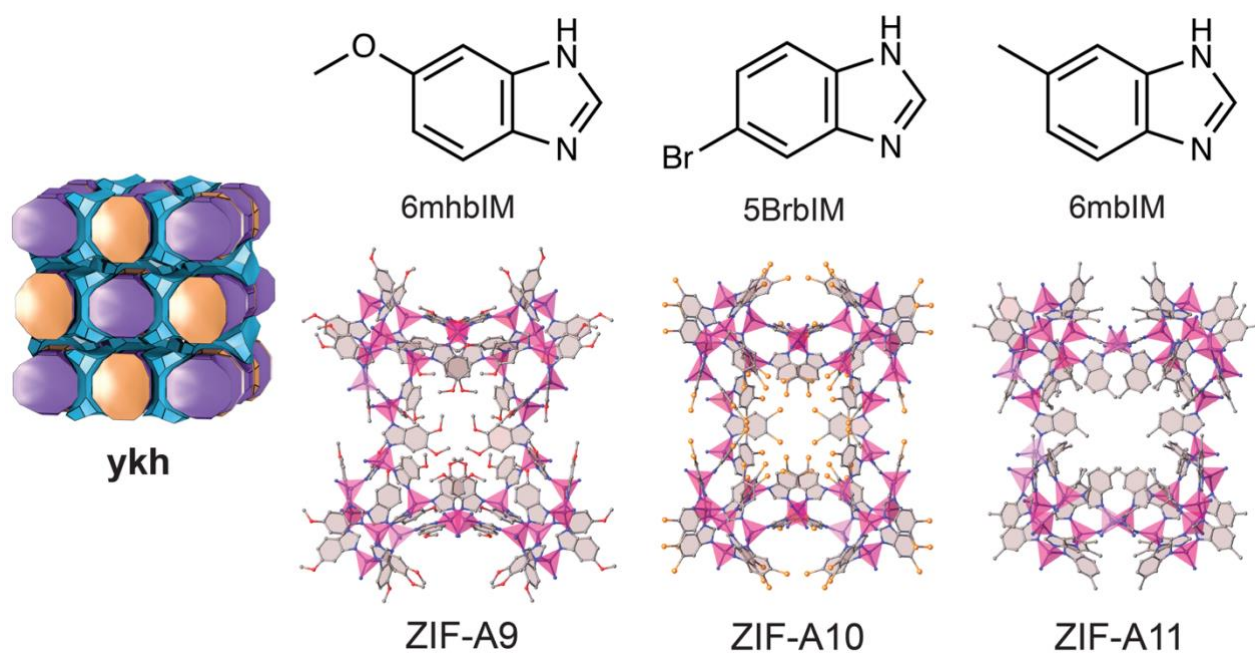


Figure 5.36. Tile representation of ykh topology and the corresponding ZIF structures (ZIF-A9, ZIF-A10, and ZIF-A11).

