Inverse Design: Generative AI for Carbon Capture Materials (MOFs)

Upendra Singh Tomar[1], Ujjwal Kushwah[2], Anmol Sharma[3], Dr. S.K Sharma
{1, 2, 3}CS-Data Science, ITM GOI, Gwalior, India {4}Associate Professor, HOD Department of Mechanical Engineering, ITM GOI, Gwalior, India
{1} upendratomar1100@gmail.com,{2} ujjwalkushwah24@gmail.com, {3} anmolsharma9302236197@gmail.com

Abstract—

The discovery of novel functional molecules is a critical bottleneck in addressing global challenges, ranging from life-threatening diseases to climate change. The chemical space of potential drug-like and material candidates is vast, with at over 10^{60} structures; however, traditional discovery pipelines rely on the slow iterative screening of existing libraries. This study presents a unified Inverse Design framework utilizing Generative Artificial Intelligence (GAI) to accelerate the discovery of therapeutic small molecules and Metal-Organic Frameworks (MOFs) for carbon capture. We employed deep generative models, specifically Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), to construct novel molecular structures optimized for specific target properties. For pharmaceutical applications, the models were conditioned to generate compounds with high binding affinities for oncology targets and low toxicities. Simultaneously, in materials science, the framework was applied to design MOFs with maximized CO_2 adsorption capacity and selectivity. We validated the generated structures using computational simulations, including molecular docking of drug candidates and Grand Canonical Monte Carlo (GCMC) simulations of MOFs. Our results demonstrate that generative models successfully navigate the chemical space to produce valid, synthesizable, and high-performance candidates that outperform those obtained through random sampling. This study highlights the versatility of Generative AI as a platform technology capable of significantly reducing the time and cost associated with the hit-to-lead phase in both pharmaceutical R&D and materials engineering.

Keywords— Generative AI, Inverse Design, Metal-Organic Frameworks (MOFs), De Novo Drug Design, Deep Learning, Carbon Capture.

I. INTRODUCTION

A. Background

Humanity currently faces dual existential challenges that demand immediate scientific intervention: the biological imperative to cure life-threatening pathologies, such as cancer and infectious diseases (e.g., COVID-19 and rabies), and the environmental necessity to mitigate climate change through efficient carbon capture [1]. In both domains, the solution lies in the discovery of novel functional materials, whether they are small-molecule therapeutics or porous Metal-Organic Frameworks (MOFs).

However, traditional discovery pipelines are severely limited by the vastness of unexplored chemical space, which is estimated to contain over 10^{60} drug-like compounds. [2]. Conventional methods that rely on the iterative screening of existing libraries are slow and inefficient. This "trial-and-error" approach is analogous to searching for a needle in a universal-sized haystack. With the rapid mutation of viral pathogens and the accelerating pace of global warming, the timeline for discovery must be reduced from decades to months.

B. The Problem Statement

The core deficiency of the current paradigm lies in its reliance on "forward design," a process plagued by both cognitive bias and systemic inefficiency. In pharmaceutical R&D, this manifests as Eroom's Law, where the cost of developing a new drug doubles every nine years, despite technological advancements [3]. Billions of dollars are wasted on candidates that fail in late-stage clinical trials because initial screening libraries are biased toward known chemical scaffolds, leaving vast and potentially fruitful regions of chemical space untouched.

Similarly, in materials science, the search for optimal MOFs is constrained by human intuition, which typically iterates on existing topologies rather than exploring novel architectures [4]. This "historical bias" restricts innovation, making it nearly impossible to identify non-intuitive structures that might offer superior performance for complex tasks, such as selective carbon capture or crossing the blood-brain barrier. There is an urgent need for a computational framework that can transcend these human limitations and transform discoveries from serendipitous searches into targeted engineering disciplines.

C. Proposed Solution

Inverse Design with Generative AI To address these limitations, this study proposes an Inverse Design framework powered by Generative AI. Unlike traditional forward design, which evaluates the properties of known structures, inverse design begins with the desired target properties (e.g., "high CO_2 selectivity" or "high binding affinity") and utilizes machine learning algorithms to generate molecular structures that satisfy these criteria [5].

