International Journal of
Molecular Sciences

MDPI

Review

Artificial Intelligence in Molecular Optimization: Current Paradigms and Future Frontiers

Xin Xia [1], Yajie Zhang [2], Xiangxiang Zeng [3], Xingyi Zhang [2], Chunhou Zheng [2] and Yansen Su [1,*]

[1] The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei 230601, China; xiaxin98@stu.ahu.edu.cn
[2] The Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China; yjzhang17719490727@163.com (Y.Z.); xyzhanghust@gmail.com (X.Z.); zhengch99@126.com (C.Z.)
[3] College of Computer Science and Electronic Engineering, Hunan University, Lushan Road, Changsha 410012, China; xzeng@hnu.edu.cn
* Correspondence: suyansen@ahu.edu.cn

Abstract: Molecular optimization plays a pivotal role in many domains since it holds promise for improving the properties of lead molecules. The advent of artificial intelligence (AI)-driven molecular optimization has revolutionized lead optimization workflows, which have significantly accelerated the development of drug candidates. However, AI models are also confronted with new challenges in practical molecular optimization, such as high-dimensional chemical space and data sparsity issues. This paper initially highlights the inherent benefits of molecular optimization in terms of optimizing the properties and maintaining the structural similarity of lead molecules, thereby highlighting its critical role in drug discovery. The next section systematically categorizes and analyzes existing AI-aided molecular optimization methods, comprising iterative search in discrete chemical space, end-to-end generation in continuous latent space, and iterative search in continuous latent space methods. Finally, we discuss the key challenges in AI-aided molecular optimization methods, including molecular representations, dataset selection, the properties to be optimized, and optimization algorithms, while proposing potential solutions and future research directions. In summary, this review provides a comprehensive analysis of existing representative AI-aided molecular optimization methods, thereby offering guidance for future research directions.

Keywords: molecular optimization; artificial intelligence; iterative search; end-to-end generation

Academic Editor: Bruno Rizzuti

Received: 10 April 2025
Revised: 7 May 2025
Accepted: 14 May 2025
Published: 19 May 2025

Citation: Xia, X.; Zhang, Y.; Zeng, X.; Zhang, X.; Zheng, C.; Su, Y. Artificial Intelligence in Molecular Optimization: Current Paradigms and Future Frontiers. *Int. J. Mol. Sci.* **2025**, *26*, 4878. https://doi.org/10.3390/ijms26104878

1. Introduction

In various engineering fields such as materials engineering, the chemical industry, and drug development, molecular optimization plays a pivotal role in enhancing molecular properties by modifying the structures of lead molecules [1,2]. The drug discovery pipeline comprises several critical stages, including target identification, lead compound screening, lead compound optimization, and preclinical and clinical validation, which is a time-consuming and expensive process with a high failure rate [3]. The inherent complexity of human pathophysiology, coupled with the vastness of chemical space, necessitates rigorous decision-making at each stage of the discovery process. Thus, it is imperative to employ specific techniques to accelerate the drug discovery process and enhance the success rate [4,5]. Computer-aided drug design has significantly advanced various key aspects of drug discovery through its applications in disease target prediction, virtual

screening, molecular optimization, etc. [6–8]. Among these applications, molecular optimization is one of the crucial steps in obtaining optimal drug candidates, with the aim of improving the properties of lead molecules, such as their biological activity properties and physicochemical properties, while preserving their structural features [9,10]. The strategic optimization of the unfavorable properties of lead molecules significantly increases their likelihood of success in subsequent preclinical and clinical evaluations [11,12]. Therefore, the development of efficient molecular optimization methods offers substantial potential for streamlining the drug discovery and development process.

In recent years, artificial intelligence (AI)-aided molecular optimization methods have been extensively developed, facilitating a more comprehensive exploration of the huge chemical space and holding promise for enhancing the drug discovery and development process [13–15]. For example, while conventional drug development takes around 12 years and costs USD 2.6 billion on average [16], Zhavoronkov et al. proposed a deep learning model to rapidly identify DDR1 kinase inhibitors in just 21 days, substantially reducing both time and cost [17]. In general, AI-based molecular optimization methods follow two processes: the selection or construction of appropriate chemical spaces, followed by the exploration of the space to identify target molecules. Within this framework, many combinatorial optimization techniques operating directly on discrete molecular representations, such as molecular sequences and graphs, have been proposed to optimize molecules. Furthermore, the integration of deep learning has introduced novel capabilities, enabling the construction of continuous latent spaces for chemical molecules. This advanced representation facilitates molecular optimization through continuous vector space manipulation, offering an alternative to traditional discrete optimization approaches.

However, AI-aided molecular optimization methods face both significant challenges and emerging opportunities in practical drug discovery applications. The primary limitation stems from the inherent constraints of conventional molecular representations, which bring different challenges for effective AI-driven optimization [18]. Moreover, the efficacy of these AI-driven methods is fundamentally contingent upon the availability and quality of relevant molecular datasets [19]. Well-curated datasets are essential for training optimization models and ensuring their applicability. Furthermore, practical molecular optimization must navigate the complex task of simultaneously enhancing multiple properties while maintaining critical structural constraints [20]. This multi-objective optimization problem necessitates the development of more efficient computational frameworks. In addition, the rapid advancement of AI technologies offers promising avenues for addressing these challenges. Novel machine learning architectures and optimization algorithms hold significant potential for developing more efficient and reliable molecular optimization methods [21].

In this review, we provide a comprehensive overview of AI-aided molecular optimization. We first emphasize the advantages of molecular optimization to highlight its significance in drug discovery. Then, we summarize recent AI-aided molecular optimization methods, categorizing them into two distinct paradigms: combinatorial optimization methods operating in discrete chemical space and deep learning models operating in continuous latent space. Furthermore, we compile the experimental results of various methods on the same experimental task to visually demonstrate the optimization performance of different approaches. Finally, we discuss the challenges associated with AI-aided molecular optimization in practical drug discovery and offer corresponding recommendations. Our aims are to provide insights into AI-aided molecular optimization and offer guidance for future directions in computational drug discovery.

