

Molecular Design with Artificial Intelligence: Progress and Perspectives for Small Molecules

Published as part of Chemical Reviews *special issue* “Artificial Intelligence for Chemistry”.

Masato Sumita, Shoichi Ishida, Kazuki Yoshizoe, Ryo Tamura, Kei Terayama, and Koji Tsuda*



Cite This: *Chem. Rev.* 2026, 126, 3007–3054



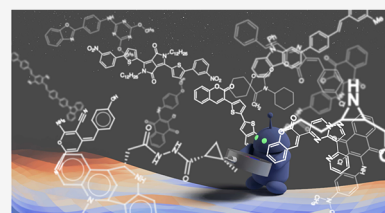
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: Progress in chemistry has been driven by the streamlining of inverse problem-solving methods. In the history of chemistry, several revolutionary technologies have led to leaps forward: the establishment of atomistic theory in the 19th century, structural analysis by spectroscopy in the 20th century, and the development of simulation by theoretical chemistry. Currently, chemistry is about to make a significant leap forward by integrating generative artificial intelligence (AI). In 2016, deep learning techniques were introduced in this domain, leading to explosive development. This paper reviews the development path, including traditional models such as variational autoencoders and more up-to-date models such as large language models and diffusion models. We also discuss how AI can have a real impact on chemistry, including the possibilities and problems associated with synthesizing AI-generated molecules.



CONTENTS

1. Introduction	3008	4.10. Monte Carlo Tree Search and Large-Scale Parallelization	3024
1.1. Inverse Problems and the Beginning of Chemistry	3008	Defining the Search Space for MCTS	3024
1.2. Observing Quantum-Mechanical World	3009	1D Representations (SMILES, SELFIES, InChI)	3024
1.3. Expectation on Simulations	3009	2D and 3D Representations	3024
1.4. AI: A New Tool for the Inverse Problems	3010	Evaluating Internal Nodes: Sampling-Based Rollouts	3025
2. Disentanglement of Molecular Generative AI	3011	Parallelization Requirements and Parallel Graph Search	3025
2.1. Components of Molecular Generative AI	3011	5. Evaluation Strategies	3025
2.2. Molecular Representations for Generative Methods	3012	5.1. Property Estimation of Generated Molecules	3025
3. Generative Strategies	3014	5.2. Synthetic Accessibility Evaluation	3026
3.1. Combinatorial Generation	3014	5.3. Single- and Multi-Objective Evaluation	3026
3.2. Autoregressive Models	3014	5.4. Benchmark	3027
3.3. Variational Autoencoders	3015	6. Application of Generative Model for Molecule Identification	3027
3.4. Generative Adversarial Networks	3016	7. Experimental Validation of AI-Designed Molecules	3029
3.5. Normalizing Flow-Based Models	3016	7.1. Drug Discovery	3029
3.6. Energy-Based Models	3017	7.2. Material Discovery	3033
3.7. Transformer and Large Language Models	3017	8. Outlook	3034
3.8. Diffusion Models	3018	Author Information	3036
3.9. Overall Perspective on Generative Models	3018	Corresponding Author	3036
4. Optimization and Search Strategies	3019		
4.1. Random Sampling	3019		
4.2. Metaheuristics	3020		
4.3. Bayesian Optimization	3020		
4.4. Reinforcement Learning	3021		
4.5. Gradient-Based Optimization	3022		
4.6. Bayesian Sampling	3022		
4.7. Exhaustive Search	3022		
4.8. Integration with Synthetic Route Planning	3023		
4.9. Integration with Quantum Computing	3023		

Received: August 11, 2025
Revised: February 4, 2026
Accepted: February 13, 2026
Published: March 1, 2026



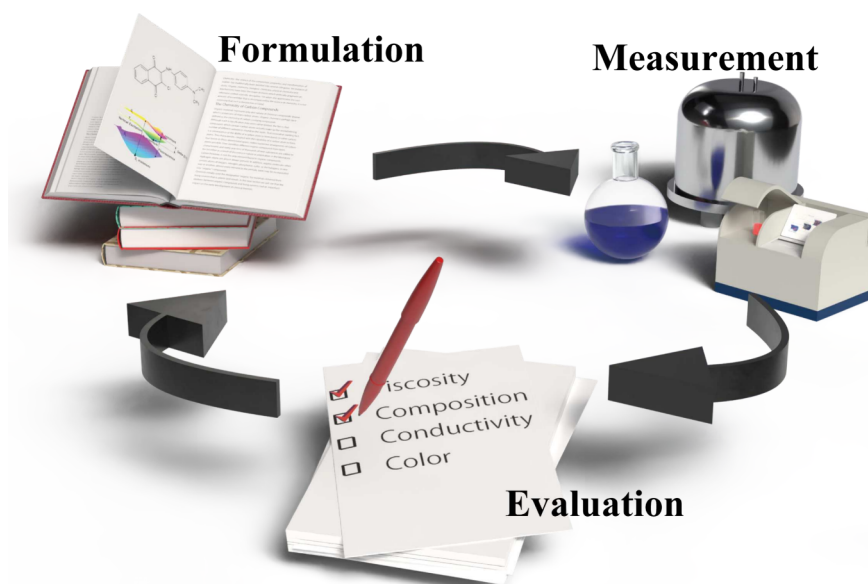


Figure 1. Conventional cycle (namely black-box optimization¹⁷) of formulation, measurement, and evaluation for solving inverse problems in chemistry. Formulation is the process of predicting the next target for the next process (measurement) based on accumulated knowledge. Measurement is the process of obtaining observables of the target matter, substance, or molecules, including their synthesis. Evaluation is the process of assessing whether the (synthesized) target satisfies the requirements.

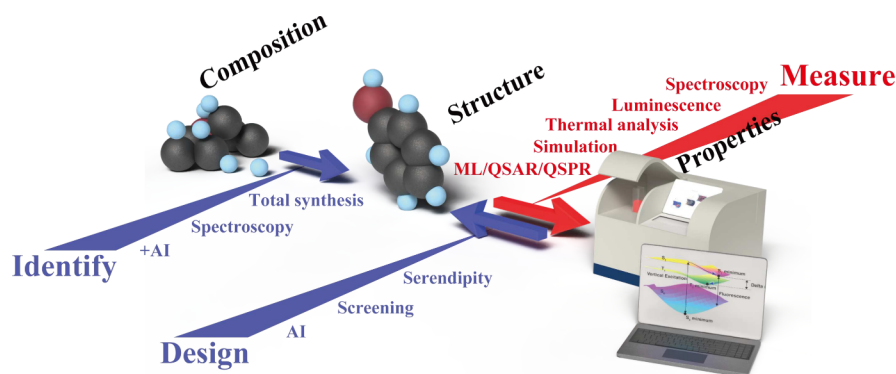


Figure 2. Inverse problems in chemistry. The direction of the arrows indicates input-to-output. The blue arrow indicates the inverse problem (observable-to-structure) to be solved by the cycle shown in Figure 1. The red arrow indicates the forward problem (structure-to-observable).

Authors	3036
Notes	3036
Biographies	3036
Acknowledgments	3037
Abbreviations	3037
References	3038

1. INTRODUCTION

First, we review the history of chemistry from the viewpoint of inverse problems, revisiting important technologies and concepts that have revolutionized chemistry. This clarifies the influence of artificial intelligence (AI) in chemistry. Chemistry enables us to deal with substances that are not as complex as biology but are complex and moderately idealized. This moderate affordability has enabled the integration of many fields and the development of chemistry. Although introducing all the important technologies in this field is not possible, we believe that we understand this major flow in the history of chemistry.

1.1. Inverse Problems and the Beginning of Chemistry

Chemistry deals with matter at the invisible molecule/atomistic level, where inverse problems (speculating an object from an observation) in seizing the molecule/atomistic world are the primary themes. The solution to this inverse problem conventionally relies on the repetition of the cycle of formulation, measurement, and evaluation, as shown in Figure 1. If recognition of matter using humans' five senses is a forward problem, finding matter that meets human demands is an inverse problem. Hence, solving the inverse problem may have benefitted humanity in the areas of cosmetics,^{1–3} dyes,^{4–9} and detergents,¹⁰ which have been used since antiquity and are the research targets of modern chemistry. A deeper understanding of these materials has led to an invisible world of molecules and atoms beyond the perception of the five human senses, which may have started as mere human interest.

The fact that matter consists of particles has been known since antiquity, and the modern understanding that chemical phenomena originate from atoms and molecules was attributed to Dalton's atomic theory¹¹ in the 19th century. In this century, without reliable methods to observe atomistic and molecular

structures, human imagination has progressed from 2D structures, as represented by Kekulé structures, to 3D structures, establishing the foundations of structural chemistry.^{12,13} An improved scientific solution to the inverse problem of identifying the molecular structure from observations (composition) obtained by combustion experiments¹⁴ was the laborious method of total synthesis¹⁵ (Figure 2). The revolution in solving the inverse problem of identifying atomistic-level molecular structures came with the arrival of quantum mechanics¹⁶ in the 20th century.

1.2. Observing Quantum-Mechanical World

In the 20th century, as theories represented by quantum mechanics¹⁶ were developed, analytical equipment at the atomistic and molecular levels advanced considerably, significantly expanding the knowledge about chemical structures. Spectroscopies that use transitions between quantized molecular energy levels,¹⁸ scattering of atoms and several other particles,¹⁹ and nuclear magnetic resonance²⁰ are powerful, making molecular structure determination possible by interpreting the obtained spectra. This became possible because the spectral peak positions shown by representative molecular substructures are empirically known,^{21–23} allowing molecular structures to be decoded by repeating the cycle shown in Figure 1 (assembling substructures) in the mind of chemist. As a result, speculation on the molecular structures makes it possible to identify molecular structure without repeating the cycle shown in Figure 1 in the real world. Although total synthesis remains the ultimate method of molecule identification,²⁴ spectroscopy has revolutionized the inverse problem of molecule identification. Deciphering a spectrum requires a specialist; spectroscopy can be used to characterize not only small molecules but also large and complex substances such as DNA and proteins.^{25–34} As atomistic-level molecular structures became evident, the synthesis of aesthetic molecules progressed,^{35–37} along with research exploring the limits of scientific molecules (e.g., challenge of synthesizing giant molecules).³⁸ Molecular structures with characteristic physical and chemical properties were clarified.

As molecule identification and isolation techniques stimulate each other, synthetic techniques have also been developed³⁹ and chemistry textbooks have become thicker, with case-by-case methods for formulating chemistry (the accumulation of chemical reactions under specific conditions and substituent effects, etc.)^{21,22}

In addition to structural analysis instruments that revolutionized molecule identification, various tools such as UV–vis spectroscopy and luminescence, redox, and thermal analyses have been designed based on quantum mechanics and thermodynamics.²³ The development of measurement instruments has made the accurate determination of molecular properties possible; therefore, attention has been paid to the control of molecular properties (functions) subject to molecular structure. By the mid-20th century, nanotechnology had gradually gained prominence and focused on the development of functional molecules. Attempts to design functional materials from molecular units grew rapidly, with examples including drug delivery systems,^{40–42} dye-sensitized solar cells,^{43–57} several sensors,^{58–60} organic electronics^{61,62} and so forth.^{63–67} Thus, the importance of solving the inverse problem has increased, not only in identifying molecular structures from several spectra but also in optimizing (designing) molecular structures for molecular functions (properties) with the cycle shown in Figure

1. However, as atomistic-level molecular structures resulting in properties have become apparent, their high diversity has surfaced as an obstacle to blocking free molecular design. Similar to the aforementioned molecule-identification problem, it may be possible to design a molecular structure with desired molecular properties by iteratively repeating the cycle shown in Figure 1 in the mind of a researcher. Unfortunately, current molecular design technique has not yet reached this stage, and this longing has increased the volume of chemistry textbooks. This is predominantly due to the great diversity at the atomistic levels. Therefore, the high diversity forces chemists to rely on trial and error for the optimization of molecular properties, imposing the arduous task of repeating the cycle of Figure 1. Naturally, experimental costs also increase. To avoid diversity and develop functional molecules as efficiently as possible, researchers have optimized (derivative development) molecules that already function,^{58,60,63–65,68–77} rather than exploring unknown molecules. Arguably, this becomes a local solution search (searching for a subspace), possibly missing superior molecules. To include as many candidate substances as possible, the optimization of the function of the molecule by repeating cycle of Figure 1 must be accelerated. Although search methods to solve inverse problems (from molecular properties to structure) have been proposed,^{78–80} these methods were not versatile and practical. Rather than developing the method for solving inverse problems, chemists' efforts focus on accelerating the cycle by streamlining the measurement process, thereby expanding knowledge of the formulation. Hence, a forward approach, such as a deeper understanding and rapid processes leading to measurement, is dominant. Consequently, theories find new uses as simulation tools that can help chemists avoid costly experimental trial and error.

1.3. Expectation on Simulations

Along with innovations in structural analysis technology in the 20th century, dramatic improvements in computer performance have had various impacts on society. This led to significant developments in chemistry. The field of quantum chemistry (QC)⁸¹ benefited most from this. Although quantum mechanics revealed the equation required to understand the properties of matters and molecules, it was too complex to solve. Thus, QC enables the application of quantum mechanics to chemistry by introducing various approximations. Valence bond methods,^{82–84} based on Heitler–London theory,⁸⁵ which were developed focusing on chemical bonding formation, face limitations in application to experimentally interesting molecules owing to increasing bonding states.⁸⁶ Alternatively, the Hartree–Fock (HF) method (molecular orbital method)⁸⁷ sacrifices a clear description of chemical bonds. Owing to this sacrifice, various *ab initio* calculations as post-HF theories have been developed, including the perturbation theory for improved wave function and internal energy accuracy,⁸⁸ multiconfiguration wave function theory for thermal and photochemical reactions,^{89–91} and methods using coupled cluster expansion,^{92–94} emphasizing experimental value reproducibility. However, developments based on molecular orbitals have shown signs of a deadlock. Regardless of the reproducibility, the cost is too high to be incorporated into the inverse problem cycle. By incorporating density functional theory (DFT) based on electron density,^{95–97} QC balances the reproducibility of experimental values and computational cost. Many theories have been developed to reproduce the observations measured by the aforementioned instruments^{98–102} to reduce computational

time and accuracy.^{103–105} To further reduce the computational cost, several simulation methods have been developed as alternatives to QC, such as force fields/parametric potentials for classical molecular dynamics^{105,106} including recent ML-assisted force fields,^{107,108} tight-binding DFT,^{105,109} and computational estimation tools like quantitative structure–activity/property relationship models (QSAR/QSPR),^{110–114} and recent ML prediction models.¹¹⁵ However, the atomistic simulations (particularly for QC) have been extensively used to speculate on complementary areas where experimental observations cannot be obtained for deeper observation and understanding.^{116–124} The original purpose of the simulations should have been to accelerate the cycle shown in Figure 1 through a deeper understanding and inference of chemical phenomena. Owing to the prevalence of several simulation packages^{109,125–130} and their visualization tools,^{131–135} simulations are commonly used in functional molecule development as formulation tools for selecting target molecules for synthesis (measurement) in modern experimental chemistry laboratories (in silico cycle of Figure 1).¹³⁶ However, because the simulations require an input chemical structure, they are inherently applicable only to the forward problems shown in Figure 2. Operational limitations are apparent in the number of candidate molecules (the diversity of molecules) that must be examined. Consequently, the simulations themselves do not have an exploratory function and are not a tool for dealing with molecular diversity. The high diversity of molecules (even one molecule) at the atomistic level hinders the control of their molecular properties.

The number of available elements in an observable universe is limited.¹³⁷ However, these ensembles construct numerous molecules and substances.^{138,139} From a mathematical standpoint, the set of all possible molecules and substances is referred to as chemical space. In recent years, chemical space has attracted considerable attention due to advances in data science. However, the scope of chemical space is too broad to be comprehensively addressed in this review. The definition of the axes that create a chemical space has been discussed, including reaction-based hypergraphs¹⁴⁰ and substance similarity;¹⁴¹ however, no clear answer has been obtained. Attempts to understand the subspaces under constraints have led to the construction of molecular and material databases.^{142–153} One subspace is almost impossible for researchers to grasp it owing to its large size. Even when limited to potential organic molecules such as drugs, an estimated 10^{60} molecules exist in this space.¹⁵⁴ However, according to PubChem,¹⁵⁵ only 10^8 molecules have been successfully synthesized or isolated as realistic molecules by researchers. Although the number of molecules in the Earth's environment is postulated to be limited, researchers clearly have access to remarkably few molecules. Molecules with innovative functions may undeniably exist beyond the range of molecules accessible to researchers, and exploring unknown territories is undoubtedly important. Simulations had been expected to become an alternative search tool to expensive experiments; however, this was an excessive expectation.

1.4. AI: A New Tool for the Inverse Problems

With the recent advent of generative AI, in silico chemical space searching with simulation is approaching a new phase. Simulations are inherently measurement tools because they require an observation target as the input, as shown in Figure 2. Hence, preparing the input molecules in advance is necessary for use as tools in the design process. A straightforward method

involves preparing a database of targets, computing their properties, and evaluating them. This approach is called high-throughput screening^{156,157} and is shown in Figure 3(a).

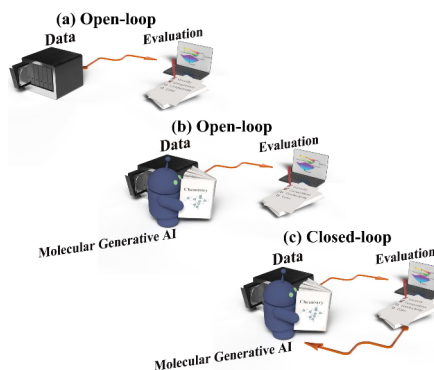


Figure 3. In silico molecular design through (a) open-loop screening and (b) open-loop with generative AI. (c) Closed-loop with generative AI (one-shot prediction). The measurement and evaluation of Figure 1 are unified to computational evaluation. The screening data do not inform the targets; hence, no formulation system is observed. Generative AI in (b) extracts features as a probabilistic model and generates molecules based on the data. On the other hand, generative AI in (c) works as a formulation that can decide the next target based on the previous evaluation. The closed-loop system for the inverse problem is made possible owing to generative AI.

Recently, several studies have identified functional molecules or polymers by screening virtual/real libraries using on-demand simulations, machine learning prediction models, or QSAR/QSPR.^{158–181} Rather than an exhaustive search in a database, such as screening, a molecular generative AI system based on a probabilistic model that proposes the molecular structure of a candidate for a desired property is expected to improve the search efficiency, as shown in Figure 3(b). This probabilistic model represents an assumed distribution obtained by fitting to the distribution of the training data. Naturally, its performance depends heavily on the choice of axes (features), the assumed probabilistic model, and the scope of the training data; therefore, the accessible search space remains confined within the database. This means that probabilistic model-based molecular generative AI is essentially limited to exploring known chemical scaffolds. In other words, the search space is not fundamentally expanded beyond that of conventional molecular design. Moreover, constructing such a generative AI system first requires a database that is not significantly different from the one used for the traditional screening. Therefore, closing the loop to achieve an in silico design in real terms is necessary, as shown in Figure 1, which can improve the probabilistic distribution of targets by repeating the cycle. To enable effective exploration in this loop, as shown in Figure 3(c), various optimization strategies play important roles, including Bayesian optimization (BO), reinforcement learning (RL), and evolutionary or genetic algorithms as will be described later.

The recent emergence of generative AI has impacted society as a possible alternative to human creativity. The efficiency of the inverse problem-solving capability in chemistry relies on how rarely the real world is used. As shown in Figure 2, two dominant inverse problems exist in chemistry: the identification of the molecular structure and molecular design. The former has successfully formulated the observable spectra–structure correlation to some degree, although identification of macro-molecule and species dependencies is challenging.^{25–34}

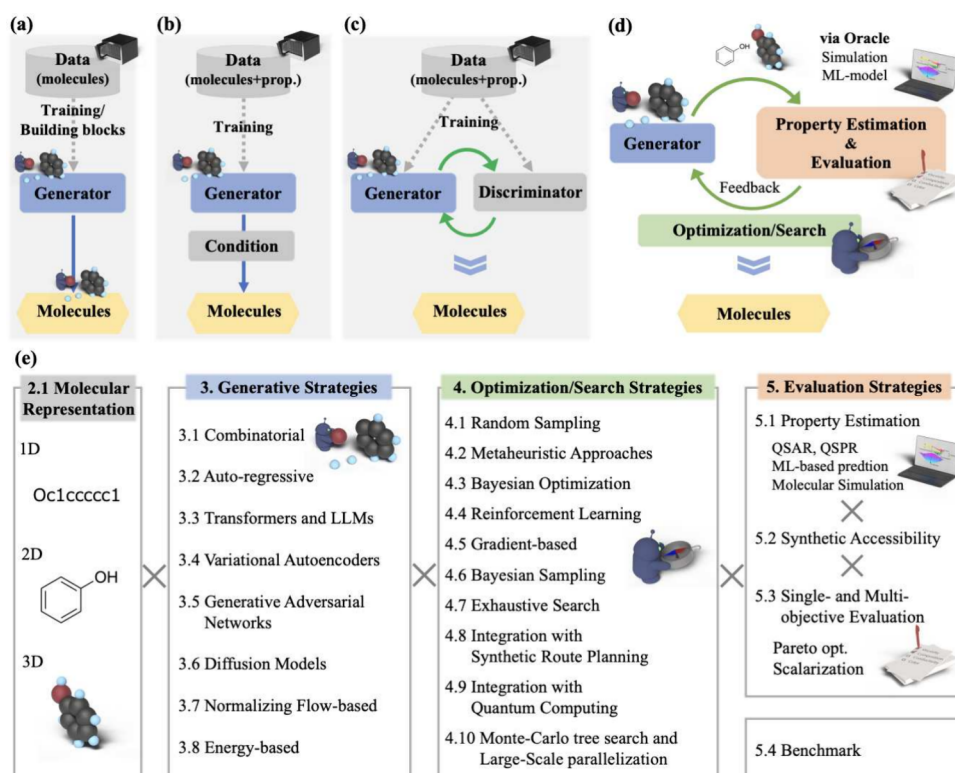


Figure 4. (a–d) Overviews of typical molecular generative AI systems. (e) Possible combination of generative, optimization, and evaluation (reward) strategies for molecular design. For each description, see the numbered section.

However, the latter is hindered by the diversity of chemical space, as mentioned previously. The chemical space is not sufficiently simple to be explored in the brain of the researcher; it requires an external brain. We can postulate that molecular generative AI can revolutionize these inverse problems in chemistry. Since 2016, generative AI algorithms for molecular design^{182–191} have been extensively developed and applied to explore the property space, which is a chemical space projected using evaluation tools such as simulations (oracles). Although generative AI has been leveraged to build molecular libraries for screening,^{176–179} it can generate molecules and search in the property space, as shown in Figure 1. This technology has brought about innovation to the inverse problems in chemistry that have been challenged since antiquity. The application of generative AI to chemistry is an integration of wisdom of researchers that transcends the boundaries of various disciplines. This review disentangles molecular generative AI; describes the surroundings that could accelerate, help, or automate solving inverse problems in chemistry; introduces real AI-designed molecules; and discusses future directions and limitations.

2. DISENTANGLEMENT OF MOLECULAR GENERATIVE AI

2.1. Components of Molecular Generative AI

Over the past decade, molecular generating AI, particularly generative approaches based on deep neural networks, has advanced significantly. Using these approaches, numerous molecular generative AI systems that can design molecules with desired properties have been proposed. Along with this, high-quality review papers summarizing these methods have been published.^{182–191} However, molecular generative AI systems may still be too complex and confusing for researchers

unfamiliar with ML, AI, and cheminformatics. To demystify molecular generative AI, we decompose it into its basic components, such as generative methods and optimization/search techniques, and provide a detailed explanation of these methods.

Figure 4(a)–(d) show the four basic approaches to computational molecular design. Figure 4(a) shows the simplest molecular design approach that builds a molecular generative model that learns structural patterns from a molecular data set and then generates molecules probabilistically. In addition, molecules can be constructed combinatorially after creating molecular fragments from a data set. Notably, this approach does not necessarily aim to directly optimize specific molecular properties. Figure 4(b) shows a conditional generative model that learns the distribution $p_\theta(x|c)$ when condition c (e.g., the target property) is given. Molecules that satisfy this condition can be generated by providing a specific condition (target value). This framework enables the design of molecules with the desired properties. Furthermore, to generate molecules more flexibly, a framework that trains a molecule generator and discriminator to evaluate the properties of the generated molecules simultaneously, as shown in Figure 4(c), has also been proposed. Generative adversarial network (GAN)-based methods are included in this approach. In the above approaches, the training data, which may contain property data, are fixed; thus, the generated molecules are essentially similar to the training data. As shown in Figure 4(d), molecular design is possible by combining a molecular generative model with feedback based on molecular evaluation using an external evaluation method, often called an oracle. Examples of oracles include molecular simulations, experiments, and prediction models based on ML, such as QSAR/QSPR models. In this approach, various optimization algorithms or search techniques are employed to

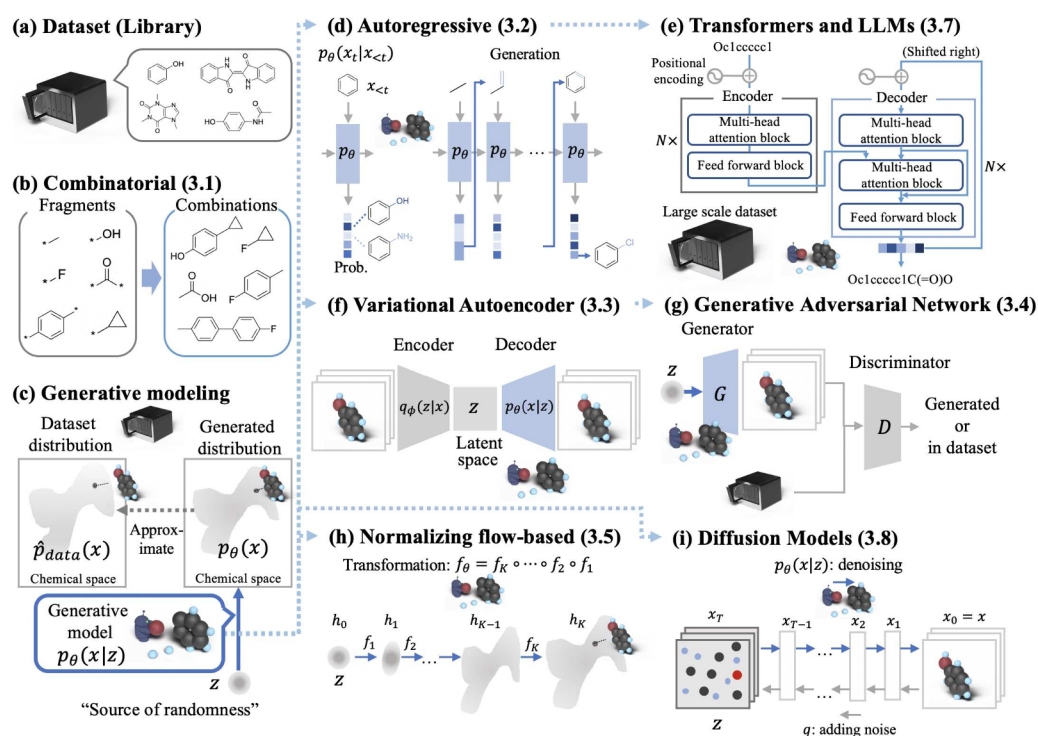


Figure 6. Overview of representative generative approaches introduced in Section 3. (a) Preparing the molecule library in advance. (b) Possible combinatorial generation based on the molecular fragments. (c) Probabilistic generative models, which can be classified as (d) autoregressive models, (e) transformer and LLMs, (f) variational autoencoders, (g) generative adversarial networks, (h) normalizing flow-based models, and (i) diffusion models. For each description, see numbered sections.

define atoms and bonds as “characters” in advance and represent molecules as strings (a sentence) by combining characters according to a certain grammar, including stereochemical information such as chirality. SMILES is the most commonly used representation and has relatively high readability. However, it has the drawback of easily generating syntactically invalid or chemically invalid strings. SELFIES solves these problems and has recently attracted attention as a representation method in which any string can be mapped onto a valid molecular structure. In addition, various variants of SMILES, such as DeepSMILES,¹⁹⁶ BigSMILES,¹⁹⁷ and t-SMILES,¹⁹⁸ have been proposed to address the shortcomings and challenges of SMILES. InChI¹⁹⁹ and IUPAC names²⁰⁰ are also string representations of molecules. By treating molecules as strings (sentences), it is possible to learn and generate them using autoregressive models and transformers, including LLMs, and then use them for generation. Recently, LLM-based approaches that utilize semantically enriched textual descriptions as molecular representations have also emerged, further broadening the scope of string-based molecular representations.²⁰¹

Molecular fingerprints or descriptors are occasionally used as another type of 1D representation, as shown in Figure 5(a). Molecular fingerprints basically represent a given molecule as a list indicating the presence or number of local structures (i.e., a vector of numbers). Various fingerprints have been proposed, including atom pair fingerprints²⁰² and extended connectivity fingerprints (ECFP).²⁰³ Molecular descriptors are lists of relatively easy-to-calculate molecular characteristics such as molecular weight and topological polar surface area, and various methods have been proposed, such as PaDEL-Descriptor and Mordred.^{193,204,205} These fingerprints and descriptors are widely used as input features for machine learning-based property prediction, including QSARs; however, molecular design

methods that generate these 1D vectors have also been proposed. Generally, with such fingerprints or descriptors, two or more different molecules may have the same representation, resulting in an inability to uniquely identify a molecule from a given representation.

In 2D representations, the most common method is to represent molecules as graphs, as shown in Figure 5(b). The atoms and bonds of a molecule are treated as vertices and edges of a graph, respectively. Although various methods have been developed for representing molecular graphs on a computer, one of the basic methods is to use an adjacency matrix, where each row and column represent the index of a node, and the elements of the matrix represent the bond information (typically the degree) between atoms. Recently, graph neural networks (GNNs),²⁰⁶ which use molecular graphs as inputs for predicting molecular properties, have been actively researched. Most GNNs are influenced by the message-passing neural network (MPNN) framework:²⁰⁷ vertices represent their own atom information and the characteristics of their neighbors, and by exchanging information with neighboring atoms in each layer of the network, the overall characteristics of the molecule are captured. GNN was initially employed for predicting molecular properties; however, approaches that generate molecules by directly generating molecular graphs have recently been proposed, as introduced in Section 3. Challenges in graph-based molecular generation include handling molecules of different sizes, ensuring permutation invariance with respect to atomic order, and accurately encoding stereochemical features such as chirality. To overcome these challenges, numerous graph representation and generation methods have been proposed, as described in Section 3.

The most intuitive approach to represent molecules would be to express them as a list of atom positions in 3D coordinates, as

shown in Figure 5(c). Bond information is occasionally handled explicitly in 3D. Other representation methods have been proposed, such as atom distributions on 3D voxels or methods based on electron density distributions. 3D representation-based generative approaches have been developed relatively recently compared with 1D or 2D representations and are particularly suitable for diffusion models, as shown in Section 3.8. Symmetry considerations are crucial for the learning and generation 3D molecular representations. As molecular conformations are invariant under translation and rotation, E(3) (the Euclidean group in 3D) or SE(3) (the special Euclidean group in 3D) equivariant representations are typically used.²⁰⁸ Note that E(3) is also invariant under mirror transformations, which means that chirality cannot be considered adequately. The detailed representations and generative methods are introduced in Section 3.8.

3. GENERATIVE STRATEGIES

Figure 6 presents an overview of the generative approaches introduced in this section. When a list of all candidate molecules was provided as a data set (Figure 6(a)), a generative model may not be necessary. Even in this case, combinations of these data sets with the efficient search methods in Section 4.3 or exhaustive methods in Section 4.7 are useful. As shown in Figure 6(b), combinatorial methods that virtually combine molecular fragments or building blocks to generate new molecules have been studied for a long time. Although these methods are still effective, in recent years, probabilistic generative models, particularly those using deep neural networks, have attracted attention, as shown in Figure 6(c)–(i). Before introducing each generative approach, we describe the basic concepts of probabilistic generative models.

Figure 6(c) illustrates the fundamental framework of the probabilistic generative model. The goal is to develop a model that generates molecules similar to a given molecular data set $X = \{x_1, x_2, \dots, x_N\}$, which is often referred to as a training data set. More precisely, the objective of probabilistic generative modeling is to learn a parametrized probability distribution $p_\theta(x)$ that approximates the underlying probability distribution $p_{data}(x)$ that generates a given molecular data set X . Here, each $x_i \in X$ represents a molecule encoded in a certain format (e.g., SMILES, graph, 3D coordinates). In this framework, a latent variable z is introduced as the source of probabilistic generation. For molecule generation, z is randomly sampled from its prior distribution $p(z)$ (e.g., a multivariate standard normal distribution $p(z) = N(0, I)$), and then a molecule x is generated according to the conditional distribution $p_\theta(x|z)$. $p_\theta(x|z)$ is typically a neural network with a parameter θ that maps the latent space to chemical space. This generation process allows us to sample diverse molecules x corresponding to various z values; the generated distribution of the molecules is expressed as follows:

$$x \sim p_\theta(x) = \int p_\theta(x|z)p(z) dz \quad (1)$$

“Training of a generative model” is the process of optimizing the model parameters θ so that the generative distribution $p_\theta(x)$ approximates the empirical distribution $\hat{p}_{data}(x)$ based on the observed data. Once trained, $p_\theta(x)$ can be used to sample (generate) molecules that are similar to, but different from, the training data (Figure 4(a)).

This framework can be extended to conditional generation, where molecular generation is guided by external conditions c

such as target proteins, physical properties, and structural constraints (Figure 4(b)). In this setting, the model learns the conditional distribution $p_\theta(x|z, c)$, and sampling is performed as follows:

$$p_\theta(x|c) = \int p_\theta(x|z, c)p(z|c) dz \quad (2)$$

In this framework, the latent variable z is sampled based on condition c , and a molecule x is obtained using the generation distribution $p_\theta(x|z, c)$ composed of a neural network. Although this generation is a type of molecular design with the desired properties, the trained molecules and their property data are fixed.

Depending on the implementation of this generation framework with different molecular representations, various generative approaches have emerged, as described below. Some descriptions about the optimization/exploration techniques are included in the following description but see Section 4 for details.

3.1. Combinatorial Generation

Combinatorial generation divides molecules into fragments, called building blocks, and combines them according to predefined rules to construct new molecular libraries. This approach has been studied for many years, primarily for applications in drug discovery.²⁰⁹ Classical fragmentation methods include the Retrosynthetic Combinatorial Analysis Procedure (RECAP)²¹⁰ and Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS),²¹¹ which are systematic fragment-generation methods that consider synthetic feasibility. Classic fragment assembly methods include LUDI²¹² and SPROUT.²¹³ molBLOCKS²¹⁴ statistically analyzes decomposable fragments and extractable ones that are chemically and functionally meaningful. In recent years, extended methods of BRICS have been proposed to improve the diversity and utility of generated fragments, such as eMolFrag,²¹⁵ which first fragmented molecules using BRICS and then refined the fragment set by removing redundancy and preserving bonding information, and MacFrag,²¹⁶ which modified BRICS rules to extract more diverse and high-quality fragments. Graph-based approaches have also been proposed to represent molecules as graphs and construct novel molecules by employing more flexible graph fragmentation and recombination strategies, such as Compound Generator (CoG)²¹⁷ and graph-based genetic algorithm (GB-GA).²¹⁸ In particular, GB-GA allowed operations such as breaking and forming ring structures, which extends the diversity of accessible chemical space. Although these graph-based approaches do not necessarily account for synthetic feasibility, they offer the potential to generate novel structures.