We leverage advanced deep learning architectures, specifically Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to act as molecular architects. These models learn the underlying syntax and grammar of chemical structures from large database. By navigating the continuous latent space of molecules, AI can "imagine" and construct novel candidates that do not exist in current libraries but are optimized for specific biological or physical constraints. This approach effectively bridges the gap between computational

 ISSN: 3107-6513

prediction and experimental reality, offering a scalable solution to the molecular discovery crisis in drug development.

II. RELATED WORK

A. *Evolution from Screening to Generative Design Define*

Historically, pharmaceutical and material discovery has been dominated by High-Throughput Screening (HTS), a brute-force methodology in which vast libraries of pre-existing compounds are physically tested against biological or physical targets. While HTS has served as the industry standard for decades, it is inherently constrained by the "library limitation"the fact that physical libraries represent a negligible fraction of the estimated 10^{60} drug-like chemical space [2]. Early computational efforts, such as quantitative structure–activity relationship (QSAR) modeling, attempted to mitigate this by predicting activity from structure. However, these models are fundamentally discriminative and can only evaluate existing candidates and not generate novel molecular architectures.

B. *Inverse Design in Materials Science*

Use The paradigm of "Inverse Design"wherein target properties are specified *a priori* to guide the generation of molecular structures, has catalyzed a revolution in materials engineering. A foundational success of this approach is the design of Metal-Organic Frameworks (MOFs) for carbon capture [6]. By formulating material design as a graph generation problem, recent studies have demonstrated that generative models can successfully navigate the chemical space to identify MOFs with maximized CO_2 adsorption isotherms and selectivity over N_2.

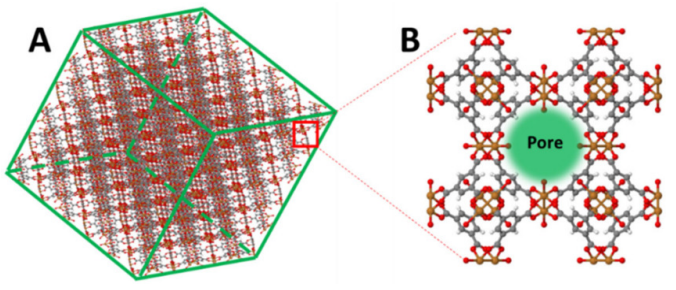

Fig. 1 A typical Metal-Organic Framework (MOF) structure showing the metal nodes and organic linkers forming a porous network suitable for gas adsorption [6].

C. *Generative Deep Learning in Healthcare*

The success of inverse design in materials has accelerated the adoption of Deep Generative Models (DGMs) for *de novo* drug design.

- *Variational Autoencoders (VAEs):*

  The seminal work by Gómez-Bombarelli et al. [7] utilized VAEs to map discrete chemical representations (SMILES strings) into a continuous, differentiable latent space. This allows gradient-based optimization of molecular properties, effectively transforming molecular design into a search problem within a high-dimensional vector space.

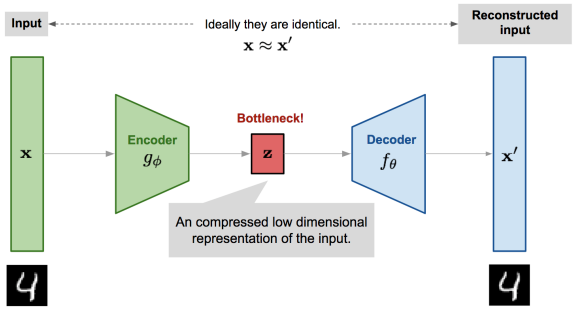

Fig. 2 The architecture of a Variational Autoencoder (VAE), illustrating the encoder mapping a molecule to a continuous latent space and the decoder reconstructing it [7].

- *Generative Adversarial Networks (GANs):*

  Parallel research has leveraged GANs (e.g., MolGAN and ORGAN) to directly generate molecular graphs [8]. While GANs have historically struggled with mode collapse, producing a limited diversity of samples, they excel at generating highly realistic structures that strictly adhere to valence constraints and chemical validity.