2. Definition of Molecular Optimization

In the drug discovery process, molecular optimization represents a critical stage subsequent to the lead molecule screening stage, which focuses on the structural refinement of promising lead molecules to enhance their properties. Therefore, molecular optimization methods can optimize specific properties of a given molecule, leading to molecules with enhanced properties. For example, Jin et al. [22] established a benchmark optimization task that requires improving molecules with quantitative estimation of drug-likeness (QED) values ranging from 0.7 to 0.8 to achieve QED scores exceeding 0.9 while maintaining a structural similarity value larger than 0.4. It is worth noting that, compared to de novo molecular generation, molecular optimization beginning with the lead molecule can shorten the search process for finding target molecules. The definition of molecular optimization is formulated as follows:

Definition 1. *Given a lead molecule x, its associated properties are $p_1(x), \cdots, p_m(x)$, and the goal of molecular optimization is to generate a molecule y with properties $p_1(y), \cdots, p_m(y)$, satisfying*

$$\begin{cases} p_i(y) \succ p_i(x), i = 1, 2, \cdots, m, \\ sim(x, y) > \delta, \end{cases} \tag{1}$$

where $p_i(y) \succ p_i(x)$ indicates that $p_i(y)$ is better than $p_i(x)$. p_i represents a molecular property, which can encompass various physicochemical and pharmacological properties, such as QED, bioactivity, and synthetic accessibility. $sim(x, y)$ is the similarity between x and y, and δ is the threshold of similarity. A frequently used molecular similarity metric is the Tanimoto similarity [23] of Morgan fingerprints [24], which is shown in Equation (2):

$$sim(x, y) = \frac{fp(x) \cdot fp(y)}{|fp(x)|^2 + |fp(y)|^2 - fp(x) \cdot fp(y)}, \tag{2}$$

where fp represents the Morgan fingerprints of the molecule. A fundamental consideration in molecular optimization is the necessity of maintaining structural similarity between the optimized molecule and its lead compound. This similarity constraint serves a dual purpose. First, it effectively delineates the chemical space to be explored around the lead molecule, thereby enhancing search efficiency [25]. Second, it preserves crucial structural features that are essential for maintaining desirable physicochemical and biological properties [26]. The significance of structural similarity is reflected in its incorporation into numerous benchmark molecular optimization tasks. For example, one widely adopted benchmark task involves optimizing the penalized logP of molecules while maintaining a Tanimoto similarity larger than 0.4 [22]. Another extensively studied benchmark task aims to improve biological activity against the dopamine type 2 receptor (DRD2) while preserving a structural similarity value greater than 0.4 [22].

3. Current Approaches and Barriers

AI-aided molecular optimization methods typically involve two fundamental steps: (1) the construction of an implicit chemical space and (2) the implementation of an optimization approach to find the desired molecules within the implicit chemical space. Existing AI-aided molecular optimization methods can be broadly classified based on their operational spaces: discrete chemical spaces and continuous latent spaces. For discrete chemical space approaches, molecules are represented through discrete structural representations, such as molecular sequences or graph-based structures, enabling direct structural modifications. Conversely, continuous latent space methods employ encoder–decoder frameworks to transform molecules into continuous vector representations, facilitating optimization in a differentiable space. To systematically organize these methods, this section categorizes

these methods based on the constructed chemical spaces and the employed optimization algorithms. For an enhanced comparative analysis, Figure 1 shows the workflows of various AI-based molecular optimization methods, and Table 1 provides a comprehensive summary of representative AI-based molecular optimization methods, along with their molecular representations, data types, and optimization objectives.

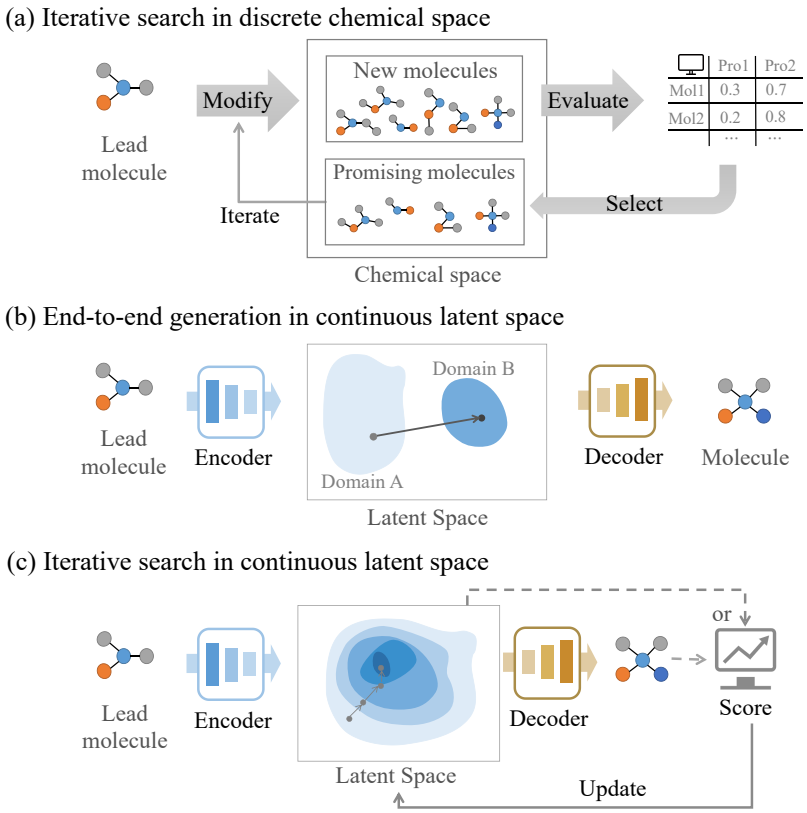

Figure 1. The workflows of artificial intelligence models for molecular optimization.

3.1. Molecular Optimization in Discrete Chemical Spaces

Molecular optimization methods operating in discrete chemical spaces employ direct structural modifications based on discrete representations, such as SMILES [27], SELFIES [28], and molecular graphs [29] (where nodes represent atoms, and edges represent chemical bonds). These methods typically explore the discrete chemical space through the following process: First, they generate a set of novel molecular structures through structural modifications, and then they select promising molecules for subsequent iterative optimization, as illustrated in Figure 1a. These methods can be primarily classified into genetic algorithm (GA)-based methods and reinforcement learning (RL)-based methods.

3.1.1. GA-Based Molecular Optimization Methods

Genetic algorithms (GAs) are heuristic optimization approaches that show competitive optimization performance to explore chemical spaces globally and locally. GA-based optimization methods begin with an initial population and generate new molecules through crossover and mutation operations. Then, molecules with high fitness are selected in the new population to guide the evolution process [30]. Some GA-based molecular optimization methods only mutate molecules while maintaining structural similarity. For example, STONED [31] generates offspring molecules by applying a random mutation on the SELFIES strings of molecules, which finds molecules with better properties. However, the absence of a crossover operator limits the global exploration of the vast chemical space. MolFinder [32] integrates crossover and mutation in the SMILES-based chemical space,

which enables both global search and local search. All of the aforementioned methods aggregate multiple properties into a single fitness function to guide the evolution, which requires predefined weights for multiple properties. In comparison, GB-GA-P [33] employs two Pareto-based genetic algorithms on molecular graphs, thereby enabling multi-objective molecular optimization to identify a set of Pareto-optimal molecules with enhanced properties.

In short, GAs have gained widespread adoption in molecular optimization due to their inherent flexibility, robustness, and ability to explore chemical space without requiring extensive training datasets. However, the efficacy of GA-based molecular optimization methods depends on the population size and the number of evolutionary generations, since repeated evaluations of molecular properties can be costly.

Table 1. Representative molecular optimization methods and their categories, molecular representations, data types, and optimization objectives.