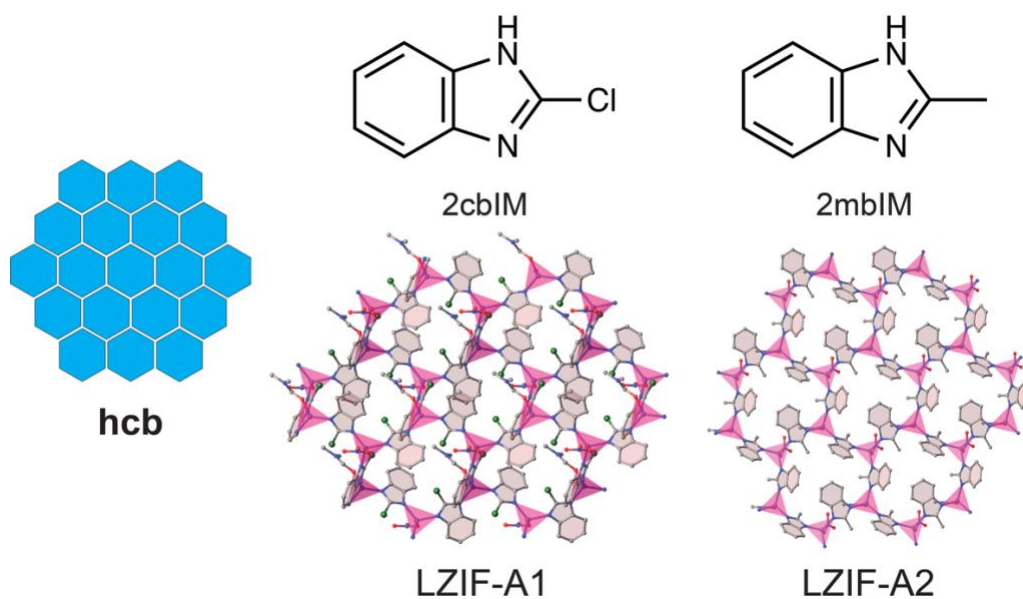


Figure 5.37. Tile representation of hcb topology and the corresponding LZIF structures (LZIF-A1 and LZIF-A2).

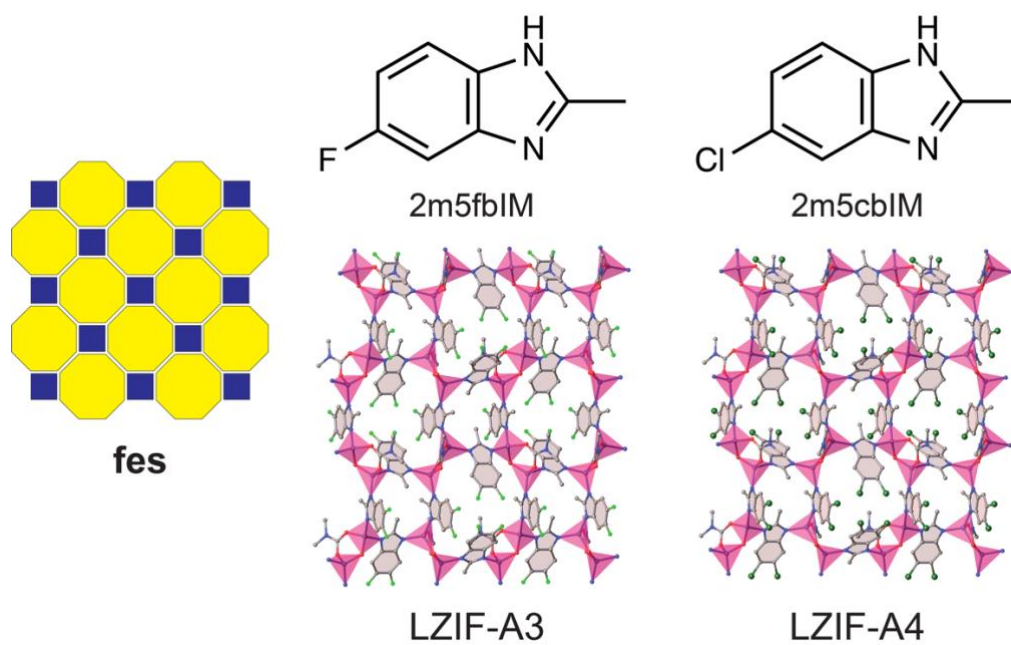


Figure 5.38. Tile representation of fes topology and the corresponding LZIF structures (LZIF-A3 and LZIF-A4).