3.2. Autoregressive Models

Autoregressive (AR) models are a class of generative models that generate structured data as a sequence of decisions, where each decision depends on the previous ones. For molecular generation, as shown in Figure 6(d), this is often framed as modeling the conditional probability of each molecular component, e.g., a character in a 1D string representation or an atom and its position, given the components generated so far. Formally, for a molecule represented as a sequence $x = (x_1, x_2, \dots, x_T)$, the probability distribution of the molecules is formulated as

$$p_{\theta}(x) = \prod_{t=1}^T p_{\theta}(x_t | x_{<t}) \quad (3)$$

Although the latent variables z do not appear explicitly in this framework, the probabilistic nature of each conditional step enables stochastic generation. In most implementations, a special end-of-sequence token is included, and the generation terminates when this token is sampled.

Recurrent neural networks (RNNs) are a foundational class of autoregressive models for processing sequential data that were initially formalized in the early 1990s. To address the limitations of modeling long-range dependencies, variants such as long short-term memory (LSTM) networks^{219,220} and gated recurrent units (GRUs)²²¹ have been introduced and have become widely used, particularly in the context of natural language processing.^{222,223} Segler et al.²²⁴ first demonstrated that molecules represented by SMILES can be learned and generated using RNNs, leading to the development of various molecule-generation methods. To improve the diversity and robustness, the effectiveness of randomized SMILES²²⁵ and training that considers invalid SMILES²²⁶ have been reported. Grisoni et al.²²⁷ proposed a bidirectional RNN (BIMODAL) that generates molecules in both directions from a central position to overcome the unidirectional limitations of standard RNNs. Flam-Shepherd et al.²²⁸ demonstrated that SELFIES-based RNN models outperformed graph-based generative models in approximating molecular distributions. Conditional autoregressive models can utilize information linked to molecules for generation. For example, Xu et al.²²⁹ proposed a conditional RNN model that integrated 3D protein-binding pocket information and achieved target-specific generation. Drug-target interAction-based Generation oF novel biologically active molecules (DRAGONFLY)²³⁰ extended this approach by incorporating protein–ligand interaction networks to further improve ligand design.

Beyond 1D strings, AR models have been extended to 2D graph representations, in which molecules are constructed by sequentially adding atoms and bonds. At each step, atoms or bonds are sampled based on the previously constructed partial graph, and the generation continues until the termination condition is satisfied. Li et al.²³¹ introduced a graph autoregressive model that parametrized sequential graph construction using message-passing neural networks. MolRNN²³² employed a similar conditional generative approach and simultaneously optimized multiple molecular properties in drug design–related tasks. Popova et al.²³³ developed MolecularRNN by incorporating valence constraints and REINFORCE²³⁴-based optimization, thereby improving the validity and controllability.

In a 3D setting, autoregressive models atomically generate molecules in 3D space. G-SchNet,²³⁵ based on the SchNet architecture,²³⁶ was a symmetry-adapted 3D autoregressive generative model that sampled atom types and coordinates while respecting rotational invariance and Euclidean geometric constraints. Luo et al.²³⁷ proposed a 3D generative framework that formulates a probabilistic density over atom occurrences in 3D space and uses an autoregressive sampler to generate structures within protein-binding pockets. In addition, Pocket2Mol²³⁸ and Fragment-based LigAnd Generation framework (FLAG)²³⁹ have enabled 3D molecular generation by incorporating target–protein information. More recently, Quetzal²⁴⁰ has been introduced as a scalable 3D autoregressive generative model that builds molecules atom-by-atom in 3D by

combining a causal transformer for atom types with a diffusion-based module for coordinates.

3.3. Variational Autoencoders

Variational autoencoders (VAEs) are a class of generative models that learn a continuous latent space to enable diverse molecular generation and smooth optimization.^{241,242} This latent space provides a low-dimensional representation of abstract molecular features, in which structurally similar molecules are often embedded close together (Figure 6(f)). The VAE is a probabilistic model that generates molecules using this latent space. It maps an input molecule x to a latent variable z using encoder $q_{\phi}(z|x)$ and reconstructs the molecule from z using decoder $p_{\theta}(x|z)$. These encoder and decoder are typically implemented using neural networks and are trained to reconstruct input molecules. The training objective is to maximize the marginal log-likelihood $\log p_{\theta}(x)$, as defined in eq 1. However, this quantity is generally intractable to compute directly. Therefore, the model is optimized by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(x) = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - \text{KL}[q_{\phi}(z|x) || p(z)] \quad (4)$$

Here, the first term quantifies the reconstruction accuracy of the molecule from the latent variable, whereas the second term is the Kullback–Leibler divergence, which regularizes the encoder distribution $q_{\phi}(z|x)$ to remain close to the prior $p(z)$, typically chosen as a standard Gaussian distribution. After training, novel molecules can be generated by sampling $z \sim p(z)$ and decoding with $x \sim p_{\theta}(x|z)$.

Chemical VAE (CVAE)²⁴³ was one of the earliest and most representative methods for generating molecules, demonstrating the potential to optimize molecular properties in the latent space. However, as this model employed SMILES strings, it often generated syntactically invalid molecules. To overcome this limitation, GrammarVAE²⁴⁴ introduced context-free grammar constraints to improve syntactic validity. Syntax-directed variational autoencoder (SD-VAE)²⁴⁵ further enhanced syntactic validity by incorporating attribute grammars.

To handle molecular structures flexibly, numerous VAE-based approaches that consider graph and 3D structures have been proposed. Junction Tree Variational Autoencoder (JT-VAE)²⁴⁶ and its extension, Variational Junction Tree Encoder-Decoder (VJTNN),²⁴⁷ represented molecules as junction trees of molecular fragments, facilitating the generation of larger compounds. GraphVAE²⁴⁸ directly modeled molecular graphs using the VAE framework, and Constrained Graph Variational Autoencoder (CGVAE)²⁴⁹ introduced structural constraints to improve chemical validity. NeVAE²⁵⁰ improved the representational flexibility by addressing issues of permutation invariance and variable graph sizes. RELATION (REceptor–LigAnd interaction)²⁵¹ incorporated 3D pharmacophore features derived from protein–ligand complexes into a conditional VAE framework to capture interaction-relevant structural information. 3DMolNet²⁵² generated 3D molecular conformations by leveraging the charge, distance, and bond matrices. LiGAN²⁵³ employed a conditional VAE to generate 3D molecules conditioned on receptor binding sites, enabling structure generation directly tailored to protein–ligand interactions.

Various methods have been proposed to efficiently explore the latent spaces. Guo et al.²⁵⁴ introduced Property Controllable VAE (PCVAE), in which the latent space was structured to align specific dimensions with molecular properties, enabling

controlled manipulation via latent traversal. MDVAE (Monotonic Disentangled VAE)²⁵⁵ enhanced interpretability and controllability by enforcing monotonic relationships between latent variables and target properties. Generative tensorial reinforcement learning (GENTRL)²⁵⁶ combined reinforcement learning with the VAE for efficient de novo molecule generation. Tripp et al.²⁵⁷ introduced a weighted resampling scheme to improve sampling efficiency and optimization performance in latent space exploration. Query-based molecule optimization (QMO)²⁵⁸ employed zeroth-order gradient estimation in a query-based gradient descent framework for multiobjective property optimization.

3.4. Generative Adversarial Networks

Generative adversarial networks (GANs)²⁵⁹ are classified as deep generative models composed of two neural networks: a generator G that attempts to produce data samples (e.g., molecules) and a discriminator D that attempts to distinguish between real and generated samples (Figure 4(c) and Figure 6(g)). Through adversarial training, in which the generator and discriminator are optimized in a competitive manner, the generator progressively learns to create outputs that resemble the training data. Unlike VAEs, GANs do not require an explicit likelihood function, making them suitable for learning sharper data distributions, although training stability often remains a challenging task.²⁶⁰ Formally, the objective of this training is expressed as

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))] \quad (5)$$

where $D(x) \in (0, 1)$ denotes the discriminator's estimated probability that x comes from the training data rather than being generated, $p_{\text{data}}(x)$ is the data distribution, and $p(z)$ is a prior distribution (typically Gaussian) from which latent variables are sampled.

Initial attempts of GANs to molecular generation focused on molecular fingerprints rather than on explicit molecular structures. Kadurin et al.²⁶¹ proposed one of the earliest adversarial autoencoder (AAE) models for molecule generation, in which the model generated MACCS (Molecular ACCESS System) keys as outputs. This approach was extended in druGAN (drug GAN),²⁶² which incorporated property conditioning into the AAE framework to generate fingerprint vectors with desirable molecular attributes. Moreover, Bian et al.²⁶³ developed dcGAN (deep convolutional GAN), which applied convolutional architectures to various fingerprints such as MACCS and ECFP, demonstrating the feasibility of generating chemically relevant fingerprints. However, translating these generated fingerprints into valid molecular structures remained a nontrivial challenge.

To directly generate structured molecules, later studies investigated string- and graph-based GANs. Objective-Reinforced GAN (ORGAN)²⁶⁴ combined RNN-based sequence generation with adversarial training and reinforcement learning. Objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC),²⁶⁵ combined GAN and reinforcement learning with task-specific reward function, enhancing the validity and diversity of the generated molecules. Extensions such as RANC (Reinforced Adversarial Neural Computer)²⁶⁶ and ATNC (Adversarial Threshold Neural Computer)²⁶⁷ replaced the RNN generator in ORGAN-style models with a Differentiable Neural Computer (DNC), enhancing the model's ability to handle longer SMILES

sequences and improving molecular diversity. To improve the training stability, LatentGAN²⁶⁸ and the model proposed by Abbasi et al.²⁶⁹ decoupled the generative process by first training an autoencoder and then applying a GAN in the latent space to explore property-optimized molecules. Among the graph-based approaches, MolGAN (Molecular GAN)²⁷⁰ generated molecular graphs using a neural generator and applied reinforcement learning to optimize the chemical properties. Mol-CycleGAN²⁷¹ adapted the CycleGAN framework²⁷² to translate structurally distinct molecular sets, enabling lead optimization under similarity constraints. DEFactor (Differentiable Edge Factorization-based probabilistic graph generation)²⁷³ introduced a differentiable graph decoder that enabled the generation of variable-sized graphs with improved efficiency.

GAN-based 3D molecular structure-generation methods have also been proposed. EDMNets²⁷⁴ generated distance matrices that were invariant to translation and rotation. LiGANN²⁷⁵ employed a GAN-based architecture to generate 3D ligand shapes complementary to protein-binding pockets.

3.5. Normalizing Flow-Based Models

Normalizing flow models construct invertible mappings between simple base distributions and complex data distributions, enabling exact likelihood computations and efficient sampling (Figure 6(i)). In molecular generation, these models offer an alternative to variational approaches such as VAEs, which only approximate likelihoods by learning bijective transformations from a latent space to the space of molecular representations. Because the transformations are bijective with tractable Jacobians, flows provide exact log-likelihoods and exact latent-variable inference via the change-of-variables formula, allowing for analytical computation and facilitating latent-space optimization. Foundational flow models include nonlinear independent component estimation (NICE),²⁷⁶ real-valued nonvolume preserving (Real NVP),²⁷⁷ and Glow,²⁷⁸ which have inspired applications for discrete and structured data, including molecules. Formally, a normalizing flow defines a bijective and differentiable transformation $f_\theta = f_K \circ \dots \circ f_1$, where each f_k is an invertible transformation, and $z \sim p(z) = \mathcal{N}(0, I)$. The log-likelihood of x can then be expressed using the change-of-variables formula:

$$\log p_\theta(x) = \log p(z) - \sum_{k=1}^K \log \left| \det \frac{\partial f_k}{\partial h_{k-1}} \right| \quad (6)$$

where $h_0 = z$, $h_k = f_k(h_{k-1})$, and $x = h_K$. The sampling proceeds by drawing $z \sim \mathcal{N}(0, I)$ and applying $x = f_\theta(z)$. Conditional generation can be incorporated by conditioning each transformation f_k on the auxiliary information c , leading to a conditional distribution $p_\theta(x|c)$.

Recent advances in flow-based molecular generation have extended this framework to include various structural representations. GraphNVP,²⁷⁹ MoFlow,²⁸⁰ and graph residual flow (GRF)²⁸¹ were flow-based models that learned invertible transformations over graph-structured data. GraphAF²⁸² formalized generation as a sequential (autoregressive) flow and demonstrated reinforcement learning–based property optimization, and GraphDF²⁸³ likewise employed reinforcement learning to fine-tune a pretrained discrete-flow model for optimizing specified molecular properties. Hierarchical architectures such as molecular graph flow (MolGrow)²⁸⁴ and MolHF²⁸⁵ introduced multilevel control over structural features. FastFlows²⁸⁶ adopted SELFIES representations for efficient

training of small data sets. In 3D structure generation, E(n) Equivariant Normalizing Flows (E-NFs)²⁸⁷ enabled equivariant modeling of conformations through continuous-time flows. These advances highlight the adaptability of normalizing flows to various molecular representations and design tasks.

3.6. Energy-Based Models

Energy-based models (EBMs) define an energy function $E_\theta(x)$ over molecular structures, where lower energy values correspond to higher probabilities. Instead of directly modeling a properly normalized probability distribution $p_\theta(x)$, EBMs define an unnormalized density proportional to $\exp(-E_\theta(x))$, which defines the probability distribution as $p_\theta(x) = \frac{1}{Z_\theta} \exp(-E_\theta(x))$, where Z_θ is the partition function. This framework, which is rooted in statistical physics and was first formalized for machine learning through models such as Hopfield networks and Boltzmann machines,^{288,289} has generated renewed interest in the development of scalable learning and sampling algorithms.^{290,291} Formally, given a target distribution $p^*(x)$ representing desirable molecular structures or properties, EBMs aim to learn an energy function $E_\theta(x)$ such that configurations with lower energies are more likely under the target distribution $p^*(x)$. Sampling is typically performed using Langevin dynamics:

$$x_{t+1} = x_t - \frac{\eta}{2} \nabla_x E_\theta(x_t) + \sqrt{\eta} \cdot \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I) \quad (7)$$

which approximates the sampling from the Boltzmann distribution $p_\theta(x) \propto \exp(-E_\theta(x))$. Conditional generation can also be incorporated by extending the energy function to include auxiliary variables $E_\theta(x, c)$, where c encodes conditions such as target properties or biological constraints.

Although still relatively underexplored compared with other generative paradigms, EBMs have been applied to molecular graph generation. GraphEBM²⁹² introduced a permutation-invariant energy function for molecular graphs and employed Langevin dynamics-based sampling²⁹³ to generate molecules corresponding to low-energy (i.e., high-probability) regions. Pang et al.²⁹⁴ introduced an energy-based prior over the latent space of a SMILES-based generator, enabling efficient sampling while preserving chemical validity. TagMol (Target-specific Generation of Molecules),²⁹⁵ which was developed for target-specific drug discovery, combined a protein encoder, ligand generator, and energy network to evaluate and guide ligand generation toward high binding affinity. Further development of molecule generation methods using EBMs is expected. Although EBMs and diffusion-based generative models (Section 3.8) share conceptual connections, both rely on gradients of log-density (i.e., scores) to guide sampling through Langevin-type dynamics,^{296,297} we describe diffusion models separately in the following section because they constitute a distinct and rapidly developing class of generative approaches in molecular design.

3.7. Transformer and Large Language Models

The transformer architecture proposed by Vaswani et al.²⁹⁸ in 2017 revolutionized sequence modeling techniques for natural language processing. Unlike previous RNNs and LSTMs, which process sequence data sequentially and generate the next output based on the previous token, the transformer uses a self-attention mechanism to simultaneously calculate relationships between any token, enabling effective learning of long-range dependencies and parallelization of computations. By leveraging this structural advantage, transformers are highly suited for self-

supervised learning with large-scale text data, giving rise to large language models (LLMs) such as generative pretraining transformer (GPT) series.^{299,300} These models first learn the statistical structure of a language using vast amounts of unlabeled data, and then adapt to various downstream tasks through fine-tuning and prompt design. Inspired by these technical achievements, research on the application of transformers and LLM frameworks in the field of chemistry has gained significant attention.^{190,301,302}

Transformers have been widely used to predict molecular properties. For example, SMILES Transformer,³⁰³ Molecular Transformer,³⁰⁴ SMILES-BERT,³⁰⁵ MolBERT,³⁰⁶ and SELFIES-TED³⁰⁷ were pretrained on 1D molecular strings (e.g., SMILES) using transformer architectures and utilized for downstream tasks such as reaction-prediction or property prediction. ChemBERTa,³⁰⁸ Chemformer,³⁰⁹ and X-MOL³¹⁰ demonstrated that large-scale pretraining on millions to hundreds of millions of molecules improved prediction performance. Polymer-specific models such as polyBERT³¹¹ and TransPolymer³¹² have also been reported. The application of transformers in molecular graph representations has also been actively explored. Graphormer³¹³ and GROVER (Graph Representation frOm self-superVised mESSAGE passing tRansformer)³¹⁴ were trained on molecular graph data sets consisting of several million to ten million compounds and have shown superior performance compared with conventional GNNs. GraphTrans (Graph Transformer)³¹⁵ combines GNN-based preprocessing of atom positions and bonding structures with transformer models, thereby enabling high-accuracy structural representations.

Transformers have been employed in an autoregressive manner for molecular generation, where molecules are constructed by sequentially generating token sequences (e.g., SMILES) one character at a time (Figure 6(e)). MolGPT³¹⁶ generated SMILES in an autoregressive manner using a transformer decoder, optionally biasing generations toward desired properties via conditioning signals. MolGen³¹⁷ enabled task-agnostic and diverse molecule generation through large-scale pretraining on SELFIES. TransVAE³¹⁸ introduced a Transformer-based variational autoencoder to perform generation in latent space. He et al.^{319,320} proposed transformer-based models that used matched molecular pairs or similarity-based rules to learn structural conversions and improve diversity. Transformer models that incorporate 3D information have also been explored. For instance, Molformer³²¹ combined atom- and motif-level graph representations with spatial and semantic attention to achieve high accuracy in structure prediction and generation.

Conditional molecular generation based on the molecular properties or target proteins has also been actively explored. Grechishnikova et al.³²² developed a transformer-based translation model from target protein sequences to drug-like molecules. cMolGPT³²³ applied a transformer architecture to generate target-specific molecules by using attention-based conditioning. Multiconstraint molecular generation (MCMG),³²⁴ Regression Transformer,³²⁵ and conditional molecular generation net (CMGN)³²⁶ extended conditional generation by incorporating reinforcement learning, sequence-based property encoding, or fragment/property recovery tasks. Target-aware molecular generation (TamGen)³²⁷ introduced a GPT-based conditional generation model that leverages target-protein context to guide molecular generation. SPM (Structure–Property Multi-Modal foundation model)³²⁸ and

Llamol³²⁹ employed multimodal learning or Llama³³⁰-based language modeling to enable flexible bidirectional or context-aware generation. Aksamit et al.³³¹ integrated metaheuristic multiobjective optimization into a transformer-based SMILES model, as advanced integrations with optimization techniques. Ai et al. proposed MTMol-GPT,³³² which mimicked expert-designed strategies for goal-oriented molecular generation using GAIL (generative adversarial imitation learning).³³³

3.8. Diffusion Models

Diffusion models are a class of generative models that have recently attracted attention in the field of molecular design owing to their ability to directly generate 3D molecular structures (Figure 6(h)). These models formulate the generation as a gradual denoising process that transforms a noisy 3D conformation into structured molecular data. Diffusion models were first formalized as denoising diffusion probabilistic models (DDPM).³³⁴ Since then, they have been widely applied in various fields. Examples of image generation include Stable Diffusion³³⁵ and Imagen.³³⁶ In biomolecular modeling, they have been used in systems such as RoseTTAFold All-Atom³³⁷ and AlphaFold 3.³³⁸ This framework has also been generalized through continuous-time formulations using stochastic differential equations (SDEs), such as EEGSDE (equivariant energy-guided SDE),³³⁹ offering a unified perspective on score-based generative modeling.²⁹⁷

Formally, we first define a forward process $q(x_t|x_{t-1})$ that incrementally perturbs the data through a Markov chain of Gaussian transitions for discrete timesteps $t = 1, \dots, T$.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (8)$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a normal distribution with mean μ and covariance Σ , β_t controls the variance of the added noise at each step, and I is the identity matrix. The reverse process $p_\theta(x_{t-1}|x_t)$ is parametrized by a neural network trained to estimate the denoised distribution at each time step. A new sample x_0 is generated by first drawing a noise vector $x_T \sim \mathcal{N}(0, I)$, which can be interpreted as a latent variable z analogous to those in other generative models, and then iteratively sampling from the learned reverse transitions as follows:

$$x_T \xrightarrow{p_\theta(x_{T-1}|x_T)} x_{T-1} \xrightarrow{p_\theta(x_{T-2}|x_{T-1})} \dots \xrightarrow{p_\theta(x_1|x_2)} x_1 \xrightarrow{p_\theta(x_0|x_1)} x_0 \quad (9)$$

resulting in a structured sample x_0 , which represents the generated molecule. This formulation is naturally extended to conditional generation by incorporating auxiliary information c , such as target protein structures or desired properties, into the model. The resulting conditional distribution $p_\theta(x|c)$ enables molecule generation guided by specific structural or property requirements.

The initial applications of diffusion models to molecular generation focused on the direct modeling of 3D molecular structures while preserving symmetries, particularly E(3)/SE(3)-equivariance. The E(3) Equivariant Diffusion Model (EDM)³⁴⁰ demonstrated that molecular geometries could be generated in 3D space by treating atoms as equivariant point clouds, preserving rotational and translational symmetry. EEGSDE³³⁹ extended this approach by incorporating energy-based guidance using stochastic differential equations, thereby improving the physical plausibility of the generated structures. Geometric Latent Diffusion Models (GEOLDM)³⁴¹ improved

validity and scalability by performing diffusion in a latent space that encoded molecular geometry with equivariant representations. MDM³⁴² introduced dual equivariant encoders to encode interatomic relations and a distributional controlling variable to enhance exploration, improving 3D generation for larger and more diverse molecules.

Based on these advances, several studies have integrated structural or functional constraints into the diffusion process to enable target-specific or property-conditioned molecular generation. For example, TargetDiff³⁴³ introduced 3D representations of protein-binding pockets as conditions, guiding ligand generation toward spatial compatibility with the binding site. DiffBP (target-aware molecular Diffusion model for Protein Binding)³⁴⁴ enhanced structure-conditioned molecular generation by incorporating geometric constraints from protein targets into the diffusion process. DiffSBDD³⁴⁵ employed an SE(3)-equivariant conditional diffusion model and demonstrated inpainting and iterative oracle-driven improvement, facilitating scaffold hopping and property tuning. Huang et al.³⁴⁶ further extended this approach by integrating both generation and lead optimization into a dual-model framework. MOOD (Molecular Out-Of-distribution Diffusion),³⁴⁷ while not explicitly leveraging protein structures, enabled exploration beyond the training distribution by guiding diffusion toward molecules with the desired predicted properties.

Diffusion models have also been adapted to specialized molecular design tasks, including substructure completion, guided optimization, and multiobjective generation. DiffLinker³⁴⁸ generated linkers that bridge predefined fragments, whereas DiffDec³⁴⁹ performed R-group decoration with the awareness of pocket geometry. DiffInt³⁵⁰ incorporated explicit hydrogen-bond (protein–ligand) interaction guidance to capture relevant intermolecular interactions. GaUDI (guided diffusion model)³⁵¹ introduced a flexible post hoc guidance framework that decouples sampling and optimization, supporting single or multiobjective design. DiffSMol³⁵² leveraged shape-conditioned diffusion to align generated molecules with reference ligand geometries. CMD-GEN (Coarse-grained and Multidimensional Data-driven molecular generation)³⁵³ generated selective inhibitors by integrating pharmacophore-guided sampling with structure-based filtering. These developments underscore the versatility of diffusion models for addressing a wide range of molecular design scenarios.

3.9. Overall Perspective on Generative Models

Selecting an appropriate generative method for a specific molecular design task is a difficult question. At present, the suitable method depends on the task, the molecular representation, and also on how well the model works together with the optimization strategies introduced in Section 4. Therefore, there is no single generative model that is universally best. Historically, 1D representations have been easier to handle and supported by larger data sets, and they have shown relatively stable performance, particularly for valid molecule generation. In recent years, the use of large-scale models based on transformers has further improved their capabilities. On the other hand, early 2D and 3D generative models faced several practical difficulties. For 2D graph-based models, enforcing chemical validity and learning sufficiently expressive graph features was challenging with the architectures available at that time. Early 3D generative models struggled to respect geometric constraints and to maintain physically reasonable structures. As described above, advances in graph neural networks, equivariant architectures,

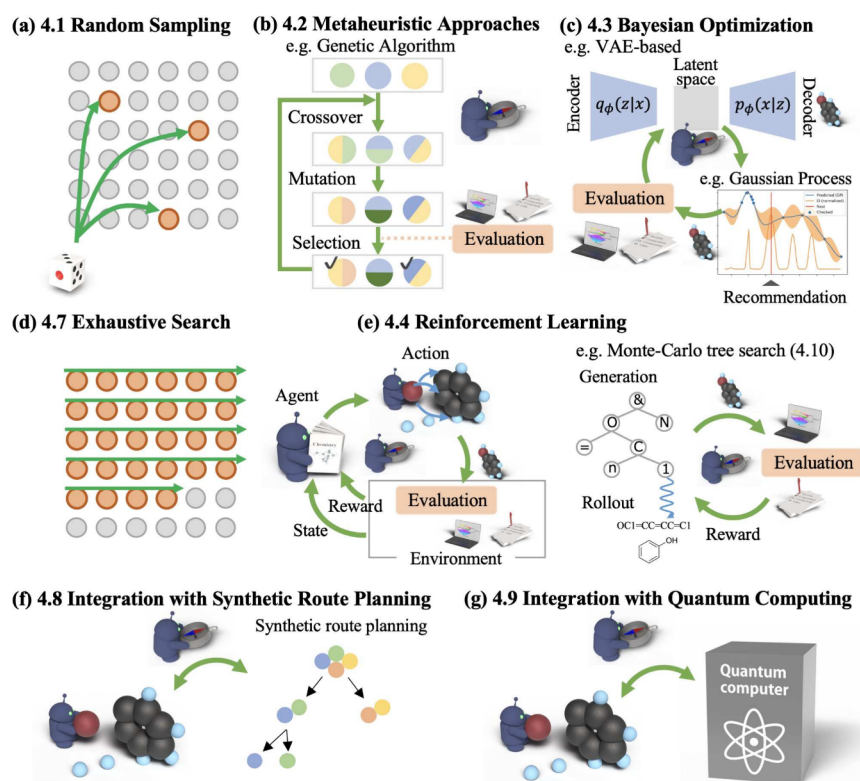


Figure 7. Representative optimization and search strategies introduced in Section 4. (a) Random sampling (each circle indicates a molecule). (b) Metaheuristic approach (each circle indicates a molecule and a circle with different colors indicates structurally mixed molecules). (c) Bayesian optimization (molecular properties are typically optimized in the latent space of VAE through Bayesian optimization). (d) Exhaustive search (all molecules in a data set are evaluated and each circle indicates a molecule). (e) Reinforcement learning. (f) Integration with synthetic route planning. (g) Integration with quantum computing. For further details, refer to their respective sections.

and various geometric generative models, including diffusion-based models, have greatly improved the quality and robustness of 2D and 3D molecular generation. In particular, 3D-based models have become promising for tasks such as generating molecules that bind to a target protein. As a way to compare generative methods, benchmark studies often provide a useful basis. Details of benchmarks are summarized in Section 5.4. Also, Section 7 provides examples of molecular generation cases that were experimentally synthesized, which can also help readers understand the practical behavior of different generative models.

4. OPTIMIZATION AND SEARCH STRATEGIES

Molecular design that combines generators and evaluators, shown in Figure 4(d), is often formulated as a black-box optimization problem.¹⁷ Black-box optimization refers to the problem of finding inputs that minimize or maximize an objective function whose gradient is unavailable or whose evaluation is costly. In molecular design, the objective function maps the generated molecules to their properties. If the objective function is differentiable or convex, efficient gradient-based methods and fast algorithms can be employed. However, molecular property estimation is typically complex and nonlinear, and when relying on simulation-based evaluation, the computational cost becomes high, resulting in a black-box optimization problem. Therefore, efficient search and optimization algorithms are required to identify molecules with the desired properties and tune the parameters of molecule generators to produce more promising molecules.

Figure 7 presents an overview of the representative optimization and search strategies described in the following section. The simplest strategy is random sampling (Section 4.1), which leverages probabilistic generative models to generate molecules repeatedly until favorable candidates are found (Figure 7(a)). Metaheuristic algorithms (Section 4.2), such as evolutionary and genetic algorithms (Figure 7(b)), as well as particle swarm optimization, are also widely used. Additionally, Bayesian optimization (Figure 7(c) and Section 4.3), a representative strategy for global optimization, is effective for exploring the latent space and searching within combinatorially generated molecules. Furthermore, optimization approaches that modify the generative model itself or efficiently explore the chemical space using reinforcement learning (Figure 7(e) and Section 4.4) have been actively studied to enhance the generation of desirable molecules. Property optimization has also been achieved through gradient-based methods (Section 4.5), Bayesian approaches (Section 4.6), and exhaustive search (Figure 7(d) and Section 4.7). Section 4.8 introduces approaches that simultaneously perform molecular design and synthetic route planning (Figure 7(f)). Recently, the application of quantum computing (Section 4.9) has also been investigated to enable the search for an extremely large molecular space (Figure 7(g)).

4.1. Random Sampling

Random sampling or random search is a strategy for generating molecules by drawing samples from a probabilistic generative model. In probabilistic generative models, random sampling is commonly used to assess the quality and diversity of learned molecular distributions.^{224,228} Although random sampling is

conceptually simple, it serves as a fundamental strategy in molecular design and is frequently used as a baseline for evaluating advanced generation and optimization methods computed on samples emitted by the models.^{354,355} Promising molecules can also be discovered via a random search, which uniformly selects and evaluates candidates from a preconstructed or combinatorially generated data set, as described in Section 3.1. Furthermore, the conditional generative model $p_{\theta}(x|c)$, as shown in Figure 4(b), allows random sampling to be guided by a specific condition c , thereby enabling the generation of molecules with desired properties.

4.2. Metaheuristics

Metaheuristics refer to a broad class of optimization strategies that iteratively improve candidate solutions using simple modification rules or evolutionary operations without relying on gradient information. These strategies are particularly effective in black-box optimization settings, where the objective function is nondifferentiable, noisy, or computationally expensive.

Among various metaheuristic algorithms, evolutionary algorithms (EAs),³⁵⁶ particularly genetic algorithms (GAs), have been widely used in molecular design. In GA-based molecular designs, as illustrated in Figure 7(b), molecules are treated as individuals in a population and iteratively modified via selection, crossover, and mutation operations to produce candidates with desired properties. Early applications included rule-based structural generation³⁵⁷ and graph-based crossover strategies.²¹⁷ The GA has also been applied to the design of organic photovoltaic materials.³⁵⁸ The ACSESS (Algorithm for Chemical Space Exploration with Stochastic Search) framework³⁵⁹ introduced a diversity-driven GA-based strategy for molecular exploration, which was applied in various case studies.³⁶⁰ Kanal et al.³⁶¹ also proposed an efficient goal-directed GA strategy.

Independent of the development of deep generative models, research on EA-based approaches is ongoing. For example, Yoshikawa et al. proposed ChemGE,³⁶² which applied grammatical evolution to optimize SMILES strings directly. Jensen introduced GB-GA (graph-based genetic algorithm),²¹⁸ a graph-based evolutionary algorithm approach that enforced valence bond-order constraints. These approaches demonstrated that the EA/GA-based strategies could achieve levels of diversity and efficiency comparable to deep generative models,³⁶³ thereby encouraging various extensions. For example, GA-D³⁶⁴ integrated a discriminator network with SELFIES-based GA to promote diversity while steering property optimization. MolFinder³⁶⁵ introduced slicing- and replacement-based crossover and mutation operations directly to SMILES strings for molecular design. EvoMol³⁶⁶ demonstrated that EA-based algorithm could be used to optimize a wide range of molecular properties, including SAScore, highest occupied molecular orbital (HOMO) energies, and lowest unoccupied molecular orbital (LUMO) energies.

Building on these foundational studies, more complex strategies for enhancing the practical performance have been proposed. ChemistGA³⁶⁷ combines a genetic algorithm with deep learning by redefining crossover via a neural network to bias offspring toward synthetically accessible molecules, improving synthesizability and success rates. GEGL (genetic expert-guided learning)³⁶⁸ combined a SMILES-based RNN generator with GB-GA to guide molecular optimization and progressively retrained the RNN using high-scoring molecules

via imitation learning, improving generation performance. EAs have also been applied to multiobjective optimization (MOO) problems.³⁶⁶ GENERA³⁶⁹ achieved a successful MOO based on Pareto optimization, and EDM (evolutionary design methodologies)³⁷⁰ introduced a hybrid strategy that applied GA in molecular fingerprint space and translated optimized vectors into SMILES using an RNN. Blaschke et al.³⁷¹ reported the practical optimization of mechanosensitive molecules using GA. Recently, MOLLEO (Molecular Language-Enhanced Evolutionary Optimization)³⁷² has employed an LLM to implement crossover and mutation operations within a GA framework.

In addition to evolutionary algorithms, other metaheuristic approaches have been explored for molecular design. JANUS³⁷³ applied parallel tempering to coordinate GA-based searches at different levels of selection pressure, in combination with STONED (superfast traversal, optimization, novelty, exploration and discovery)-based mutation,³⁷⁴ enabling broader exploration of the chemical space. Particle swarm optimization (PSO), which updates particles in a continuous vector space according to heuristics inspired by collective behavior,³⁷⁵ has also been employed in molecular design. Winter et al.³⁷⁶ applied PSO to the latent space of a VAE to optimize its molecular properties. ChemMORT (Chemical Molecular Optimization, Representation and Translation)³⁷⁷ extended PSO to multiobjective settings for generating candidates with favorable absorption, distribution, metabolism, elimination, and toxicity (ADMET) profiles.

In summary, metaheuristic algorithms offer flexible and diverse tools for molecular designs. Additionally, other metaheuristic algorithms such as tabu search and ant colony optimization have the potential to provide unique solutions to specific molecular design problems. Although metaheuristics are effective in molecular design, several common challenges are associated with their applications. Specifically, these include the difficulty of tuning numerous parameters to achieve an optimal search performance and the lack of a theoretical guarantee that the algorithm will converge to a global optimum. Additionally, striking a balance between maintaining molecular diversity and ensuring convergence toward the desired properties often poses practical challenges.