Category	Model	Molecular Representation	Data Type	Optimization Objective	Citation
Iterative search in discrete space	STONED	SELFIES	Unpaired	Multi-property	[31]
	MolFinder	SMILES	Unpaired	Multi-property	[32]
	GB-GA-P	Graph	Unpaired	Multi-property	[33]
	GCPN	Graph	Unpaired	Single-property	[34]
	MolDQN	Graph	Unpaired	Multi-property	[35]
End-to-end generation in continuous space	CMG	SMILES	Paired	Multi-property	[36]
	T&S polish	Graph	Paired	Multi-property	[25]
	Mol-CycleGAN	Graph	Unpaired	Single-property	[37]
	UGMMT	SMILES	Unpaired	Single property	[38]
	IPCA	SMILES	Unpaired	Multi-property	[39]
	GPMO	SMILES	Paired	Multi-property	[40]
	VJTNN	Graph	Paired	Single-property	[22]
	SCVAE	Graph	Paired	Single-property	[41]
	Modef	Graph	Paired	Multi-property	[42]
	CFOM	SMILES	Unpaired	Single-property	[43]
	TamGen	SMILES	Unpaired	Single-property	[44]
Iterative search in continuous space	QMO	SMILES	Unpaired	Multi-property	[45]
	DST	Graph	Unpaired	Multi-property	[46]
	LIMO	SELFIES	Unpaired	Multi-property	[47]
	InversionGNN	Graph	Unpaired	Multi-property	[48]
	MOMO	SMILES	Unpaired	Multi-property	[49]
	DecompOpt	3D	Unpaired	Multi-property	[50]
	GCDM	3D	Unpaired	Multi-property	[51]
	Retmol	SMILES	Unpaired	Multi-property	[52]
	MO-LSO	Graph	Unpaired	Multi-property	[53]
	Prompt-MolOpt	SMILES	Paired	Multi-property	[54]
	Drugassist	SMILES	Unpaired	Multi-property	[55]

3.1.2. RL-Based Molecular Optimization Methods

Reinforcement learning (RL) [56] is a machine learning paradigm used to address decision-making problems, and it has shown potential in optimizing molecular properties by designing states, actions, and rewards. Most RL-based molecular optimization methods operate on molecular graphs. For example, GCPN [34] formalizes molecular optimization as a Markov decision process, which modifies molecules by adding atoms or fragments and connecting them with bonds. Additionally, GCPN uses a policy network to predict actions, which integrates molecular properties and an adversarial loss as rewards to update the policy gradients. In comparison, MolDQN [35] directly applies actions to the molecular graph and ensures molecular validity by chemically valid actions. MolDQN trains a Deep Q network to estimate the rewards, which enables it to discover molecules with enhanced properties.

Overall, RL-based molecular optimization methods facilitate active exploration of the chemical space beyond the training data. These approaches typically define the molecular modification process as a Markov process that performs sequential modifications to refine the molecular structure. However, the iterative modification can be inefficient due to the large number of available substructures in chemical space.

3.1.3. Analysis of Molecular Optimization Methods in Discrete Space

GA-based molecular optimization methods exhibit strong flexibility and broad applicability. By leveraging population-based parallel search mechanisms, GAs can explore a wider chemical space while reducing the risk of converging to local optima. These methods only require molecular property evaluators to compute the fitness function, significantly reducing the dependence on labeled data. Consequently, they demonstrate superior task scalability and can be flexibly applied to various quantifiable molecular optimization scenarios. In contrast, RL-based molecular optimization methods maximize the global reward through interactions between actions and reward environments. However, RL typically requires a large number of iterations to converge, meaning that obtaining high-quality optimized molecules often demands substantial computational resources. It is worth noting that molecular optimization based on discrete representations has certain limitations. It relies on expert-designed modification rules. The generated molecular structures could be chemically invalid. The exploration efficiency of chemical space is relatively limited.

3.2. Molecular Optimization in Continuous Latent Spaces

The rapid advancement of deep learning (DL) techniques has opened up new opportunities for molecular optimization. DL-based molecular optimization methods leverage the powerful nonlinear representation capabilities of deep neural networks to extract complex chemical knowledge from extensive molecular datasets, thereby facilitating the construction of continuous latent spaces. These methods typically employ an encoder–decoder framework, where an encoder transforms discrete molecules into continuous latent space, which enables them to efficiently modify the continuous vector of the lead molecule to obtain new vectors, and a decoder maps these new vectors back to discrete chemical space to obtain novel molecular structures with enhanced properties. In this subsection, we categorize molecular optimization methods in continuous latent spaces into end-to-end generation methods and iterative search methods.

3.2.1. End-to-End Generation Methods

End-to-end generation molecular optimization methods typically employ a deep learning architecture comprising an encoder–decoder framework (Figure 1b). These methods directly generate optimized molecular structures as output through the input of a lead

molecule, and they can be further classified into translation-based methods and conditional generation-based methods. Translation-based methods learn the translation rules from matched molecular pairs or sets, which enables the model to map input lead molecules to their optimized structures. Conditional generation-based methods integrate additional condition features (e.g., target properties or structural constraints) with the lead molecule to guide the generation of novel molecular structures with desired properties.

Translation-based methods. Inspired by the conceptual analogy between molecular optimization and translation tasks in natural language processing, many translation-based molecular optimization methods have been proposed to facilitate the transformation of lead molecules into target molecules [57]. For example, CMG [36] treats molecular optimization as a sequence-to-sequence translation problem, and it employs a Transformer framework with two constraint networks to generate structurally similar molecules based on SMILES. This approach relies heavily on molecular sequence representation and the Transformer architecture. In comparison, Graph Polish [25] adopts molecular graph representation, which translates lead molecules to optimized molecules through two modules, i.e., a pre-labeling module and a translation module. To be specific, the pre-labeling module identifies the optimization centers and label branches in the molecules, while the translation module trains a deep neural network from the labeled molecules to translate the target molecules. The graph-based approach emphasizes the structural integrity and topological features of molecules.

While most translation-based molecular optimization methods rely on paired molecules for supervised learning, several unsupervised translation-based methods have been developed to address the challenge of limited paired data. For example, Mol-CycleGAN [37] leverages the CycleGAN framework in the latent space of the JT-VAE codec [58], which divides the training data into low- and high-property domains to facilitate adversarial learning across these two domains. Similarly, UGMMT [38] employs CycleGAN to learn the translation rules based on molecular SMILES representations. Although both Mol-CycleGAN and UGMMT can translate lead molecules to target molecules with improved properties, their optimization capabilities are limited to a single molecular property. In comparison, IPCA [39] extends UGMMT by introducing an integrated poly-cycle architecture that concurrently optimizes multiple properties. This approach translates molecules through a shared latent embedding space and a central decoder, thereby allowing for the optimization of two properties. Additionally, translation-based methods also face challenges such as exposure bias, where the generation of molecules depends on the previously predicted outputs [59]. To mitigate this problem, GPMO [40] integrates contrastive learning into the Transformer framework to translate desired molecules while reducing exposure bias.