5.4 References

- (1) Park, K. S.; Ni, Z.; Côté, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O’Keeffe, M.; Yaghi, O. M. Exceptional Chemical and Thermal Stability of Zeolitic Imidazolate Frameworks. *Proc. Natl. Acad. Sci.* **2006**, *103* (27), 10186–10191.
- (2) Phan, A.; Doonan, C. J.; Uribe-Romo, F. J.; Knobler, C. B.; O’Keeffe, M.; Yaghi, O. M. Synthesis, Structure, and Carbon Dioxide Capture Properties of Zeolitic Imidazolate Frameworks. *Acc. Chem. Res.* **2010**, *43* (1), 58–67.
- (3) Noh, K.; Lee, J.; Kim, J. Compositions and Structures of Zeolitic Imidazolate Frameworks. *Isr. J. Chem.* **2018**, *58* (9–10), 1075–1088.
- (4) Wang, H.; Pei, X.; Kalmutzki, M. J.; Yang, J.; Yaghi, O. M. Large Cages of Zeolitic Imidazolate Frameworks. *Acc. Chem. Res.* **2022**, *55* (5), 707–721.
- (5) Zheng, Z.; Rong, Z.; Nguyen, H. L.; Yaghi, O. M. Structural Chemistry of Zeolitic Imidazolate Frameworks. *Inorg. Chem.* **2023**, *62* (51), 20861–20873.
- (6) Gropp, C.; Canossa, S.; Wuttke, S.; Gándara, F.; Li, Q.; Gagliardi, L.; Yaghi, O. M. Standard Practices of Reticular Chemistry. *ACS Cent. Sci.* **2020**, *6* (8), 1255–1273.
- (7) Glasby, L. T.; Cordiner, J. L.; Cole, J. C.; Moghadam, P. Z. Topological Characterization of Metal–Organic Frameworks: A Perspective. *Chem. Mater.* **2024**, *36* (19), 9013–9030.
- (8) Guo, Z.; Zhang, Z.; Sun, J. Topological Analysis and Structural Determination of 3D Covalent Organic Frameworks. *Adv. Mater.* **2024**, *36* (19), 2312889.
- (9) Tan, J. C.; Bennett, T. D.; Cheetham, A. K. Chemical Structure, Network Topology, and Porosity Effects on the Mechanical Properties of Zeolitic Imidazolate Frameworks. *Proc. Natl. Acad. Sci.* **2010**, *107* (22), 9938–9943.
- (10) Evans, J. D.; Coudert, F.-X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29* (18), 7833–7839.
- (11) Bryant, M. J.; Maloney, A. G. P.; Sykes, R. A. Predicting Mechanical Properties of Crystalline Materials through Topological Analysis. *CrystEngComm* **2018**, *20* (19), 2698–2704.
- (12) Wang, H.; Dong, X.; Lin, J.; Teat, S. J.; Jensen, S.; Cure, J.; Alexandrov, E. V.; Xia, Q.; Tan, K.; Wang, Q.; Olson, D. H.; Proserpio, D. M.; Chabal, Y. J.; Thonhauser, T.; Sun, J.; Han, Y.; Li, J. Topologically Guided Tuning of Zr-MOF Pore Structures for Highly Selective Separation of C6 Alkane Isomers. *Nat. Commun.* **2018**, *9* (1), 1745.
- (13) Shevchenko, A. P.; Alexandrov, E. V.; Golov, A. A.; Blatova, O. A.; Duyunova, A. S.; Blatov, V. A. Topology *versus* Porosity: What Can Reticular Chemistry Tell Us about Free Space in Metal–Organic Frameworks? *Chem. Commun.* **2020**, *56* (67), 9616–9619.
- (14) Chen, Y.; Hu, W.; Li, X.; Wang, C.; Yang, W. Machine Learning Prediction of Transition-State Energies in Small-Pore Zeolites Combining Acidity and Topology Descriptors. *J. Phys. Chem. C* **2024**, *128* (9), 3770–3779.