4.3. Bayesian Optimization

Bayesian optimization (BO) is a type of sequential model-based global optimization (SMBO) commonly used for black-box optimization problems³⁷⁸ (Figure 7(c)). In BO, a surrogate model, typically Gaussian process regression,³⁷⁹ is first constructed to predict the distribution of property values for candidate molecules. An acquisition function is then used to score the candidates based on the surrogate model's predictions, and the candidate with the highest score is selected for evaluation. Examples of commonly used acquisition functions include the probability of improvement (PI),³⁸⁰ expected improvement (EI),³⁸¹ upper confidence bound (UCB),³⁸² and Thompson sampling.³⁸³ The selected candidate is then evaluated via simulations or experiments, and the surrogate model is updated using new data, enabling more accurate recommendations in subsequent iterations. Although the metaheuristic algorithms described in Section 4.2 are generally fast and flexible, their convergence to the global optimum is sensitive to hyperparameter settings. By contrast, BO is particularly well suited for global optimization tasks, particularly when candidate evaluations are computationally or experimentally expensive. In recent years, BO has gained increasing

attention in various fields, including chemistry, as an efficient strategy for experimental and molecular design.^{384–388}

In the context of molecular design, particularly following the development of VAE-based generative methods that enable the construction of latent spaces for molecules, numerous approaches applying BO to these latent spaces have been reported. In the first VAE applications,²⁴³ Gómez-Bombarelli et al. constructed a model that predicted chemical properties from the latent space using Gaussian process and reported that drug-like molecules could be efficiently explored based on this model. Although BO was not explicitly performed in the study, Griffiths et al.³⁸⁹ subsequently introduced BO and proposed a constrained BO method (Constrained BO-VAE) that added constraints to avoid the “dead region,” an area where VAE decoding was prone to failure. BO-based latent space-optimization methods have been widely used in various applications. For example, Truong et al.³⁹⁰ proposed a design method combining latent space based on JT-VAE with BO and gradient optimization and reported the discovery of VEGFR-2 (vascular endothelial growth factor receptor 2) inhibitors. Matsukiyo et al.³⁹¹ integrated a transformer-based VAE with BO to develop a method for designing new inhibitor and activator candidates for specific target proteins.

Applying BO to existing large-scale databases or generated molecular libraries is also an effective approach for discovering promising molecules. Mehta et al.³⁹² applied BO to libraries comprising millions of molecules from the ZINC database and proposed a (Machine learning framework for Enhanced Molecular Screening) framework, which discovered high-performance molecules by evaluating only a small fraction of the data set. Graff et al.³⁹³ combined a surrogate model with BO to efficiently discover top-scoring docking candidates from a virtual library of 100 million molecules.

Because BO performance depends on the choice of kernel functions and molecular feature representation, research on appropriate features and kernels for molecules has also progressed. Korovina et al. developed ChemBO,³⁹⁴ which incorporated graph kernels and an optimal-transport kernel into Gaussian processes to capture graph structures and enable efficient optimization. BOSS (Bayesian Optimization over String Spaces)³⁹⁵ applied BO directly to raw string representations using string kernels and genetic algorithms, thus avoiding the need for latent space projections and allowing efficient optimization under syntactic constraints.

4.4. Reinforcement Learning

Reinforcement learning (RL)^{396,397} is a framework for learning optimal action–selection strategies, known as policies, using trial-and-error interactions with the environment to achieve long-term objectives (Figure 7(e)). A policy defines the action to be taken in a given state. In RL, this policy is iteratively refined based on accumulated experience. A key challenge in learning the policy in RL is balancing between exploration, attempting to discover untried options, and exploitation, choosing options that have been found to be good based on past experience. Exploration alone can be inefficient, whereas pure exploitation may overlook potentially better alternatives; therefore, dynamically learning this trade-off is essential for successful RL. Traditionally, RL has been based on methods such as value iteration and policy iteration from dynamic programming, as well as approximation methods such as temporal-difference (TD) learning.³⁹⁶ In recent years, deep reinforcement learning has emerged, which uses deep neural networks as function

approximators. For example, Deep Q-Network (DQN)³⁹⁸ approximates Q-values, which represent the values of state-action pairs, using neural networks, whereas proximal policy optimization (PPO)³⁹⁹ directly learns policies using gradient-based updates. Modern RL has demonstrated remarkable performance in complex sequential decision-making tasks such as video-game playing (e.g., Atari),⁴⁰⁰ board games (e.g., Go),^{401,402} and robotics.^{403–405}

RL can be applied naturally to molecular design. In this context, the sequential construction or modification of molecular structures can be formulated as a decision-making process in which actions such as adding or replacing atoms or bonds are chosen based on the current molecular structure (i.e., the state). The evaluated properties of the generated molecules serve as rewards. In the following section, we introduce RL-based approaches on three types of molecular representations: 1D strings, 2D molecular graphs, and 3D structures.

Methods using RL based on 1D strings, such as SMILES and SELFIES, have been actively developed. In these approaches, molecular generation models are typically implemented using RNNs or transformers, in which the generator outputs a probability distribution over the next token, which acts as a “stochastic policy.” This generative policy can be optimized using policy-gradient methods, particularly REINFORCE,²³⁴ to encourage the generation of molecules with the desired properties. In 2017, Olivecrona et al. proposed REINVENT,⁴⁰⁶ a method that fine-tunes a pretrained RNN generator by assigning rewards based on the properties of the generated molecules. Following this work, REINVENT2⁴⁰⁷ and REINVENT4⁴⁰⁸ further refined the RL framework, improving both the performance and usability as practical tools for molecular design. Also, ReLeaSE (Reinforcement Learning for Structural Evolution)⁴⁰⁹ combined an RNN-based molecular generator with the REINFORCE algorithm to optimize molecular properties, while the property evaluation is performed through a prediction model. Recently, transformer-based architectures have also been introduced as molecular generators in RL-based molecular designs.^{410,411}

As an alternative approach to string-based molecular generation, Monte Carlo tree search (MCTS),⁴¹² originally developed for game-playing AI such as Go,⁴¹³ has also been applied. In ChemTS⁴¹⁴ an RNN-based SMILES generator was combined with MCTS to construct molecules by sequentially expanding partial SMILES strings as a tree structure. The generated molecules were then evaluated, and the rewards served as feedback to guide the optimization process. MCTS-based exploration can be accelerated through large-scale parallelization^{415,416} and has been employed to design dyes and fluorescent molecules using computationally intensive evaluations, such as QC calculations.^{417–419} The details of MCTS and its parallelization are described in Section 4.10. Furthermore, several extensions of the MCTS-based approach have been developed, including applications in drug discovery using docking simulations,⁴²⁰ multiobjective optimization tasks,^{421–423} and an improved version, ChemTSv2.⁴²⁴

In RL using 2D graph representations, structural modifications in molecular graphs are regarded as “actions,” enabling exploration that faithfully reflects chemical structures. In the graph convolutional policy network (GCPN),⁴²⁵ such graph-level actions were optimized using PPO, whereas a GAN-based discriminator served as a reward function to evaluate and improve the molecular properties. Similarly, MolDQN⁴²⁶ employed DQN to learn actions such as bond addition and

removal on molecular graphs, optimizing the structure based on reward feedback. Other approaches have also applied MCTS directly to graph operations. For instance, GB-GM-MCTS (graph-based generative model with Monte Carlo tree search)²¹⁸ explored sequences of effective graph transformations to optimize 2D molecular structures. RationaleRL⁴²⁷ addressed multiobjective molecular optimization by first identifying rationales, which are substructures responsible for specific properties, using MCTS. These rationales were then integrated into a merged graph based on the maximum common substructure, and subsequent molecule generation was performed using VAE. VGAE-MCTS⁴²⁸ combined a graph-based VAE with MCTS to explore and optimize molecular structures more effectively.

In RL methods that handle 3D molecular structures, the spatial configuration of atoms and bonds is treated directly as a state. MolGym⁴²⁹ represented molecules as 3D point clouds and learned the policy of adding both the next atom and its spatial position using PPO. In this framework, SchNet²³⁶ was employed as a state representation invariant to spatial symmetries, such as rotation and translation. In a subsequent study, COVA-RIANT⁴³⁰ explicitly incorporated structural symmetries to improve the stability and chemical validity of the generated structures. DeepLigBuilder⁴³¹ sequentially generated ligand molecules that fit into protein-binding pockets by combining MCTS with a policy network (L-Net), using docking scores and other metrics as rewards to optimize 3D structures.

More advanced methods that incorporate RL have also been developed. One of the challenges of RL is its limited exploratory efficiency. Several hybrid approaches have been proposed to address this issue. For example, some methods combined conventional RL with hill-climbing-based exploration techniques,^{432,433} while others incorporated active learning to more effectively guide exploration.⁴³⁴ Encouraging agents to discover novel molecules has been demonstrated to improve sample efficiency in molecular design tasks.^{435–437} Furthermore, more integrated approaches have emerged, such as methods that address multiobjective drug design by combining multiple MCTS-based exploration strategies⁴³⁸ or incorporating reinforcement learning into adversarial training frameworks.⁴³⁹ Moreover, Generative Flow Networks (GFlowNets)⁴⁴⁰ have been proposed as a reinforcement learning-inspired generative framework that learns to sample candidates proportionally to their rewards, thereby promoting broader exploration and preventing overconcentration on only a few high-scoring solutions. In molecular design tasks, GFlowNets-based methods have demonstrated improved diversity and sample efficiency, especially in multiobjective optimization settings, and are increasingly being developed for controllable generation based on desired molecular properties.^{441–443} In summary, RL provides a flexible framework that is well-suited to various requirements in molecular design and is expected to play an important role in the future.

4.5. Gradient-Based Optimization

If one can construct a differentiable function that maps a continuous representation of molecules to the property space, efficient exploration using gradient-based methods becomes theoretically possible. Several gradient-based methods using latent spaces obtained by VAE have been proposed. Kondratyev et al.⁴⁴⁴ proposed a method for efficiently exploring molecules with desired HOMO values by first constructing a latent space using JT-VAE, then training a neural network to predict HOMO

values from latent vectors, and finally applying gradient descent in the latent space to find the optimized molecules. LIMO (Latent Inverse Molecular Optimizer)⁴⁴⁵ also used VAE, but instead of directly predicting properties from the latent space, it first decoded the latent vector into a molecular structure and then applied a separate property prediction model to the decoded molecule. This allowed for the construction of an end-to-end differentiable model that connects the latent vector to the predicted property, enabling the efficient optimization of the latent vector using gradient-based methods, such as the Adam optimizer.⁴⁴⁶ Fu et al. proposed DST (Differentiable Scaffolding Tree),⁴⁴⁷ which transformed molecular structures into scaffold trees with locally differentiable operations and backpropagates gradients from the property predictor based on GNN, thereby enabling efficient structure optimization. Therrien et al.⁴⁴⁸ proposed a GNN-based generation method that constructed a differentiable function with structural and chemical constraints and directly optimized the structure based on gradient methods. Although gradient-based methods are efficient exploration techniques, they are based on local exploration, which often leads to practical challenges in finding global solutions or exploring diverse molecules. In scenarios where property data can be incrementally obtained via simulations or experiments, a practical challenge arises in balancing the model retraining with the continued application of gradient-based optimization.

4.6. Bayesian Sampling

Bayesian sampling approaches to molecular generation and optimization utilize a forward model, a predictive model that estimates molecular properties from molecules, to address the inverse problem of generating structures that satisfy a desired set of properties. Typically, this is achieved by employing Markov Chain Monte Carlo (MCMC)⁴⁴⁹ or Sequential Monte Carlo (SMC)⁴⁵⁰ methods to sample molecular structures according to their posterior likelihood, given the target properties.

Ikebata et al.⁴⁵¹ proposed a method that combined ML-based property prediction with SMC to efficiently sample molecules with favorable properties in the SMILES space. Miyao et al.^{452,453} constructed a feature space based on monotonously changing molecular descriptors and employed a Gaussian mixture model and cluster-wise linear regression to generate molecular structures that satisfied the desired properties as an inverse QSPR/QSAR problem. Recently, graph-based MCMC methods have been proposed for structural optimization. MIMOSA (Multi-constraint MOleculE SAMpling)⁴⁵⁴ trained GNNs to score and prioritize candidate molecular editing operations, such as addition, deletion, or substitution, and performed multiobjective optimization using Gibbs sampling to balance property scores and structural similarity. MARS (MARkov molecuLAR Sampling)⁴⁵⁵ further improved sampling efficiency and structural diversity by incorporating learned proposal distributions into the editing process. Zhang et al.⁴⁵⁶ developed a framework for concurrently designing molecules and their synthetic reaction networks using SMC-based probabilistic sampling to identify candidates that satisfied multiple constraints, including molecular structure, synthetic feasibility, and reaction yield.

4.7. Exhaustive Search

Exhaustive search, or brute-force search, refers to approaches that comprehensively evaluate all molecules in a preconstructed candidate set, as shown in Figure 7(d). Such data sets, which are often referred to as virtual libraries, can be generated using combinatorial methods or deep generative models, as described

in Section 3. Although exhaustive search may not be classified as an optimization or search algorithm in a strict sense, it remains a practically useful and widely adopted strategy for inverse design. A typical example is virtual screening (VS),⁴⁵⁷ where each molecule in a large library is assessed using computational methods such as docking^{458–461} or QC calculations. Recent advances in high-performance computing, including the use of supercomputers and GPUs, have enabled large-scale VS studies that were previously computationally infeasible. In the field of drug discovery, VS using docking calculations has been performed on molecular libraries of hundreds of millions to billions of molecules, and hit compounds have been successfully identified.^{462–464} In materials science, many studies have reported that large-scale screening using QC calculations has led to the discovery of promising candidate materials.¹⁶³ Although conceptually naive and computationally demanding, VS constitutes a valid approach, particularly in settings where the exhaustive evaluation of a large candidate library is computationally feasible.

4.8. Integration with Synthetic Route Planning

In practical applications of molecular generative AI, the actual value of the designed molecules depends not only on their estimated properties but also on their synthetic feasibility. Without the capability for synthesis, even the most promising candidates identified by AI cannot be experimentally realized. In this section and in Section 5.2, we survey the current approaches for incorporating synthetic feasibility into AI-driven molecular generation (Figure 7(f)). Two typical categories of synthesizability have been incorporated into AI-driven molecular generation. The first approach integrates molecular designs with synthetic route planning, whereas the other employs synthetic feasibility metrics or prediction models as rewards during molecular optimization. Both approaches have their advantages and limitations. Integrating molecular design with synthetic route planning tends to yield more synthetically feasible molecules; however, these methods are often less adaptable to other molecule-generation frameworks because the algorithms for molecular design and synthetic route planning are tightly coupled. Although synthetic feasibility metrics or prediction models are less capable of directly promoting synthesizable molecules, they are highly amenable to integration with various molecule-generation frameworks.

To integrate molecular generation with synthetic route planning, most methods first implement a strategy for constructing synthetic routes, either top-down or bottom-up, from latent spaces or molecular embeddings. The molecules are then designed to possess the desired properties via an optimization process that utilizes latent spaces and/or molecular embeddings. Bradshaw et al.⁴⁶⁵ proposed MoleculeChef, a generative model that first generates a set of reactants from a molecule embedding and then generates a product by simulating their reaction using a reaction prediction model called Molecular Transformer.³⁰⁴ Bradshaw et al.⁴⁶⁶ developed a serialization method to construct multistep synthetic routes based on directed acyclic graphs (DAGs) and proposed an RNN model that constructs a multistep synthetic route starting from building blocks obtained from the latent space to generate the final product. For molecular optimization, they referred to the RNN model with a fixed latent space as DoG-Gen and fine-tuned the model using a hill-climbing algorithm³⁵⁴ to generate DAGs with more desirable properties. Gao et al.⁴⁶⁷ proposed SynNet, which frames multistep retrosynthetic route generation

as a Markov decision process conditioned on a target molecular embedding and explores a chemical space composed of reaction templates and building blocks. Because the target embedding conditions the generation, optimizing this embedding enables SynNet to produce molecules with improved properties, along with their corresponding retrosynthetic pathways. In SynNet, a genetic algorithm was used to perform optimization. SynMolOpt⁴⁶⁸ extends SynNet by preparing property-specific functional reaction templates to enhance molecular optimization tasks. Another method that generates molecules while constructing multistep synthetic routes is the cascaded variational autoencoder (casVAE).⁴⁶⁹ The casVAE consists of two junction tree-based VAEs, a product VAE and a reaction VAE, and is trained on reaction trees of molecules constructed by Retro*.⁴⁷⁰ This architecture enables the model to generate molecules while reproducing reaction trees similar to those constructed by Retro*. In addition to latent space-based methods, Bayesian methods combined with single-step forward synthesis prediction models, such as ChemBO³⁹⁴ and Seq-Stack-Reaction,⁴⁵⁶ have been proposed. Similarly, methods such as reaction-driven objective reinforcement (REACTOR)⁴⁷¹ and policy gradients for forward synthesis (PGFS),⁴⁷² which combine RL with single-step forward synthesis prediction models, have been introduced. Furthermore, SyntheMol was proposed as a method for exploring the vast combinatorial chemical space defined by building blocks and well-validated chemical reactions using MCTS.⁴⁷³ This section reviews the various molecular generative AIs that integrate molecular design and synthetic route planning. The synthetic feasibility metrics and prediction models are described in Section 5.2.

4.9. Integration with Quantum Computing

Molecule generation and optimization techniques incorporating quantum circuits have been proposed (Figure 7(g)). Kao et al.⁴⁷⁴ implemented a molecular generation model by integrating a variational quantum circuit (VQC) into the conventional GAN-based MolGAN framework. Specifically, they replaced the noise generator, molecular generator, and molecular discriminator components of MolGAN with VQCs. Validation using a quantum simulator suggested that higher-performance models could be achieved, particularly in the discriminator part. Additionally, Vakili et al.⁴⁷⁵ employed the quantum circuit born machine⁴⁷⁶ to generate initial samples for molecule generation. Their experiments utilized IBM quantum device, and by combining it with a long short-term memory (LSTM) network, they successfully generated KRAS inhibitors.

Since candidate molecular structures can lead to a combinatorial explosion, attempts have also been made to address molecule generation using quantum annealing or Ising machines specialized for combinatorial optimization problems. Tu čs et al.⁴⁷⁷ used a D-Wave quantum annealer to generate peptides. By combining a black-box optimization method based on quantum annealing, which is known as FMQA (Factorization Machine with Quantum Annealing),⁴⁷⁸ with a binary variational autoencoder, they successfully synthesized peptides with antimicrobial activity and nonhemolytic properties. Furthermore, Mao et al.⁴⁷⁹ applied a similar approach using GPU-based Ising machines for molecule generation. Gao et al.⁴⁸⁰ employed an IBM quantum device using the QAOA (Quantum Approximate Optimization Algorithm)⁴⁸¹ framework to optimize molecular structures. These efforts indicate that molecule generation techniques leveraging quantum computing are expected to continue advancing in the future.

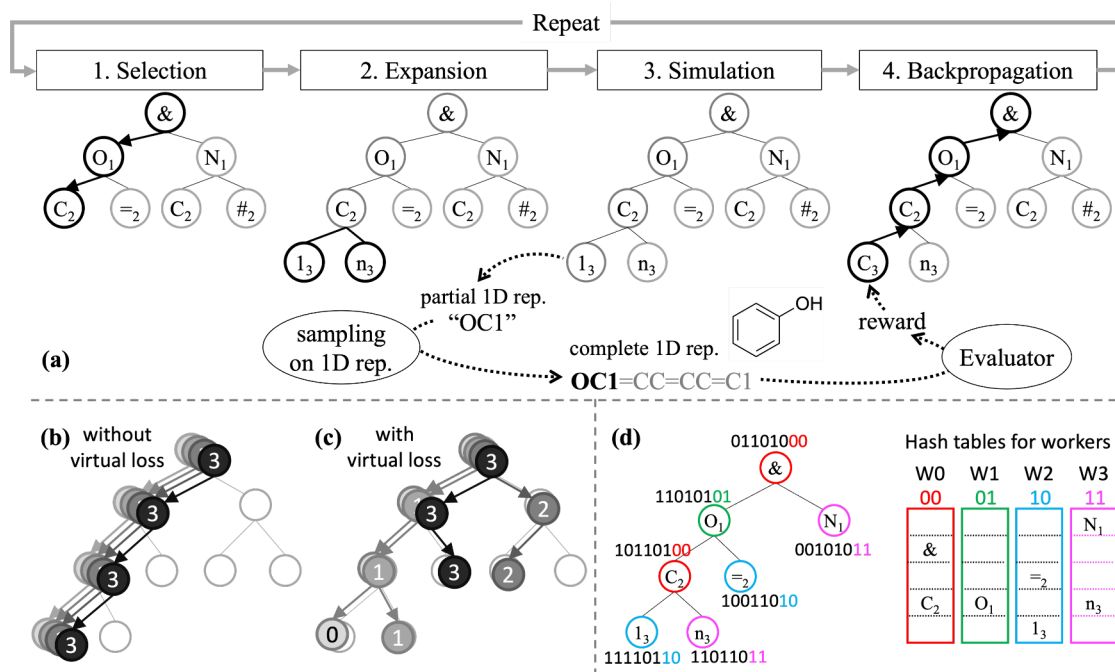


Figure 8. One cycle of Monte Carlo Tree Search (MCTS) and its parallelization with Virtual Loss and a Hash-Driven method. (a) Four steps of MCTS applied to a 1D representation (SMILES). 1. Selection: The most promising node is identified. 2. Expansion: A new node is added to the tree if the visit count of the current node exceeds a predefined threshold. 3. Simulation (Rollout): A sequence is generated by sampling, and the associated reward is computed. 4. Backpropagation: The reward information is propagated back through the path to update the nodes. (b) When four iterations of MCTS are run in parallel, they normally follow the same high-value path, eliminating any parallel speed-up. (c) Virtual loss assigns a temporary penalty to nodes that an iteration has just selected; this lowers their score for the other iterations and spreads the search over different subtrees. (d) In hash-driven parallelization, a hash function assigns every node to a *home* worker. Each worker stores the statistics of the assigned nodes in a local hash table, so the search graph is automatically partitioned and the computational load is balanced across workers.

4.10. Monte Carlo Tree Search and Large-Scale Parallelization

In this section, we survey and classify various optimization and search algorithms. Because some methods naturally pair with specific generative strategies, whereas others do not, declaring a single method as universally superior is challenging. Nevertheless, when we focus solely on the search performance—disregarding the choice of generative strategy—the MCTS family of algorithms consistently demonstrates the strongest exploratory power. In particular, for large-scale decision-making problems involving vast and complex search spaces, few techniques rival the empirical performance of MCTS in large, combinatorial spaces.

MCTS was originally developed for the game of Go and gained widespread recognition through AlphaGo, the first program to defeat a human world champion.^{401,413,482} Although MCTS is classified as a graph-search algorithm, its theoretical foundations overlap significantly with RL, and it often excels when integrated with the evaluation systems (e.g., simulations and ML models) that guide the search. MCTS has a broad range of applications, with researchers quickly adapting it to domains far beyond games.^{412,483} Several studies have successfully applied MCTS to molecular design.^{414,422,427,428}

From the standpoint of molecular design, MCTS presents two principal challenges.

- **Nontrivial graph representation:** Defining an appropriate graph representation of the search space is not straightforward.

- **Limited parallelism:** Compared with simpler optimization techniques, MCTS is more challenging to parallelize efficiently.

In the following subsections, we introduce solutions proposed in previous studies for each of these challenges.

Defining the Search Space for MCTS. The ease with which a search space can be defined for MCTS depends strongly on its underlying molecular representation. In 1D representation formats such as SMILES or SELFIES, a tree can be generated almost mechanically by appending characters one at a time to a growing string. By contrast, in 2D (e.g., graphs) or 3D (e.g., 3D coordinates) representations, one must first decide what constitutes a node or an edge, a design choice that is far from trivial. Existing studies have addressed the companion problem of evaluating internal nodes by employing stochastic *rollouts*: partially specified strings or graphs are completed using a sequence (or graph-generating) model and then assessed using an external property predictor.^{414,484}

1D Representations (SMILES, SELFIES, InChI). Because 1D molecular representations are literal character strings, they allow direct hierarchical decomposition of the chemical space.^{194,195,199} Figure 8(a) shows a standard four-step MCTS cycle applied to SMILES. Each node encodes a partial SMILES string, and each edge adds one character (atom or bond) to the prefix. For example, the leftmost node at depth 4, labeled l_3 , represents all strings beginning with “OC1.” Starting from the root (all valid SMILES), successive edges naturally define nested subsets of molecules, yielding an exhaustive yet well-organized search space that supports efficient exploration and exploitation.

2D and 3D Representations. 2D and 3D molecular representations carry far richer structural details but require

complicated tree construction. A child node may correspond to the addition of an atom, bond, ring closure, or even a full conformer, and each choice alters both the branching factor and search speed. Recent studies, such as RationaleRL⁴²⁷ and VGAE-MCTS,⁴²⁸ have shown that graph-based MCTS is feasible and can yield competitive results. However, every study introduces its own customized node/edge definitions. As no consensus exists on a graph-based scheme, each new project must revisit this design step, which complicates the application of MCTS in 2D/3D molecular discovery.

Evaluating Internal Nodes: Sampling-Based Rollouts.

Scoring a node that represents *many* molecules is equally challenging. String-based approaches typically sample one or more complete SMILES/SELFIES strings from the current prefix using an autoregressive model, estimate them using an evaluator, and backpropagate the obtained reward, mirroring the simulation phase of classical MCTS.^{414,422} This procedure updates the value estimates of the nodes and steers subsequent simulations toward more promising regions in the chemical space.

Parallelization Requirements and Parallel Graph Search. The performance gains of modern computing stem more from increased core counts and accelerators (e.g., GPUs) than from individual CPU core speeds. To leverage these resources, the algorithms must be explicitly designed for parallel execution. Although many methods reviewed in this section are amenable to straightforward parallelization, MCTS remains particularly difficult to parallelize efficiently.

Figure 8(a) shows a single MCTS iteration. Because each iteration relies on the statistics collected from previous iterations, naive parallel execution is infeasible. In practice, the effective parallelization of graph-search algorithms almost always requires structural modification of the underlying algorithm.

Several techniques have been developed to enable large-scale parallel MCTS. The two notably effective strategies are as follows:

- *Virtual Loss* temporarily penalizes paths already being explored, discouraging other threads/processes from selecting the same node.
- *Hash-driven parallelization* assigns ownership of nodes by a hash function and distributes the nodes among the hash table of each worker (e.g., process).

Figure 8(b) and (c) illustrate the concept of parallelizing MCTS using virtual loss by running four iterations in parallel. Without virtual loss, all four iterations explored the same paths, making parallelization ineffective. Virtual loss introduces a small penalty to the paths currently under exploration by other iterations, thereby encouraging diversification across threads or processes. In the two-player game settings where MCTS first succeeded, rewards are derived from game outcomes; therefore, adding virtual loss acts as an explicit penalty during selection. Although this method is widely credited to early Go implementations, the earliest published descriptions appear in a study on tree parallelization for MCTS.⁴⁸⁵

Figure 8(d) illustrates hash-driven parallelization. A hash function is defined over graph nodes and determines a node's *home* worker, which stores its entry in a local hash table. In the example, four workers each maintain a hash table, distributing ownership of the graph nodes among them.

This scheme introduces inter-worker messages when an edge crosses worker boundaries; however, provided the hash is

sufficiently uniform, it yields excellent load balance, which is historically the most challenging part of a parallel graph search. Hash-based work distribution has been shown to scale A* search up to thousands of cores.^{486,487} Similarly, this technique has been successfully applied to MCTS, achieving massive parallelization.⁴¹⁵ In molecular design, a recent implementation achieved near-linear speed-ups on approximately 1,000 workers.⁴¹⁶

5. EVALUATION STRATEGIES

To design desirable molecules using optimization and search strategies, estimating the properties of the generated molecules and evaluating their overall quality are important. Here, we introduce basic strategies for estimating molecular properties, evaluating synthetic accessibility, which is essential for practical molecular design, and integrating methods for estimated target properties based on a multiobjective optimization framework. Finally, we present benchmark tasks for evaluating the performance of molecular generative AIs.

5.1. Property Estimation of Generated Molecules

Various methods are available for estimating the properties of molecules designed using molecular generative AI systems. Representative approaches include experimental synthesis and measurement, molecular simulations, and ML-based predictions, such as QSAR and QSPR. However, owing to methodological constraints and limitations, not all methods are applicable to the generated molecules. Additionally, these estimation methods often involve trade-offs between accuracy and cost. Therefore, when evaluating the generated molecules, selecting the most suitable method by considering applicability, accuracy, and (computational) cost is essential.

Experimental syntheses and measurements provide the most reliable estimates of molecular properties; however, they are generally time-consuming and costly. Although examples of molecular design combined with Bayesian optimization have been reported,^{387,388} common practice is to design molecules computationally, select several promising candidates, and then measure their properties experimentally. Molecular property predictions based on ML models^{114,115,305,488} are extremely fast. However, the prediction accuracy strongly depends on factors such as the amount and diversity of training data, as well as the architecture of the prediction model.^{489–491} It is important to note that ML-based predictions are generally accurate in regions where the training data are dense (interpolation); however, their accuracy often decreases significantly when applied to sparse or out-of-distribution regions (extrapolation). Molecular simulations often provide accurate estimations and enable the exploration of new structures, although they generally require higher computational costs than ML-based predictions. Various simulation methods are available, such as docking simulations,^{458–461} molecular dynamics simulations,^{492–494} and quantum chemical calculations.^{109,125,128–130} Although the scope, accuracy, and cost vary depending on the method used, simulations based on physicochemical backgrounds theoretically have the potential to estimate properties even for molecules different from known ones. Therefore, simulations are a key tool in discovering new molecules outside of the training data set. Additionally, to prevent human intervention from becoming a bottleneck in the discovery process, it is crucial to automate the integration of these simulations into the molecular design closed loop. Several frameworks have been developed to support this integration, including the atomistic

simulation environment (ASE),⁴⁹⁵ QChASM (Quantum Chemistry Automation and Structure Manipulation),⁴⁹⁶ and QCForever.^{497,498}

5.2. Synthetic Accessibility Evaluation

As for methods estimating synthetic feasibility, the synthetic accessibility score (SAScore)⁴⁹⁹ has long been widely used owing to its simplicity and ease of computation. SAScore is a statistical method that estimates synthetic accessibility based on the frequency of molecular fragments occurring in one million molecules in the PubChem database,⁵⁰⁰ with penalties applied to rare fragments and complex structural features, yielding an ease-of-synthesis rating from 1 (easy) to 10 (hard). Following the same concept as the SAScore, the ChEMBL-likeness score (CLScore)⁵⁰¹ and natural product (NP)-likeness score⁵⁰² have also been proposed, where the data sets used to calculate fragment scores were replaced with the ChEMBL⁵⁰³ database and NP data sets,^{503,504} respectively. As an extension of the SAScore, BR-SAScore⁵⁰⁵ has been developed, in which fragment scoring is modified to incorporate reaction templates and building block information used by retrosynthetic AI (Retro*⁴⁷⁰ in this study), enabling the score to better reflect the feasibility of synthetic route planning using such AI models. More recently, various ML-based prediction models have been proposed to predict synthetic feasibility-related characteristics. The synthetic complexity score (SCScore)⁵⁰⁶ is a neural network trained on reactant–product pairs from 12 million single-step reactions in Reaxys and is designed to predict a synthetic complexity index that correlates with the estimated number of reaction steps required to synthesize a given molecule, outputting values from 1 (simple) to 5 (complex). Synthetic Bayesian accessibility (SYBA)⁵⁰⁷ is based on a Bernoulli naïve Bayes classifier trained to distinguish easily synthesizable molecules obtained from the ZINC15 database and hard-to-synthesize molecules generated by the Nonpher algorithm. For the same task as SYBA, graph attention-based assessments of synthetic accessibility (GASA)⁵⁰⁸ and DeepSA⁵⁰⁹ have been proposed as classifiers based on graph neural network and chemical language model architectures, respectively. Retrosynthetic accessibility score (RAScore)⁵¹⁰ is an ML-based prediction model, including neural networks and XGBoost, trained on 200,000 molecules from the ChEMBL database to predict whether the synthetic route for a given molecule can be solved using synthetic route-planning AI (specifically AiZynthfinder in this study). Unlike conventional synthesizability scores, MolSkill⁵¹¹ employs a preference-learning approach to model the intuitive judgments of medicinal chemists by training based on pairwise comparisons of molecules. The trained model is expected to capture medicinal chemistry intuitions that are difficult to encode using traditional synthetic feasibility-related scores. Focused synthesizability score (FSscore)⁵¹² is a model based on the concepts of SCScore and MolSkill. FSscore is first pretrained on a reaction data set and can be further fine-tuned with a small amount of human-labeled data to fit specific chemical spaces. The choice of appropriate synthetic feasibility scores is highly case-dependent and varies with the molecule-generation model used and the scope of the target chemical space. No single metric can be generally considered optimal, as each metric has specific strengths and limitations; thus, users should choose metrics with an understanding of their strengths and limitations. ML-based prediction models require caution because their performance degrades for molecules outside the applicability domains. To mitigate this issue, statistical metrics, such as SAScore, may

offer more stable performance, whereas fine-tuned models, such as FSscore, can be useful when the chemical space of interest to be explored using molecule generation models is relatively well anticipated. Another approach involves using the results of synthetic route planning AIs^{470,513–519} directly as rewards. However, this often results in high computational costs. Recent methods, such as Saturn,⁵²⁰ aim to alleviate this issue by employing more sample-efficient language-based generative models, such as Mamba.⁵²¹

5.3. Single- and Multi-Objective Evaluation

In practical molecular design, multiple target properties often need to be optimized. For example, in drug discovery, it is necessary to simultaneously consider not only the activity against a target but also properties such as selectivity,⁴²¹ synthetic accessibility,⁵²² and ADMET properties.^{488,523} Therefore, a multiobjective optimization (MOO) becomes critically important.^{524,525}

A representative approach for MOO is Pareto optimization.⁵²⁶ This method seeks a set of nondominated solutions, where no solution is inferior to all other solutions for all target properties. Pareto optimization is particularly useful when designing molecules that must balance multiple—often conflicting—properties. Typically, they are used when designing molecules that improve multiple properties through a trade-off relationship. Numerous applications of molecular design and exploration based on Pareto optimization have been reported.^{217,369,422,527–529} One of its advantages is that it only requires specifying which properties to optimize and the desired direction of improvement without the need to predefine importance weights for each property.

On the other hand, the scalarization approach is also widely used.^{243,376,421,426,454,530} In this approach, the estimated values of multiple targets are combined into a single score. Typically, based on the predefined weights assigned to each property, weighted sums,^{243,426,454,530} weighted averages,³⁷⁶ weighted products, or geometric means^{421,423,524} are used. Although scalarization requires predefined weights, it offers the practical advantage of enabling optimization using a single scalar value. A weighted product or geometric mean is effective in situations where molecules are desired that achieve certain values across all properties, as the total score becomes zero if any individual property is zero.⁵²⁴

Furthermore, curiosity-driven exploration⁵³¹ has attracted attention as a method for designing novel and diverse molecules without setting clear optimization targets. For example, in random goal exploration, a single point is selected in the property space and molecules with properties predicted to be similar to that point are chosen. For example, in random goal exploration, a point in the property space is randomly selected and molecules estimated to exhibit properties close to this point are selected or generated.^{532,533} This has facilitated the discovery and generation of diverse molecules.⁵³⁴ In addition, BLOX (BoundLess Objective-free eXploration),⁵³⁵ a method that actively explores molecules that deviate from trends in the property space without limiting the search space, has also been proposed and evaluated for finding exceptional photofunctional molecules.

When estimating a molecular property using ML-based prediction models, the problem of extrapolation becomes significant and unintended optimization, often referred to as reward hacking,^{536,537} may occur during the molecular design cycle.^{538–540} As a solution, generating predictions in the

Table 1. Molecule Identification Using Molecular Generative AI: Input Spectra, Output, Algorithm, and Molecular Representation Are Tabulated

Input Spectra	Output	Algorithm/AI	Representation	Reference
MS	whole molecule	Spec2Mol	SMILES	576
	whole molecule	LSTM		
MS	whole molecule	MSNovelist	SMILES	577
Rotational	functional group	MLP	SMILES	578
	formula			
IR, MS	functional group	autoencoder+MLP	SMARTS	579
IR, composition	whole molecule	transformer	SMILES	580
¹ H NMR	whole molecule	NMR-TS	SMILES	418
Composition				
¹³ C-/ ¹ H NMR	whole molecule	CNN	Graph	581
Composition				
¹³ C NMR, IR	whole molecule	DeepSPInN	Graph	582,583
Composition				
¹³ C-/ ¹ H NMR, IR	whole molecule	Spectro	SELFIES	584

extrapolation region can be avoided by utilizing the applicability domain,^{541–545} which defines reliable regions of predictions, and prediction uncertainty.^{489–491} Even in multiobjective optimization, a method for adjusting the reliability of multiple prediction models has been proposed.⁴²³ As described above, various strategies exist for integrating multiple evaluation criteria into the molecular design, and selecting an appropriate approach according to a specific problem is crucial.