In summary, translation-based molecular optimization methods learn transition rules from matched molecular pairs or sets, which enables end-to-end optimization by directly generating optimized molecules from input lead molecules. However, these methods face notable limitations, particularly the scarcity of molecular data that simultaneously satisfy the multiple property conditions required for effective model training. Furthermore, although transformation rules can be inferred from matched molecular sets categorized by low and high property values, the lack of explicit structural guidance may impede the optimization process.

Conditional generation-based methods. Several molecular optimization methods generate molecules with enhanced properties by integrating the features of lead molecules with specific conditions, such as the structures or properties of the target molecules. For example, VJTNN [22] employs a graph message passing network to encode both the molecular graphs and junction trees of paired molecules, i.e., the lead molecule and its

corresponding target molecule. The features of the target molecule are extracted as conditions, which are subsequently fused with the latent vector of the lead molecule to generate new molecules. Later on, SCVAE [41] leverages the graph alignment for paired molecules, which incorporates structural similarity as a condition during the decoding process to produce target molecules. However, both VJTNN and SCVAE require the encoding and decoding of entire molecular graphs, which introduces significant learning challenges due to computational complexity and data requirements.

In contrast, Modef [42] simplifies this process by encoding only the differences between paired molecular graphs as conditional inputs. This approach not only reduces the number of parameters but also minimizes the amount of training data required, thereby enhancing computational efficiency and scalability. CFOM [43] decomposes the lead molecule into a molecular core and molecular chains. Utilizing a core encoder and a chains generator, CFOM generates novel chains, which are subsequently attached to the core to produce new molecules with enhanced properties. Furthermore, in recent years, several studies have used the structure of target proteins to generate target-aware molecules. For example, the TamGen framework [44] processes the geometric data of amino acids to generate protein representations while simultaneously incorporating molecular SMILES to derive molecular embeddings. The protein representation is subsequently utilized as a conditional to output optimized molecules by a compound decoder.

Conditional generation-based molecular optimization methods generate optimized molecules by incorporating specific conditions on properties or structures. These methods generate high-quality molecules by leveraging the conditions to guide the optimization process. However, a notable limitation of these approaches is the prerequisite of obtaining the target conditions prior to model training.

3.2.2. Iterative Search Methods

Iterative search-based molecular optimization methods in continuous latent space typically explore the space through step-by-step optimization to identify superior molecular continuous vectors (Figure 1c). When iterative search-based methods generate a set of molecules, these molecules are evaluated and selected to update the molecular continuous vector or to retrain the generator model for iterative optimization. There are several representation iterative search-based methods, which are introduced below.

For example, QMO [45] decouples the molecular representation learning and the guided search processes by using a pre-trained encoder–decoder framework. This framework evaluates molecular properties in discrete space and approximates gradients in continuous space by a model-independent zero-gradient descent method. However, the accuracy of the approximated gradients can significantly impact the search process. There are some methods that compute gradients based on the property values predicted in continuous latent spaces. For example, DST [46] trains graph neural networks on molecular differentiable scaffold tree representation to predict properties, which updates the scaffold tree of molecules by propagating local derivatives. In addition, LIMO [47] integrates a SELFIES-based VAE with a property prediction network, which facilitates rapid gradient-based optimization. InversionGNN [48] is a sample-efficient, dual-path graph neural network (GNN)-based framework designed for multi-objective molecular optimization. In its direct prediction path, InversionGNN leverages a GNN to extract knowledge from differentiable molecular scaffolding trees, enabling accurate property prediction. Subsequently, it employs gradient-based Pareto optimization to approximate molecules along the Pareto front.

There are several iterative search-based methods that explore the continuous latent space without updating the gradient. For example, MOMO [49], combines a pre-trained

encoder–decoder with a Pareto-based evolutionary algorithm to collaboratively evolve molecules between implicit space and discrete space. DecompOpt [50] uses diffusion models to capture molecular grammar in a data-driven manner, and it integrates iterative optimization to generate molecules with desired properties. Similarly, Morhead et al. [51] developed a geometry-complete diffusion model (GCDM), which learns the essential geometric properties of 3D molecules, enabling the generation of valid 3D molecular structures. The GCDM achieves property-guided 3D molecular optimization by iteratively accepting the generated molecules as intermediate states.

There are also existing methods that iteratively update the database for search-based optimization. For example, Retmol [52] samples high-quality molecules from a predefined retrieval dataset, which are combined with the lead molecule to obtain optimized molecules. The generated molecules are dynamically added to the retrieval database for iterative optimization. MO-LSO [53] employs an iterative weighted retraining strategy, which progressively refines the generative model to generate desired molecules. Specifically, MO-LSO performs Pareto ranking on the training molecules and assigns weights for these molecules based on their ranks. Then, it trains the generative model based on the weighted dataset to produce enhanced molecules. The newly generated molecules are ranked to update the training set, which is further used to refine the generative model.

In addition, recently, the advent of large language models (LLMs) has spurred their application in molecular optimization. For example, Prompt-MolOpt [54] integrates large language models (LLMs) with Transformer architectures to enhance molecular optimization capabilities. It employs an iterative fragment-based optimization strategy; i.e., at each step, a single molecular substructure is modified, which has demonstrated potential in multi-property molecular optimization. DrugAssist [55] is an interactive molecular optimization framework that iteratively refines molecular structures through human–AI dialogue. After an optimized molecule is generated by DrugAssist, its properties are evaluated. If the molecule meets the predefined property requirements, the process terminates. If not, DrugAssist retrieves molecules from the database that are the most structurally similar to the lead molecule and satisfy the property constraints, guiding further optimization until the desired molecular properties are achieved.

In summary, iterative search-based molecular optimization methods mitigate the reliance of deep learning models on extensive training data by incrementally identifying molecules with enhanced properties through a step-by-step optimization process. However, this iterative nature inherently renders these methods more computationally intensive and time-consuming.

3.2.3. Analysis of Molecular Optimization Methods in Continuous Space

Compared to molecular optimization in discrete chemical space, the construction of continuous chemical spaces enables a more efficient and smooth exploration of high-quality molecules [60]. Among continuous-space approaches, end-to-end generation methods offer faster optimization speeds. Once the models are trained on relevant datasets, they can perform batch molecular optimization in a single forward step. However, the optimization capability of end-to-end methods heavily depends on the quality of the training data and the training process [61]. When the training data are limited, the optimized molecules tend to exhibit relatively lower property values. Iterative search in continuous space has emerged as a popular new paradigm in molecular optimization in recent years. These methods require fewer labeled target molecules for training and can search for better molecules through iterative property evaluation. However, they typically involve multiple rounds of iterative optimization for a single lead molecule, resulting in longer optimization times.