- (15) Wells, A. F. *Three-Dimensional Nets and Polyhedra*; Wiley monographs in crystallography; Wiley: New York, 1977.
- (16) Delgado-Friedrichs, O.; O’Keeffe, M. Identification of and Symmetry Computation for Crystal Nets. *Acta Crystallogr. A* **2003**, *59* (4), 351–360.
- (17) O’Keeffe, M.; Hyde, B. G. *Crystal Structures: Patterns and Symmetry*, Dover edition.; Dover Publications, Inc: Mineola, New York, 2020.
- (18) Ockwig, N. W.; Delgado-Friedrichs, O.; O’Keeffe, M.; Yaghi, O. M. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Acc. Chem. Res.* **2005**, *38* (3), 176–182.
- (19) O’Keeffe, M.; Yaghi, O. M. Deconstructing the Crystal Structures of Metal–Organic Frameworks and Related Materials into Their Underlying Nets. *Chem. Rev.* **2012**, *112* (2), 675–702.
- (20) O’Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.* **2008**, *41* (12), 1782–1789.
- (21) Kalmutzki, M. J.; Hanikel, N.; Yaghi, O. M. Secondary Building Units as the Turning Point in the Development of the Reticular Chemistry of MOFs. *Sci. Adv.* **2018**, *4* (10), eaat9180.
- (22) Friedrichs, O. D.; Dress, A. W. M.; Huson, D. H.; Klinowski, J.; Mackay, A. L. Systematic Enumeration of Crystalline Networks. *Nature* **1999**, *400* (6745), 644–647.
- (23) Blatov, V. A.; Delgado-Friedrichs, O.; O’Keeffe, M.; Proserpio, D. M. Three-Periodic Nets and Tilings: Natural Tilings for Nets. *Acta Crystallogr. A* **2007**, *63* (5), 418–425.
- (24) Bruker AXS Inc. APEX4, 2021.
- (25) Bruker AXS Inc. SAINT, 2020.
- (26) Bruker AXS Inc. SADABS, 2016.
- (27) Sheldrick, G. M. *SHELXT* – Integrated Space-Group and Crystal-Structure Determination. *Acta Crystallogr. Sect. Found. Adv.* **2015**, *71* (1), 3–8.
- (28) Sheldrick, G. M. A Short History of *SHELX*. *Acta Crystallogr. A* **2008**, *64* (1), 112–122.
- (29) Sheldrick, G. M. Crystal Structure Refinement with *SHELXL*. *Acta Crystallogr. Sect. C Struct. Chem.* **2015**, *71* (1), 3–8.
- (30) Dolomanov, O. V.; Bourhis, L. J.; Gildea, R. J.; Howard, J. A. K.; Puschmann, H. *OLEX2*: A Complete Structure Solution, Refinement and Analysis Program. *J. Appl. Crystallogr.* **2009**, *42* (2), 339–341.
- (31) Van Der Sluis, P.; Spek, A. L. BYPASS: An Effective Method for the Refinement of Crystal Structures Containing Disordered Solvent Regions. *Acta Crystallogr. A* **1990**, *46* (3), 194–201.
- (32) Blatov, V. A.; Shevchenko, A. P.; Proserpio, D. M. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Cryst. Growth Des.* **2014**, *14* (7), 3576–3586.

- (33) Shevchenko, A. P.; Shabalin, A. A.; Karpukhin, I. Yu.; Blatov, V. A. Topological Representations of Crystal Structures: Generation, Analysis and Implementation in the *TopCryst* System. *Sci. Technol. Adv. Mater. Methods* **2022**, *2* (1), 250–265.
- (34) Ramsden, S. J.; Robins, V.; Hyde, S. T. Three-Dimensional Euclidean Nets from Two-Dimensional Hyperbolic Tilings: Kaleidoscopic Examples. *Acta Crystallogr. A* **2009**, *65* (2), 81–108.
- (35) Delgado-Friedrichs, O. The Gavrog Project. <http://www.gavrog.org/>.
- (36) Li, K.; Olson, D. H.; Seidel, J.; Emge, T. J.; Gong, H.; Zeng, H.; Li, J. Zeolitic Imidazolate Frameworks for Kinetic Separation of Propane and Propene. *J. Am. Chem. Soc.* **2009**, *131* (30), 10368–10369.
- (37) Tian, Y.; Yao, S.; Gu, D.; Cui, K.; Guo, D.; Zhang, G.; Chen, Z.; Zhao, D. Cadmium Imidazolate Frameworks with Polymorphism, High Thermal Stability, and a Large Surface Area. *Chem. – Eur. J.* **2010**, *16* (4), 1137–1141.
- (38) Yang, Q.-F.; Cui, X.-B.; Yu, J.-H.; Lu, J.; Yu, X.-Y.; Zhang, X.; Xu, J.-Q.; Hou, Q.; Wang, T.-G. A Series of Metal–Organic Complexes Constructed from in Situ Generated Organic Amines. *CrystEngComm* **2008**, *10* (11), 1534.
- (39) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478–2601.

Chapter 6

Outlook

This dissertation has presented AIRES as an autonomous platform for reticular materials discovery by integrating high-throughput synthesis, automated optical microscopy with computer vision, single-crystal X-ray diffraction, machine learning prediction, and algorithmic decision-making. The results demonstrate that such a system can move beyond passive automation and function as an iterative discovery engine. At the same time, the present implementation represents only one stage within a broader vision of digital materials discovery.