5.4. Benchmark

Several benchmarking platforms have been developed to evaluate molecular generative AIs, including GuacaMol,³⁵⁴ Molecular Sets (MOSES),³⁵⁵ Therapeutics Data Commons (TDC),⁵⁴⁶ practical molecular optimization (PMO),³⁶³ and TARTARUS.⁵⁴⁷ These platforms provide standardized metrics and data sets for assessing the performance of distribution learning tasks and goal-directed optimizations, facilitating fair comparisons across different AI systems. Distribution learning tasks evaluate how well AI reproduces the distribution of a given data set, whereas goal-directed optimization tasks assess their ability to generate molecules that satisfy predefined objectives. Broadly, GuacaMol and MOSES provide general-purpose evaluation tasks for benchmarking molecule-generation models. GuacaMol supports both distribution learning tasks and goal-directed optimization, whereas MOSES focuses on distribution learning tasks. Newer platforms such as TDC, PMO, and TARTARUS offer evaluation tasks that are more closely aligned with real-world applications of molecule-generation models. RediscMol was proposed as a benchmark for evaluating models in fine-tuning scenarios, providing data sets for both pertaining and fine-tuning.⁵⁴⁸ A structure-based drug design (SBDD) Benchmark has been developed⁵⁴⁹ to assess the performance of 3D structure-based molecular generative AIs. Although most benchmark tasks are primarily designed for drug design, several platforms, such as TARTARUS and the polymer-focused benchmark introduced by Yue et al.,⁵⁵⁰ have been extended to materials science. By leveraging these benchmarks, we can systematically compare molecular generative AIs from various perspectives related to molecule-generation tasks. However, these benchmark results should be interpreted with caution. Each molecule-generation model typically involves its own set of hyperparameters, which should be tuned for each specific task to achieve optimal performance. Because the benchmarks are generally performed under fixed configurations, the reported

results may not fully capture the optimal performance of each model. Further information on each benchmark category can be found in original publications and other review articles.^{185,354,355,363,546,547}

6. APPLICATION OF GENERATIVE MODEL FOR MOLECULE IDENTIFICATION

In chemistry, understanding the world at the atomistic level is an inverse problem, as illustrated in Figure 2. Directly observing the molecular structure represented by the 3D atom arrangement is impossible, as is the case with software that visualizes simulation results. The various quantized energy levels of a molecule absorb and emit specific photons as the molecule transitions between energy levels. The spectra obtained by observing these photons contains information on various molecular structures. Currently, this method is also important for observing and proving molecular dynamics.^{551–553} On the other hand, the mass spectrum obtained by ionizing an object and detecting its mass-to-charge ratio (m/z) provides not only information on the mass of the observed object but also a wealth of information on its structure.⁵⁵⁴ Although spectroscopy/spectrometry has revolutionized modern chemistry, these spectra impose typical inverse problems on chemists who decipher molecular structures based on empirical knowledge. Because chemists know from experience that atomic/fractional species appear at certain positions in the spectrum, they can construct molecules in their minds. This may be easy for simple molecules; however, as the size of the molecule increases, this becomes tedious. Total synthesis can sometimes reveal errors in spectral identification.¹⁵ Although computer-assisted structure determination has been attempted for many years,⁵⁵⁵ ML/AI has opened a new phase in this field.⁵⁵⁶ ML prediction models for UV–Vis,^{557–559} vibrational,^{560–567} and NMR^{568–573} spectra of a given molecule have been developed as an alternative tool for the forward problem in Figure 2. These spectral prediction model tools will help build in silico databases that are useful for screening strategies for molecule identification by mass spectrometry (MS).^{574,575} However, they could not retrieve molecules from databases. By replacing molecular properties with molecular spectra, a generative AI system for molecular design becomes a molecule identifier. Generative AI is revolutionizing molecule identification, which is an inherent inverse problem in chemistry. Here, we present several research results, summar-

ized in Table 1, that challenge the identification of molecules from spectrum/multispectra data.

Litsa et al.⁵⁷⁶ attempted to predict molecular structures using tandem MS (MS/MS) with generative AI. They constructed a generative AI (Spec2Mol) composed of CNN as the encoder and GRU as the decoder, which were trained using the NIST MS/MS spectrum library and 135 million molecules curated from the PubChem^{155,500}/ZINC⁵⁸⁵ data set. Although they evaluated the molecules generated by their generative AI from the viewpoint of molecular similarity and physicochemical properties, a perfect match was difficult to obtain using MS/MS spectra alone. This is mainly because the MS spectra are highly instrument-dependent, and the fragmentation of molecules is also highly stochastic. Strave et al.⁵⁷⁷ also applied a similar strategy for the identification of molecules from a MS/MS spectrum. Their generative AI, MSNovelist, is an encoder-decoder RNN model trained with a data set curated from HMDB,⁵⁸⁶ COCONUT,⁵⁸⁷ and DSSTox.⁵⁸⁸ MSNovelist receives mass spectrum data via SIRIUS,⁵⁸⁹ which can transfer a mass spectrum to a molecular formula and structural fingerprint. Hence, MSNovelist is not a rigorous identifier from a mass spectrum based on the accuracy of SIRIUS, which successfully determines the molecular formula for small compounds with a low error rate (<10%). Molecule identification by MS, which induces stochastic molecular fragmentation and is highly instrument-dependent, will require the construction of a large database and integration of its search technology with deep learning in the future. MS has the disadvantage of destroying the observed objects; however, spectroscopic analysis, which does not destroy objects, is a major measurement method. Similarly, attempts have been made to automate the identification of molecules from spectra obtained using various spectroscopic methods with generative AI.

The degrees of freedom of a molecule include translational, rotational, vibrational, and electronic states. The width between their quantized energy states increases in this order. The energy level of rotation corresponds to the energy band of microwaves. The detection of microwaves emitted by the rotational energy of molecules is indispensable for identifying interstellar molecules.⁵⁹⁰ The identification of unknown molecules using microwave spectroscopy is often performed by comparing the experimentally obtained spectral parameters with those obtained by a human's intuitive calculation of the electronic structure of the molecule. MacCarthy et al.⁵⁷⁸ introduced generative AI to abolish the need for human operators. They trained multilayer perceptrons (MLPs) to identify functional groups, LSTM for a SMILES string¹⁹⁴ with computational data of the molecules curated from PubChem^{155,500} and systematically generated by the Open Molecule Generator (OMG).⁵⁹¹ Although they successfully generated the functional groups and chemical formulas that the molecules were thought to possess, they did not succeed in identifying the entire molecule in SMILES.¹⁹⁴

The internal degree of a molecule corresponds to the molecular vibration quantized to the energetic region corresponding to infrared light (4000–400 cm⁻¹).²³ Infrared (IR) spectroscopy is a relatively inexpensive and useful technique for functional group identification, as the functional groups of molecules absorb near 1500 cm⁻¹.^{21,23} In recent years, IR spectroscopy has been widely used to observe molecular dynamics using pulsed lasers.⁵⁹² Fine et al.⁵⁷⁹ developed generative AI that can identify functional groups from an IR spectrum coupled with an MS spectrum. They constructed a

generative AI consisting of MLP and an autoencoder to remove redundant information and noise from the data and trained it using 7393 compounds and the corresponding standardized spectra obtained from the NIST web book.⁵⁹³ Without specialized knowledge, the generative AI accurately identified the functional groups of the molecule from the IR and MS spectra. In chemistry laboratories, identifying molecules using multiple spectral data sets is common practice. Therefore, such attempts are conducive to reducing the workload of chemists. In addition, attempts have been made to extract potential information contained in the spectra.

Vibrational spectra are typically used to identify only the moieties of a molecule. As vibrational spectra contain information about the internal degrees of freedom of a molecule, they inherently include information about the entire molecular structure. Hence, it is possible to obtain the molecular structure from the vibrational spectrum. Although vibrational spectra are cumbersome and impossible for humans to obtain the whole molecular structure, AI may be able to capture the properties of the molecular structure from the vibrational spectrum. Alberts et al.⁵⁸⁰ demonstrated the potential use of AI for extracting information from an IR spectrum. They constructed a generative AI based on an autoregressive encoder–decoder transformer trained using computational IR spectra and the chemical formulas of over 600,000 molecules curated from PubChem.^{155,500} Based on this generative AI, the AI was fine-tuned using experimental data obtained from the NIST database.⁵⁹⁴ Their AI showed a top-10% precision of 69 correct predictions for 79% of the experimental data. These results indicate that the introduction of generative AI to chemistry decisively expands the usage of spectral data. This is also true for NMR spectroscopy, which has traditionally been a powerful method for molecule identification.

NMR spectra were obtained by placing molecules containing magnetic nuclei (¹H, ¹³C, and ¹⁹F) in an electromagnetic field and observing the frequency at which the resonance occurred. Because the nuclear magnetic moment is shielded by the electronic magnetic field induced by an external magnetic field, the same element resonates at varying frequencies. Therefore, NMR spectroscopy can capture the state of nuclei owing to the influence of the local electronic structure and is one of the most important structural characterization techniques in chemistry. Empirical knowledge on the chemical shift of a certain nucleus, which depends on the surroundings, has also accumulated. On the basis of this knowledge, chemists have deduced the structures of these molecules. Zhang et al.⁴¹⁸ attempted to automate the identification of molecules using the ¹H NMR spectrum. They developed NMR-TS based on ChemTS,⁴¹⁴ in which MCTS and RNN were fused to build and complete the molecules represented in SMILES.¹⁹⁴ NMR-TS attempts to find a suitable molecule whose NMR spectrum agrees with the input NMR spectrum by comparing the input spectrum with the DFT-simulated NMR spectra of the generated molecules within the Wasserstein distance^{595,596} within the composition constraint. They successfully identified molecules with 66% accuracy. NMR-TS uses QC calculations to evaluate the generated molecules. However, this method is computationally expensive. Furthermore, ChemTS can only search in the direction of growing molecules represented in SMILES, which is inefficient owing to the nature of ChemTS.

Although the closed-loop search shown in Figure 1(c) is considered excellent for searching unknown molecules, it is inefficient. Constructing a generative AI in Figure 1(b) can

Table 2. AI-Generated Molecules with Demonstrated Activity in Biological Assays: Target Proteins, Algorithm/AI, and Molecular Representations Are Tabulated^a

Target	Index in Figure 9	Algorithm/AI	Representation	Reference
RXR	(a)	RNN	SMILES	612,613
LXR α	(b)	LSTM	SMILES	614
JAK1	(c)	GraphGMVAE	Graph	615
TNIK	(d)	Chemistry42	–	616
Nurr1	(e)	CLM	SMILES	617
PI3K γ	(f)	CLM	SMILES	618
SARS-CoV-2	(g)	CogMol	SMILES	619
PARP1/2	(h)	CMD-GEN	3D	353
A. baumannii	(i)	SyntheMol	Fragment	473
DDR1	(j)	GENTRL	SMILES	256
DDR1	(k)	REINVENT	SMILES	620
H ⁺ , K ⁺ -ATPase	(l)	Deep Quartet	SMILES	621
CXCR4	(m)	SAC	Fragment	622
JAK3	(n)	AAE	SMILES	623
MEK1 and mTOR	(o)	POLYGON	SMILES	624
ADORA2A and PDE4D	(p)	DualFASMI-FRA	Fragment	625
		DualTransORGAN	SMILES	

^aStructural formula of molecules manifested by generative AI designs are summarized in Figure 9.

predict molecular structures with a one-shot using NMR spectra as the input, as described in the identification of molecules from the MS and IR spectra, is possible. Huang et al.⁵⁸¹ constructed a generative AI based on a CNN that received ¹³C NMR, ¹H NMR, and chemical formulas as inputs and predicted the probabilities of 957 substructures that might be included. They were trained using a CNN with computational spectral data of molecules curated from the GDB13.¹⁵⁰ By assembling the probable substructures, their generative AI successfully predicted the molecular structures with a top-10 accuracy of more than 95% for the experimental data. Using only one NMR spectrum, the generative AI identified the molecules with more than 80% accuracy. However, using multiple spectra improves accuracy. Indeed, using multiple spectra for molecular identification is common practice in chemistry laboratories.

NMR spectra provide information about magnetic nuclei, and vibrational spectra provide information about the bonds that link atoms. Hence, their combination is chemically plausible for obtaining the entire molecular structure. Devata et al.^{582,583} developed a generative AI called DeepSPIN, which identifies molecules by receiving IR and ¹³C NMR spectra as inputs. They trained a message passing neural network (MPNN) using computational IR and ¹³C NMR spectra to featurize a molecule. DeepSPIN builds molecules using a rule-based MCTS with MPNN and shows a high top-1 accuracy of 86.47%. Their model was based on the QM9-NMR data set,⁵⁹⁷ which included molecules with up to nine heavy atoms. The accuracy of molecule identification was reduced to 64.63% for molecules with nine heavy atoms.

The molecule identification assisted by generative AI introduced above receives numerical data of spectra as input. The attempt of using images of spectra has also been conducted. Chako et al.⁵⁸⁴ tried to identify a molecule from images of IR and ¹³C-¹H NMR spectra. However, signals of NMR spectra are too spiky for the accurate detection. Instead, they utilized the attribution text of NMR spectra. Consequently, they built generative AI, Spectro, that can identify molecules from multiple-modal input, combining an encoder consisted of MLP for functional group identification from IR spectrum image (j-IR-vis) and LLM for attribution texts of NMR spectra,

and a decoder (RNN). They trained their generative AI with NMR data (obtained from NMRium⁵⁹⁸) and IR image (obtained from NIST⁵⁹³) of 6,833 molecules. Spectro has successfully identified molecules with an accuracy of 93%.

7. EXPERIMENTAL VALIDATION OF AI-DESIGNED MOLECULES

In this section, we present reports on experimental validation of AI-designed molecules for their function as drugs or materials. To date, few molecules have been verified as molecular materials. Therefore, we present examples of functional molecules designed by generative AI with the aim of drug development.

7.1. Drug Discovery

Historically, drug discovery has relied on empirical strategies and systematic experimental screening, which are often driven by human intuition and constrained by limited insights into molecular mechanisms. With advances in atomistic-level understanding, the rational design of ligand molecules targeting disease-related proteins has become a central paradigm. Consequently, molecular simulations are expected to play a crucial role in evaluating protein–ligand binding affinity,^{599–607} although they still fall short of capturing the full complexity of in vivo systems. Conventional theory-driven and empirical approaches have inherent limitations when dealing with biological systems characterized by numerous interdependent features and nonlinear interactions. In this context, AI techniques are expected to provide an alternative paradigm for identifying latent patterns and relationships in complex biological data. AI-based drug design is still in its early stages, and the increasing sophistication of AI models underscores the fundamental difficulty of accurately modeling and predicting biological systems.

Generative AI based on probabilistic models has been successful in identifying molecules based on their spectra. This is because the correlation between molecular structures and spectra has been elucidated to the degree formulated in some literature; thus, the main challenge lies in handling the structural complexity of molecules. However, molecular design is not simple. In contrast to spectral data obtained under ideal conditions, drug development must account for multifactorial biological processes in vivo. Therefore, small-molecule drug discovery has emerged as a primary focus for leveraging generative AI to accelerate early stage R&D.^{608–611} The development of molecular generative AIs extends beyond merely proposing molecular structures: it involves the development of a huge generative AI platform that verifies the feasibility of the generated molecules during or after

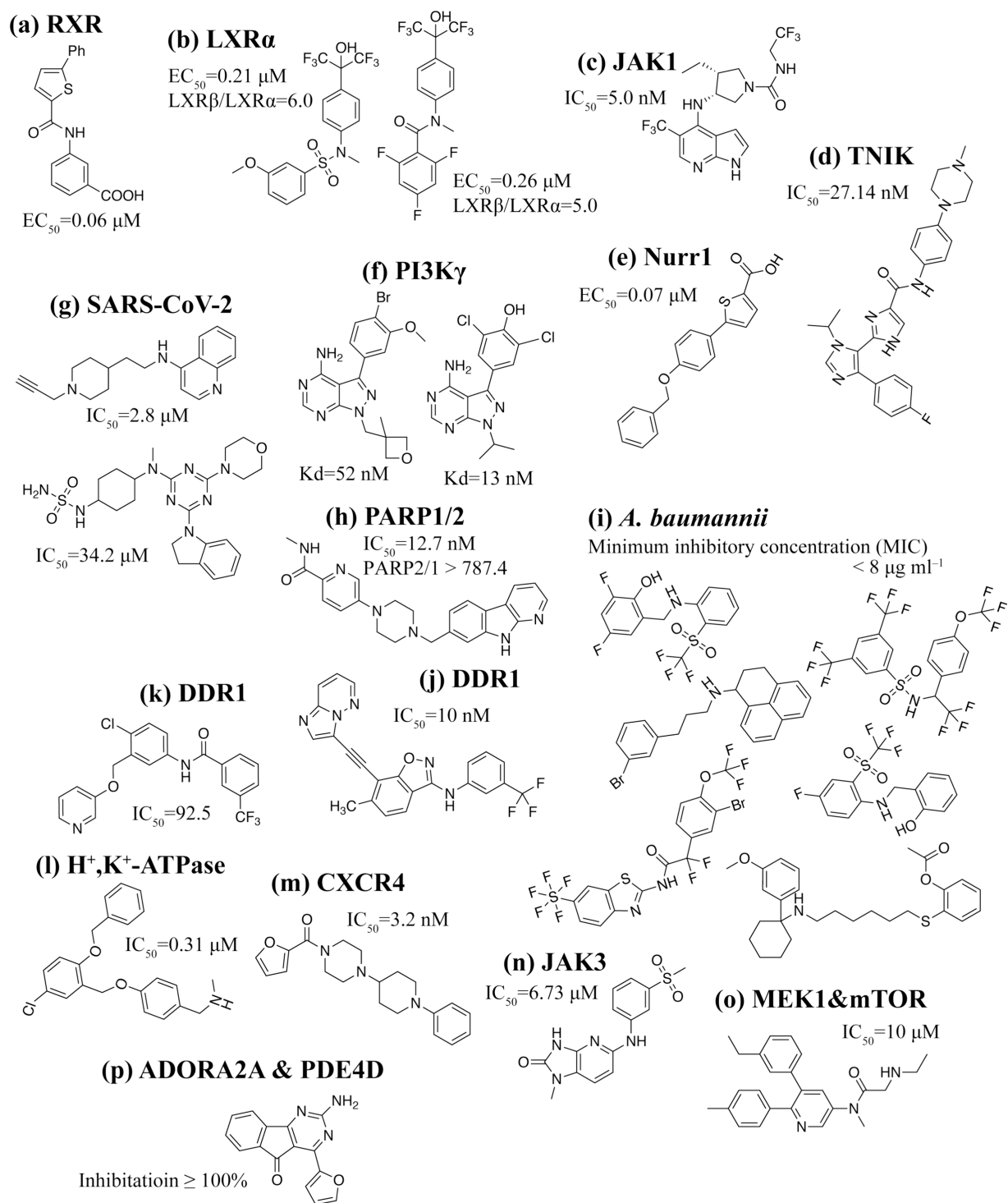


Figure 9. Synthesized AI-designed molecules for experimentally validated biological activity with their target proteins and performances.

generation, including retrosynthesis analysis and the validation of protein–ligand binding affinity via simulation, as well as the assessment of drug-likeness-related properties. However, even with such a large and complex generative AI platform, drug development remains difficult. In Table 2, we present studies on AI-designed molecules that exhibit biological activity in experimental settings.

Retinoid X receptors (RXRs) are involved in retinoid signaling pathways.⁶²⁶ Abnormal RXR signaling affects neuronal stress and neuroinflammatory networks in several neuropathological conditions. Merk et al.^{612,613} designed natural product-mimicking modulators for RXRs using generative AI based on an RNN.^{220,627} Their approach involved fine-tuning a pretrained model using 25 fatty acid mimetics⁶¹²

or six natural products⁶¹³ known to possess biological activity, which served as templates. They first trained RNN with 541,555 bioactive compounds (K_D , K_i , IC/EC_{50} values $< 1 \mu M$) extracted from the ChEMBL22 database. Subsequently, the model was fine-tuned using template molecules. In a previous study,⁶¹² 1000 molecules were generated using the fine-tuned model, and subsequently five molecules were selected based on the predicted modulatory effects using SPiDER⁶²⁸ and synthetic feasibility considerations. Among these, two were identified as RXR agonists, and the compound shown in Figure 9(a) shows $EC_{50} = 0.06 \pm 0.02 \mu M$ for RXR γ . In a later study,⁶¹³ the authors investigated the effect of the number of templates used for fine-tuning on the performance of de novo molecule generation. Fine-tuning

with a single template yielded low structural diversity and a high rate of invalid outputs (75%) among the 1000 generated molecules. By contrast, using three or six templates reduced the invalid rate to 21% and increased the proportion of unique molecules to 49% and 74%, respectively. These findings indicate that a data set for fine-tuning should be carefully designed to generate isofunctional new chemical entities effectively.

Recently, the integration of generative AIs and robotic synthesis systems has been explored to accelerate drug discovery. Grisoni et al.⁶¹⁴ developed a modular molecular design pipeline comprising generative AI based on LSTM,²²⁰ synthesizability analysis based on retrosynthesis route prediction, and a microfluidic platform. They trained their generative AI with commercially available synthetic compounds assembled from Asinex,⁶²⁹ ChemBridge screening compound collection,⁶³⁰ Enamine Advanced and HTS collection,⁶³¹ and Specs screening compounds,⁶³² and fine-tuned the model with the 40 agonist molecules for LXR α that are effective with a median effective concentration (EC_{50}) < 0.5 μ M. During generation, the molecules were sampled based on the predicted LXR activity, scaffold diversity, and pharmacophore similarity to the compounds used for fine-tuning. Through the search for the retrosynthesis route analysis based on 17 virtual reaction schemes, they finally succeeded in synthesizing 25 molecules using their automatic synthesis module. In vitro bioactive validation of these molecules showed that two molecules among them had significant potency and selectivity between LXR α and LXR β . One (Figure 9(b)) shows that EC_{50} = 0.21 μ M to LXR α and high selectivity to LXR β with 6.0, and the other shows that EC_{50} = 0.26 μ M to LXR α and high selectivity to LXR β with 5.0. Although structurally similar molecules often exhibit similar pharmacological effects,¹⁷⁷ even small changes in molecular scaffolds can lead to dramatic differences in biological activity. Yu et al.⁶¹⁵ developed a generative scaffold hopping model (GraphGMVAE) for designing Janus kinase 1 (JAK1) inhibitors. GraphGMVAE consists of an encoder comprising dual message-passing neural networks (DMPNN), a hidden space based on a Gaussian mixture distribution for each cluster of scaffolds, and a GRU-based decoder. They trained GraphGMVAE using ZINC database compounds (6.8 million) and 210 K bioactive molecules (pIC_{50} > 5) selected from ChEMBL database. The generated molecules were filtered using their own prioritization pipelines based on molecular filters, pharmacophore alignment, and molecular simulations, resulting in seven molecules for the experimental assay. All seven compounds showed inhibitory activity against JAK1, and one molecule shown in Figure 9(c) demonstrated significant activity (IC_{50} = 5.0 nM) in comparison with the reference compound (IC_{50} = 45 nM).

A notable example of AI-driven drug discovery has advanced to clinical trials. Using a generative AI-driven approach, Ren et al.⁶¹⁶ designed INS018_055, a small-molecule inhibitor targeting TRAF2- and NCK-interacting kinase (TNIK) for antifibrotic therapy. They designed TNIK inhibitors using the Chemistry42 platform⁶³³ with the assumption that the TNIK binding site is an ATP-binding site. Compounds initially selected from the designed inhibitors, based on their synthetic accessibility, novelty, and drug-like properties, were synthesized and screened. Although the early hits exhibited suboptimal ADME profiles, lead optimization resulted in INS018_055 in Figure 9(d), which demonstrated efficacy in ameliorating fibrotic processes in vitro and in vivo across several disease models.

Ballarotto et al.⁶¹⁷ used a chemical language model (CLM) to design agonists targeting orphan nuclear receptor-related 1 (Nurr1), which has been implicated as a therapeutic target for neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease. They pretrained the CLM on 365 K molecules selected from ChEMBL and fine-tuned it using a previously reported Nurr1 agonist with a demonstrated affinity (EC_{50} = 0.4 μ M, K_d = 0.7 μ M).⁶³⁴ Using beam search sampling,⁶³⁵ they generated molecules and selected six for synthesis and experimental validation based on their high sampling frequency and structural similarity to the affinity Nurr1 agonist. Among these, one compound (Figure 9(e)) exhibited potent agonistic activity, with an EC_{50} of 0.07 μ M and K_d of 0.14 μ M. Moret et al.⁶¹⁸ developed a hybrid CLM for de novo design of ligands targeting phosphoinositide 3-kinase gamma (PI3K γ). To enable the CLM to learn the chemical features relevant to

drug-like molecules, the CLM was first pretrained on 839,674 compounds extracted from the U.S. patent database. Subsequently, transfer learning was employed to tailor the CLM specifically to the chemical space of PI3K γ inhibitors. For fine-tuning, 46 PI3K γ inhibitors with IC_{50} $\leq$ 100 nM were selected from the Drug Target Commons (DTC) database. Using the fine-tuned CLM, molecular generation was performed to fit PI3K γ inhibitors, followed by the synthesis of two candidate molecules. These molecules exhibited a high affinity for PI3K γ , with measured K_d values of 52 nM and 13 nM, respectively (see Figure 9(f)), validating the ability of the model to generate bioactive ligands within the desired potency range.

Chenthamarakshan et al.⁶¹⁹ constructed a generative AI framework called CogMol, which is based on VAE trained on a large data set of chemical SMILES representations of molecules, protein sequences, and available protein–ligand binding affinities. They used CogMol to generate small-molecule inhibitors targeting the receptor-binding domain (RBD) of the spike (S) protein and the main protease (M^{pro}) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The generated molecules were prioritized based on docking simulations, physicochemical properties, and retrosynthetic feasibility using the IBM RXN platform.^{517,636} Four compounds were selected and synthesized for each target. Experimental validation demonstrated a 50% success rate for both RBD and M^{pro} . In particular, one of the RBD-targeting compounds shown in Figure 9(g) exhibited potent neutralization activity, with an IC_{50} of 2.8 μ M, whereas the compound against M^{pro} showed inhibitory activity, with an IC_{50} of 34.2 μ M.

Most of the aforementioned generative AIs focus solely on optimizing molecular properties, such as protein-binding affinity and drug likeness, without explicitly considering the three-dimensional (3D) structures of the generated molecules. In contrast, structure-based generative AIs that directly generate 3D conformations tailored to fit within the binding pocket of a target protein have emerged. A notable example is the work of Zou et al.,³⁵³ who proposed coarse-grained and multidimensional data-driven molecular generation (CMD-GEN), which is a hierarchical framework that combines a diffusion model with a transformer encoder–decoder architecture. CMD-GEN effectively generated pocket-specific and drug-like 3D molecules by decomposing the molecular generation process into three stages: pharmacophore point sampling, molecular structure generation, and conformational alignment. They selected eight candidate molecules generated by CMD-GEN targeting PARP1/2, synthesized them, and evaluated their inhibitory activities in vitro. Among them, one compound shown in Figure 9(h) potently inhibited PARP1, with an IC_{50} of 12.7 nM, and demonstrated remarkable selectivity over PARP2, with a 787.4-fold difference in potency.

Generative AIs that rely on probabilistic distributions learned from training data tend to generate molecules close to the high-density regions of the training data distribution.⁶¹³ Although some strategies, such as a VAE equipped with scaffold-hopping mechanisms, have been proposed to improve structural diversity,⁶¹⁵ the generated molecules are inherently constrained by the underlying distribution of the training data set. To address this limitation, RL has been incorporated into generative frameworks to guide molecular generation toward user-defined objectives. By optimizing a reward function, such as binding affinity, synthesizability, or specific properties, RL enables the exploration of the chemical space beyond what is represented in the original training data distribution. Swanson et al.⁴⁷³ designed antibiotics targeting *Acinetobacter baumannii* (*A. baumannii*) using SyntheMol, a generative AI that assembles molecules from predefined building blocks via MCTS. The molecule-generation process was guided by three different property prediction models to estimate the antibacterial activity. To ensure synthetic feasibility and practicality, SyntheMol searched within the restricted subset of Enamine Readily Accessible (REAL) space, consisting of 29.6 billion molecules built from 138,085 building blocks using 13 well-validated chemical reactions. After applying filters for structural diversity, novelty, and high predicted activity, 58 candidate molecules were synthesized and experimentally validated. Among these, six molecules exhibited potent growth inhibition against *A. baumannii* as single agents, as shown in Figure 9(i). Discoidin domain receptor 1 (DDR1), a collagen receptor

with tyrosine kinase activity, is predominantly expressed in the epithelial cells and has been implicated in the regulation of tissue fibrosis.⁶³⁷ Zhavoronkov et al.²⁵⁶ reported the rapid discovery of DDR1 inhibitors using a generative AI called GENTRL, which was pretrained on the ZINC database and further fine-tuned with known DDR1 kinase inhibitors. GENTRL integrates the VAE, tensor decomposition, and REINFORCE²³⁴ algorithm for reinforcement learning, thereby enabling goal-directed molecule generation. Among the 30,000 molecules generated using GENTRL, six were selected for synthesis based on their synthetic feasibility. Two of the six compounds exhibited potent DDR1 inhibition in an enzymatic kinase assay, with IC_{50} values of 10 and 21 nM (one of them is shown in Figure 9(j)). One of the two lead compounds was further evaluated in a rodent model and showed favorable pharmacokinetic properties. This work demonstrates that de novo molecular design using generative AI can lead to experimentally validated lead candidates within less than two months, highlighting its potential to dramatically accelerate early stage drug discovery. Yoshimori et al.⁶²⁰ employed REINVENT^{407,408} to design DDR1 inhibitors that satisfied a pharmacophore model comprising eight features, as defined by LigandScout 4.4.⁶³⁸ Nine compounds were selected from the generated molecules for synthesis, based on pharmacophore fit, predicted binding affinity, and synthetic feasibility. Among them, one compound, which has also been reported in PubChem,^{155,500} exhibited potent inhibitory activity against DDR1 with $IC_{50} = 92.5$, as shown in Figure 9(k). Abe et al.⁶²¹ developed a generative AI framework called Deep Quartet, which integrates REINVENT⁴⁰⁷ with pharmacophore-based scoring. Deep Quartet was trained using SMILES-formatted molecules from ChEMBL⁶³⁹ and was optimized to generate molecules that satisfied the spatial pharmacophore constraints derived from the crystal structures of the gastric proton pump, H^+ , K^+ -ATPase. From the generated molecules, they synthesized a compound (DQ-06) that exhibited potent inhibitory activity ($IC_{50} = 0.7 \pm 0.04 \mu M$), outperforming the reference compound SCH28080 ($IC_{50} = 1.97 \pm 0.12 \mu M$). Furthermore, high-resolution cryo-electron microscopy (cryo-EM) revealed the binding pose of DQ-06, which enabled structure-guided optimization. This led to the development of the chloro derivative DQ-18 with further improved potency ($IC_{50} = 0.31 \pm 0.02 \mu M$), as shown in Figure 9(l). Jiang et al.⁶²² proposed an RL-based framework for de novo molecular design with an emphasis on synthetic feasibility, applying the model to design active molecules targeting the inflammation-related receptor CXCR4 chemokine receptor type 4 (CXCR4). The model employs the soft actor–critic (SAC) algorithm,⁶⁴⁰ which constructs molecules through the iterative selection of reaction templates and building blocks. To simultaneously optimize binding affinity and synthetic feasibility, the reward function incorporates both molecular docking scores against CXCR4, calculated using AutoDock Vina,⁴⁶⁰ and synthetic accessibility estimated through forward synthesis planning formulated as a Markov decision process (MDP). To optimize the affinity to the receptor and the synthesizability of the generated molecules, they built a reward that evaluated the affinity to CXCR4 through docking simulation using AutoDock Vina⁴⁶⁰ and synthesis planning using the Markov decision process (MDP). After five independent molecular-generation campaigns, 20 candidate molecules were selected for synthesis and biological evaluation. Among them, one compound exhibited a high binding affinity for CXCR4, with $IC_{50} = 3.2 \pm 0.2$ nM, as shown in Figure 9(m). Polykovskiy et al.⁶²³ proposed an entangled conditional adversarial autoencoder (ECAAE), which is an improved variant of the adversarial autoencoder (AAE), for de novo molecular generation conditioned on specific properties. To design selective JAK3 inhibitors, the ECAAE model was trained using known JAK2 and JAK3 inhibitors extracted from the ChEMBL database.⁶⁴¹ Among the 300,000 generated molecules, candidates were prioritized through a multistage filtering process, including a series of filters, molecular docking, prediction of side effects and chemical properties, and molecular dynamics simulations, which resulted in 100 molecules. One compound shown in Figure 9(n) was synthesized and evaluated in vitro; it exhibited potent activity against JAK3 ($IC_{50} = 6.73 \mu M$), whereas it remained inactive against JAK2 ($IC_{50} = 17.58$ mM).

Most small-molecule drugs interact with multiple biological targets, often resulting in off-target effects that can be either beneficial or detrimental. While such polypharmacological activity is considered a liability owing to potential side effects, interest in intentional multitarget modulation that can improve therapeutic efficacy, reduce adverse effects, and delay the emergence of resistance is growing. The rational design of single compounds capable of modulating multiple targets remains a challenge, mainly owing to the difficulty in achieving balanced potency, selectivity, and favorable pharmacokinetic properties across structurally and functionally diverse proteins. Munson et al.⁶²⁴ introduced POLYGON, a framework integrating VAE with RL, to enable the de novo design of dual-target inhibitors, and validated it on the MEK1 and mTOR kinase pair. POLYGON was trained on a large set of molecules from ChEMBL⁵⁰³ and target-specific affinity data from BindingDB and Pharos.^{642–644} From POLYGON-generated top-100 molecules, 32 compounds were synthesized and experimentally validated. One compound exhibited dual inhibitory activity, with IC_{50} values of approximately $10 \mu M$ for both MEK1 and mTOR, as shown in Figure 9(o). Yasuda et al.⁶²⁵ developed a dual-target generative AI framework consisting of DualTransORGAN⁶⁴⁵ and DualFASMI-FRA⁶⁴⁶ and validated it by designing compounds targeting both adenosine A2a receptor (ADRA2A) and phosphodiesterase 4D (PDE4D). DualFASMI-FRA is a fragment-based structure generator that assembles substructures using a GA and provides a conservative yet pragmatic approach to molecular generation. By contrast, Dual-TransORGAN is a generative model based on GAN with a transformer encoder and decoder architecture that captures the semantic features of molecules and generates chemically plausible structures via RL with a stochastic policy gradient. Target-specific activity was predicted using QSAR models constructed with random forest regressors trained using the ChEMBL database.⁵⁰³ Ten compounds were selected from the AI-generated molecules for synthesis and in vitro experimental validation. Among these, one compound exhibited potent inhibitory activity against both ADRA2A and PDE4D, along with high target selectivity for a panel of 39 human proteins, as shown in Figure 9(p).