3.3. Optimization Performance Comparison

To systematically evaluate the optimization performance of different AI-based molecular optimization methods, this study selected three representative benchmark tasks and integrated results from the literature with partially reproduced experimental data. The experimental designs for the three optimization tasks are as follows:

Task 1: PlogP optimization task

The objective of Task 1 is to maximize the penalized logP (PlogP) property value of molecules while maintaining a Tanimoto similarity of at least 0.4 with the lead molecules. The experiment uses the benchmark dataset constructed by Jin et al. [22], which consists of 800 molecules with low PlogP values selected from the ZINC database. The evaluation metric for this task is the average PlogP improvement of the optimized molecules.

Task 2: QED optimization task

The goal of Task 2 is to improve the QED of molecules while preserving a similarity value of at least 0.4 with the lead molecules. The test set proposed by Jin et al. [22]contains 800 molecules with QED values ranging from 0.7 to 0.8. The evaluation metric for this task is the optimization success rate, defined as the proportion of lead molecules whose QED is improved to above 0.9 while maintaining a similarity value larger than 0.4 among all tested lead molecules.

Task 3: Multi-property optimization task

Task 3 involves optimizing four properties: QED, synthetic accessibility (SA), the estimated inhibition score against the glycogen synthase kinase-3β target (GSK3β inhibition), and the estimated inhibition score against the c-Jun N-terminal kinase-3 target (JNK3 inhibition). The evaluation metric is the average property score (APS) of the top 100 generated molecules.

Figure 2 presents a performance comparison of different molecular optimization methods across three benchmark tasks. To ensure experimental fairness, we adopted uniform evaluation criteria and clearly state the source of each result. In Figure 2a, the average property improvement values of twelve methods in the PlogP optimization task are displayed. In the figure, the results for MolFinder, GB-GA-P, and MOMO were obtained by our reproduction of the experiments under the same oracle call settings, while the other results were extracted from the original publications. The results show that the iterative search methods achieved superior PlogP improvement while maintaining molecular similarity. Figure 2b compares eight methods in the QED optimization task. The MolFinder and GB-GA-P results came from our reproduction experiments with matched oracle calls, and the others were sourced from the original papers. The iterative search methods again showed better performance in QED optimization. Figure 2c presents the results of five methods on Task 3, all of which are cited from the original publications. Among these methods, InverseGNN exhibited the best comprehensive optimization across all four properties. Notably, due to architectural differences between the models, the training datasets varied across the methods (detailed in the original references). All reproduction experiments strictly followed the hyperparameter settings recommended in the original papers.

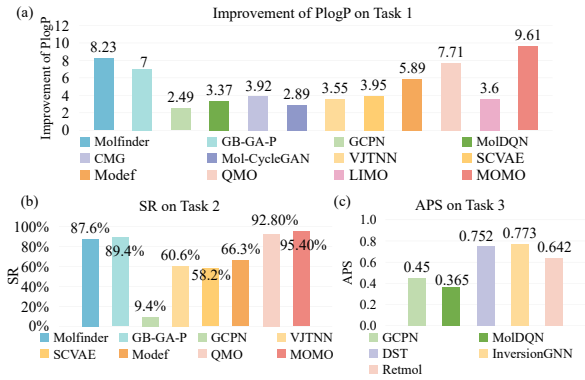

Figure 2. The performance of some existing methods on three optimization tasks. (**a**) The average PlogP improvement on Task 1. (**b**) The success rate (SR) on Task 2. (**c**) The average property score (APS) on Task 3.

4. Crucial Considerations and Future Opportunities

4.1. Reasonable Molecular Representation

Molecular optimization relies on several widely used molecular representations, the quality of which significantly influences the performance of molecular optimization methods. In this section, we outline the key characteristics of ideal molecular representations and provide a detailed analysis of the existing representations employed in molecular optimization.

4.1.1. Informative Molecular Representations

An ideal molecular representation should be highly informative, which will enable optimization methods to capture abundant molecular features. Molecular sequence representations, such as SMILES and SELFIES, have been extensively utilized in drug design due to their simplicity and interpretability. Cheng et al. [62] introduced Group SELFIES, a fragment-based molecular representation method designed to effectively capture chemical motifs and structural flexibility through string encoding. However, these representations often lack detailed structural information, leading to significant structural variations, even with minor sequence changes [63]. In contrast, graph-based representations offer greater robustness by efficiently encoding chemical interatomic connectivity. Despite this advantage, molecular graphs fail to capture certain critical features, such as the bond angles between atoms. Molecular image representations provide richer information by incorporating bond angles and positional information [64], yet they still fall short of fully representing molecules, which are inherently 3D quantum mechanical objects [65]. While 3D representations include spatial information that better reflects molecular geometry, the added complexity of 3D data significantly increases the computational challenges associated with model learning [66].

4.1.2. Modifiable Molecular Representations

Ideal molecular representations should facilitate easy modifications while preserving the chemical validity of molecules. The molecular SMILES representation is simple and easy to learn; however, it does not guarantee molecular validity, as minor changes in a single character can lead to significant structural modifications [34]. Although some researchers have attempted to address this issue by incorporating grammatical constraints into the encoding–decoding process, they still struggle to ensure the validity of the generated molecules [67]. In contrast, SELFIES [28] achieves 100% validity by separating information related to branches and rings, but it falls short when representing complex, crystalline, and large molecules [68]. Graph-based molecular optimization methods can generate molecules with high validity; however, their modeling processes are more com-

plex than those of sequence-based methods [69]. Image-based molecular optimization encounters challenges in generating valid structures from optimized images. Additionally, 3D molecular representations face difficulties in capturing translation, rotation, and reflection invariance.

4.2. Appropriate Datasets

The availability of appropriate datasets poses a significant challenge in the development of effective molecular optimization methods. In this subsection, we first categorize the types of molecular data used across different optimization methods. Then, we review widely adopted molecular datasets and discuss key considerations associated with data acquisition and utilization.

4.2.1. Types of Molecular Datasets

Existing molecular optimization methods typically rely on two types of datasets: paired and unpaired molecular datasets. Unpaired molecular optimization methods often require two sets of molecules with low and high properties or a single set of molecules exhibiting low properties [37,39]. In comparison, paired molecular optimization methods necessitate datasets consisting of numerous molecular pairs, where each pair includes two similar molecules with distinct property values, one with a low property and the other with a high property.

The molecular datasets employed in molecular optimization methods are often sourced from public databases [70,71], for example, the ZINC database [72], which provides 3D molecular structures for virtual screening applications; the ChEMBL database [73], specializing in bioactive molecules with drug-like properties; the QM9 dataset [74], encompassing small organic molecules with quantum chemical properties; and the GDB-13 database [75], which is the largest publicly available repository of small organic molecules. For unpaired molecular datasets, the molecules in these databases can be filtered based on property values, such as low or high QED values. For paired molecular datasets, both the property values of the molecules and their structural similarity must be considered. For example, Jin et al. [22] constructed a paired QED dataset selected from ZINC, in which one molecule in each pair had a QED value between 0.7 and 0.8, while the other had a QED value between 0.9 and 1, with a similarity score exceeding 0.4. Table 2 provides statistics and descriptions of the databases commonly used in molecular optimization tasks.