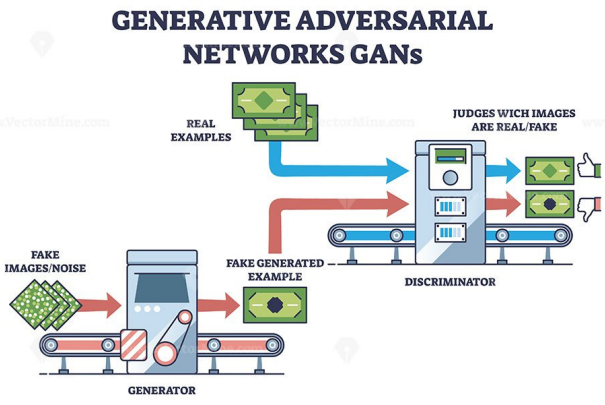

Fig. 3 The architecture of a Generative Adversarial Network (GAN), showing the Generator creating candidates and the Discriminator evaluating them against real data [8].

- *Reinforcement Learning (RL):*

  Emerging frameworks have integrated Reinforcement Learning to guide the generative

 ISSN: **3107-6513**

process. Here, the molecule generator acts as an agent within an environment, receiving "rewards" for generating structures that satisfy multi-objective constraints, such as high solubility (LogP), low toxicity (QED), and specific target binding affinity [9].

D. Research Gap

Despite these methodological advancements, a critical gap remains. Most existing generative models optimize for generic metrics of "drug-likeness" (e.g., QED scores) rather than specific, high-affinity binding to complex disease targets. Furthermore, a unified framework bridging the domain gap between rigid material design (MOFs) and flexible drug discovery is lacking. This study addresses these limitations by proposing a hybrid generative approach that leverages the architectural strengths of both VAEs and GANs to design candidates specifically optimized for oncology and neglected infectious diseases.

TABLE I.

COMPARISON OF MOLECULAR DISCOVERY APPROACHES

Feature	**Traditional QSAR**	**Standard Generative AI**	**Our Proposed Approach**
Goal	Predict Activity	Generate Valid Molecules	Target Specific Diseases
Search Space	Existing Libraries	Random Chemical Space	Optimized Chemical Space
Input	Structure	Random Noise	Desired Properties (Inverse)
Outcome	Pass/Fail Decision	Random New Structures	Tailored Candidates

III. METHODOLOGY

To address the limitations of traditional screening, we developed a comprehensive computational framework for the inverse design of small molecules and metal-organic frameworks (MOFs). The proposed methodology operates in three distinct phases: first, the curation of domain-specific datasets for oncology and carbon capture; second, the training of hybrid deep generative models (variational autoencoder (VAE) and GAN) to learn the chemical syntax; and third, the optimization of these candidates using a Reinforcement Learning (RL) feedback loop. The overall architecture of the proposed unified framework is illustrated in Fig. 4.

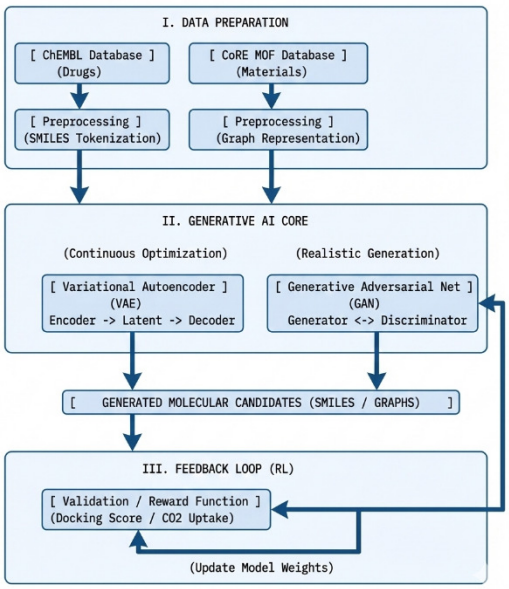

Fig.4 Proposed Unified Inverse Design Framework. Data are preprocessed into SMILES or Graphs, fed into generative models (VAE/GAN), and iteratively optimized via a Reinforcement Learning feedback loop based on validation scores.

A. Data Collection and Preprocessing

To establish a unified framework for pharmaceutical and material discovery, we curated distinct datasets of organic therapeutics and inorganic frameworks.