The lack of consensus on an appropriate distance metric between molecules in the chemical space makes it difficult to determine whether the generated molecules fall within or outside the applicability domain of the training data. Therefore, the RL algorithm is indispensable for exploring chemical space. However, when a reward function is based on ML models, such as QSPR/QSAR to avoid an excessive computational load in the simulation, exploration is still limited to the applicability domain covered by the training data, even if the probability distribution of the features in the training data is changed by an RL algorithm. Hence, some approaches integrate on-demand simulations, such as molecular docking simulations, into the scoring process to overcome the constraints of the training data. Korablyov et al.⁶⁴⁷ developed a molecular design framework called LAMBDZERO, which integrates generative active learning with RL for the de novo design of small molecules with potential binding affinity to a target protein while minimizing reliance on computationally expensive simulations. In LAMBDZERO, molecules are constructed by sequentially assembling building blocks. To avoid exhaustive docking simulations, the framework employed an approximate Bayesian surrogate model to predict the docking scores of the generated molecules. Synthetic accessibility was assessed using RetroGNN, and drug-likeness was quantified using the QED score, both of which were incorporated as soft constraints in the exploration process. High-scoring molecules predicted by the surrogate model were periodically evaluated using docking simulations with AutoDock Vina,⁴⁶⁰ and the resulting scores were used to update the surrogate model. Through this iterative strategy, LAMBDZERO significantly reduced the number of docking simulation calls required to reach high-affinity candidates. LAMBDZERO was used to design small-molecule inhibitors of soluble epoxide hydrolase 2 (sEH). Thirty-five analogs based on LAMBDZERO-generated molecules were synthesized and evaluated by medicinal chemists, of which 24 exhibited in vitro inhibitory activity. Notably, a scaffold containing N-(4,6-di(pyrrolidin-1-yl)quinazolin-2-yl)-N-methylbenzamide was reported to be novel and not found among previously known sEH inhibitors. An RL algorithm that works in

tandem with on-demand simulation tools is indispensable for material development.

7.2. Material Discovery

Many of the restrictions surrounding drug development have been eliminated by developing functional molecules for materials. In addition, while affinity (pharmacophore) for proteins is the dominant target in the development of drug molecules, functional molecules as materials vary widely according to their functions, that is, light-relevant, electronic, and other physical properties. Hence, in contrast to drug molecules, materials that exploit the diversity of molecules must be developed. While the difficulty in drug development is the molecular design under severe constraints, a large search space also makes the design of material molecules difficult.

Generative AI is expected to be helpful in chemical space exploration, including molecular diversity. Although some proof-of-concept studies, including experimental validation, have been conducted for the AI-driven development of photofunctional molecules,^{417,419} electret,⁶⁴⁸ and CO₂ capture,⁶⁴⁹ examples of real synthesis and verification of new molecules using generative AI are scarce in the materials field. It is easy to imagine that generative AI based on a probabilistic model falls outside the training data. RL-based generative AI is necessary. Successful examples of validating generative AI-designed molecules are summarized in Table 3.

Table 3. Real Functional Molecules Designed by Generative AI: Target Properties, Algorithm/AI, and Molecular Representations Are Tabulated^a

Target	Index in Figure 10	Algorithm/AI	Representation	Reference
Light emission	(a)	ChemTS	SMILES	650
Low-viscosity	(b)	REINVENT	SMILES	651
Light absorption	(c)	eGEGl	SMILES	652
Light emission	(d)	Bayesian optimization	Fragment	653

^aStructural formulas of molecules designed by generative AI are summarized in Figure 10.

Derivatives of a molecular skeleton that exhibit certain properties show relatively continuous changes with the addition or exchange of substituents. Therefore, the search for derivatives is efficient for chemists from the viewpoint of cost, and the various effects of substituents have been summarized in typical textbooks.^{21,22} Therefore, designing molecules with certain properties from scratch is an unthinkable task. Among the functional molecules, luminescent molecules are the most typical. The representative motifs of substances used as fluorescent probes or for bioimaging are almost entirely settled: rhodamine,^{68–70} GFP,^{654,655} and coumarin.⁶⁰ If we continue searching for these basic skeletons, the strategy for obtaining correlations with the structure may be correct. However, a dramatic change in performance could also be expected when the basic skeletons are changed. Therefore, as demonstrated in drug discovery, the hopping search method was introduced to dramatically change the basic skeleton.⁶¹⁵ By contrast, dramatic changes in the skeleton make establishing correlations difficult; therefore, using a prediction model for scoring is inappropriate. If the goal is to develop a new skeleton, a simulation based on a deductive approach is the only way to proceed. Sumita et al.⁶⁵⁰ used this strategy to achieve honesty. Using massive parallel search-enabled ChemTS,⁴¹⁴ designed fluorescent molecules with on-demand QC computations at the B3LYP/3–21G level. They experimentally validated eight product molecules, including one new molecule, and showed that they could design fluorescent molecules from scratch in 75% of cases. The AI-designed fluorescent molecule with ab initio QC is shown in Figure 10 (a), with a fluorescence length of 550 nm and a quantum yield of 0.007.

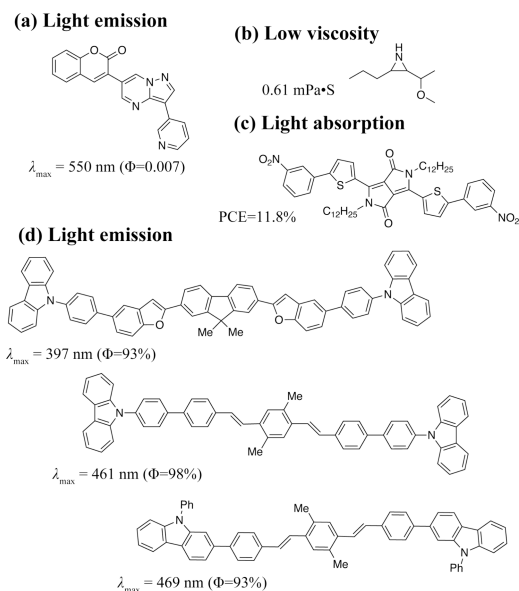


Figure 10. Synthesized AI-designed molecules for experimental validation as materials with target properties and values indicated.

There are good examples that show what happens when prediction models are used as a molecule evaluation system. Matsuzawa et al.⁶⁵¹ designed low-viscosity molecules that can significantly improve the efficiency of electrochemical devices, such as batteries and capacitors. They selected 10,000 molecules from the GDB-17 chemical structure database and calculated the viscosity using molecular dynamics (MD) simulations. In addition, they calculated the HOMO energy of each molecule using DFT, using it as an indicator of molecular stability. Based on these data, a machine learning model was constructed. Using these models, molecular generation was performed via an RL approach (REINVENT)⁴⁰⁸ using SMILES strings¹⁹⁴ to identify chemical structures that minimize the viscosity while maintaining a sufficiently low HOMO level. Among the generated molecules, a novel compound, 2-(1-methoxyethyl)-3-propylaziridine (Figure 10(b)), was synthesized, and its viscosity was evaluated. The experimentally measured viscosity of this molecule was 0.61 mPa · s, which was found to be in close agreement with the MD-simulated viscosity of 0.58 mPa · s. Although the HOMO and viscosity distributions of the generated molecules clearly indicate a failure to extrapolate from the training data, their results show that interpolated searches within the training data can also find good candidates.

Similarly, Tan et al.⁶⁵² designed nonfullerene acceptors (NFAs) in organic photovoltaics (OPVs). They aimed to discover new molecules with better light-absorption properties and a lower cost than fullerene-based materials using generative AI with the evaluation of prediction model. They employed an evolutionary algorithm-based generative AI, eGEGl,^{368,656} trained using a database of NFAs⁶⁵⁷ to generate molecules by combining fragments. According to the modified Scherbar model, the open-circuit voltage (V_{OC}) can be predicted based on the gap between the donor's HOMO and acceptor's LUMO and V_{OC} results in power conversion efficiency (PCE) estimation. Therefore, they also created a directed message-passing neural network (D-MPNN) prediction model¹¹⁵ that can predict the gap from a SMILES string. Their D-MPNN was trained using NFA molecules curated from an OPV.^{657,658} Seven molecules were synthesized from the generated candidates. All of them were experimentally confirmed to exhibit a high PCE, indicating strong potential for application in high-efficiency organic photovoltaic devices. The molecule with the highest PCE (11.8%) is shown in Figure 10(c).

The closed loop shown in Figure 3(c) is incorporated into the cycle in Figure 1 as a formulation process for developing real-world materials. Ideally, the introduction of robots into the measurement process (including synthesis) would create a self-driving laboratory, as shown schematically in Figure 11. Automated molecular synthesis technology

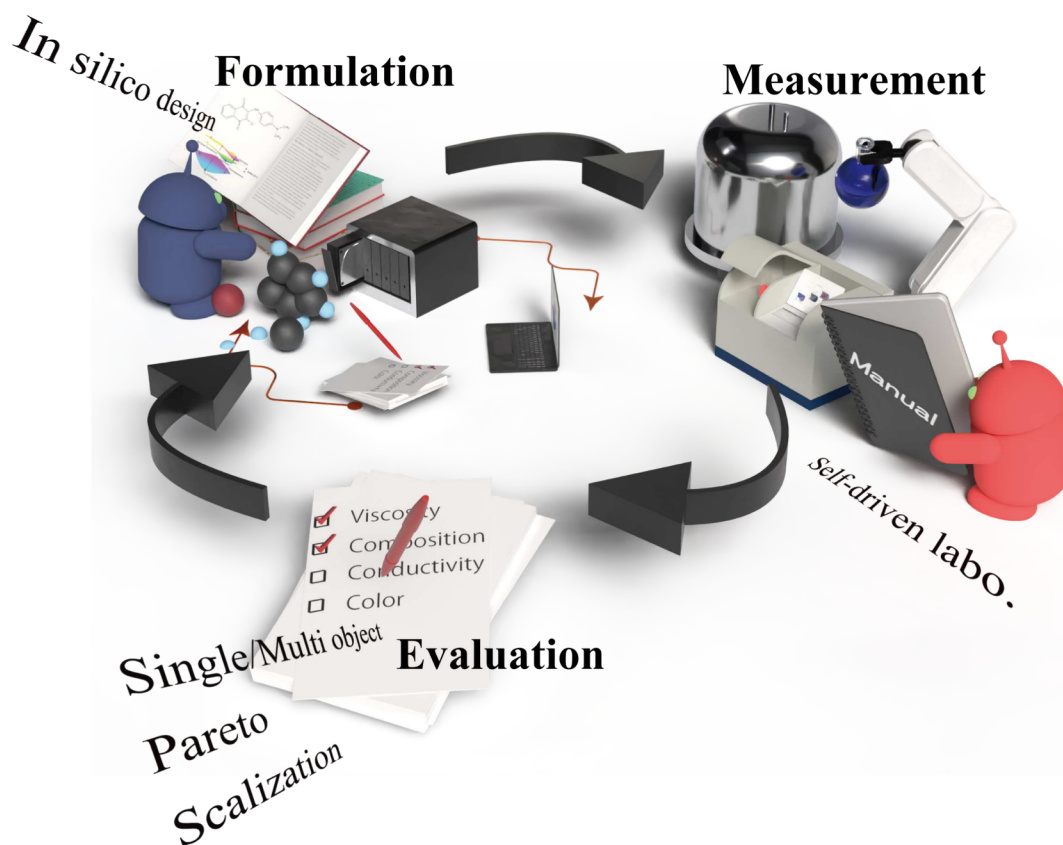


Figure 11. Iterative cycle for inverse problem whose formulation in Figure 1 is replaced with in silico molecular design, as shown in Figure 3(c). The formulation and measurement processes are facilitated by molecular generative AI and AI-assisted robots, respectively. Recent ML/AI technology based on probabilistic models presented in Section 3 will greatly help to digitalize the contents or concepts written in textbooks or manuals. Incorporating robots into the measurement process will result in a self-driving laboratory.

gies and self-driving laboratories for molecular design are also advancing. Recent advancements in self-driving laboratory technologies have been summarized elsewhere.^{659,660} Hence, this review does not address many of these issues. However, an important example of the ultimate form of AI-implemented self-driving laboratories was presented. Strieth-Kalthoff et al.⁶⁵³ developed organic laser emitters using a cycle of AI-based formulations and real-world synthesis and evaluation, as shown in Figure 1. They assembled a formulation system that combined a GP prediction model using molecular descriptors, including physical quantities from DFT calculations, with Bayesian optimization. The search area is limited to chemicals that can be divided into core, bridge, and cap substructures that can be joined via the Suzuki–Miyaura reaction. Specifically, a fragment library capable of generating over 150,000 candidate molecules was created using 32 caps, 30 bridges, and 161 core building blocks. The molecules selected by the formulation system were synthesized using a combination of various automated experimental setups around the world, such as Chemputer,⁶⁶¹ Chemspeed, and an end-to-end automated screening platform.⁶⁶² Finally, 21 novel molecules were synthesized. Owing to the purification limit, three of them, shown in Figure 10 (d), were confirmed to exhibit more than $1.5\text{--}1.9\ \mu\text{J cm}^{-2}$ amplified spontaneous emission (ASE). Thus, automated synthesis has become an important technology for molecular design, and various molecules are expected to be developed with the advancement of automated synthesis technologies.

8. OUTLOOK

We provided an overview of the generative AI algorithms used for designing functional molecules and introduced several applications for molecular identification, which represents a major class of inverse problems in chemistry. While our review

focused on generative AI for small-molecule design in Section 7, the algorithms introduced in Section 3 are equally applicable to the design of supramolecules,^{663,664} metal complexes,^{665–667} polymers,⁶⁶⁸ proteins,^{669,670} and inorganic materials.⁶⁷¹ Although the development of molecular generative AI based on probabilistic models will continue to evolve, their limitations are simultaneously becoming clear, and overblown expectations should be avoided.

Vast number of molecular generative AI systems have been developed, depending on the type and availability of modules for molecular representation, generation, optimization, and evaluation. While the appropriate combination of these modules is crucial, molecule-generation modules based on probabilistic models are fundamentally designed to generate molecules similar to those in the training data, as discussed in Section 3. Thus, although it may be possible to find molecules that dramatically improve performance by chance, the algorithms are not designed to actively find them. This fact naturally limits the use of molecular generative AI.

Spectroscopy is based on electromagnetic radiation with well-defined detection ranges, and its underlying mechanisms are largely understood. This makes spectral analysis well suited for generative AI. Most molecular generative AI models that perform identification based on molecular spectra require additional information such as composition/chemical formulas as inputs. This indicates that molecular identification is difficult without narrowing the search space in advance. As compositional analysis is mostly destructive to the target molecule, it is best to avoid this, if possible. Therefore, narrowing the search

space using multiple spectra would be more ideal. In chemistry laboratories, molecules are not typically identified from a single spectrum but by combining multiple spectra. Therefore, the development of a generative AI that identifies molecules from multiple spectra will likely become mainstream. The use of molecular generative AI built from the learning data of spectra and molecules through simulations has become standard. Although there are examples of successful molecular identification using experimental spectra as inputs for generative AI based on simulation data, building training data sets from experimental data would be better. Therefore, developing methodological techniques that feed experimental data as training data is important.⁶⁷² Applications to supramolecules such as proteins and polymers are also expected.^{673,674} Ultimately, molecular identification via generative AI is a pivotal theme in chemistry and a cornerstone technology for the realization of self-driving chemistry laboratories.

Molecular generative AI embodies the long-awaited automation of molecular design. Since its emergence, generative AI has garnered significant attention in drug discovery, with several reports already documenting the experimental validation and clinical trials of AI-designed molecules. Some studies specifically highlight that generative AI can substantially shorten drug development timelines.⁶¹⁶ However, generative AI has yet to succeed in developing molecules that significantly outperform conventional drugs. This stagnation is likely because molecular generation is constrained by the scope of training data, synthetic accessibility, and various pharmacological restrictions. Nevertheless, even when exploration is confined to the range of existing training data, AI can uncover superior molecules that researchers may have previously overlooked. This underscores the fact that the quality and curation of training data are the primary drivers of successful molecular development.

Beyond simple probabilistic models, some methods employ reinforcement learning (RL) to update model distributions based on the scores of generated molecules. However, this approach offers limited benefit if the scores are provided by conventional prediction models. As reported by Matsuzawa et al.,⁶⁵¹ RL does not enable extrapolation beyond the inherent limits of the scoring function. Consequently, even when integrating RL into generative AI, the ability to identify novel molecules remains constrained by the training data used to build the underlying prediction models. It is fundamentally paradoxical to search for statistical outliers by relying on statistical prediction models.

We also introduced approaches that employ simulations rather than prediction models to circumvent the limited applicability domain of model-based molecular scoring. In these cases, the search space is governed by the training data used for molecule generation and the inherent capabilities of the molecular simulation. Among various simulations, quantum chemistry (QC) calculations offer the best generalization performance for generative AI. A key reason for the success of QC is its ability to numerically represent chemical phenomena by calculating internal molecular energy relative to nuclear variations. Indeed, fluorescent molecules have been generated *de novo* by integrating the advantages of QC calculations with RL-based generative AI. If a simulation clarifies the underlying mechanisms of a phenomenon and its processes can be automated, then molecular generative AI can, in principle, design any molecule. While some studies define inverse molecular design as the task of correlating structures with target properties,⁶⁶⁴ we argue that explicitly determining structure–

property correlations is unnecessary when screening-based methods can address the inverse problem directly; this task can be effectively delegated to computers.

The integration of molecular generative AI and QC calculations represents a milestone in the evolution of the theoretical chemistry. Once confined to explaining and speculating on chemical phenomena, theoretical calculations are now poised to lead the next revolution in materials development. However, simulations based on electronic structure theory are often computationally expensive. The hierarchical evaluation systems proposed by Korablyov et al.,⁶⁴⁷ which combine prediction models and simulations, offer a potential solution to this cost. Nevertheless, the high computational demand of QC still directly increases the time required to evaluate each candidate, hindering extensive searches. Therefore, parallel exploration becomes crucial for efficiently navigating chemical space under demanding evaluation regimes. In Section 4.10, we present virtual loss- and hash-driven techniques that enable massive parallelization of Monte Carlo Tree Search (MCTS). Both techniques are expected to significantly expand the searchable area, even when utilizing computationally intensive evaluation systems like QC.

Simulation appears to be an essential tool for overcoming the limitations of the applicability domain in molecular generative AI. Consequently, the automation of simulations, along with improvements in their accuracy and speed, has become increasingly important. However, automating simulations poses significant challenges. This involves more than just streamlining tasks that traditionally require human intervention and chemical intuition, such as conformational searching,⁶⁷⁵ managing initial structure dependency, or defining convergence criteria in molecular dynamics. Furthermore, current molecular simulations are far from perfect, and many theoretical aspects require fundamental refinement. To address this, data-driven improvements in computational efficiency and accuracy have been implemented for specific density functional theory (DFT) calculations.^{676–680} Research has also focused on developing force fields and potentials where QC calculations are regarded as the ground truth, further highlighting the necessity of ensuring QC accuracy. Ultimately, although future improvements in computer performance are anticipated, the development of simulation algorithms and the clarification of underlying phenomena remain vital for enriching the feature space required by molecular generative AI.

Given the emergence of molecular generative AI for molecular identification and design, its integration with robotics is poised to realize self-driving chemistry laboratories, as schematically illustrated in Figure 11. The development of fluorescent molecules by Strieth-Kalthoff et al.⁶⁵³ and pharmaceuticals by Grisoni et al.⁶¹⁴ clearly signals this trend. However, perhaps due to constraints in synthetic accessibility and training data—both of which significantly restrict the searchable space—molecules that fundamentally advance the frontiers of science have yet to be developed. Current chemical reaction technologies encompass only about 10^8 known chemicals.¹⁵⁵ Even if AI were capable of performing molecular design utilizing every one of these reactions, our existing technology remains insufficient for truly exploring the vast expanse of chemical space. Consequently, for the time being, self-driving chemistry laboratories should be regarded as a field of technological development, distinct from pure scientific discovery. This “detour” into technological optimization is worthwhile, provided we do not lose sight of the underlying science. Ultimately, basic research aimed at

transcending current synthesis paradigms—which rely heavily on repetitive solvent-based reactions and fractionation—is essential for the next leap in molecular discovery. While human leadership is essential for the true expansion of technology, the integration and supplementation of established knowledge offer a powerful catalyst for advancement, where probabilistic generative models can play a pivotal role.

The previously fragmented fields of chemistry are rapidly integrating as informatics becomes deeply embedded in chemical research. Consequently, the breadth and depth of knowledge required for researchers are becoming increasingly diverse and vast. Furthermore, chemistry is a visually oriented discipline and even a single subfield can exceed the capacity of an individual to master. Large Language Models (LLMs) currently offer a promising means of managing specialized knowledge alongside visual information. As schematically shown in Figure 11, LLMs could effectively utilize instrument manuals to guide experimental procedures. Furthermore, the integration of LLMs with emerging algorithmic approaches allows for the prediction of multiple features and properties of chemical substances and address inverse problems. Such a well-trained computational framework has the potential to provide a comprehensive platform, moving beyond the development of individual predictive or generative models for specific molecular properties. By enhancing their reliability through integration with external databases, LLMs could be incorporated into self-driving chemistry laboratories, contributing to diverse areas such as molecular design, automated synthesis, image recognition, and safety data management.

However, it is important to remember that probabilistic models, including LLMs, are fundamentally dependent on their training data. Currently, chemical data tends to be concentrated on derivatives of high-performing molecules, and training on such biased data sets will inevitably yield biased AI responses. Therefore, researchers must carefully curate databases tailored to their specific objectives. Furthermore, as previously mentioned, one must not forget to validate the predictions made by these models, because these models often appear more authoritative than their actual capabilities warrant. Predicting the future solely based on the past is inherently impossible; thus, critical judgments that shape the future must remain the responsibility of human researchers. Even if today's generative AI had existed when Teflon and various plastics were first introduced, it likely would not have predicted the unprecedented environmental impacts described in "Silent Spring."⁶⁸¹ To utilize AI safely and effectively, we must strive even harder to cultivate a higher degree of imagination and creativity.

AUTHOR INFORMATION

Corresponding Author

Koji Tsuda — Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, Chuo-ku, Tokyo 103-0027, Japan; Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan; MolNavi LLC, Yokohama, Kanagawa 220-0072, Japan; orcid.org/0000-0002-4288-1606; Email: tsuda@k.u-tokyo.ac.jp

Authors

Masato Sumita — RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, Chuo-ku, Tokyo

103-0027, Japan; MolNavi LLC, Yokohama, Kanagawa 220-0072, Japan; orcid.org/0000-0002-3506-1028

Shoichi Ishida — Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa 230-0045, Japan; MolNavi LLC, Yokohama, Kanagawa 220-0072, Japan; orcid.org/0000-0002-5638-3579

Kazuki Yoshizoe — Research Institute for Information Technology, Kyushu University, Nishi-ku, Fukuoka 819-0395, Japan; MolNavi LLC, Yokohama, Kanagawa 220-0072, Japan; orcid.org/0000-0001-7715-1868

Ryo Tamura — Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan; Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan; RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, Chuo-ku, Tokyo 103-0027, Japan; orcid.org/0000-0002-0349-358X

Kei Terayama — Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa 230-0045, Japan; RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, Chuo-ku, Tokyo 103-0027, Japan; MDX Research Center for Element Strategy and School of Life Science and Technology, Institute of Science Tokyo, Yokohama, Kanagawa 226-8501, Japan; MolNavi LLC, Yokohama, Kanagawa 220-0072, Japan; orcid.org/0000-0003-3914-248X

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.chemrev.5c00689>

Notes

The authors declare no competing financial interest.

Biographies

Masato Sumita is Research Scientist at the Center for Advanced Intelligence Project, RIKEN. He received a Doctor of Science in 2008 from the University of Tsukuba. His research interests include quantum chemistry and application of machine learning for chemistry and materials sciences.

Shoichi Ishida is Visiting Lecturer at the Graduate School of Medical Life Science, Yokohama City University. He earned his doctoral degree in Pharmaceutical Sciences from Kyoto University in 2021. His research interests include the application of cheminformatics and machine learning techniques to drug discovery and materials science, as well as the development of related tools and software.

Kazuki Yoshizoe received his Ph.D. in computer science from the University of Tokyo, where he specialized in graph-search algorithms. He has served as a researcher at the Japan Science and Technology Agency (JST), an assistant professor at the University of Tokyo, and a unit leader at RIKEN. He is currently a professor at Kyushu University, where he heads the High-Performance Computing Division of the Research Institute for Information Technology. His research interests include graph-search algorithms, parallel computing, and machine learning, with applications to games, combinatorial optimization, decision-making problems, and molecular design.

Ryo Tamura is Team Leader at Center for Basic Research on Materials (CBRM) in National Institute for Materials Science (NIMS). He obtained his Ph.D. in 2012 from Graduate School of Science, the University of Tokyo. He has worked in the field of materials science, in particular, his research interests focus on materials informatics and automated materials explorations.

Kei Terayama is Associate Professor at the Graduate School of Medical Life Science, Yokohama City University. He received a Doctor of Human and Environmental Studies degree in 2016 from Kyoto University. His research interests include developing and applying machine learning and computer vision techniques for chemistry, materials sciences, biology, and underwater monitoring.

Koji Tsuda received B.E., M.E., and Ph.D degrees from Kyoto University, Japan, in 1994, 1995, and 1998, respectively. Subsequently, he joined former Electrotechnical Laboratory (ETL), Tsukuba, Japan, as Research Scientist. When ETL was reorganized as AIST in 2001, he joined newly established Computational Biology Research Center, Tokyo, Japan. In 2000–2001, he worked at GMD FIRST in Berlin, Germany, as Visiting Scientist. In 2003–2004 and 2006–2008, he worked at Max Planck Institute for Biological Cybernetics, Tuebingen, Germany, first as Research Scientist and later as Project Leader. Currently, he is Professor at Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, the University of Tokyo. He is also affiliated with National Institute of Material Science (NIMS) and RIKEN Center for Advanced Intelligence Project. His current interests involve machine learning, computational biology and computational materials science. His recent work includes the use of machine learning for automatic design of organic and inorganic materials and biopolymers such as peptides and proteins.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Shinsuke Ishihara, Dr. Naoya Suzuki, Dr. Takehiro Fujita, Dr. Yasuyuki Nakamura, and Dr. Masanobu Naito for the topics discussed in this review. K. Tsuda is supported by JST CREST JPMJCR21O2 and ERATO JPMJER1903. K. Terayama is supported by JST FOREST JPMJFR232U, JST BOOST JPMJBY24F0, JSPS KAKENHI JP24H00473, and MEXT JPMXP1020230120. K. Yoshizoe is supported by JSPS KAKENHI JP20H04251 and JP20H05963. R. Tamura is supported by JST PRESTO JPMJPR24T8.

ABBREVIATIONS

1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
AAE	Adversarial Autoencoder
ADME	Absorption, Metabolism, Distribution, and Excretion
ADRA2A	Adrenoceptor Alpha 2A
AI	Artificial Intelligence
AR	Auto-Regressive
ASE	Amplified Spontaneous Emission
BERT	Bidirectional Encoder Representations from Transformers
BIMODAL	Bidirectional Recurrent Neural Networks
BO	Bayesian Optimization
BOSS	Bayesian Optimization Over String Spaces
BRICS	Breaking of Retrosynthetically Interesting Chemical Substructures
casVAE	Cascaded Variational Autoencoder
CLM	Chemical Language Model
CLscore	ChEMBL-Likeness Score
CMGN	Conditional Molecular Generation Net
CoG	Compound Generator
CVAE	Chemical Variational Autoencoder
CXCR4	C-X-C Chemokine Receptor Type 4
D-MPNN	Directed Message Passing Neural Network
DAG	Directed Acyclic Graphs

DDPM	Denosing Diffusion Probabilistic Model
DDR1	Discoidin Domain Receptor 1
DFT	Density Functional Theory
DMPNN	Dual Message Passing Neural Networks
DNC	Differentiable Neural Computer
DQN	Deep Q-Network
DRAGONFLY	Drug-Target Interactome-Based Generation of Novel Biologically Active Molecules
DST	Differentiable Scaffolding Tree
DTC	Drug Target Commons
E(3)	Euclidean Group in 3D
EA	Evolutionary Algorithm
EBM	Energy-Based Models
EC ₅₀	Median Effective Concentration
ECFP	Extended-Connectivity Fingerprint
EDM	E(3) Equivariant Diffusion Model
EI	Expected Improvement
FLAG	Fragment-Based Ligand-Generation Framework
FMQA	Factorization Machine with Quantum Annealing
FSscore	Focused Synthesizability Score
GA	Genetic Algorithm
GAIL	Generative Adversarial Imitation Learning
GAN	Generative Adversarial Network
GASA	Graph Attention-Based Assessment of Synthetic Accessibility
GB-GA	Graph-Based Genetic Algorithm
GB-GM-MCTS	Graph-Based Generative Model with Monte Carlo Tree Search
GCPN	Graph Convolutional Policy Network
GdMC	Gradient-driven Molecule Construction
GENTRL	Generative Tensorial Reinforcement Learning
GNN	Graph Neural Network
GPT	Generative Pre-Training
GraphTrans	Graph Transformer
GROVER	Graph Representation from Self-Supervised Message Passing Transformer
GRU	Gated Recurrent Units
HF	Hartree–Fock
HOMO	Highest Occupied Molecular Orbital
IC ₅₀	Median Inhibitory Concentration
IR	Infrared Absorption Spectrum
JAK1	Janus Kinase 1
L-Net	Ligand Neural Network
LIMO	Latent Inverse Molecular Optimizer
LLM	Large Language Model
LSTM	Long Short-Term Memory
LUMO	Lowest Unoccupied Molecular Orbital
M ^{Pro}	Main Protease
MCMC	Markov Chain Monte Carlo
MCMG	Multi-Constraint Molecular Generation
MCTS	Monte Carlo Tree Search
MD	Molecular Dynamics
MDP	Markov Decision Process
ML	Machine Learning
MLP	Multilayer Perceptron
MOO	Multi-Objective Optimization
MOSES	Molecular Sets
MPNN	Message Passing Neural Network
MS	Mass Spectrometry
NFA	Non-Fullerene Acceptors
NP	Natural Product

Nurr1	Nuclear Receptor Related 1
OMG	Open Molecule Generator
OPV	Organic Photovoltaics
PARP	Poly [ADP-ribose] Polymerase
PCE	Power Conversion Efficiency
PDE4D	Phosphodiesterase 4D
PGFS	Policy Gradient for Forward Synthesis
PI	Probability of Improvement
PI3K γ	Phosphatidylinositol 3-Kinase Gamma
pIC ₅₀	Negative Log of Median Inhibitory Concentration Value
PMO	Practical Molecular Optimization
PPO	Proximal Policy Optimization
PSO	Particle Swarm Optimization
QAOA	Quantum Approximate Optimization Algorithm
QC	Quantum Chemistry
QED	Quantitative Estimate of Drug Likeness
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RAscore	Retrosynthetic Accessibility Score
RBD	Receptor-Binding Domain
REACTOR	Reaction-Driven Objective Reinforcement
RECAP	Retrosynthetic Combinatorial Analysis Procedure
ReLeaSE	Reinforcement Learning for Structural Evolution
RL	Reinforcement Learning
RNN	Recurrent Neural Networks
RXR	Retinoid X Receptor
SAC	Soft Actor-Critic
SARS	CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2
SAScore	Synthetic Accessibility Score
SCScore	Synthetic Complexity Score
SDE	Stochastic Differential Equations
SDS	Safety Data Sheet
SE(3)	Special Euclidean Group in 3D
sEH	Soluble Epoxide Hydrolase
SELFIES	Self-Referencing Embedded Strings
SMBO	Sequential Model-Based Global Optimization
SMC	Sequential Monte Carlo
SMILES	Simplified Molecular Input Line Entry System
SPMM	Structure–Property Multimodal Foundation Model
SYBA	Synthetic Bayesian Accessibility
TamGen	Target-Aware Molecular Generation
TD	Temporal Difference
TDC	Therapeutics Data Commons
TNIK	TRAF2- and NCK-interacting Kinase
UV	Vis Ultraviolet–Visible
VAE	Variational Autoencoder
VQC	Variational Quantum Circuit
VS	Virtual Screening

REFERENCES

- (1) Beck, L.; Caffy, I.; Delqu -Koli , E.; Moreau, C.; Dumoulin, J.-P.; Perron, M.; Guichard, H.; Jeammet, V. Absolute dating of lead carbonates in ancient cosmetics by radiocarbon. *Commun. Chem.* **2018**, *1*, 1–7.
- (2) McMullen, R. L.; Dell’Acqua, G. History of Natural Ingredients in Cosmetics. *Cosmetics* **2023**, *10*, 71.
- (3) Parish, L. C.; Crissey, J. T. Cosmetics: A historical review. *Clin Dermatol.* **1988**, *6*, 1–4.
- (4) Alegbe, E. O.; Uthman, T. O. A review of history, properties, classification, applications and challenges of natural and synthetic dyes. *Heliyon* **2024**, *10*, No. e33646.
- (5) Decelles, C. The story of dyes and dyeing. *J. Chem. Educ.* **1949**, *26*, 583.
- (6) Hagan, E.; Poulin, J. Statistics of the early synthetic dye industry. *Herit. Sci.* **2021**, *9*, 33.
- (7) Levey, M. Dyes and dyeing in ancient Mesopotamia. *J. Chem. Educ.* **1955**, *32*, 625.
- (8) Luo, Y.; Zhang, X. Effects of yellow natural dyes on handmade Daqian paper. *Herit. Sci.* **2021**, *9*, 1–10.
- (9) Yadav, S.; Tiwari, K. S.; Gupta, C.; Tiwari, M. K.; Khan, A.; Sonkar, S. P. A brief review on natural dyes, pigments: Recent advances and future perspectives. *Results Chem.* **2023**, *5*, 100733.
- (10) Levey, M. The early history of detergent substances: A chapter in Babylonian chemistry. *J. Chem. Educ.* **1954**, *31*, 521.
- (11) Dalton, J. *A New System of Chemical Philosophy*; Sothis Press: 1808.
- (12) Rocke, A. J. It Began with a Daydream: The 150th Anniversary of the Kekul  Benzene Structure. *Angew. Chem., Int. Ed.* **2015**, *54*, 46–50.
- (13) Franc , M. A molecule with a ring to it. *Nat. Chem.* **2015**, *7*, 6–7.
- (14) Proctor, S.; Lovera, S.; Tomich, A.; Lavallo, V. Searching for the Truth: Elemental Analysis-A Powerful but Often Poorly Executed Technique. *ACS Cent. Sci.* **2022**, *8*, 874–876.
- (15) Nicolaou, K. C.; Snyder, S. A. Chasing Molecules That Were Never There: Misassigned Natural Products and the Role of Chemical Synthesis in Modern Structure Elucidation. *Angew. Chem., Int. Ed.* **2005**, *44*, 1012–1044.
- (16) Dirac, P. *The Principles of Quantum Mechanics*; Oxford University Press: London, 1958.
- (17) Terayama, K.; Sumita, M.; Tamura, R.; Tsuda, K. Black-Box Optimization for Automated Discovery. *Acc. Chem. Res.* **2021**, *54*, 1334.
- (18) E. Bright Wilson, J.; Decius, J. C.; Cross, P. C. *Molecular Vibrations The Theory of Infrared and Raman Vibrational Spectra*; Dover Publications, Inc.: 1955.
- (19) Chaikin, P.; Lubensky, T. *Principles of Condensed Matter Physics*; Cambridge University Press: 2003.
- (20) Giunta, C. J.; Mainz, V. V. *Pioneers of Magnetic Resonance*; ACS Symposium Series; American Chemical Society: 2020; Vol. 1349; pp 3–20.
- (21) Pine, S. H. *Organic Chemistry*; McGraw-Hill Book Co.: 1987.
- (22) Vollhardt, P.; Schore, N. *Organic Chemistry: Structure and Function*; W. H. Freeman and Company: New York, 2009.
- (23) Atkins, P.; Paula, J. D.; Keeler, J. *Atkins’ physical chemistry*; Oxford University Press: 2025.
- (24) Maier, M. E. Structural revisions of natural products by total synthesis. *Nat. Prod. Rep.* **2009**, *26*, 1105.
- (25) Crowfoot, D. A Review of Some Recent X-ray Work on Protein Crystals. *Chem. Rev.* **1941**, *28*, 215–228.
- (26) Crouse, D. T. X-ray Diffraction and the Discovery of the Structure of DNA. A Tutorial and Historical Account of James Watson and Francis Crick’s Use of X-ray Diffraction in Their Discovery of the Double Helix Structure of DNA. *J. Chem. Educ.* **2007**, *84*, 803.
- (27) Inokuma, Y.; Yoshioka, S.; Ariyoshi, J.; Arai, T.; Hitora, Y.; Takada, K.; Matsunaga, S.; Rissanen, K.; Fujita, M. X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature* **2013**, *495*, 461–466.
- (28) Gronenborn, A. M. Small, but powerful and attractive: 19F in biomolecular NMR. *Structure* **2022**, *30*, 6–14.
- (29) Hekstra, D. R. Emerging Time-Resolved X-Ray Diffraction Approaches for Protein Dynamics. *Annu. Rev. Biophys.* **2023**, *52*, 255–274.
- (30) Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021**, *93*, 1866–1879.