Table 2. Common datasets and statistics for molecular optimization tasks.

Dataset	Description	Amount	Website
ZINC	Free database of commercially available compounds for virtual screening	>750,000,000	https://zinc15.docking.org/ (accessed on 9 April 2025)
ChEMBL	A manually curated database of bioactive molecules with drug-like properties	2,300,000	https://www.ebi.ac.uk/chembl/ (accessed on 9 April 2025)
PubChem	Largest collection of freely accessible chemical information	119,000,000	https://pubchem.ncbi.nlm.nih.gov/ (accessed on 9 April 2025)
MOSES	Benchmark platform for training process of standardized molecular generation model	1,940,000	https://github.com/molecularsets/moses (accessed on 9 April 2025)

Table 2. *Cont.*

Dataset	Description	Amount	Website
QM9	Molecules with up to 9 heavy atoms	133,885	http://quantum-machine.org/datasets/ (accessed on 9 April 2025)
GDB-13	Small organic molecules database	977,468,314	https://gdb.unibe.ch/downloads/ (accessed on 9 April 2025)
GDB-17	Small organic molecules database	50,000,000	https://gdb.unibe.ch/downloads/ (accessed on 9 April 2025)
QED Pairs	Similar molecule pairs with low and high QED values	88,000	https://github.com/wengong-jin/\iclr19-graph2graph (accessed on 9 April 2025)
PlogP Pairs	Similar molecule pairs with low and high PlogP values	99,000	
Drd2 Pairs	Similar molecule pairs with low and high Drd2 values	34,000	

4.2.2. Challenges and Suggestions in Obtaining Datasets

Data quality. High-quality molecular datasets are crucial for enhancing the performance of molecular optimization methods. However, the quality of molecules is affected by potential errors [76]. To address these issues, data reduction and cleaning techniques are useful for providing reusable and trustworthy data, and they have been employed in drug design to improve data quality [71,77]. For example, Papadatos et al. [78] discussed several molecular data management strategies applied to the ChEMBL database, including enhancing data integrity, flagging outliers, and adding annotations.

Data quantity. The scarcity of molecular data impedes the performance of drug design methods, particularly for novel or poorly studied diseases [79,80]. In the pharmaceutical industry, data related to drug and lead candidates are often confidential due to intellectual property protections. To address the challenges posed by limited data, several techniques can be employed. First, data augmentation can efficiently expand the training dataset [81]. Second, meta-learning frameworks facilitate knowledge transfer from tasks with abundant information to those with limited data [82]. Furthermore, privacy-preserving computational methods, such as secure multi-party computing [83], federated learning [84], and differential privacy [85], can be employed to jointly train a model from multiple parties without disclosing the original molecular data [86].

Imbalanced data. Molecular optimization methods also encounter the challenge of imbalanced data. For a given protein target, the majority of molecules may exhibit inactivity, with only a small fraction demonstrating activity. Several strategies have been proposed to mitigate this issue. First, resampling and oversampling techniques can be employed at the data level to adjust the proportion of active and inactive molecules [87]. Second, deep learning-based molecular optimization methods can integrate the imbalanced training loss to enhance learning efficacy from datasets with imbalances [88].

4.3. Optimization Properties

4.3.1. Common Molecular Properties

Non-biological activity properties. Non-biological activity properties are often derived from molecular structures, which can be directly assessed using publicly available tools such as MOSES [71], RDKit [89], TDC [90], and ADMET [91]. Several important

non-biological activity properties in molecular optimization are described in Table 3, including QED, logP, PlogP, SA, and similarity. These molecular properties can be calculated by RDkit.

Biological activity properties. In practical drug development, biological activity properties are crucial for assessing the activity, inhibition, and binding affinity of molecules to disease targets. Biological activity properties are typically estimated by trained prediction models. Table 3 presents several commonly used activities, i.e., DRD2 activity, GSK3β inhibition, and JNK3 inhibition, which have been integrated on the Therapeutics Data Commons (TDC) platform [90].

Table 3. Common non-biological activity molecular properties for optimization.

Properties	Descriptions
Quantitative estimate of drug-likeness (QED) [92]	A comprehensive index that quantifies the drug-likeness of a molecule as a value between 0 and 1, calculated by combining eight physical descriptors.
Octanol–water partition coefficients (LogP) [93]	A metric assessing the dissolution and diffusion of molecules in the human body through their combined water and lipid solubility, reflecting the membrane absorption capacity.
Penalized logP (PlogP) [58]	The logarithm of the partition ratio of the solute between octanol and water minus the synthetic accessibility score and the number of long cycles.
Synthetic accessibility (SA) [94]	Quantification of the difficulty of synthesizing small molecules in the laboratory on a scale ranging from 1 to 10, where a lower score indicates easier synthesis.
Similarity [23]	Similarity between the lead molecule and the optimized molecule. Tanimoto similarity is widely employed in existing molecular optimization studies due to its computational efficiency.
DRD2 activity [95]	The predicted biological activity score against the dopamine receptor D2 target.
GSK3β inhibition [96]	The estimated inhibition score against the glycogen synthase kinase-3 target.
JNK3 inhibition [96]	The estimated inhibition score against the c-Jun N-terminal kinase-3 target.
1SYH [97]	The docking score of a molecule and an ionotropic glutamate receptor that is associated with neurological and psychiatric diseases.
6Y2F [97]	The docking score of a molecule and the main protease of SARS-CoV-2 that is responsible for the translation of the viral RNA of the SARS-CoV-2 virus.
4LDE [97]	The β2-adrenoceptor GPCR receptor that spans the cell membrane and binds adrenaline, a hormone that mediates muscle relaxation and bronchodilation.

Moreover, the interaction between molecules and proteins is crucial for practical protein–ligand design, which is typically evaluated by docking scores obtained from molecular simulation docking platforms [98]. For example, Nigam et al. [97] established three benchmark tasks to optimize the docking scores of molecules with target proteins, including the 1SYH, 6Y2F, and 4LDE proteins (Table 3). It is worth noting that practical drug development often involves numerous other disease-related targets that require consideration. When assessing the biological activity of molecules against novel targets, the property values can be obtained through laboratory experiments. Additionally, the biological activity properties can be predicted by surrogate models trained on molecular data with known values of biological properties. To incorporate the docking scores between molecules and new targets, these scores can be simulated using docking platforms such as AutoDock [98], based on the structures of the protein and the molecule.

4.3.2. Multi-Property Optimization.

As for the aforementioned properties, practical molecular optimization must simultaneously balance multiple conflicting properties. For example, a drug candidate must exhibit desirable drug-likeness, demonstrate effective interactions with disease targets, and possess synthetic feasibility. Furthermore, practical molecular optimization often incorporates further constraints, such as adherence to specific molecular descriptor thresholds or compliance with predefined structural rules [99]. Consequently, molecular optimization is an inherently constrained multi-objective optimization problem that encompasses various objectives and constraints. For multi-objective optimization problems, since multiple properties are in conflict with each other, there is no single molecule with the highest value of all properties but rather a set of Pareto molecules with different preferences for various properties [100]. Moreover, the introduction of additional constraints renders certain regions of the chemical space infeasible, thereby increasing the complexity of the exploration process. Figure 3 visually contrasts the search processes for single-objective, multi-objective, and constrained multi-objective optimization.