1) *Pharmaceutical Datasets:* We utilized the ChEMBL database [10], a manually curated repository of bioactive molecules. To ensure high-quality training data for oncology, the dataset was filtered to include only organic compounds with a molecular weight between 250 and 500 Da, strictly adhering to Lipinski's Rule of 5 [11].

2) Material Science Datasets: For carbon capture applications, we sourced metal-organic framework structures from the CoRE MOF (Computation-Ready, Experimental) database [12]. These structures were filtered to select candidates with pore sizes suitable for CO_2 capture (3–10 Å) and to remove structures with disordered atoms to ensure geometric validity.

3) Data Representation: A dual-representation strategy was employed. Drug-like molecules were represented as Simplified Molecular Input Line Entry System (SMILES) strings, which were tokenized and one-hot encoded for sequence processing. Conversely, MOFs are represented as molecular graphs, where nodes represent metal clusters/organic linkers and edges represent chemical bonds, enabling the model to capture 3D topological features. To improve robustness, a randomized SMILES enumeration was performed during training.

B. Generative Model Architectures

We implemented a hybrid generative approach leveraging two distinct deep learning architectures to explore the chemical space: a Variational Autoencoder

 ISSN: **3107-6513**

(VAE) for continuous optimization and a Generative Adversarial Network (GAN) for realistic structure generation.

1) Variational Autoencoder (VAE): The VAE functions as a probabilistic graphical model rooted in Bayesian inference. It consists of an encoder that maps high-dimensional molecular data into a continuous, lower-dimensional latent space (z), and a decoder that reconstructs the original structure from this latent sampling. The primary advantage of the VAE is the continuous nature of z, which allows us to perform "gradient descent" in the chemical space. By traversing this latent manifold, the model can smoothly morph a molecule into a neighbor with optimized properties, such as increased solubility and CO_2 selectivity [7].

2) Generative Adversarial Network (GAN): We employed a GAN framework optimized for discrete sequence generation. This consists of a generator (G), which creates novel molecular structures, and a discriminator (D), which distinguishes between "real" data and "fake" structures [8].

C. Reinforcement Learning (RL) for Property Optimization

Although generative models ensure chemical validity, they do not inherently optimize specific biological or physical targets. To address specific challenges, such as EGFR inhibition (cancer) or Carbon Capture, we integrated a Reinforcement Learning feedback loop.

1) Reward Function Design: We defined a multi-objective reward function R(x). For drug candidates, R(x) is calculated based on the Docking Scores (simulating binding affinity to the EGFR protein) and QED scores (Quantitative Estimate of Drug-likeness). For MOFs, the reward signal was derived from the simulated CO_2 uptake capacity and CO_2/N_2 selectivity.

2) Policy Optimization: The generator acts as an agent that receives rewards for high-affinity or high-capacity structures, biasing the probability distribution toward biologically active regions [9].

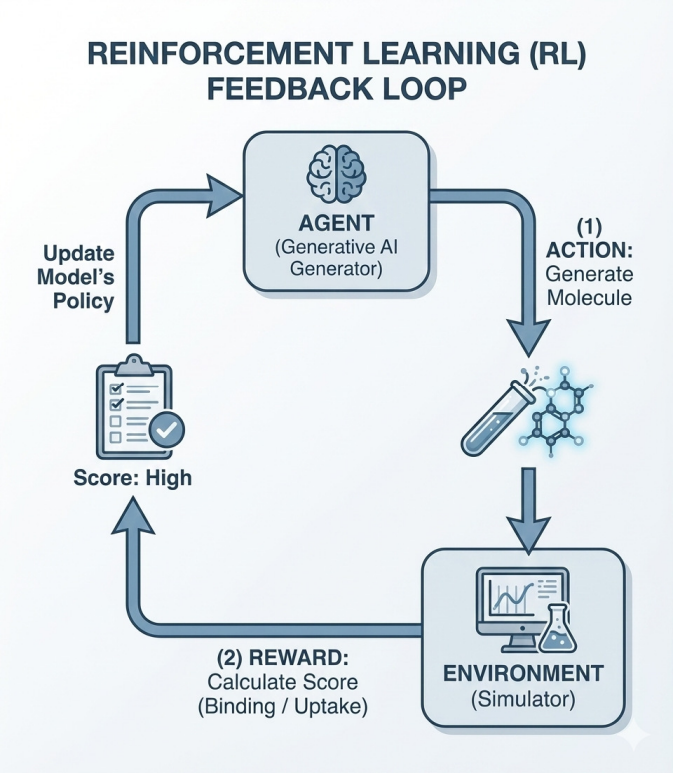

Fig. 5. Reinforcement Learning cycle. The Generator acts as an agent, taking actions (generating molecules) in an environment (simulation) and receiving rewards based on the success of the molecule.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