- (31) Meisburger, S. P.; Thomas, W. C.; Watkins, M. B.; Ando, N. X-ray Scattering Studies of Protein Structural Dynamics. *Chem. Rev.* **2017**, *117*, 7615–7672.
- (32) Purslow, J. A.; Khatiwada, B.; Bayro, M. J.; Venditti, V. NMR Methods for Structural Characterization of Protein-Protein Complexes. *Front. Mol. Biosci.* **2020**, *7*. DOI: 10.3389/fmolb.2020.00009
- (33) Yachi, Y.; Yoshii, Y.; Matsuya, Y.; Mori, R.; Oikawa, J.; Date, H. Track Structure Study for Energy Dependency of Electrons and X-rays on DNA Double-Strand Break Induction. *Sci. Rep.* **2019**, *9*, 17649.
- (34) He, W.; Yu, Y.; Iizuka, K.; Takezawa, H.; Fujita, M. Supramolecular coordination cages as crystalline sponges through a symmetry mismatch strategy. *Nat. Chem.* **2025**, *17*, 653–662.
- (35) Bally, T.; Masamune, S. Cyclobutadiene. *Tetrahedron* **1980**, *36*, 343.
- (36) Maier, G. Tetrahedrane and Cyclobutadiene. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 309.
- (37) Reekie, T. A.; Williams, C. M.; Rendina, L. M.; Kassiou, M. Cubanes in Medicinal Chemistry. *J. Med. Chem.* **2019**, *62*, 1078–1095.
- (38) Zhang, B.; Wepf, R.; Fischer, K.; Schmidt, M.; Besse, S.; Lindner, P.; King, B. T.; Sigel, R.; Schurtenberger, P.; Talmon, Y.; Ding, Y.; Kröger, M.; Halperin, A.; Schlüter, A. D. The Largest Synthetic Structure with Molecular Precision: Towards a Molecular Object. *Angew. Chem., Int. Ed. Engl.* **2011**, *50*, 737–740.
- (39) Woodward, R. B. Recent advances in the chemistry of natural products. *Pure Appl. Chem.* **1968**, *17*, 519–547.
- (40) Hoffman, A. S. The origins and evolution of “controlled” drug delivery systems. *J. Controlled Release* **2008**, *132*, 153–163.
- (41) Izawa, H.; Kawakami, K.; Sumita, M.; Tateyama, Y.; Hill, J. P.; Ariga, K. β -Cyclodextrin-crosslinked alginate gel for patient-controlled drug delivery systems: regulation of host-guest interactions with mechanical stimuli. *J. Mater. Chem. B* **2013**, *1*, 2155.
- (42) Vargason, A. M.; Anselmo, A. C.; Mitragotri, S. The evolution of commercial drug delivery technologies. *Nat. Biomed. Eng.* **2021**, *5*, 951–967.
- (43) Boschloo, G.; Hagfeldt, A. Characteristics of the iodide/triiodide redox mediator in dye-sensitized solar cells. *Acc. Chem. Res.* **2009**, *42*, 1819–26.
- (44) Brogdon, P.; Cheema, H.; Delcamp, J. H. Near-Infrared-Absorbing Metal-Free Organic, Porphyrin, and Phthalocyanine Sensitizers for Panchromatic Dye-Sensitized Solar Cells. *ChemSusChem* **2018**, *11*, 86–103.
- (45) Chen, K.-S.; Liu, W.-H.; Wang, Y.-H.; Lai, C.-H.; Chou, P.-T.; Lee, G.-H.; Chen, K.; Chen, H.-Y.; Chi, Y.; Tung, F.-C. New Family of Ruthenium-Dye-Sensitized Nanocrystalline TiO₂ Solar Cells with a High Solar-Energy-Conversion Efficiency. *Adv. Funct. Mater.* **2007**, *17*, 2964–2974.
- (46) Duncan, W. R.; Prezhdo, O. V. Theoretical studies of photoinduced electron transfer in dye-sensitized TiO₂. *Annu. Rev. Phys. Chem.* **2007**, *58*, 143–84.
- (47) Grätzel, M. Dye-sensitized solar cells. *J. Photochem. Photobiol. C: Photochem. Rev.* **2003**, *4*, 145–153.
- (48) Grätzel, M. Mesoscopic Solar Cells for Electricity and Hydrogen Production from Sunlight. *Chem. Lett.* **2005**, *34*, 8–13.
- (49) Grätzel, M. Recent advances in sensitized mesoscopic solar cells. *Acc. Chem. Res.* **2009**, *42*, 1788–98.
- (50) Hagfeldt, A.; Boschloo, G.; Sun, L.; Kloo, L.; Pettersson, H. Dye-sensitized solar cells. *Chem. Rev.* **2010**, *110*, 6595–663.
- (51) Hagfeldt, A.; Grätzel, M. Light-Induced Redox Reactions in Nanocrystalline Systems. *Chem. Rev.* **1995**, *95*, 49–68.
- (52) Hara, K.; Sato, T.; Katoh, R.; Furube, A.; Yoshihara, T.; Murai, M.; Kurashige, M.; Ito, S.; Shinpo, A.; Suga, S.; Arakawa, H. Novel conjugated organic dyes for efficient dye-sensitized solar cells. *Adv. Funct. Mater.* **2005**, *15*, 246–252.
- (53) Lin, L.-Y.; Chen, Y.-H.; Huang, Z.-Y.; Lin, H.-W.; Chou, S.-H.; Lin, F.; Chen, C.-W.; Liu, Y.-H.; Wong, K.-T. A low-energy-gap organic dye for high-performance small-molecule organic solar cells. *J. Am. Chem. Soc.* **2011**, *133*, 15822–15825.
- (54) Rahman, S.; Haleem, A.; Siddiq, M.; Hussain, M. K.; Qamar, S.; Hameed, S.; Waris, M. Research on dye sensitized solar cells: recent advancement toward the various constituents of dye sensitized solar cells for efficiency enhancement and future prospects. *RSC Adv.* **2023**, *13*, 19508–19529.
- (55) Robertson, N. Optimizing Dyes for Dye-Sensitized Solar Cells. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 2338–2345.
- (56) Tsubomura, H.; Matsumura, M.; Nomura, Y.; Amamiya, T. Dye sensitized zinc oxide: aqueous electrolyte: platinum photocell. *Nature* **1976**, *261*, 402–403.
- (57) Kranthiraja, K.; Saeki, A. Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells. *Adv. Funct. Mater.* **2021**, *31*, 1–11.
- (58) Duke, R. M.; Veale, E. B.; Pfeffer, F. M.; Kruger, P. E.; Gunnlaugsson, T. Colorimetric and fluorescent anion sensors: an overview of recent developments in the use of 1,8-naphthalimide-based chemosensors. *Chem. Soc. Rev.* **2010**, *39*, 3936–3953.
- (59) Yeung, M. C.-L.; Yam, V. W.-W. Luminescent cation sensors: from host-guest chemistry, supramolecular chemistry to reaction-based mechanisms. *Chem. Soc. Rev.* **2015**, *44*, 4192–4202.
- (60) Cao, D.; Liu, Z.; Verwilt, P.; Koo, S.; Jangjili, P.; Kim, J. S.; Lin, W. Coumarin-Based Small-Molecule Fluorescent Chemosensors. *Chem. Rev.* **2019**, *119*, 10403–10519.
- (61) Torsi, L.; Magliulo, M.; Manoli, K.; Palazzo, G. Organic field-effect transistor sensors: A tutorial review. *Chem. Soc. Rev.* **2013**, *42*, 8612–8628.
- (62) Jou, J.-H.; Kumar, S.; Agrawal, A.; Li, T.-H.; Sahoo, S. Approaches for fabricating high efficiency organic light emitting diodes. *J. Mater. Chem. C* **2015**, *3*, 2974–3002.
- (63) Senge, M. O.; Sergeeva, N. N.; Hale, K. J. Classic highlights in porphyrin and porphyrinoid total synthesis and biosynthesis. *Chem. Soc. Rev.* **2021**, *50*, 4730–4789.
- (64) Nikolaou, V.; Nikoloudakis, E.; Ladomenou, K.; Charalambidis, G.; Coutsolelos, A. G. Porphyrins—valuable pigments of life. *Front. Chem. Biol.* **2024**, *2*, 1346465.
- (65) Hiroto, S.; Miyake, Y.; Shinokubo, H. Synthesis and Functionalization of Porphyrins through Organometallic Methodologies. *Chem. Rev.* **2017**, *117*, 2910–3043.
- (66) Guo, Z.; Park, S.; Yoon, J.; Shin, I. Recent progress in the development of near-infrared fluorescent probes for bioimaging applications. *Chem. Soc. Rev.* **2014**, *43*, 16.
- (67) Shaath, N. A. Ultraviolet filters. *Photochem. Photobiol. Sci.* **2010**, *9*, 464–469.
- (68) Grimm, J. B.; Tkachuk, A. N.; Xie, L.; Choi, H.; Mohar, B.; Falco, N.; Schaefer, K.; Patel, R.; Zheng, Q.; Liu, Z.; Lippincott-Schwartz, J.; Brown, T. A.; Lavis, L. D. A general method to optimize and functionalize red-shifted rhodamine dyes. *Nat. Methods* **2020**, *17*, 815–821.
- (69) Ohno, H.; Sasaki, E.; Yamada, S.; Hanaoka, K. Recent advances in Si-rhodamine-based fluorescent probes for live-cell imaging. *Org. Biomol. Chem.* **2024**, *22*, 3099–3108.
- (70) Zeng, S.; Liu, X.; Kafuti, Y. S.; Kim, H.; Wang, J.; Peng, X.; Li, H.; Yoon, J. Fluorescent dyes based on rhodamine derivatives for bioimaging and therapeutics: recent progress, challenges, and prospects. *Chem. Soc. Rev.* **2023**, *52*, 5607–5651.
- (71) Ilina, K.; MacCuaig, W. M.; Laramie, M.; Jeouty, J. N.; McNally, L. R.; Henary, M. Squaraine Dyes: Molecular Design for Different Applications and Remaining Challenges. *Bioconjugate Chem.* **2020**, *31*, 194–213.
- (72) Gao, M.; Kwaria, D.; Norikane, Y.; Yue, Y. Visible-light-switchable azobenzenes: Molecular design, supramolecular systems, and applications. *Nat. Sci.* **2023**, *3*. DOI: 10.1002/ntls.20220020
- (73) Aleotti, F.; Petropoulos, V.; Van Overeem, H.; Pettini, M.; Mancinelli, M.; Pecorari, D.; Maiuri, M.; Medri, R.; Mazzanti, A.; Preda, F.; Perri, A.; Polli, D.; Conti, I.; Cerullo, G.; Garavelli, M. Engineering Azobenzene Derivatives to Control the Photoisomerization Process. *J. Phys. Chem. A* **2023**, *127*, 10435–10449.
- (74) Waldeck, D. H. Photoisomerization dynamics of stilbenes. *Chem. Rev.* **1991**, *91*, 415–436.

- (75) Villarán, D.; Wezenberg, S. J. Stiff-Stilbene Photoswitches: From Fundamental Studies to Emergent Applications. *Angew. Chem., Int. Ed. Engl.* **2020**, *59*, 13192–13202.
- (76) Lui, A. C. W.; Pow, K. C.; Lin, N.; Lam, L. P. Y.; Liu, G.; Godwin, I. D.; Fan, Z.; Khoo, C. J.; Tobimatsu, Y.; Wang, L.; Hao, Q.; Lo, C. Regioselective stilbene O-methylations in Saccharinae grasses. *Nat. Commun.* **2023**, *14*. DOI: 10.1038/s41467-023-38908-5
- (77) Sepehri, S.; Khedmati, M.; Yousef-Nejad, F.; Mahdavi, M. Medicinal chemistry perspective on the structure-activity relationship of stilbene derivatives. *RSC Adv.* **2024**, *14*, 19823–19879.
- (78) Gani, R.; Brignole, E.A. Molecular Design of Solvents for Liquid Extraction based on UNIFAC. *Fluid Phase Equilibria* **1983**, *13*, 331.
- (79) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188–195.
- (80) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595–10599.
- (81) Lowe, J. P. *Quantum Chemistry*; Academic Press: London, 1993.
- (82) Cooper, D. L., Ed. *Valence Bond Theory*; Elsevier: 2002.
- (83) Dunning, T. H.; Xu, L. T.; Takeshita, T. Y.; Lindquist, B. A. Insights into the Electronic Structure of Molecules from Generalized Valence Bond Theory. *J. Phys. Chem. A* **2016**, *120*, 1763–1778.
- (84) Dunning, T. H.; Xu, L. T.; Cooper, D. L.; Karadakov, P. B. Spin-Coupled Generalized Valence Bond Theory: New Perspectives on the Electronic Structure of Molecules and Chemical Bonds. *J. Phys. Chem. A* **2021**, *125*, 2021–2050.
- (85) Heitler, W.; London, F. Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Z. Phys.* **1927**, *44*, 455–472.
- (86) Pauncz, R. *Spin Eigenfunctions*; Springer Science+Business Media: New York, 1979.
- (87) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publishing: 1996.
- (88) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618.
- (89) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E.M. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-CI Approach. *Chem. Phys.* **1980**, *48*, 157.
- (90) Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.
- (91) Nakano, H. Quasidegenerate Perturbation Theory with Multiconfigurational Self-Consistent-field Reference Functions. *J. Chem. Phys.* **1993**, *99*, 7983.
- (92) Bartlett, R. J.; Musial, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291–352.
- (93) Lyakh, D. I.; Musial, M.; Lotrich, V. F.; Bartlett, R. J. Multireference Nature of Chemistry: The Coupled-Cluster View. *Chem. Rev.* **2012**, *112*, 182–243.
- (94) Bartlett, R. J. Perspective on Coupled-cluster Theory. The evolution toward simplicity in quantum chemistry. *Phys. Chem. Chem. Phys.* **2024**, *26*, 8013–8037.
- (95) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (96) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for density functional theory. *Chem. Rev.* **2012**, *112*, 289–320.
- (97) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.
- (98) Furche, F.; Burke, K. *Chapter 2 Time-Dependent Density Functional Theory in Quantum Chemistry*; Elsevier: Amsterdam, 2005; Vol. 1; p 19–30.
- (99) Lacombe, L.; Maitra, N. T. Non-adiabatic approximations in time-dependent density functional theory: progress and prospects. *npj Comput. Mater.* **2023**, *9*, 124.
- (100) Schirmer, J. Review of the foundations of time-dependent density-functional theory (TDDFT). *Phys. Chem. Chem. Phys.* **2025**, *27*, 4992–5005.
- (101) Miller, K. J. Calculation of the Molecular Polarizability Tensor. *J. Am. Chem. Soc.* **1990**, *112*, 8543.
- (102) Send, R.; Furche, F. First-order nonadiabatic couplings from time-dependent hybrid density functional response theory: Consistent formalism, implementation, and performance. *J. Chem. Phys.* **2010**, *132*, 044107.
- (103) Obara, S.; Saika, A. Efficient Recursive Computation of Molecular Integrals over Cartesian Gaussian functions. *J. Chem. Phys.* **1986**, *84*, 3963–3974.
- (104) Vandevondele, J.; Borštnik, U.; Hutter, J. with millions of atoms in the condensed phase. Linear scaling self-consistent field calculations with millions of atoms in the condensed phase. *J. Chem. Theory Comput.* **2012**, *8*, 3565–3573.
- (105) Jensen, F. *Introduction to Computational Chemistry 3rd ed.*; Wiley: 2017.
- (106) Rapaport, D. C. *The Art of Molecular Dynamics Simulation*; Cambridge, UK, 2010.
- (107) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **2017**, *121*, 511–522.
- (108) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142–10186.
- (109) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
- (110) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- (111) Hansch, C.; Fujita, T. $p - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
- (112) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
- (113) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- (114) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1044.
- (115) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
- (116) Domcke, W.; Yarkony, D. R.; Köppel, H., Eds. *Conical Intersections: Electronic Structure, Dynamics & Spectroscopy*; World Scientific Publishing Co. Pte. Ltd.: Singapore, 2004.
- (117) Leung, K. DFT modelling of explicit solid-solid interfaces in batteries: methods and challenges. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10412–10425.
- (118) Takahashi, O.; Sumita, M. Characterization of the hyperline of D1/D0 conical intersections between the maleic acid and fumaric acid anion radicals. *J. Chem. Phys.* **2004**, *121*, 7030–1.
- (119) Sumita, M.; Saito, K. Ab initio study on one-way photoisomerization of the maleic acid and fumaric acid anion radical system as a model system of their esters. *J. Phys. Chem. A* **2006**, *110*, 12276–81.
- (120) Sumita, M.; Ryazantsev, N.; Saito, K. Acceleration of the Z to E photoisomerization of penta-2, 4-dieniminium by hydrogen out-of-plane motion: theoretical study on a model system of retinal protonated Schiff base w. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6406–6414.
- (121) Sumita, M.; Tanaka, Y.; Ikeda, M.; Ohno, T. Charged and Discharged States of Cathode/Sulfide-Electrolyte Interfaces in All-Solid-State Lithium-Ion Batteries. *J. Phys. Chem. C* **2016**, *120*, 13332–13339.
- (122) Sumiya, M.; Sumita, M.; Asai, Y.; Tamura, R.; Uedono, A.; Yoshigoe, A. Dynamic Observation and Theoretical Analysis of Initial O₂ Molecule Adsorption on Polar and m-Plane Surfaces of GaN. *J. Phys. Chem. C* **2020**, *124*, 25282–25290.

- (123) Sumiya, M.; Sumita, M.; Tsuda, Y.; Sakamoto, T.; Sang, L.; Harada, Y.; Yoshigoe, A. High reactivity of H₂O vapor on GaN surfaces. *Sci. Technol. Adv. Mater.* **2022**, *23*, 189–198.
- (124) Sumita, M.; Takahashi, H.; Sugita, T.; Tsuda, K. Theoretical Insight into the Origin of Dielectric and Magnetoelectric FeCo Alloy-Metal Fluoride Insulators. *J. Phys. Chem. C* **2024**, *128*, 1386–1392.
- (125) Frisch, M. J. et al. *Gaussian16*, Revision C.01; Gaussian Inc.: Wallingford, CT, 2016.
- (126) Barca, G. M. J.; et al. Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **2020**, *152*, 154102.
- (127) Neese, F. Software update: the ORCA program system, version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1606.
- (128) Aprà, E.; et al. NWChem: Past, present, and future. *J. Chem. Phys.* **2020**, *152*, 184102.
- (129) Kühne, T. D.; et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103.
- (130) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.
- (131) Dennington, R.; Keith, T. A.; Millam, J. M. *GaussView Version 6*. 2019; Semichem Inc.: Shawnee Mission, KS.
- (132) Momma, K.; Izumi, F. VESTA: a three-dimensional visualization system for electronic and structural analysis. *J. Appl. Crystallogr.* **2008**, *41*, 653–658.
- (133) Schrödinger, L.; DeLano, W. PyMOL. <https://www.pymol.org/>, (accessed 2025–12–07).
- (134) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput. Aided Mol. Des.* **2000**, *14*, 123–134.
- (135) Sharma, R.; Zeller, M.; Pavlovic, V. I.; Huang, T. S.; Lo, Z.; Chu, S.; Zhao, Y.; Phillips, J. C.; Schulten, K. Speech/Gesture Interface to a Visual-Computing Environment. *IEEECGA* **2000**, *20*, 29–37.
- (136) Izawa, H.; Nishino, S.; Sumita, M.; Akamatsu, M.; Morihashi, K.; Ifuku, S.; Morimoto, M.; Saimoto, H. Novel 1,8-Naphthalimide derivative with an open space for an anion: Unique fluorescence behaviour depending on the binding anion's electrophilic property. *Chem. Commun.* **2015**, *51*, 8596–8599.
- (137) Vopson, M. M. Estimation of the information contained in the visible matter of the universe. *AIP Adv.* **2021**, *11*, 105317.
- (138) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.
- (139) Restrepo, G. Chemical space: limits, evolution and modelling of an object bigger than our universal library. *Digit. Discovery* **2022**, *1*, 568–585.
- (140) Llanos, E. J.; Leal, W.; Luu, D. H.; Jost, J.; Stadler, P. F.; Restrepo, G. Exploration of the chemical space and its three historical regimes. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 12660–12665.
- (141) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: 1991.
- (142) Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135–6146.
- (143) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- (144) Bowman, J. M.; Qu, C.; Conte, R.; Nandi, A.; Houston, P. L.; Yu, Q. The MD17 datasets from the perspective of datasets for gas-phase “small” molecule potentials. *J. Chem. Phys.* **2022**, *156*, 240901.
- (145) Grosso, B. F.; Davies, D. W.; Zhu, B.; Walsh, A.; Scanlon, D. O. Accessible chemical space for metal nitride perovskites. *Chem. Sci.* **2023**, *14*, 9175–9185.
- (146) Gaulton, A.; Kale, N.; van Westen, G. J. P.; Bellis, L. J.; Bento, A. P.; Davies, M.; Hersey, A.; Papadatos, G.; Forster, M.; Wege, P.; Overington, J. P. A large-scale crop protection bioassay data set. *Sci. Data* **2015**, *2*, 150032.
- (147) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1*, 1–15.
- (148) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679.
- (149) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 1–7.
- (150) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- (151) Joung, J. F.; Han, M.; Jeong, M.; Park, S. Experimental database of optical properties of organic compounds. *Sci. Data* **2020**, *7*. DOI: 10.1038/s41597-020-00634-8
- (152) Ma, R.; Luo, T. PIM: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684–4690.
- (153) Mondal, A.; Paterson, L.; Cho, J.; Lin, K.-H.; van der Zee, B.; Wetzelaer, G.-J. A. H.; Stankevych, A.; Vakhnin, A.; Kim, J.-J.; Kadashchuk, A.; Blom, P. W. M.; May, F.; Andrienko, D. Molecular library of OLED host materials—Evaluating the multiscale simulation workflow. *Chem. Phys. Rev.* **2021**, *2*, 031304.
- (154) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.
- (155) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.
- (156) Broach, J. R.; Thorner, J. High-throughput screening for drug discovery. *Nature* **1996**, *384*, 14–16.
- (157) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191–201.
- (158) Reymond, J. L.; Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **2012**, *3*, 649–657.
- (159) Afzal, M. A. F.; Haghighatlari, M.; Ganesh, S. P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *J. Phys. Chem. C* **2019**, *123*, 14610–14618.
- (160) Song, S.; Chen, F.; Wang, Y.; Wang, K.; Yan, M.; Zhang, Q. Accelerating the discovery of energetic melt-castable materials by a high-throughput virtual screening and experimental approach. *J. Mater. Chem. A* **2021**, *9*, 21723–21731.
- (161) Gurnani, R.; Shukla, S.; Kamal, D.; Wu, C.; Hao, J.; Kuenneth, C.; Aklujkar, P.; Khomane, A.; Daniels, R.; Deshmukh, A. A.; Cao, Y.; Sotzing, G.; Ramprasad, R. AI-assisted discovery of high-temperature dielectrics for energy storage. *Nat. Commun.* **2024**, *15*, 6107.
- (162) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185–6192.
- (163) Gómez-Bombarelli, R.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120–1127.
- (164) Li, P.; Wang, Z.; Li, W.; Yuan, J.; Chen, R. Design of Thermally Activated Delayed Fluorescence Materials with High Intersystem Crossing Efficiencies by Machine Learning-Assisted Virtual Screening. *J. Phys. Chem. Lett.* **2022**, *13*, 9910–9918.
- (165) Perkinson, C. F.; Tabor, D. P.; Einzinger, M.; Sheberla, D.; Utzat, H.; Lin, T.-A.; Congreve, D. N.; Bawendi, M. G.; Aspuru-Guzik, A.; Baldo, M. A. Discovery of blue singlet exciton fission molecules via a

high-throughput virtual screening and experimental approach. *J. Chem. Phys.* **2019**, *151*, 121102.

(166) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613–1623.

(167) Martsinovich, N.; Troisi, A. High-Throughput Computational Screening of Chromophores for Dye-Sensitized Solar Cells. *J. Phys. Chem. C* **2011**, *115*, 11781–11792.

(168) Li, S.; Chung, Y. G.; Snurr, R. Q. High-Throughput Screening of Metal-Organic Frameworks for CO₂ Capture in the Presence of Water. *Langmuir* **2016**, *32*, 10368–10376.

(169) Han, S.; Huang, Y.; Watanabe, T.; Dai, Y.; Walton, K. S.; Nair, S.; Sholl, D. S.; Meredith, J. C. High-Throughput Screening of Metal-Organic Frameworks for CO₂ Separation. *ACS Comb. Sci.* **2012**, *14*, 263–267.

(170) Witman, M.; Ling, S.; Anderson, S.; Tong, L.; Stylianou, K. C.; Slater, B.; Smit, B.; Haranczyk, M. In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. *Chem. Sci.* **2016**, *7*, 6263–6272.

(171) Wu, J.; et al. Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science* **2024**, *386*, 1256–1264.

(172) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2012**, *4*, 83–89.

(173) Ma, R.; Zhang, H.; Xu, J.; Sun, L.; Hayashi, Y.; Yoshida, R.; Shiomi, J.; Wang, J.-x.; Luo, T. Machine learning-assisted exploration of thermally conductive polymers based on high-throughput molecular dynamics simulations. *Mater. Today Phys.* **2022**, *28*, 100850.

(174) Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **2018**, *148*, 241735.

(175) Shu, Y.; Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **2015**, *142*, 104104.

(176) Nanjo, S.; Arifin, S.; Maeda, H.; Hayashi, Y.; Hatakeyama-Sato, K.; Himeno, R.; Hayakawa, T.; Yoshida, R. SPACIER: on-demand polymer design with fully automated all-atom classical molecular dynamics integrated into machine learning pipelines. *npj Comput. Mater.* **2025**, *11*, 16.

(177) Button, A.; Merk, D.; Hiss, J. A.; Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nat. Mach. Intell.* **2019**, *1*, 307–315.

(178) Zheng, S.; Tan, Y.; Wang, Z.; Li, C.; Zhang, Z.; Sang, X.; Chen, H.; Yang, Y. Accelerated rational PROTAC design via deep learning and molecular simulations. *Nat. Mach. Intell.* **2022**, *4*, 739–748.

(179) Park, J.; Sorourifar, F.; Muthyala, M. R.; Houser, A. M.; Tuttle, M.; Paulson, J. A.; Zhang, S. Zero-Shot Discovery of High-Performance, Low-Cost Organic Battery Materials Using Machine Learning. *J. Am. Chem. Soc.* **2024**, *146*, 31230–31239.

(180) Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *JACS Au* **2021**, *1*, 427–438.

(181) Chakraborty, S.; Almog, I.; Gershoni-Poranne, R. COMPAS-4: A Data Set of (BN)₁ Substituted Cata-Condensed Polybenzenoid Hydrocarbons-Data Analysis and Feature Engineering. *J. Chem. Inf. Model.* **2025**, *65*, 5508–5517.

(182) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

(183) 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*, 828–849.

(184) Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discovery Today* **2021**, *26*, 2707–2715.

(185) Du, Y.; Fu, T.; Sun, J.; Liu, S. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. *arXiv preprint* **2022**, DOI: 10.48550/ARXIV.2203.14500, (accessed 2025–12–07).

(186) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.

(187) Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; Wei, L. Deep generative models in de novo drug molecule generation. *J. Chem. Inf. Model.* **2024**, *64*, 2174–2194.

(188) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chem. Sci. *J. Am. Chem. Soc.* **2023**, *145*, 8736–8750.

(189) Xu, W. Current Status of Computational Approaches for Small Molecule Drug Discovery. *J. Med. Chem.* **2024**, *67*, 18633–18636.

(190) Luong, K.-D.; Singh, A. Application of transformers in cheminformatics. *J. Chem. Inf. Model.* **2024**, *64*, 4392–4409.

(191) Du, Y.; Jamasb, A. R.; Guo, J.; Fu, T.; Harris, C.; Wang, Y.; Duan, C.; Liò, P.; Schwaller, P.; Blundell, T. L. Machine learning-aided generative molecular design. *Nat. Mach. Intell.* **2024**, *6*, 589–604.

(192) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **2020**, *12*, 56.

(193) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1603.

(194) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(195) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. learn.: sci. technol.* **2020**, *1*, 045024.

(196) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv preprint* **2018**, DOI: 10.26434/chemrxiv.7097960.v1, (accessed 2025–12–07).

(197) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523–1531.

(198) Wu, J.-N.; Wang, T.; Chen, Y.; Tang, L.-J.; Wu, H.-L.; Yu, R.-Q. t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nat. Commun.* **2024**, *15*, 4993.

(199) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **2015**, *7*, 1–34.

(200) 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*, 1692–1701.

(201) Lee, N.; Laghuvarapu, S.; Park, C.; Sun, J. Vision Language Model is NOT All You Need: Augmentation Strategies for Molecule Language Models. *arXiv preprint* **2024**, DOI: 10.48550/ARXIV.2407.09043, (accessed 2025–12–07).

(202) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(203) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(204) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.

(205) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4.

(206) 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*.

(207) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *International conference on machine learning*. 2017; pp 1263–1272.

- (208) Schütt, K. T.; Saucedo, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, DOI: 10.1063/1.5019779
- (209) Faulon, J.-L.; Bender, A. *Handbook of chemoinformatics algorithms*; CRC Press: 2010.
- (210) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- (211) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using ‘Drug-Like’ Chemical Fragment Spaces. *ChemMedChem*. **2008**, *3*, 1503–1507.
- (212) Böhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593–606.
- (213) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: a program for structure generation. *J. Comput. Aided Mol. Des.* **1993**, *7*, 127–153.
- (214) Ghersi, D.; Singh, M. molBLOCKS: decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics* **2014**, *30*, 2081–2083.
- (215) Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break down in order to build up: decomposing small molecules for fragment-based drug design with e molfrag. *J. Chem. Inf. Model.* **2017**, *57*, 627–631.
- (216) Diao, Y.; Hu, F.; Shen, Z.; Li, H. MacFrag: segmenting large-scale molecules to obtain diverse fragments with high qualities. *Bioinformatics* **2023**, *39*, btad012.
- (217) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087.
- (218) Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572.
- (219) Hochreiter, S.; Schmidhuber, J. LSTM can solve hard long time lag problems. *Adv. Neural Inf. Process. Syst.* **1996**, *9*.
- (220) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735–1780.
- (221) Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, 2014; pp 1724–1734.
- (222) Graves, A. Generating Sequences With Recurrent Neural Networks. *arXiv preprint* **2014**, DOI: 10.48550/ARXIV.1308.0850, (accessed 2025–12–07).
- (223) Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; Bengio, S. Generating Sentences from a Continuous Space. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany, 2016; pp 10–21.
- (224) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.
- (225) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Raymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **2019**, *11*, 1–13.
- (226) Skinnider, M. A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **2024**, *6*, 437–448.
- (227) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional molecule generation with recurrent neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 1175–1183.
- (228) Flam-Shepherd, D.; Zhu, K.; Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nat. Commun.* **2022**, *13*, 3293.
- (229) Xu, M.; Ran, T.; Chen, H. De novo molecule design through the molecular generative model conditioned by 3D information of protein binding sites. *J. Chem. Inf. Model.* **2021**, *61*, 3240–3254.
- (230) Atz, K.; Cotos, L.; Isert, C.; Håkansson, M.; Focht, D.; Hilleke, M.; Nippa, D. F.; Iff, M.; Ledergerber, J.; Schiebroke, C. C. Prospective de novo drug design with deep interactome learning. *Nat. Commun.* **2024**, *15*, 3408 others.
- (231) Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; Battaglia, P. Learning Deep Generative Models of Graphs. *arXiv preprint* **2018**, DOI: 10.48550/ARXIV.1803.03324, (accessed 2025–12–07).
- (232) Li, Y.; Zhang, L.; Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *J. Cheminform.* **2018**, *10*, 1–24.
- (233) Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. MolecularRNN: Generating realistic molecular graphs with optimized properties. *arXiv preprint* **2019**, DOI: 10.48550/ARXIV.1905.13372, (accessed 2025–12–07).
- (234) Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256.
- (235) Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
- (236) Schütt, K.; Kindermans, P.-J.; Saucedo, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
- (237) Luo, S.; Guan, J.; Ma, J.; Peng, J. A 3D generative model for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 6229–6239.
- (238) Peng, X.; Luo, S.; Guan, J.; Xie, Q.; Peng, J.; Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. *International Conference on Machine Learning*. 2022; pp 17644–17655.
- (239) Zhang, Z.; Min, Y.; Zheng, S.; Liu, Q. Molecule generation for target protein binding with structural motifs. *The eleventh international conference on learning representations*. 2023.
- (240) Cheng, A. H.; Sun, C.; Aspuru-Guzik, A. Scalable autoregressive 3d molecule generation. *arXiv preprint* **2025**, DOI: 10.48550/ARXIV.2505.13791, (accessed 2025–12–07).
- (241) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint* **2013**, DOI: 10.48550/ARXIV.1312.6114, (accessed 2025–12–07).
- (242) Rezende, D. J.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *International conference on machine learning*. 2014; pp 1278–1286.
- (243) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.
- (244) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. *Proceedings of the 34th International Conference on Machine Learning*. 2017; pp 1945–1954.
- (245) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. *International Conference on Learning Representations*. 2018.
- (246) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *International conference on machine learning*. 2018; pp 2323–2332.
- (247) Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. Learning Multimodal Graph-to-Graph Translation for Molecule Optimization. *International Conference on Learning Representations*. 2019.
- (248) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *Artificial Neural Networks and Machine Learning - ICANN 2018*. **2018**, 11139, 412–422.
- (249) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained graph variational autoencoders for molecule design. *Adv. Neural Inf. Process. Syst.* **2018**.