To effectively address multi-property optimization challenges, Pareto-based optimization has emerged as a robust framework. This approach avoids the need for assumptions regarding the relative importance of properties and generates a diverse set of Pareto-optimal molecules [101]. Such molecules have been successfully employed in various methods for screening and identifying desired candidates [60].

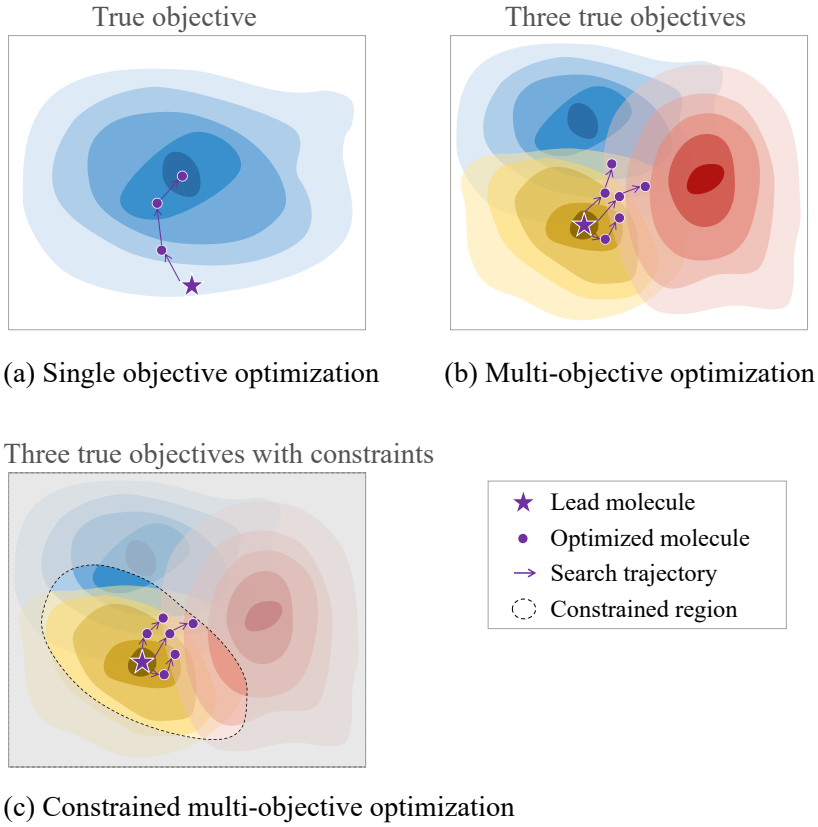

Figure 3. Comparison of the search processes of single-objective, multi-objective, and constrained multi-objective molecular optimization. The darker the colored region, the higher the property. (**a**) Single-objective optimization searches for regions with a high objective value. (**b**) Multi-objective optimization searches for molecules with different trade-offs between three objectives (taking three objectives as an example). (**c**) In constrained multi-objective optimization, the search space is limited by constraints (in gray), and it aims to search for molecules with multiple desired objectives in the constrained region.

4.3.3. Challenges in Practical Molecular Optimization Tasks

In real-world drug design, candidate molecules must simultaneously satisfy multiple critical property requirements, including the specific binding affinity to target proteins, synthetic feasibility, good solubility, appropriate blood–brain barrier permeability, low toxicity, and minimal side effects. Beyond these quantifiable optimization objectives, molecular structures must also meet stringent drug-like constraints, such as compliance with Lipinski's Rule of Five, the avoidance of structural alerts, and the maintenance of a reasonable molecular weight. These multiple requirements make molecular optimization an inherently complex constrained multi-objective optimization problem. Notably, these properties often exhibit intricate interrelationships and trade-offs, making it extremely challenging to identify a single molecule that is optimal across all attributes. Instead, the solution typically involves identifying a set of Pareto-optimal solutions that represent the best possible compromises among competing properties.

For the multi-property optimization challenges, Pareto-based optimization methods have demonstrated significant advantages and are considered the most robust multi-property optimization frameworks. These methods do not require predefined weights of relative importance among properties and can generate a diverse set of candidate molecules. Furthermore, in practical applications, the number of properties to be optimized may exceed four, and additional properties may be introduced over time. In these scenarios, the chemical space containing the desired molecules shrinks considerably, which increases the difficulty of optimization. Traditional Pareto optimization methods often struggle to generate and select high-quality molecules under multi-property optimization. To address these challenges, intelligent generation strategies can be designed to dynamically adjust optimization priorities based on the optimization status of each property, and more efficient selection strategies can also be developed to identify high-quality candidates. For dynamic molecular optimization problems, progressive optimization frameworks offer flexible solutions to accommodate emerging optimization requirements. For instance, DyMol [102] initiates optimization with a single objective and incorporates additional objectives over time, which decomposes complex multi-objective problems into manageable sub-problems for incremental optimization.

In drug design practice, the synthetic feasibility of generated molecules serves as a critical determinant of whether an optimization model can be practically applied in real-world development. Current AI-driven molecular optimization methods predominantly employ synthetic accessibility (SA) [94] scores to evaluate the synthesis of molecules. However, a significant gap persists between these computational scores and the actual synthetic feasibility. To generate molecules that can be reliably synthesized in laboratory settings, on the one hand, it may be useful to train sophisticated deep generative models on extensive databases of known synthesizable molecules to learn synthetic rules; on the other hand, domain-specific knowledge and expert experience from medicinal chemistry should be incorporated during model training to compensate for the limitations of purely data-driven approaches.

In the field of computer-aided drug discovery, structure-based molecular design plays a pivotal role in the development of therapeutic agents for specific diseases. This approach enables the efficient design of candidate molecules with a high binding affinity to target proteins by exploring vast chemical spaces. Notably, leveraging protein–ligand interaction information can significantly enhance the pharmacological activity of generated molecules. For example, PMDM [103] incorporates protein structural information as generation constraints, which establishes a conditional equivariant diffusion model that simultaneously considers both local atomic interactions and global molecular dynamics. The practical utility of PMDM was demonstrated through case studies on two critical

drug targets. In a CDK2 lead optimization study, the researchers synthesized molecules generated by PMDM and validated their significantly improved inhibitory activity against CDK2 through in vitro assays. These results demonstrate the practical value of structure-based molecular generation technology in real-world drug development scenarios.

4.4. Optimization Algorithms

To enhance molecular optimization, various emerging techniques can be employed to design novel and effective optimization methods. In this subsection, we introduce several promising techniques that provide valuable guidelines for future molecular optimization.