A first direction for future development is the tighter integration of generative modeling and computational screening with the experimental platform. In the current workflow, machine learning models predict crystallization outcomes within a defined candidate space. In the future, this candidate space could itself be proposed algorithmically. Generative models could suggest new linkers, linker combinations, or precursor chemistries, while computational screening could eliminate infeasible or non-promising candidates before experimental execution. In the broader digital discovery cycle, this corresponds to the design layer, where user-defined goals are translated into candidate structures through generative models, quantum-level geometry optimization, and machine learning-based synthesizability prediction. In such a framework, experiment would no longer begin with a manually curated list of molecules, but with a model-guided proposal of chemically meaningful, synthesizable targets.

A second direction concerns the continued expansion of the experimental platform itself. AIRES currently demonstrates the value of automating reaction preparation, imaging, and structural verification, but future systems should be more modular, adaptable, and scalable. This includes integration with additional equipment, such as powder X-ray diffraction, infrared and Raman spectroscopy, thermal analysis, and liquid- or solid-handling systems with broader chemical compatibility. Equally important is increasing the scale and flexibility of parallel experimentation. Some workflows will continue to rely on 96-well plate formats, but future autonomous laboratories should also be capable of larger-scale batch or flow-based experimentation, as well as flexible transfer between instruments. In the broader digital cycle, this corresponds to a more advanced synthesis layer, where automated synthesis, multimodal characterization, and machine learning-assisted data analysis are all tightly connected.

A third direction is extension to other classes of materials and more complex synthesis spaces. The present work uses single-linker Zn-ZIFs as a rigorous demonstration system, but the same closed-loop methodology can be extended to multi-linker ZIFs, other MOFs, COFs, hybrid crystalline solids, or even broader classes of functional materials. In such systems, however, the characterization problem becomes more difficult. Single crystals are not always obtainable, and the outcome may consist of mixed

phases, poorly crystalline solids, or metastable intermediates. As a result, future autonomous discovery platforms will require more sophisticated and multi-modal characterization strategies. SCXRD will remain crucial whenever definitive structures can be obtained, but it should be complemented by PXRD, spectroscopy, microscopy, and other measurements that together provide a more complete description of the reaction outcome. This is especially important because, as the chemistry becomes more complex, the challenge is no longer only to detect success, but to define and label success in a scientifically rigorous way.

A fourth direction concerns the feature representations and machine learning models that underlie prediction and decision-making. This dissertation shows that physically informed descriptors, especially those encoding linker geometry, electronic structure, and prenucleation interactions, are critical for effective learning in low-data regimes. This will likely remain true even as the platform expands. At the same time, as larger and more diverse datasets become available, there will be increasing opportunity to use neural network models, graph-based architectures, and representation-learning approaches that can infer useful chemical features directly from structure. Such models may become especially valuable when moving to systems with greater compositional diversity, more flexible building units, or richer characterization outputs. The long-term opportunity is not to replace chemical intuition with black-box models, but to combine physics-informed features, domain knowledge, and data-driven architectures within a unified predictive framework.

A fifth direction is to connect discovery more directly with optimization, scale-up, and deployment. AIRES currently identifies crystalline products and solves their structures, but the discovery of a material is only the beginning of its technological life cycle. Experimentally validated structures can serve as structural anchors for downstream development, including scale-up synthesis, phase verification, crystallinity optimization, and property testing. In the full digital discovery cycle, this corresponds to the transition from discovery to scaling, where material and form factor, device configuration, field-data collection, and eventual deployment are considered. The solved crystal structure is especially valuable here because it provides a definitive reference for subsequent bulk characterization, including PXRD-based phase identification and purity assessment during scale-up.

Finally, the future of autonomous discovery should be considered not only in terms of laboratory efficiency, but also in terms of societal deployment. A mature digital discovery ecosystem would connect user-defined needs, generative design, autonomous synthesis, structural verification, scaling, and real-world application into one continuous loop. In this vision, autonomous platforms do not simply produce more data; they shorten the time between scientific question and usable material. Such systems could accelerate the development of porous materials for gas separation, catalysis, sensing, and other applications of societal relevance. Realizing this vision will require not only better models and instruments, but also standardized datasets, interoperable software, robust labeling strategies, and careful consideration of how discovered materials move from academic proof-of-concept to manufacturable and deployable technologies.