- (250) Samanta, B.; De, A.; Jana, G.; Gómez, V.; Chattaraj, P.; Ganguly, N.; Gomez-Rodriguez, M. Nevae: A deep generative model for molecular graphs. *AAAI* **2019**, *33*, 1110.
- (251) Wang, M.; Hsieh, C.-Y.; Wang, J.; Wang, D.; Weng, G.; Shen, C.; Yao, X.; Bing, Z.; Li, H.; Cao, D.; Hou, T. Relation: A deep generative model for structure-based de novo drug design. *J. Med. Chem.* **2022**, *65*, 9478–9492 others.
- (252) Nesterov, V.; Wieser, M.; Roth, V. 3DMolNet: A Generative Network for Molecular Structures. *arXiv preprint* **2020**, DOI: 10.48550/ARXIV.2010.06477, (accessed 2025–12–07).
- (253) Ragoza, M.; Masuda, T.; Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **2022**, *13*, 2701–2713.
- (254) Guo, X.; Du, Y.; Zhao, L. Property controllable variational autoencoder via invertible mutual dependence. *International Conference on Learning Representations*. 2020.
- (255) Du, Y.; Guo, X.; Shehu, A.; Zhao, L. Interpretable molecular graph generation via monotonic constraints. *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. 2022; pp 73–81.
- (256) Zhavoronkov, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040.
- (257) Tripp, A.; Daxberger, E.; Hernández-Lobato, J. M. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020.
- (258) Hoffman, S. C.; Chenthamarakshan, V.; Wadhawan, K.; Chen, P.-Y.; Das, P. Optimizing molecules using efficient queries from property evaluations. *Nat. Mach. Intell.* **2022**, *4*, 21–31.
- (259) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
- (260) Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv preprint* **2017**, DOI: 10.48550/ARXIV.1701.07875, (accessed 2025–12–07).
- (261) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **2017**, *8*, 10883.
- (262) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.
- (263) Bian, Y.; Wang, J.; Jun, J. J.; Xie, X.-Q. Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors. *Mol. Pharmaceutics* **2019**, *16*, 4451–4460.
- (264) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv preprint* **2017**, DOI: 10.48550/ARXIV.1705.10843, (accessed 2025–12–07).
- (265) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *ChemRxiv preprint* **2017**, DOI: 10.26434/chemrxiv.5309668.v3, (accessed 2025–12–07).
- (266) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.
- (267) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial threshold neural computer for molecular de novo design. *Mol. Pharmaceutics* **2018**, *15*, 4386–4397.
- (268) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminform.* **2019**, *11*, 1–13.
- (269) Abbasi, M.; Santos, B. P.; Pereira, T. C.; Sofia, R.; Monteiro, N. R.; Simões, C. J.; Brito, R. M.; Ribeiro, B.; Oliveira, J. L.; Arrais, J. P. Designing optimized drug candidates with generative adversarial network. *J. Cheminform.* **2022**, *14*, 40.
- (270) De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models* **2018**.
- (271) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchol, M. Mol-CycleGAN: a generative model for molecular optimization. *J. Cheminform.* **2020**, *12*, 2.
- (272) Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*. 2017; pp 2223–2232.
- (273) Assouel, R.; Ahmed, M.; Segler, M. H.; Saffari, A.; Bengio, Y. Defactor: Differentiable edge factorization-based probabilistic graph generation. *arXiv preprint* **2018**, DOI: 10.48550/ARXIV.1811.09766, (accessed 2025–12–07).
- (274) Hoffmann, M.; Noé, F. Generating valid Euclidean distance matrices. *arXiv preprint* **2019**, DOI: 10.48550/ARXIV.1910.03131, (accessed 2025–12–07).
- (275) Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. From target to drug: generative modeling for the multimodal structure-based ligand design. *Mol. Pharmaceutics* **2019**, *16*, 4282–4291.
- (276) Dinh, L.; Krueger, D.; Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint* **2014**, DOI: 10.48550/ARXIV.1410.8516, (accessed 2025–12–07).
- (277) Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using real nvp. *arXiv preprint* **2016**, DOI: 10.48550/ARXIV.1605.08803, (accessed 2025–12–07).
- (278) Kingma, D. P.; Dhariwal, P. Glow: Generative flow with invertible 1×1 convolutions. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
- (279) Madhawa, K.; Ishiguro, K.; Nakago, K.; Abe, M. GraphNVP: An Invertible Flow Model for Generating Molecular Graphs. *arXiv preprint* **2019**, DOI: 10.48550/ARXIV.1905.11600, (accessed 2025–12–07).
- (280) Zang, C.; Wang, F. Moflow: an invertible flow model for generating molecular graphs. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020; pp 617–626.
- (281) Honda, S.; Akita, H.; Ishiguro, K.; Nakanishi, T.; Oono, K. Graph Residual Flow for Molecular Graph Generation. *arXiv preprint* **2019**, DOI: 10.48550/ARXIV.1909.13521, (accessed 2025–12–07).
- (282) Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; Tang, J. GraphAF: A Flow-based Autoregressive Model for Molecular Graph Generation. *International Conference on Learning Representations*. 2020.
- (283) Luo, Y.; Yan, K.; Ji, S. Graphdf: A discrete flow model for molecular graph generation. *International Conference on Machine Learning*. 2021; pp 7192–7203.
- (284) Kuznetsov, M.; Polykovskiy, D. MolGrow: A graph normalizing flow for hierarchical molecular generation. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021; pp 8226–8234.
- (285) Zhu, Y.; Ouyang, Z.; Liao, B.; Wu, J.; Wu, Y.; Hsieh, C.-Y.; Hou, T.; Wu, J. MolHF: A Hierarchical Normalizing Flow for Molecular Graph Generation. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023; pp 5002–5010.
- (286) Frey, N. C.; Gadepally, V.; Ramsundar, B. Fastflows: Flow-based models for molecular graph generation. *arXiv preprint* **2022**, DOI: 10.48550/ARXIV.2201.12419, (accessed 2025–12–07).
- (287) Garcia Satorras, V.; Hoogeboom, E.; Fuchs, F.; Posner, I.; Welling, M. E (n) equivariant normalizing flows. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4181–4192.
- (288) Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554–2558.
- (289) Hinton, G. E. Learning distributed representations of concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*. 1986.
- (290) Du, Y.; Mordatch, I. Implicit generation and modeling with energy based models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

- (291) Nijkamp, E.; Hill, M.; Zhu, S.-C.; Wu, Y. N. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Adv. Neural Inf. Process. Syst.* **2019**, 32.
- (292) Liu, M.; Yan, K.; Oztekin, B.; Ji, S. GraphEBM: Molecular Graph Generation with Energy-Based Models. *Energy Based Models Workshop-ICLR 2021*. 2021.
- (293) Welling, M.; Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011; pp 681–688.
- (294) Pang, B.; Han, T.; Wu, Y. N. Learning Latent Space Energy-Based Prior Model for Molecule Generation. *arXiv preprint 2020*, DOI: 10.48550/ARXIV.2010.09351, (accessed 2025–12–07).
- (295) Li, J.; Beaudoin, C.; Ghosh, S. Energy-based generative models for target-specific drug discovery. *Front. Mol. Med.* **2023**, 3, 1160877.
- (296) Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **2019**, 32.
- (297) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*. 2021.
- (298) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 30.
- (299) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; others Improving language understanding by generative pre-training. 2018; https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed 2025–12–07).
- (300) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, 33, 1877–1901.
- (301) Bran, A. M.; Schwaller, P. *Drug Development Supported by Informatics*; Springer Nature Singapore: 2024; pp 143–163.
- (302) Wang, R.; Ji, Y.; Li, Y.; Lee, S.-T. Applications of Transformers in Computational Chemistry: Recent Progress and Prospects. *J. Phys. Chem. Lett.* **2025**, 16, 421–434.
- (303) Honda, S.; Shi, S.; Ueda, H. R. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. *arXiv preprint 2019*, DOI: 10.48550/ARXIV.1911.04738, (accessed 2025–12–07).
- (304) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, 5, 1572–1583.
- (305) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*. 2019; pp 429–436.
- (306) Li, J.; Jiang, X. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wirel. Commun. Mob. Comput.* **2021**, 2021, 7181815.
- (307) Priyadarsini, I.; Takeda, S.; Hamada, L.; Brazil, E. V.; Soares, E.; Shinohara, H. SELFIES-TED: A Robust Transformer Model for Molecular Representation using SELFIES. **2025**; <https://openreview.net/forum?id=uPj9oBH80V>.
- (308) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint 2020*, DOI: 10.48550/ARXIV.2010.09885, (accessed 2025–12–07).
- (309) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. learn.: sci. technol.* **2022**, 3, 015022.
- (310) Xue, D.; Zhang, H.; Chen, X.; Xiao, D.; Gong, Y.; Chuai, G.; Sun, Y.; Tian, H.; Wu, H.; Li, Y.; Liu, Q. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *Sci. Bull.* **2022**, 67, 899–902.
- (311) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, 14, 4099.
- (312) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. *npj Comput. Mater.* **2023**, 9, 64.
- (313) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* **2021**, 34, 28877–28888.
- (314) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Adv. Neural Inf. Process. Syst.* **2020**, 33, 12559–12571.
- (315) Wu, Z.; Jain, P.; Wright, M.; Mirhoseini, A.; Gonzalez, J. E.; Stoica, I. Representing long-range context for graph neural networks with global attention. *Adv. Neural Inf. Process. Syst.* **2021**, 34, 13266–13279.
- (316) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **2022**, 62, 2064–2076.
- (317) Fang, Y.; Zhang, N.; Chen, Z.; Guo, L.; Fan, X.; Chen, H. Domain-Agnostic Molecular Generation with Chemical Feedback. *The Twelfth International Conference on Learning Representations*. 2024.
- (318) Dollar, O.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **2021**, 12, 8362–8372.
- (319) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular optimization by capturing chemist's intuition using deep neural networks. *J. Cheminform.* **2021**, 13, 1–17.
- (320) He, J.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Patronov, A.; Bjerrum, E. J.; Engkvist, O. Transformer-based molecular optimization beyond matched molecular pairs. *J. Cheminform.* **2022**, 14, 18.
- (321) Wu, F.; Radev, D.; Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023; pp 5312–5320.
- (322) Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci. Rep.* **2021**, 11, 321.
- (323) Wang, Y.; Zhao, H.; Sciabola, S.; Wang, W. cMolGPT: a conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules* **2023**, 28, 4430.
- (324) Wang, J.; Hsieh, C.-Y.; Wang, M.; Wang, X.; Wu, Z.; Jiang, D.; Liao, B.; Zhang, X.; Yang, B.; He, Q. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach. Intell.* **2021**, 3, 914–922.
- (325) Born, J.; Manica, M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mach. Intell.* **2023**, 5, 432–444.
- (326) Yang, M.; Sun, H.; Liu, X.; Xue, X.; Deng, Y.; Wang, X. CMGN: a conditional molecular generation net to design target-specific molecules with desired properties. *Briefings in bioinformatics* **2023**, 24, bbad185.
- (327) Wu, K.; Xia, Y.; Deng, P.; Liu, R.; Zhang, Y.; Guo, H.; Cui, Y.; Pei, Q.; Wu, L.; Xie, S. TamGen: drug design with target-aware molecule generation through a chemical language model. *Nat. Commun.* **2024**, 15, 9360.
- (328) Chang, J.; Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nat. Commun.* **2024**, 15, 2323.
- (329) Dobberstein, N.; Maass, A.; Hamaekers, J. Llamol: a dynamic multi-conditional generative transformer for de novo molecular design. *J. Cheminform.* **2024**, 16, 73.
- (330) Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint 2023*, DOI: 10.48550/ARXIV.2307.09288, (accessed 2025–12–07).
- (331) Aksamit, N.; Hou, J.; Li, Y.; Ombuki-Berman, B. Integrating transformers and many-objective optimization for drug design. *BMC Med.* **2024**, 25, 208.

- (332) Ai, C.; Yang, H.; Liu, X.; Dong, R.; Ding, Y.; Guo, F. MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. *PLoS Comput. Biol.* **2024**, *20*, No. e1012229.
- (333) Ho, J.; Ermon, S. Generative adversarial imitation learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
- (334) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
- (335) Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022; pp 10684–10695.
- (336) Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
- (337) Krishna, R.; Wang, J.; Ahern, W.; Sturmels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **2024**, *384*, No. eadl2528, others.
- (338) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500 others.
- (339) Bao, F.; Zhao, M.; Hao, Z.; Li, P.; Li, C.; Zhu, J. Equivariant Energy-Guided SDE for Inverse Molecular Design. *The Eleventh International Conference on Learning Representations*. 2023.
- (340) Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant Diffusion for Molecule Generation in 3D. *Proceedings of the 39th International Conference on Machine Learning*. 2022; pp 8867–8887.
- (341) Xu, M.; Powers, A. S.; Dror, R. O.; Ermon, S.; Leskovec, J. Geometric latent diffusion models for 3d molecule generation. *International Conference on Machine Learning*. 2023; pp 38592–38610.
- (342) Huang, L.; Zhang, H.; Xu, T.; Wong, K.-C. MDM: Molecular Diffusion Model for 3D Molecule Generation. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 5105–5112.
- (343) Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; Ma, J. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. *The Eleventh International Conference on Learning Representations*. 2023.
- (344) Lin, H.; Huang, Y.; Zhang, O.; Ma, S.; Liu, M.; Li, X.; Wu, L.; Wang, J.; Hou, T.; Li, S. Z. DiffBP: Generative diffusion of 3d molecules for target protein binding. *Chem. Sci.* **2025**, *16*, 1417–1431.
- (345) Schneuing, A.; Harris, C.; Du, Y.; Didi, K.; Jamsab, A.; Igashov, I.; Du, W.; Gomes, C.; Blundell, T. L.; Lio, P. Structure-based drug design with equivariant diffusion models. *Nat. Comput. Sci.* **2024**, *4*, 899–909 others.
- (346) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.-C.; Zhang, H. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **2024**, *15*, 2657 others.
- (347) Lee, S.; Jo, J.; Hwang, S. J. Exploring chemical space with score-based out-of-distribution generation. *International Conference on Machine Learning*. 2023; pp 18872–18892.
- (348) Igashov, I.; Stärk, H.; Vignac, C.; Schneuing, A.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3D-conditional diffusion model for molecular linker design. *Nat. Mach. Intell.* **2024**, *6*, 417–427.
- (349) Xie, J.; Chen, S.; Lei, J.; Yang, Y. DiffDec: structure-aware scaffold decoration with an end-to-end diffusion model. *J. Chem. Inf. Model.* **2024**, *64*, 2554–2564.
- (350) Sako, M.; Yasuo, N.; Sekijima, M. DiffInt: A diffusion model for structure-based drug design with explicit hydrogen bond interaction guidance. *J. Chem. Inf. Model.* **2025**, *65*, 71–82.
- (351) Weiss, T.; Mayo Yanes, E.; Chakraborty, S.; Cosmo, L.; Bronstein, A. M.; Gershoni-Poranne, R. Guided diffusion for inverse molecular design. *Nat. Comput. Sci.* **2023**, *3*, 873–882.
- (352) Chen, Z.; Peng, B.; Zhai, T.; Adu-Ampratwum, D.; Ning, X. Generating 3D small binding molecules using shape-conditioned diffusing models with guidance. *Nat. Mach. Intell.* **2025**, *7*, 758.
- (353) Zou, Y.; Guo, T.; Fu, Z.; Guo, Z.; Bo, W.; Yan, D.; Wang, Q.; Zeng, J.; Xu, D.; Wang, T.; Chen, L. A structure-based framework for selective inhibitor design and optimization. *Commun. Biol.* **2025**, *8*, 422.
- (354) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.
- (355) Polykovskiy, D. et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*. DOI: 10.3389/fphar.2020.565644
- (356) Katoch, S.; Chauhan, S. S.; Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126.
- (357) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **1994**, *18*, 833–844.
- (358) O’Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational design and selection of optimal organic photovoltaic materials. *J. Phys. Chem. C* **2011**, *115*, 16200–16210.
- (359) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.
- (360) Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. Strategy to discover diverse optimal molecules in the small molecule universe. *J. Chem. Inf. Model.* **2015**, *55*, 529–537.
- (361) Kanal, I. Y.; Hutchison, G. R. Rapid computational optimization of molecular properties using genetic algorithms: searching across millions of compounds for organic photovoltaic materials. *arXiv preprint* **2017**, DOI: 10.48550/ARXIV.1707.02949, (accessed 2025–12–07).
- (362) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431–1434.
- (363) Gao, W.; Fu, T.; Sun, J.; Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21342–21357.
- (364) Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *International Conference on Learning Representations*. 2020.
- (365) Kwon, Y.; Lee, J. MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES. *J. Cheminform.* **2021**, *13*, 24.
- (366) Leguy, J.; Cauchy, T.; Glavatskikh, M.; Duval, B.; Da Mota, B. EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *J. Cheminform.* **2020**, *12*, 1–19.
- (367) Wang, J.; Wang, X.; Sun, H.; Wang, M.; Zeng, Y.; Jiang, D.; Wu, Z.; Liu, Z.; Liao, B.; Yao, X.; Hsieh, C.-Y.; Cao, D.; Chen, X.; Hou, T. ChemistGA: A Chemical Synthesizable Accessible Molecular Generation Algorithm for Real-World Drug Discovery. *J. Med. Chem.* **2022**, *65*, 12482–12496.
- (368) Ahn, S.; Kim, J.; Lee, H.; Shin, J. Guiding deep molecular optimization with genetic exploration. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12008–12021.
- (369) Lamanna, G.; Delre, P.; Marcou, G.; Saviano, M.; Varnek, A.; Horvath, D.; Mangiatordi, G. F. GENERA: a combined genetic/deep-learning algorithm for multiobjective target-oriented de novo design. *J. Chem. Inf. Model.* **2023**, *63*, 5107–5119.
- (370) Kwon, Y.; Kang, S.; Choi, Y.-S.; Kim, I. Evolutionary design of molecules based on deep learning and a genetic algorithm. *Sci. Rep.* **2021**, *11*, 17304.
- (371) Blaschke, M.; Pauly, F. Designing mechanosensitive molecules from molecular building blocks: a genetic algorithm-based approach. *J. Chem. Phys.* **2023**, *159*. DOI: 10.1063/5.0155012

- (372) Wang, H.; Skreta, M.; Ser, C.-T.; Gao, W.; Kong, L.; Strieth-Kalthoff, F.; Duan, C.; Zhuang, Y.; Yu, Y.; Zhu, Y.; Du, Y.; Aspuru-Guzik, A.; Neklyudov, K.; Zhang, C. Efficient Evolutionary Search Over Chemical Space with Large Language Models. *arXiv preprint* **2025**, DOI: 10.48550/ARXIV.2406.16976, (accessed 2025-12-07).
- (373) Nigam, A.; Pollice, R.; Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digit. Discovery* **2022**, *1*, 390–404.
- (374) Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. d. P.; Aspuru-Guzik, A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **2021**, *12*, 7079–7090.
- (375) Eberhart, R.; Kennedy, J. Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks*. 1995; pp 1942–1948.
- (376) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016–8024.
- (377) Yi, J.-C.; Yang, Z.-Y.; Zhao, W.-T.; Yang, Z.-J.; Zhang, X.-C.; Wu, C.-K.; Lu, A.-P.; Cao, D.-S. ChemMORT: an automatic ADMET optimization platform using deep learning and multi-objective particle swarm optimization. *Brief. Bioinform.* **2024**, *25*, bbae008.
- (378) 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*, 148–175.
- (379) Williams, C. K.; Rasmussen, C. E. *Gaussian processes for machine learning*; MIT Press: Cambridge, MA, 2006.
- (380) Kushner, H. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *J. Basic Eng.* **1964**, *86*, 97–106.
- (381) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455–492.
- (382) Srinivas, N.; Krause, A.; Kakade, S. M.; Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory* **2012**, *58*, 3250–3265.
- (383) Chapelle, O.; Li, L. An empirical evaluation of thompson sampling. *Adv. Neural Inf. Process. Syst.* **2011**, *24*.
- (384) Ju, S.; Shiga, T.; Feng, L.; Hou, Z.; Tsuda, K.; Shiomi, J. Designing nanostructures for phonon transport via Bayesian optimization. *Phys. Rev. X* **2017**, *7*, 021024.
- (385) Yamashita, T.; Sato, N.; Kino, H.; Miyake, T.; Tsuda, K.; Oguchi, T. Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **2018**, *2*, 013803.
- (386) Motoyama, Y.; Tamura, R.; Yoshimi, K.; Terayama, K.; Ueno, T.; Tsuda, K. Bayesian optimization package: PHYSBO. *Comput. Phys. Commun.* **2022**, *278*, 108405.
- (387) Murakami, Y.; Ishida, S.; Demizu, Y.; Terayama, K. Design of antimicrobial peptides containing non-proteinogenic amino acids using multi-objective Bayesian optimization. *Digit. Discovery* **2023**, *2*, 1347–1353.
- (388) Amamoto, Y.; Koganemaru, C.; Kojio, K.; Takahara, A.; Yamamoto, S.; Okazawa, K.; Tsuji, Y.; Aritake, T.; Terayama, K. A machine learning approach to designing and understanding tough, degradable polyamides. *npj Comput. Mater.* **2025**, *11*, 198.
- (389) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **2020**, *11*, 577–586.
- (390) Truong, G.-B.; Pham, T.-A.; To, V.-T.; Le, H.-S. L.; Van Nguyen, P.-C.; Trinh, T.-C.; Phan, T.-L.; Truong, T. N. Discovery of Vascular Endothelial Growth Factor Receptor 2 Inhibitors Employing Junction Tree Variational Autoencoder with Bayesian Optimization and Gradient Ascent. *ACS omega* **2024**, *9*, 47180–47193.
- (391) Matsukiyo, Y.; Yamanaka, C.; Yamanishi, Y. De novo generation of chemical structures of inhibitor and activator candidates for therapeutic target proteins by a transformer-based variational autoencoder and bayesian optimization. *J. Chem. Inf. Model.* **2024**, *64*, 2345–2355.
- (392) Mehta, S.; Laghuvarapu, S.; Pathak, Y.; Sethi, A.; Alvalla, M.; Priyakumar, U. D. Memes: Machine learning framework for enhanced molecular screening. *Chem. Sci.* **2021**, *12*, 11710–11721.
- (393) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **2021**, *12*, 7866–7881.
- (394) Korovina, K.; Xu, S.; Kandasamy, K.; Neiswanger, W.; Poczos, B.; Schneider, J.; Xing, E. ChemBO: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. 2020; pp 3393–3403.
- (395) Moss, H.; Leslie, D.; Beck, D.; Gonzalez, J.; Rayson, P. Boss: Bayesian optimization over string spaces. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15476–15486.
- (396) Sutton, R. S.; Barto, A. G.; others *Reinforcement learning: An introduction*; MIT Press Cambridge, 1998; Vol. 1.
- (397) Sutton, R. S.; Barto, A. G. *Reinforcement learning: An introduction*; MIT Press, 2018.
- (398) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with Deep Reinforcement Learning. *arXiv preprint* **2013**, DOI: 10.48550/ARXIV.1312.5602, (accessed 2025-12-07).
- (399) Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv preprint* **2017**, DOI: 10.48550/ARXIV.1707.06347, (accessed 2025-12-07).
- (400) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
- (401) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489.
- (402) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359.
- (403) Gu, S.; Holly, E.; Lillicrap, T.; Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. *2017 IEEE international conference on robotics and automation (ICRA)*. 2017; pp 3389–3396.
- (404) Hwangbo, J.; Lee, J.; Dosovitskiy, A.; Bellicoso, D.; Tsounis, V.; Koltun, V.; Hutter, M. Learning agile and dynamic motor skills for legged robots. *Sci. Robot.* **2019**, *4*, No. eaau5872.
- (405) Andrychowicz, O. M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A. Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* **2020**, *39*, 3–20 others.
- (406) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, DOI: 10.1186/s13321-017-0235-x
- (407) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 5918–5922.
- (408) Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. REINVENT 4: Modern AI-driven generative molecule design. *J. Cheminform.* **2024**, *16*, 20.
- (409) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, No. eaap7885.
- (410) Mazuz, E.; Shtar, G.; Shapira, B.; Rokach, L. Molecule generation using transformers and policy gradient reinforcement learning. *Sci. Rep.* **2023**, *13*, 8799.
- (411) He, J.; Tibo, A.; Janet, J. P.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Evaluation of reinforcement learning in transformer-based molecular design. *J. Cheminform.* **2024**, *16*, 95.
- (412) Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S. A survey of monte carlo tree search methods. *IEEE*

Transactions on Computational Intelligence and AI in games **2012**, *4*, 1–43.

(413) Coulom, R. Efficient selectivity and backup operators in Monte-Carlo tree search. *International conference on computers and games*. 2006; pp 72–83.

(414) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.

(415) Yoshizoe, K.; Kishimoto, A.; Kaneko, T.; Yoshimoto, H.; Ishikawa, Y. Scalable distributed monte-carlo tree search. *Proceedings of the International Symposium on Combinatorial Search*. 2011; pp 180–187.

(416) Yang, X.; Aasawat, T.; Yoshizoe, K. Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design. *International Conference on Learning Representations*. 2021.

(417) Sumita, M.; Yang, X.; Ishihara, S.; Tamura, R.; Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.* **2018**, *4*, 1126–1133.

(418) Zhang, J.; Terayama, K.; Sumita, M.; Yoshizoe, K.; Ito, K.; Kikuchi, J.; Tsuda, K. NMR-TS: de novo molecule identification from NMR spectra. *Sci. Technol. Adv. Mater.* **2020**, *21*, 552–561.

(419) Fujita, T.; Terayama, K.; Sumita, M.; Tamura, R.; Nakamura, Y.; Naito, M.; Tsuda, K. Understanding the evolution of a de novo molecule generator via characteristic functional group monitoring. *Sci. Technol. Adv. Mater.* **2022**, *23*, 352–360.

(420) Ma, B.; Terayama, K.; Matsumoto, S.; Isaka, Y.; Sasakura, Y.; Iwata, H.; Araki, M.; Okuno, Y. Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *J. Chem. Inf. Model.* **2021**, *61*, 3304–3313.

(421) Yoshizawa, T.; Ishida, S.; Sato, T.; Ohta, M.; Honma, T.; Terayama, K. Selective inhibitor design for kinase homologs using multiobjective Monte Carlo tree search. *J. Chem. Inf. Model.* **2022**, *62*, 5351–5360.

(422) Suzuki, T.; Ma, D.; Yasuo, N.; Sekijima, M. Mothra: Multiobjective de novo molecular generation using monte carlo tree search. *J. Chem. Inf. Model.* **2024**, *64*, 7291–7302.

(423) Yoshizawa, T.; Ishida, S.; Sato, T.; Ohta, M.; Honma, T.; Terayama, K. A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nat. Commun.* **2025**, *16*, 2409.

(424) Ishida, S.; Aasawat, T.; Sumita, M.; Katouda, M.; Yoshizawa, T.; Yoshizoe, K.; Tsuda, K.; Terayama, K. ChemTSv2: Functional molecular design using de novo molecule generator. *WIREs Comput. Mol. Sci.* **2023**, *13*, No. e1680.

(425) You, J.; Liu, B.; Ying, Z.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.

(426) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **2019**, *9*, 10752.

(427) Jin, W.; Barzilay, R.; Jaakkola, T. Multi-objective molecule generation using interpretable substructures. *International conference on machine learning*. 2020; pp 4849–4859.

(428) Iwata, H.; Nakai, T.; Koyama, T.; Matsumoto, S.; Kojima, R.; Okuno, Y. VGAE-MCTS: a new molecular generative model combining the variational graph auto-encoder and monte carlo tree search. *J. Chem. Inf. Model.* **2023**, *63*, 7392–7400.

(429) Simm, G.; Pinsler, R.; Hernández-Lobato, J. M. Reinforcement learning for molecular design guided by quantum mechanics. *International Conference on Machine Learning*. 2020; pp 8959–8969.

(430) Simm, G. N.; Pinsler, R.; Csányi, G.; Hernández-Lobato, J. M. Symmetry-Aware Actor-Critic for 3D Molecular Design. *International Conference on Learning Representations*. 2021.

(431) Li, Y.; Pei, J.; Lai, L. Structure-based de novo drug design using 3D deep generative models. *Chem. Sci.* **2021**, *12*, 13664–13675.

(432) Neil, D.; Segler, M.; Guasch, L.; Ahmed, M.; Plumbley, D.; Sellwood, M.; Brown, N. Exploring deep recurrent models with reinforcement learning for molecule design. *ICLR 2018 Workshop*. **2018**.

(433) Thomas, M.; O'Boyle, N. M.; Bender, A.; De Graaf, C. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. *J. Cheminform.* **2022**, *14*, 68.

(434) Dodds, M.; Guo, J.; Löhr, T.; Tibo, A.; Engkvist, O.; Janet, J. P. Sample efficient reinforcement learning with active learning for molecular design. *Chem. Sci.* **2024**, *15*, 4146–4160.

(435) Thiede, L. A.; Krenn, M.; Nigam, A.; Aspuru-Guzik, A. Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning. *Mach. learn.: sci. technol.* **2022**, *3*, 03S008.

(436) Chadi, M.-A.; Mousannif, H.; Aamouche, A. Curiosity as a Self-Supervised Method to Improve Exploration in De novo Drug Design. *2023 International Conference on Information Technology Research and Innovation (ICITRI)*. 2023; pp 151–156.

(437) Park, J.; Ahn, J.; Choi, J.; Kim, J. Mol-AIR: Molecular Reinforcement Learning with Adaptive Intrinsic Rewards for Goal-Directed Molecular Generation. *J. Chem. Inf. Model.* **2025**, *65*, 2283–2296.

(438) Chen, S.; Xie, J.; Ye, R.; Xu, D. D.; Yang, Y. Structure-aware dual-target drug design through collaborative learning of pharmacophore combination and molecular simulation. *Chem. Sci.* **2024**, *15*, 10366–10380.

(439) Lu, F.; Li, M.; Min, X.; Li, C.; Zeng, X. De novo generation of dual-target ligands using adversarial training and reinforcement learning. *Brief. Bioinform.* **2021**, *22*, bbab333.

(440) Bengio, E.; Jain, M.; Korablyov, M.; Precup, D.; Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems* **2021**, *34*, 27381–27394.

(441) Zhu, Y.; Wu, J.; Hu, C.; Yan, J.; Hou, T.; Wu, J. Sample-efficient multi-objective molecular optimization with glowntes. *Advances in Neural Information Processing Systems* **2023**, *36*, 79667–79684.

(442) Roy, J.; Bacon, P.-L.; Pal, C.; Bengio, E. Goal-conditioned glowntes for controllable multi-objective molecular design. *arXiv preprint arXiv:2306.04620* **2023**, DOI: 10.48550/ARXIV.2306.04620, (accessed 2025–12–07).

(443) Cretu, M.; Harris, C.; Igashov, I.; Schneuing, A.; Segler, M.; Correia, B.; Roy, J.; Bengio, E.; Liò, P. Synflownet: Design of diverse and novel molecules with synthesis constraints. *arXiv preprint* **2024**, DOI: 10.48550/ARXIV.2405.01155, (accessed 2025–12–07).

(444) Kondratyev, V.; Dryzhakov, M.; Gimadiev, T.; Slutskiy, D. Generative model based on junction tree variational autoencoder for HOMO value prediction and molecular optimization. *J. Cheminform.* **2023**, *15*, 11.

(445) Eckmann, P.; Sun, K.; Zhao, B.; Feng, M.; Gilson, M. K.; Yu, R. Limo: Latent inceptionism for targeted molecule generation. *Proc. Mach. Learn. Res.* **2022**, *162*, 5777.

(446) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of International Conference on Learning Representations*. 2015.

(447) Fu, T.; Gao, W.; Xiao, C.; Yasonik, J.; Coley, C. W.; Sun, J. Differentiable scaffolding tree for molecular optimization. *Proceedings of the 10th International Conference on Learning Representations*. 2022.

(448) Therrien, F.; Sargent, E. H.; Voznyy, O. Using GNN property predictors as molecule generators. *Nat. Commun.* **2025**, *16*, DOI: 10.1038/s41467-025-59439-1

(449) Liu, J. S.; Liu, J. S. *Monte Carlo strategies in scientific computing*; Springer, 2001; Vol. 10.

(450) Del Moral, P.; Doucet, A.; Jasra, A. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2006**, *68*, 411–436.

(451) Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des.* **2017**, *31*, 379–391.

(452) Miyao, T.; Kaneko, H.; Funatsu, K. Ring-System-Based Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol. Inform.* **2014**, *33*, 764–778.