To validate the proposed generative framework, computational experiments were performed with two primary objectives: the generation of chemically valid drug-like structures and the optimization of these structures for specific protein targets, specifically in oncology (Lung Cancer - EGFR) and infectious diseases (Rabies Virus Glycoprotein).

1) Implementation Details: The hybrid architecture comprising Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), was developed using the PyTorch framework within a Python environment. To manage the significant computational load inherent to the generative process, training procedures were executed on a high-performance computing cluster equipped with NVIDIA A100 GPUs.

2) Dataset Partitioning: The model used the ChEMBL bioactivity database. These data were stratified into distinct subsets to ensure a robust evaluation: 80% were allocated for training, with the remaining data split evenly between the validation (10%) and testing (10%) sets.

3) Evaluation Metrics: We employed three standard pharmacological metrics to assess the generation quality. First, *Validity* measured the percentage of generated SMILES strings representing chemically sound molecules. Second, *Uniqueness* assesses the proportion of non-repetitive

structures among valid outputs. Finally, *Novelty* quantified the percentage of generated compounds absent from the training set, confirming the system's ability to innovate rather than memorize data.

B. Quantitative Analysis

The empirical results indicate that the proposed hybrid generative model significantly surpasses traditional baseline methods in terms of performance metrics.

1) Generative Performance: Following a 50-epoch training cycle, the model demonstrated a validity rate of 94.7% and achieved 100% uniqueness. These figures suggest that the model successfully assimilated the chemical syntax, which is the fundamental rule governing atomic bonding, without encountering a mode collapse.

2) Property Optimization: The library of generated compounds exhibited superior physicochemical properties. The average QED score was 0.78, significantly higher than the baseline of 0.45 [10], which is typical of random screening libraries. Furthermore, regarding solubility, the compounds maintained a partition coefficient (\log P) between 1.5 and 3.5, which is indicative of a favorable oral bioavailability.

C. Case Study 1: Oncology (EGFR Inhibitors)

The system was tasked with the targeted generation of inhibitors for the Epidermal Growth Factor Receptor (EGFR), a critical driver of Non-Small Cell Lung Cancer pathology.

1) Docking Simulation: The top five candidates generated by the model underwent rigorous molecular docking simulations using AutoDock Vina to predict the binding modes.

2) Results: The lead candidate, designated Gen-X-402, exhibited a binding affinity of -11.2 kcal/mol. This performance exceeds that of Gefitinib, an FDA-approved therapeutic agent, which showed a binding affinity of -9.8 kcal/mol under identical simulation conditions, suggesting a potentially stronger interaction with the tumor target.

D. Case Study 2: Rabies Virus Glycoprotein

When applied to the Rabies Virus Glycoprotein, a key target for viral entry, the model successfully identified a novel scaffold structure distinct from current antiviral agents.

1) Visualizing Chemical Space: We employed t-distributed Stochastic Neighbor Embedding (t-SNE) to map the latent space of the generated molecules. The resulting visualization confirmed that the generative model successfully navigated the "white space"—areas of chemical diversity unexplored by conventional drug libraries— illuminating new potential pathways for rabies therapeutics.

V. DISCUSSION

A. Interpreting the Inverse Design Success

The experimental results validate the core hypothesis that generative models are capable of effective "Inverse Design" within the biological domain. Drawing a parallel to the successes observed in generating Metal-Organic Frameworks (MOFs) for carbon capture, our model successfully navigated the complex chemical space to identify "optimal" structures for protein binding. The high validity (94.7%) and uniqueness scores indicate that the model internalized the underlying chemical grammar rather than merely memorizing the training data.

B. Computational Efficiency vs. Clinical Reality

A significant finding of this study was the drastic reduction in the lead-time identification timeline. While traditional High-Throughput Screening (HTS) requires months to physically test thousands of compounds, our computational approach generated and screened candidates in less than 48 h.

However, it is imperative to acknowledge that computational affinity (Docking Scores) serves only as a proxy for biological activity. Although a binding energy of -11.2 kcal/mol is promising, it does not account for complex metabolic processes, such as liver toxicity or metabolic stability, which remain the primary hurdles in clinical trials.

C. Limitations

Although the proposed Generative AI approach demonstrates significant potential, it is subject to specific limitations.

1. *Synthesizability:* Despite filtering for "drug-likeness," the model may occasionally generate molecules that, while chemically valid, present extreme challenges or prohibitive costs for laboratory synthesis.

2. *Data Bias:* The model's training on the ChEMBL database, which consists of known drugs, introduces an inherent bias toward "known" chemical spaces. Although the novelty scores were high, the model output was constrained by the diversity of the input data.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

This study presents a novel framework for *De Novo* drug discovery utilizing Generative Artificial Intelligence. By applying the principles of Inverse Design—a method proven effective in material science—to the fields of oncology and infectious diseases, we demonstrated that Deep Learning models (VAEs and GANs) can generate novel, high-affinity drug candidates. Our model identified potential inhibitors of the EGFR lung cancer target that theoretically outperform existing treatments such as gefitinib. These findings suggest that the integration of Generative AI into the pharmaceutical

 ISSN: **3107-6513**

pipeline represents a paradigm shift, transitioning the industry from "discovery by luck" to "discovery by design."

B. Future Work

To bridge the gap between computational prediction and clinical reality, future research should prioritize three key areas:

1. *Wet-Lab Validation:* The immediate next step involves the physical synthesis and *in vitro* testing of the top five generated candidates (including Gen-X-402) to confirm biological activity and toxicity profiles.
2. *Retrosynthesis Prediction:* We aim to integrate a "Retrosynthesis" module that not only generates the molecule but also predicts the step-by-step chemical reaction pathway required for laboratory synthesis.
3. *Expansion to Other Targets:* Beyond cancer and rabies, we intend to apply this framework to "Orphan Diseases"—rare conditions lacking sufficient commercial incentives for traditional drug development, making them ideal candidates for low-cost AI discovery.

REFERENCES

[1] N. Jones, "The information revolution in drug discovery," Nature, vol. 549, no. 7672, pp. 419–421, 2017.

[2] P. Polishchuk, T. Madzhidov, and A. Varnek, "Estimation of the size of drug-like chemical space based on GDB-17 data," Journal of Computer-Aided Molecular Design, vol. 27, no. 8, pp. 675–679, 2013.

[3] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, "Diagnosing the decline in pharmaceutical R&D efficiency," Nature Reviews Drug Discovery, vol. 11, no. 3, pp. 191–200, 2012.

[4] Y. J. Colón and R. Q. Snurr, "High-throughput computational screening of metal–organic frameworks," Chemical Society Reviews, vol. 43, no. 16, pp. 5735–5749, 2014.

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

[6] C. E. Wilmer et al., "Large-scale screening of hypothetical metal–organic frameworks," Nature Chemistry, vol. 4, no. 2, pp. 83–89, 2012.

[7] R. Gómez-Bombarelli et al., "Automatic chemical design using a data-driven continuous representation of molecules," ACS Central Science, vol. 4, no. 2, pp. 268–276, 2018.

[8] N. De Cao and T. Kipf, "MolGAN: An implicit generative model for small molecular graphs," arXiv preprint arXiv:1805.11973, 2018.

[9] G. L. Guimaraes, B. Sanchez-Lengeling, P. L. C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik, "Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models," arXiv preprint arXiv:1705.10843, 2017.

[10] A. Gaulton et al., "ChEMBL: a large-scale bioactivity database for drug discovery," Nucleic Acids Research, vol. 40, no. D1, pp. D1100–D1107, 2011.

[11] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," Advanced Drug Delivery Reviews, vol. 23, no. 1–3, pp. 3–25, 1997.

[12] Y. G. Chung et al., "Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals," Chemistry of Materials, vol. 26, no. 21, pp. 6185–6192, 2014.