- (453) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286–299.
- (454) Fu, T.; Xiao, C.; Li, X.; Glass, L. M.; Sun, J. Mimosa: Multi-constraint molecule sampling for molecule optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021; pp 125–133.
- (455) Xie, Y.; Shi, C.; Zhou, H.; Yang, Y.; Zhang, W.; Yu, Y.; Li, L. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. *International Conference on Learning Representations*. 2021.
- (456) Zhang, Q.; Liu, C.; Wu, S.; Hayashi, Y.; Yoshida, R. A Bayesian method for concurrently designing molecules and synthetic reaction networks. *Sci. Technol. Adv. Mater.: Methods* **2023**, *3*, 2204994.
- (457) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.
- (458) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (459) Allen, W. J.; Balias, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* **2015**, *36*, 1132–1156.
- (460) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891–3898.
- (461) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J. Cheminform.* **2021**, *13*, 43.
- (462) Lyu, J.; Wang, S.; Balias, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Alga, E.; Tolmachova, K. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224–229.
- (463) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580*, 663–668 others.
- (464) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452–459 others.
- (465) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M.; Hernández-Lobato, J. M. A Model to Search for Synthesizable Molecules. *Adv. Neural Inf. Process. Syst.* **2019**.
- (466) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M.; Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis DAGs. *Adv. Neural Inf. Process. Syst.* **2020**, 6852–6866.
- (467) Gao, W.; Mercado, R.; Coley, C. W. Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. *International Conference on Learning Representations*. 2022.
- (468) Yin, X.; Wang, X.; Wu, Z.; Li, Q.; Kang, Y.; Deng, Y.; Luo, P.; Liu, H.; Shi, G.; Wang, Z.; Yao, X.; Hsieh, C.-Y.; Hou, T. Syn-MolOpt: a synthesis planning-driven molecular optimization method using data-derived functional reaction templates. *J. Cheminform.* **2025**, *17*, 27.
- (469) Nguyen, D. H.; Tsuda, K. Generating reaction trees with cascaded variational autoencoders. *J. Chem. Phys.* **2022**, *156*. DOI: 10.1063/5.0076749
- (470) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search. *Proceedings of the 37th International Conference on Machine Learning*. 2020; pp 1608–1616.
- (471) Horwood, J.; Noutahi, E. Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. *ACS Omega* **2020**, *5*, 32984–32994.
- (472) Gottipati, S. K.; Sattarov, B.; Niu, S.; Pathak, Y.; Wei, H.; Liu, S.; Liu, S.; Blackburn, S.; Thomas, K.; Coley, C.; Tang, J.; Chandar, S.; Bengio, Y. Learning to Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning. *Proceedings of the 37th International Conference on Machine Learning*. 2020; pp 3668–3679.
- (473) Swanson, K.; Liu, G.; Catacutan, D. B.; Arnold, A.; Zou, J.; Stokes, J. M. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* **2024**, *6*, 338–353.
- (474) Kao, P.-Y.; Yang, Y.-C.; Chiang, W.-Y.; Hsiao, J.-Y.; Cao, Y.; Aliper, A.; Ren, F.; Aspuru-Guzik, A.; Zhavoronkov, A.; Hsieh, M.-H.; Lin, Y.-C. Exploring the Advantages of Quantum Generative Adversarial Networks in Generative Chemistry. *J. Chem. Inf. Model.* **2023**, *63*, 3307–3318.
- (475) Ghazi Vakili, M. Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors. *Nat. Biotechnol.* **2025**, *43*, 1954.
- (476) Gili, K.; Hibat-Allah, M.; Mauri, M.; Ballance, C.; Perdomo-Ortiz, A. Do quantum circuit Born machines generalize? *Quantum Sci. Technol.* **2023**, *8*, 035021.
- (477) Tučs, A.; Berenger, F.; Yumoto, A.; Tamura, R.; Uzawa, T.; Tsuda, K. Quantum Annealing Designs Nonhemolytic Antimicrobial Peptides in a Discrete Latent Space. *ACS Med. Chem. Lett.* **2023**, *14*, 577–582.
- (478) Kitai, K.; Guo, J.; Ju, S.; Tanaka, S.; Tsuda, K.; Shiomi, J.; Tamura, R. Designing metamaterials with quantum annealing and factorization machines. *Phys. Rev. Res.* **2020**, *2*. DOI: 10.1103/PhysRevResearch.2.013319
- (479) Mao, Z.; Matsuda, Y.; Tamura, R.; Tsuda, K. Chemical design with GPU-based Ising machines. *Digit. Discovery* **2023**, *2*, 1098–1103.
- (480) Gao, Q.; Jones, G. O.; Kobayashi, T.; Sugawara, M.; Yamashita, H.; Kawaguchi, H.; Tanaka, S.; Yamamoto, N. Quantum-Classical Computational Molecular Design of Deuterated High-Efficiency OLED Emitters. *Intell. Comput.* **2023**, *2*. DOI: 10.34133/icomputing.0037
- (481) Farhi, E.; Goldstone, J.; Gutmann, S. A Quantum Approximate Optimization Algorithm. *arXiv preprint* **2014**, DOI: 10.48550/ARXIV.1411.4028, (accessed 2025–12–07).
- (482) Kocsis, L.; Szepesvári, C. Bandit Based Monte-Carlo Planning. *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*. Berlin, Heidelberg, 2006; pp 282–293.
- (483) Świechowski, M.; Godlewski, K.; Sawicki, B.; Mańdziuk, J. Monte Carlo Tree Search: A Review of Recent Modifications and Applications. *Artif. Intell. Rev.* **2023**, *56*, 2497–2562.
- (484) Kajita, S.; Kinjo, T.; Nishi, T. Autonomous Molecular Design by Monte-Carlo Tree Search and Rapid Evaluations Using Molecular Dynamics Simulations. *Commun. Phys.* **2020**, *3*, 77.
- (485) Chaslot, G.; Winands, M.; van Den Herik, H. Parallel Monte-Carlo Tree Search. *Computers and Games* **2008**, *5131*, 60–71.
- (486) Romein, J. W.; Plaat, A.; Bal, H. E.; Schaeffer, J. Transposition Table Driven Work Scheduling in Distributed Search. *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*. 1999; pp 725–731.
- (487) Kishimoto, A.; Fukunaga, A.; Botea, A. Evaluation of a Simple, Scalable, Parallel Best-First Search Strategy. *Artif. Intell.* **2013**, *195*, 222–248.
- (488) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **2021**, *49*, W5–W14 others.
- (489) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **2019**, *10*, 7913–7922.
- (490) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **2021**, *7*, 1356–1367.
- (491) Nigam, A.; Pollice, R.; Hurley, M. F.; Hickman, R. J.; Aldeghi, M.; Yoshikawa, N.; Chithrananda, S.; Voelz, V. A.; Aspuru-Guzik, A. Assigning confidence to molecular property prediction. *Expert Opin. Drug Discovery* **2021**, *16*, 1009–1023.
- (492) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z.; Hou, T. End-point binding free energy calculation with MM/PBSA and

MM/GBSA: strategies and applications in drug design. *Chem. Rev.* **2019**, *119*, 9478–9508.

(493) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703 others.

(494) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; Van Vlijmen, H.; Tresadern, G.; De Groot, B. L. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **2020**, *11*, 1140–1152.

(495) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dulak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002 others.

(496) Ingman, V. M.; Schaefer, A. J.; Andreola, L. R.; Wheeler, S. E. QChASM: Quantum chemistry automation and structure manipulation. *WIREs Comput. Mol. Sci.* **2021**, *11*, No. e1510.

(497) Sumita, M.; Terayama, K.; Tamura, R.; Tsuda, K. QCforever: a quantum chemistry wrapper for everyone to use in black-box optimization. *J. Chem. Inf. Model.* **2022**, *62*, 4427–4434.

(498) Sumita, M.; Terayama, K.; Ishida, S.; Suga, K.; Saito, S.; Tsuda, K. Qcforever2: Advanced Automation of Quantum Chemistry Computations. *J. Comput. Chem.* **2025**, *46*, No. e70017.

(499) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*. DOI: 10.1186/1758-2946-1-8

(500) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

(501) Bühlmann, S.; Reymond, J.-L. ChEMBL-Likeness Score and Database GDBChEMBL. *Frontiers in Chemistry* **2020**, *8*. DOI: 10.3389/fchem.2020.00046

(502) Vanij Jayaseelan, K.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Med.* **2012**, *13*. DOI: 10.1186/1471-2105-13-106

(503) Mendez, D.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(504) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. *PLoS One* **2011**, *6*, No. e15939.

(505) Chen, S.; Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *J. Cheminform.* **2024**, *16*. DOI: 10.1186/s13321-024-00879-0

(506) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252–261.

(507) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminform.* **2020**, *12*. DOI: 10.1186/s13321-020-00439-2

(508) Yu, J.; Wang, J.; Zhao, H.; Gao, J.; Kang, Y.; Cao, D.; Wang, Z.; Hou, T. Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. *J. Chem. Inf. Model.* **2022**, *62*, 2973–2986.

(509) Wang, S.; Wang, L.; Li, F.; Bai, F. DeepSA: a deep-learning driven predictor of compound synthesis accessibility. *J. Cheminform.* **2023**, *15*. DOI: 10.1186/s13321-023-00771-3

(510) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021**, *12*, 3339–3349.

(511) Choung, O.-H.; Vianello, R.; Segler, M.; Stiefl, N.; Jiménez-Luna, J. Extracting medicinal chemistry intuition via preference machine learning. *Nat. Commun.* **2023**, *14*. DOI: 10.1038/s41467-023-42242-1

(512) Neeser, R. M.; Correia, B.; Schwaller, P. FScore: A Personalized Machine Learning-Based Synthetic Feasibility Score. *Chemistry-Methods* **2024**, *4*. DOI: 10.1002/cmted.202400024

(513) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed. Engl.* **2016**, *55*, 5904–5937.

(514) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(515) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **2020**, *12*. DOI: 10.1186/s13321-020-00472-1

(516) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge. *J. Chem. Inf. Model.* **2022**, *62*, 1357–1367.

(517) IBM RXN for Chemistry. <https://rxn.app.accelerate.science/rxn/>, 2025; (accessed 2025–12–07).

(518) molecule.one. <https://molecule.one/>, 2025 (accessed 2025–12–07).

(519) Tu, Z.; et al. ASKCOS: Open-Source, Data-Driven Synthesis Planning. *Acc. Chem. Res.* **2025**, *58*, 1764–1775.

(520) Guo, J.; Schwaller, P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chem. Sci.* **2025**, *16*, 6943–6956.

(521) Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *First Conference on Language Modeling*. 2024.

(522) Gao, W.; Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **2020**, *60*, 5714–5723.

(523) Ferreira, L. L.; Andricopulo, A. D. ADMET modeling approaches in drug discovery. *Drug Discovery Today* **2019**, *24*, 1157–1165.

(524) Cummins, D. J.; Bell, M. A. Integrating everything: the molecule selection toolkit, a system for compound prioritization in drug discovery. *J. Med. Chem.* **2016**, *59*, 6999–7010.

(525) Fromer, J. C.; Coley, C. W. Computer-aided multi-objective optimization in small molecule discovery. *Patterns* **2023**, *4*, 100678.

(526) Zhou, A.; Qu, B.-Y.; Li, H.; Zhao, S.-Z.; Suganthan, P. N.; Zhang, Q. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm Evol. Comput.* **2011**, *1*, 32–49.

(527) Liu, X.; Ye, K.; van Vlijmen, H. W.; Emmerich, M. T.; IJzerman, A. P.; van Westen, G. J. DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. *J. Cheminform.* **2021**, *13*, 85.

(528) Verhellen, J. Graph-based molecular Pareto optimization. *Chem. Sci.* **2022**, *13*, 7526–7535.

(529) Alberga, D.; Lamanna, G.; Graziano, G.; Delre, P.; Lomuscio, M. C.; Corriero, N.; Ligresti, A.; Siliqi, D.; Saviano, M.; Contino, M. DeLA-DrugSelf: Empowering multi-objective de novo design through SELFIES molecular representation. *Comput. Biol. Med.* **2024**, *175*, 108486 others.

(530) SV, S. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St. John, P. C. Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell.* **2022**, *4*, 720–730.

(531) Bustillo, L.; Laino, T.; Rodrigues, T. The rise of automated curiosity-driven discoveries in chemistry. *Chem. Sci.* **2023**, *14*, 10378–10384.

(532) Lehman, J.; Stanley, K. O. Abandoning objectives: Evolution through the search for novelty alone. *Evol. Comput.* **2011**, *19*, 189–223.

(533) Baranes, A.; Oudeyer, P.-Y. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems* **2013**, *61*, 49–73.

(534) Grizou, J.; Points, L. J.; Sharma, A.; Cronin, L. A curious formulation robot enables the discovery of a novel protocell behavior. *Sci. Adv.* **2020**, *6*, No. eaay4237.

- (535) Terayama, K.; Sumita, M.; Tamura, R.; Payne, D. T.; Chahal, M. K.; Ishihara, S.; Tsuda, K. Pushing property limits in materials discovery via boundless objective-free exploration. *Chem. Sci.* **2020**, *11*, 5959–5968.
- (536) Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; Dragan, A. Inverse reward design. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
- (537) Skalse, J.; Howe, N.; Krashennikov, D.; Krueger, D. Defining and characterizing reward gaming. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9460–9471.
- (538) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies* **2019**, *32*, 55–63.
- (539) Langevin, M.; Vuilleumier, R.; Bianciotto, M. Explaining and avoiding failure modes in goal-directed generation of small molecules. *J. Cheminform.* **2022**, *14*, 20.
- (540) Gendreau, P.; Turk, J.-A.; Drizard, N.; Ribeiro da Silva, V. B.; Descamps, C.; Gaston-Mathe, Y. Molecular assays simulator to unravel predictors hacking in goal-directed molecular generations. *J. Chem. Inf. Model.* **2023**, *63*, 3983–3998.
- (541) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals* **2005**, *33*, 155–173 others.
- (542) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
- (543) Hanser, T.; Barber, C.; Marchaland, J.; Werner, S. Applicability domain: towards a more formal definition. *SAR and QSAR in Environmental Research* **2016**, *27*, 865–881.
- (544) Kaneko, H.; Funatsu, K. Applicability domains and consistent structure generation. *Mol. Inform.* **2017**, *36*, 1600032.
- (545) Langevin, M.; Grebner, C.; Gussregen, S.; Sauer, S.; Li, Y.; Matter, H.; Bianciotto, M. Impact of applicability domains to generative artificial intelligence. *ACS omega* **2023**, *8*, 23148–23167.
- (546) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y. H.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.
- (547) Nigam, A.; Pollice, R.; Tom, G.; Jorner, K.; Willes, J.; Thiede, L.; Kundaje, A.; Aspuru-Guzik, A. Tartarus: A Benchmarking Platform for Realistic And Practical Inverse Molecular Design. *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023.
- (548) Weng, G.; Zhao, H.; Nie, D.; Zhang, H.; Liu, L.; Hou, T.; Kang, Y. RediscMol: Benchmarking Molecular Generation Models in Biological Properties. *J. Med. Chem.* **2024**, *67*, 1533–1543.
- (549) Sanjrani, N.; Coupry, D. E.; Pogány, P.; Palmer, D. S.; Pickett, S. D. Benchmarking 3D Structure-Based Molecule Generators. *J. Chem. Inf. Model.* **2025**, *65*, 8006–8021.
- (550) Yue, T.; Tao, L.; Varshney, V.; Li, Y. Benchmarking study of deep generative models for inverse polymer design. *Digit. Discovery* **2025**, *4*, 910.
- (551) Stavenga, D. G.; DeGrip, W., Jr, E. P., Eds. *Handbook of Biological Physics, Molecular Mechanisms in Visual Transduction*; Elsevier: Amsterdam, 2000.
- (552) Nibbering, E. T.; Fidler, H.; Pines, E. ULTRAFast CHEMISTRY: Using Time-Resolved Vibrational Spectroscopy for Interrogation of Structural Dynamics. *Annu. Rev. Phys. Chem.* **2005**, *56*, 337–367.
- (553) Buhrke, D.; Hildebrandt, P. Probing Structure and Reaction Dynamics of Proteins Using Time-Resolved Resonance Raman Spectroscopy. *Chem. Rev.* **2020**, *120*, 3577–3630.
- (554) Glish, G. L.; Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nat. Rev. Drug Discovery* **2003**, *2*, 140–150.
- (555) Elyashberg, M.; Argyropoulos, D. Computer Assisted Structure Elucidation (CASE): Current and future perspectives. *Magn. Reson. Chem.* **2021**, *59*, 669–690.
- (556) Lu, X.-Y.; Wu, H.-P.; Ma, H.; Li, H.; Li, J.; Liu, Y.-T.; Pan, Z.-Y.; Xie, Y.; Wang, L.; Ren, B.; Liu, G.-K. Deep Learning-Assisted Spectrum-Structure Correlation: State-of-the-Art and Perspectives. *Anal. Chem.* **2024**, *96*, 7959–7975.
- (557) Westermayr, J.; Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *J. Chem. Phys.* **2020**, *153*, 154112.
- (558) Mamede, R.; Pereira, F.; Aires-de Sousa, J. Machine learning prediction of UV-Vis spectra features of organic compounds related to photoreactive potential. *Sci. Rep.* **2021**, *11*, 23720.
- (559) Urbina, F.; Batra, K.; Luebke, K. J.; White, J. D.; Matsiev, D.; Olson, L. L.; Malerich, J. P.; Hupcey, M. A. Z.; Madrid, P. B.; Ekins, S. UV-adVISor: Attention-Based Recurrent Neural Networks to Predict UV-Vis Spectra. *Anal. Chem.* **2021**, *93*, 16076–16085.
- (560) Long, S.; Fu, J.; Jian, J.; Fan, Z.; Fan, Q.; Xie, F.; Zhang, Y.; Ma, J. Spectroscopy learning: A machine learning method for study diatomic vibrational spectra including dissociation behavior. *MethodsX* **2020**, *7*, 101127.
- (561) Kirchner, B.; Blasius, J.; Esser, L.; Reckien, W. Predicting Vibrational Spectroscopy for Flexible Molecules and Molecules with Non-Idle Environments. *Advanced Theory and Simulations* **2021**, *4*, 2000223.
- (562) Han, R.; Ketkaew, R.; Luber, S. A Concise Review on Recent Developments of Machine Learning for the Prediction of Vibrational Spectra. *J. Phys. Chem. A* **2022**, *126*, 801–812.
- (563) Zhang, W.; Kasun, L. C.; Wang, Q. J.; Zheng, Y.; Lin, Z. A Review of Machine Learning for Near-Infrared Spectroscopy. *Sensors* **2022**, *22*, 9764.
- (564) Chong, Y.; Huo, Y.; Jiang, S.; Wang, X.; Zhang, B.; Liu, T.; Chen, X.; Han, T.; Smith, P. E. S.; Wang, S.; Jiang, J. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, No. e2220789120.
- (565) Miotto, M.; Monacelli, L. Fast prediction of anharmonic vibrational spectra for complex organic molecules. *npj Comput. Mater.* **2024**, *10*, 240.
- (566) Grumet, M.; von Scarpatetti, C.; Bučko, T.; Egger, D. A. Delta Machine Learning for Predicting Dielectric Properties and Raman Spectra. *J. Phys. Chem. C* **2024**, *128*, 6464–6470.
- (567) Saquer, N.; Iqbal, R.; Ellis, J. D.; Yoshimatsu, K. Infrared spectra prediction using attention-based graph neural networks. *Digit. Discovery* **2024**, *3*, 602–609.
- (568) Cortés, I.; Cuadrado, C.; Hernández Daranas, A.; Sarotti, A. M. Machine learning in computational NMR-aided structural elucidation. *Front. Nat. Prod.* **2023**, *2*, 1122426.
- (569) McNoughton, A. D.; Joshi, R. P.; Knutson, C. R.; Fnu, A.; Luebke, K. J.; Malerich, J. P.; Madrid, P. B.; Kumar, N. Machine Learning Models for Predicting Molecular UV-Vis Spectra with Quantum Mechanical Properties. *J. Chem. Inf. Model.* **2023**, *63*, 1462–1471.
- (570) Kang, S.; Kwon, Y.; Lee, D.; Choi, Y.-S. Predictive Modeling of NMR Chemical Shifts without Using Atomic-Level Annotations. *J. Chem. Inf. Model.* **2020**, *60*, 3765–3769.
- (571) Yang, Z.; Chakraborty, M.; White, A. D. Predicting chemical shifts with graph neural networks. *Chem. Sci.* **2021**, *12*, 10802–10809.
- (572) Rull, H.; Fischer, M.; Kuhn, S. NMR shift prediction from small data quantities. *J. Cheminform.* **2023**, *15*, 114.
- (573) Leniak, A.; Pietruś, W.; Świdarska, A.; Kurczab, R. From NMR to AI: Do We Need 1 H NMR Experimental Spectra to Obtain High-Quality logD Prediction Models? *J. Chem. Inf. Model.* **2025**, *65*, 2924–2939.
- (574) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **2016**, *8*, 3.
- (575) De Jonge, N. F.; Louwen, J. J. R.; Chekmeneva, E.; Camuzeaux, S.; Vermeir, F. J.; Jansen, R. S.; Huber, F.; Van Der Hooft, J. J. J.

MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nat. Commun.* **2023**, *14*, 1752.

(576) Litsa, E. E.; Chenthamarakshan, V.; Das, P.; Kaviraki, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.* **2023**, *6*, 132.

(577) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **2022**, *19*, 865–870.

(578) McCarthy, M.; Lee, K. L. K. Molecule Identification with Rotational Spectroscopy and Probabilistic Deep Learning. *J. Phys. Chem. A* **2020**, *124*, 3002–3017.

(579) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral deep learning for prediction and prospective validation of functional groups. *Chem. Sci.* **2020**, *11*, 4618–4630.

(580) Alberts, M.; Laino, T.; Vaucher, A. C. Leveraging infrared spectroscopy for automated structure elucidation. *Commun. Chem.* **2024**, *7*, 268.

(581) Huang, Z.; Chen, M. S.; Woroch, C. P.; Markland, T. E.; Kanan, M. W. A framework for automated structure elucidation from routine NMR spectra. *Chem. Sci.* **2021**, *12*, 15329–15338.

(582) Sridharan, B.; Mehta, S.; Pathak, Y.; Priyakumar, U. D. Deep Reinforcement Learning for Molecular Inverse Problem of Nuclear Magnetic Resonance Spectra to Molecular Structure. *J. Phys. Chem. Lett.* **2022**, *13*, 4924–4933.

(583) Devata, S.; Sridharan, B.; Mehta, S.; Pathak, Y.; Laghuvarapu, S.; Varma, G.; Priyakumar, U. D. DeepSPIN - deep reinforcement learning for molecular structure prediction from infrared and 13 C NMR spectra. *Digit. Discovery* **2024**, *3*, 818–829.

(584) Chacko, E.; Praveen, R. S. A.; Luska, K. L.; Varagas-Hernández, R. A. Spectro: A Multi-Modal Approach for Molecule Elucidation Using IR and NMR data. *ChemRxiv preprint* **2024**, DOI: 10.26434/chemrxiv-2024-37v2j, (accessed 2025–12–07).

(585) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073.

(586) Wishart, D. S.; et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617.

(587) Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20.

(588) McEachran, A. D.; Mansouri, K.; Grulke, C.; Schymanski, E. L.; Ruttkies, C.; Williams, A. J. “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. *J. Cheminform.* **2018**, *10*, 45.

(589) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16*, 299–302.

(590) Guélin, M.; Cernicharo, J. Organic Molecules in Interstellar Space: Latest Advances. *Front. Astron. Space Sci.* **2022**, *9*, DOI: 10.3389/fspas.2022.787567

(591) Peironcelly, J. E.; Rojas-Chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. OMG: Open Molecule Generator. *J. Cheminform.* **2012**, *4*, 21.

(592) Kim, H.; Cho, M. Infrared Probes for Studying the Structure and Dynamics of Biomolecules. *Chem. Rev.* **2013**, *113*, 5817–5847.

(593) Linstrom, P. J.; Mallard, W. G., Eds. *NIST Chemistry WebBook*, NIST Standard Reference Database Number 69; 2005.

(594) NIST NIST Standard Reference Database 35. 2010; <https://www.nist.gov/srd/nist-standard-reference-database-35> (accessed 2025–12–07).

(595) Villane, C. *Optimal Transport: Old and New*; Springer-Verlag: Berlin, 2009.

(596) Rubner, Y.; Tomasi, C.; Guibas, L. The Earth Mover’s Distance as a Metric For Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121.

(597) Gupta, A.; Chakraborty, S.; Ramakrishnan, R. Revving up ¹³C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach. learn.: sci. technol.* **2021**, *2*, 035010.

(598) Patiny, L.; Jeannerat, D.; Wenk, M.; Musallam, H.; Wist, J.; Liermann, J.; Schloerer, N.; Blanos, A.; Kostro, D.; Zasso, M. Visualize, analyze and Process NMR Spectra Online. 2010; <https://www.nmr.org/> (accessed 2025–12–07).

(599) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(600) Adelusi, T. I.; Oyedele, A.-Q. K.; Boyenle, I. D.; Ogunlana, A. T.; Adeyemi, R. O.; Ukachi, C. D.; Idris, M. O.; Olaoba, O. T.; Adedotun, I. O.; Kolawole, O. E.; Xiaoxing, Y.; Abdul-Hammed, M. Molecular modeling in drug discovery. *Informatics in Medicine Unlocked* **2022**, *29*, 100880.

(601) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061.

(602) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

(603) Durrant, J. D.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biology* **2011**, *9*, 71.

(604) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(605) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.

(606) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(607) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annu. Rev. Biophys.* **2001**, *30*, 211–243.

(608) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458–465.

(609) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.

(610) Boniolo, F.; Dorigatti, E.; Ohnmacht, A. J.; Saur, D.; Schubert, B.; Menden, M. P. Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opin. Drug Discovery* **2021**, *16*, 991–1007.

(611) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers* **2021**, *25*, 1315–1360.

(612) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37*, DOI: 10.1002/minf.201700153

(613) Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **2018**, *1*, 1–9.

(614) Grisoni, F.; Huisman, B. J. H.; Button, A. L.; Moret, M.; Atz, K.; Merk, D.; Schneider, G. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **2021**, *7*, No. eabg3338.

(615) Yu, Y.; Xu, T.; Li, J.; Qiu, Y.; Rong, Y.; Gong, Z.; Cheng, X.; Dong, L.; Liu, W.; Li, J.; Dou, D.; Huang, J. A Novel Scalarized Scaffold Hopping Algorithm with Graph-Based Variational Autoencoder for Discovery of JAK1 Inhibitors. *ACS Omega* **2021**, *6*, 22945–22954.

(616) Ren, F.; et al. A small-molecule TNF inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **2025**, *43*, 63–75.

(617) Ballarotto, M.; Willems, S.; Stiller, T.; Nawa, F.; Marschner, J. A.; Grisoni, F.; Merk, D. De Novo Design of Nurr1 Agonists via Fragment-Augmented Generative Deep Learning in Low-Data Regime. *J. Med. Chem.* **2023**, *66*, 8170–8177.

(618) Moret, M.; Pachon Angona, I.; Cotos, L.; Yan, S.; Atz, K.; Brunner, C.; Baumgartner, M.; Grisoni, F.; Schneider, G. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **2023**, *14*, DOI: 10.1038/s41467-022-35692-6

- (619) Chenthamarakshan, V.; et al. Accelerating drug target inhibitor discovery with a deep generative foundation model. *Sci. Adv.* **2023**, *9*, No. eadg7865.
- (620) Yoshimori, A.; Asawa, Y.; Kawasaki, E.; Tasaka, T.; Matsuda, S.; Sekikawa, T.; Tanabe, S.; Neya, M.; Natsugari, H.; Kanai, C. Design and Synthesis of DDR1 Inhibitors with a Desired Pharmacophore Using Deep Generative Models. *ChemMedChem*. **2021**, *16*, 955–958.
- (621) Abe, K.; Ozako, M.; Inukai, M.; Matsuyuki, Y.; Kitayama, S.; Kanai, C.; Nagai, C.; Gopalasingam, C. C.; Gerle, C.; Shigematsu, H.; Umekubo, N.; Yokoshima, S.; Yoshimori, A. Deep learning driven de novo drug design based on gastric proton pump structures. *Commun. Biol.* **2023**, *6*, 1–14.
- (622) Jiang, X.; Lu, L.; Li, J.; Jiang, J.; Zhang, J.; Zhou, S.; Wen, H.; Cai, H.; Luo, X.; Li, Z.; Wang, J.; Ju, B.; Bai, R. Synthetically Feasible De Novo Molecular Design of Leads Based on a Reinforcement Learning Model: AI-Assisted Discovery of an Anti-IBD Lead Targeting CXCR4. *J. Med. Chem.* **2024**, *67*, 10057–10075.
- (623) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398–4405.
- (624) Munson, B. P.; Chen, M.; Bogosian, A.; Kreisberg, J. F.; Licon, K.; Abagyan, R.; Kuenzi, B. M.; Ideker, T. De novo generation of multi-target compounds using deep generative chemistry. *Nat. Commun.* **2024**, *15*, 3636.
- (625) Yasuda, K.; Berenger, F.; Amaike, K.; Ueda, A.; Nakagomi, T.; Hamasaki, G.; Li, C.; Otani, N. Y.; Kaitoh, K.; Tsuda, K.; Itami, K.; Yamanishi, Y. De novo generation of dual-target compounds using artificial intelligence. *iScience* **2025**, *28*, 111526.
- (626) Szanto, A.; Narkar, V.; Shen, Q.; Uray, I. P.; Davies, P. J. A.; Nagy, L. Retinoid X receptors: X-ploring their (patho)physiological functions. *Cell Death Differ.* **2004**, *11*, S126–S143.
- (627) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* **2018**, *37*, 1700111.
- (628) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4067–4072.
- (629) Asinex. <https://www.asinex.com> (accessed 2025–12–07).
- (630) ChemBridge screening compound collection. <https://chembridge.com> (accessed 2025–12–07).
- (631) Enamine advanced and HTS. <https://enamine.net/> (accessed 2025–12–07).
- (632) Specs screening compounds. <https://specs.net/> (accessed 2025–12–07).
- (633) Ivanenkov, Y. A.; Polykovskiy, D.; Bezrukov, D.; Zagribelnyy, B.; Aladinskiy, V.; Kamyra, P.; Aliper, A.; Ren, F.; Zhavoronkov, A. Chemistry42: An AI-Driven Platform for Molecular Design and Optimization. *J. Chem. Inf. Model.* **2023**, *63*, 695–701.
- (634) Vietor, J.; Gege, C.; Stiller, T.; Busch, R.; Schallmayer, E.; Kohlhof, H.; Höfner, G.; Pabel, J.; Marschner, J. A.; Merk, D. Development of a Potent Nurrl Agonist Tool for In Vivo Applications. *J. Med. Chem.* **2023**, *66*, 6391–6402.
- (635) Moret, M.; Helmstädter, M.; Grisoni, F.; Schneider, G.; Merk, D. Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence. *Angew. Chem., Int. Ed. Engl.* **2021**, *60*, 19477–19482.
- (636) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. *Chem. Sci.* **2020**, *11*, 3316–3325.
- (637) Sun, X.; et al. Tumour DDR1 promotes collagen fibre alignment to instigate immune exclusion. *Nature* **2021**, *599*, 673–678.
- (638) Wolber, G.; Langer, T. *J. Chem. Inf. Model* **2005**, *45*, 160–169.
- (639) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- (640) Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *arXiv preprint* **2018**, DOI: 10.48550/ARXIV.1801.01290, (accessed 2025–12–07).
- (641) Gaulton, A.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
- (642) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045–D1053.
- (643) Stathias, V.; et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.* **2020**, *48*, D431–D439.
- (644) Sheils, T. K.; et al. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res.* **2021**, *49*, D1334–D1346.
- (645) Li, C.; Yamanaka, C.; Kaitoh, K.; Yamanishi, Y. Transformer-based Objective-reinforced Generative Adversarial Network to Generate Desired Molecules. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 2022; pp 3884–3890.
- (646) Berenger, F.; Tsuda, K. Molecular generation by Fast Assembly of (Deep)SMILES fragments. *J. Cheminform.* **2021**, *13*, 88.
- (647) Korablyov, M.; et al. Generative Active Learning for the Search of Small-molecule Protein Binders. *arXiv preprint* **2024**, DOI: 10.48550/ARXIV.2405.01616, (accessed 2025–12–07).
- (648) Zhang, Y.; Zhang, J.; Suzuki, K.; Sumita, M.; Terayama, K.; Li, J.; Mao, Z.; Tsuda, K.; Suzuki, Y. Discovery of polymer electret material via de novo molecule generation and functional group enrichment analysis. *Appl. Phys. Lett.* **2021**, *118*, 223904.
- (649) Zhang, Z.; Kummeth, A. L.; Yang, J. Y.; Alexandrova, A. N. Inverse molecular design of alkoxides and phenoxides for aqueous direct air capture of CO₂. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2123496119.
- (650) Sumita, M.; Terayama, K.; Suzuki, N.; Ishihara, S.; Tamura, R.; Chahal, M. K.; Payne, D. T.; Yoshizoe, K.; Tsuda, K. De novo creation of a naked-eye-detectable fluorescent molecule based on quantum-chemical computation and machine learning. *Sci. Adv.* **2022**, *8*, No. eabj3906.
- (651) Matsuzawa, N. N.; Maeshima, H.; Hayashi, K.; Ando, T.; Afzal, M. A. F.; Marshall, K.; Coscia, B. J.; Browning, A. R.; Goldberg, A.; Halls, M. D.; Leswing, K.; Misra, M.; Ramezanghorbani, F.; Morisato, T. Exploring Molecules with Low Viscosity: Using Physics-Based Simulations and De Novo Design by Applying Reinforcement Learning. *Chem. Mater.* **2024**, *36*, 11706–11716.
- (652) Tan, J. D.; Ramalingam, B.; Chellappan, V.; Gupta, N. K.; Dillard, L.; Khan, S. A.; Galvin, C.; Hippalgaonkar, K. Generative Design and Experimental Validation of Non-Fullerene Acceptors for Photovoltaics. *ACS Energy Lett.* **2024**, *9*, 5240–5250.
- (653) Strieth-Kalthoff, F.; et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science* **2024**, *384*, No. eadk9227.
- (654) Tian, F.; Xu, G.; Zhou, S.; Chen, S.; He, D. Principles and applications of green fluorescent protein-based biosensors: a mini-review. *Analyst* **2023**, *148*, 2882–2891.
- (655) Zimmer, M. Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior. *Chem. Rev.* **2002**, *102*, 759–782.
- (656) Wüthrich, P.; Choong, J. J.; Yuki, S. InFrag: Using Attribution-Based Explainability to Guide Deep Molecular Optimization. *ChemRxiv preprint* **2021**, DOI: 10.26434/chemrxiv-2021-qtq8d, (accessed 2025–12–07).
- (657) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics. *Joule* **2017**, *1*, 857–870.
- (658) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project:

Large-Scale Computational Screening and Design of Organic Photo-voltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

(659) Tom, G.; et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124*, 9633–9732.

(660) Yoshikawa, N.; et al. Self-driving laboratories in Japan. *Digit. Discovery* **2025**, *4*, 1384–1403.

(661) Rauschen, R.; Guy, M.; Hein, J. E.; Cronin, L. Universal chemical programming language for robotic synthesis repeatability. *Nat. Synth.* **2024**, *3*, 488–496.

(662) Wu, T. C.; Aguilar-Granda, A.; Hotta, K.; Yazdani, S. A.; Pollice, R.; Vestfrid, J.; Hao, H.; Lavigne, C.; Seifrid, M.; Angello, N.; Bencheikh, F.; Hein, J. E.; Burke, M.; Adachi, C.; Aspuru-Guzik, A. A Materials Acceleration Platform for Organic Laser Discovery. *Adv. Mater.* **2023**, *35*. DOI: 10.1002/adma.202207070

(663) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578–1597.

(664) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **2021**, *3*, 76–86.

(665) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

(666) 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*, 9927–10000.

(667) Jin, H.; Merz, K. M. LigandDiff: de Novo Ligand Design for 3D Transition Metal Complexes with Diffusion Models. *J. Chem. Theory Comput.* **2024**, *20*, 4377–4384.

(668) Tran, H.; Gurnani, R.; Kim, C.; Pilania, G.; Kwon, H.-K.; Lively, R. P.; Ramprasad, R. Design of functional and sustainable polymers assisted by artificial intelligence. *Nat. Rev. Mater.* **2024**, *9*, 866–886.

(669) Yeh, A. H.-W.; et al. De novo design of luciferases using deep learning. *Nature* **2023**, *614*, 774–780.

(670) Listov, D.; Goverde, C. A.; Correia, B. E.; Fleishman, S. J. Opportunities and challenges in design and optimization of protein function. *Nat. Rev. Mol. Cell Biol.* **2024**, *25*, 639–653.

(671) Park, H.; Li, Z.; Walsh, A. Has generative artificial intelligence solved inverse materials design? *Matter* **2024**, *7*, 2355–2367.

(672) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem. Sci.* **2020**, *11*, 4351–4359.

(673) Hibi, Y.; Uesaka, S.; Naito, M. A data-driven sequencer that unveils latent “codons” in synthetic copolymers. *Chem. Sci.* **2023**, *14*, 5619–5626.

(674) Klukowski, P.; Riek, R.; Güntert, P. Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **2022**, *13*, 6151.

(675) Terayama, K.; Sumita, M.; Katouda, M.; Tsuda, K.; Okuno, Y. Efficient Search for Energetically Favorable Molecular Conformations against Metastable States via Gray-Box Optimization. *J. Chem. Theory Comput.* **2021**, *17*, 5419–5427.

(676) Terayama, K.; Osaki, Y.; Fujita, T.; Tamura, R.; Naito, M.; Tsuda, K.; Matsui, T.; Sumita, M. Koopmans’ Theorem-Compliant Long-Range Corrected (KTLC) Density Functional Mediated by Black-Box Optimization and Data-Driven Prediction for Organic Molecules. *J. Chem. Theory Comput.* **2023**, *19*, 6770–6781.

(677) Fiedler, L.; Shah, K.; Bussmann, M.; Cangi, A. Deep dive into machine learning density functional theory for materials science and chemistry. *Phys. Rev. Mater.* **2022**, *6*, 040301.

(678) Nagai, R.; Akashi, R.; Sugino, O. Machine-learning-based exchange correlation functional with physical asymptotic constraints. *Phys. Rev. Res.* **2022**, *4*, 013106.

(679) Kirkpatrick, J.; et al. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **2021**, *374*, 1385–1389.

(680) Yoshikawa, N.; Sumita, M. Automatic Differentiation for the Direct Minimization Approach to the Hartree-Fock Method. *J. Phys. Chem. A* **2022**, *126*, 8487–8493.

(681) Carson, R. *Silent Spring*; Penguin Modern Classics: London, 2000.