First, practical molecular optimization can be formulated as a constrained multi-objective optimization problem, which aims to find molecules with high properties while satisfying constraints. While most existing methods overlook constraints, making it challenging to generate drug-like molecules, constrained multi-objective optimization algorithms can systematically explore feasible chemical spaces to produce constrained Pareto fronts [104].

Second, molecular optimization methods can be significantly enhanced by leveraging the complementary strengths of different molecular representations. Multi-modal learning, which integrates diverse representations, enables the capture of richer implicit information, thereby improving the robustness and accuracy of optimization [105]. For instance, Luo et al. developed a method that effectively combines 2D and 3D molecular data to extract more comprehensive chemical knowledge [106]. Moreover, within real-world drug discovery pipelines, the incorporation of target protein structural or functional data is critical to computationally guide the generation of bioactive molecules exhibiting specific binding interactions [107].

Third, advanced AI techniques hold potential to enhance molecular optimization [108]. For example, active learning can reduce the labeling cost by iteratively selecting the most informative samples for model training, and it has been used to predict biological activity and target–ligand interactions [109]. Furthermore, transfer learning enables the transfer of knowledge from well-studied tasks to related but data-scarce tasks, making it particularly suitable for molecular optimization in novel diseases with limited datasets [110]. Moreover, multi-task learning can mitigate bias and overfitting by simultaneously training on different tasks in a single model [111], and it can be integrated with existing molecular optimization methods. In addition, with the advancement of large language models (LLMs), integrating LLMs with molecular optimization enables the enhanced extraction of molecular information, the robust capture of chemical rules, and effective knowledge transfer, thereby facilitating the multi-property optimization of molecules [103,112].

5. Conclusions

This review provides a comprehensive analysis of existing research in molecular optimization. We first explicitly formulate definitions of molecular optimization problems and highlight their significance in drug discovery and development. We then comprehensively categorize and analyze existing AI-aided optimization models based on the chemical spaces that they explore and the optimization methods that they employ. Furthermore, we discuss the challenges and future prospects associated with the application of molecular optimization models. Thus, this review offers potentially beneficial recommendations for the advancement of AI-based molecular optimization approaches.

Author Contributions: Conceptualization, Y.S. and X.X.; writing—original draft preparation, X.X., Y.Z., and Y.S.; writing—review and editing, X.Z. (Xingyi Zhang), X.Z. (Xiangxiang Zeng), and C.Z.; project administration, X.X. and Y.S.; funding acquisition, Y.S., X.Z. (Xingyi Zhang), X.Z. (Xiangxiang Zeng), and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by funds from the National Science Foundation of China (NSF: # 62172002, # 62202004, # 62403002, and #62322301); the University Synergy Innovation Program of Anhui Province (# GXXT-2022-035, # GXXT-2021-039); the Anhui Provincial Natural Science Foundation (# 2108085QF267 and # 2008085QF294); the University Outstanding Youth Research Project of Anhui Province (# 2022AH020010); the Project of Key Laboratory of Intelligent Computing & Signal Processing (Anhui University), Ministry of Education (# 2020A005); and the State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia Fund (SKL-HIDCA-2024-AH1).

Data Availability Statement: No primary research results, software, or code was included, and no new data were generated or analyzed as part of this review.

Acknowledgments: The authors thank the anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no competing interests.

References

1. 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. [CrossRef] [PubMed]
2. Qi, Y.; Zhao, H.; Lei, Y. Organic molecular design for high-power density sodium-ion batteries. *Chem. Commun.* **2025**, *61*, 2375–2386. [CrossRef] [PubMed]
3. Godinez, W.J.; Ma, E.J.; Chao, A.T.; Pei, L.; Skewes-Cox, P.; Canham, S.M.; Jenkins, J.L.; Young, J.M.; Martin, E.J.; Guiguemde, W.A. Design of potent antimalarials with generative chemistry. *Nat. Mach. Intell.* **2022**, *4*, 180–186. [CrossRef]
4. Caburet, J.; Boucherle, B.; Bourdillon, S.; Simoncelli, G.; Verdirosa, F.; Docquier, J.D.; Moreau, Y.; Krimm, I.; Crouzy, S.; Peuchmaur, M. A fragment-based drug discovery strategy applied to the identification of NDM-1 β-lactamase inhibitors. *Eur. J. Med. Chem.* **2022**, *240*, 114599. [CrossRef] [PubMed]
5. Wu, Y.; Fan, S.; Dong, M.; Li, J.; Kong, C.; Zhuang, J.; Meng, X.; Lu, S.; Zhao, Y.; Wu, C. Structure-guided design of CPPC-paired disulfide-rich peptide libraries for ligand and drug discovery. *Chem. Sci.* **2022**, *13*, 7780–7789. [CrossRef]
6. Huang, Z.; Weng, X.; Ou-Yang, L. GFLearn: Generalized Feature Learning for Drug-Target Binding Affinity Prediction. *IEEE J. Biomed. Health Inform.* 2025, *Online ahead of print*. [CrossRef]
7. Yang, L.; Guo, Q.; Zhang, L. AI-assisted chemistry research: A comprehensive analysis of evolutionary paths and hotspots through knowledge graphs. *Chem. Commun.* **2024**, *60*, 6977–6987. [CrossRef]
8. Li, Y.; Zhang, L.; Wang, Y.; Zou, J.; Yang, R.; Luo, X.; Wu, C.; Yang, W.; Tian, C.; Xu, H.; et al. Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor. *Nat. Commun.* **2022**, *13*, 6891. [CrossRef]
9. Zhang, J.; Wang, Q.; Wen, H.; Gerbaud, V.; Jin, S.; Shen, W. Multi-objective optimization strategy for green solvent design via a deep generative model learned from pre-set molecule pairs. *Green Chem.* **2024**, *26*, 412–427. [CrossRef]
10. Zhu, Y.; Wu, J.; Hu, C.; Yan, J.; Hou, T.; Wu, J. Sample-efficient multi-objective molecular optimization with gflownets. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 79667–79684.
11. Wang, X.; Du, Z.; Guo, Y.; Zhong, J.; Song, K.; Wang, J.; Yu, J.; Yang, X.; Liu, C.Y.; Shi, T.; et al. Computer-aided molecular design and optimization of potent inhibitors disrupting APC–Asef interaction. *Acta Pharm. Sin. B* **2024**, *14*, 2631–2645. [CrossRef]
12. Bos, P.H.; Houang, E.M.; Ranalli, F.; Leffler, A.E.; Boyles, N.A.; Eyrich, V.A.; Luria, Y.; Katz, D.; Tang, H.; Abel, R.; et al. AutoDesigner, a De Novo Design Algorithm for rapidly exploring large chemical space for lead optimization: Application to the design and synthesis of d-Amino acid oxidase inhibitors. *J. Chem. Inf. Model.* **2022**, *62*, 1905–1915. [CrossRef] [PubMed]
13. Sridharan, B.; Goel, M.; Priyakumar, U.D. Modern machine learning for tackling inverse problems in chemistry: molecular design to realization. *Chem. Commun.* **2022**, *58*, 5316–5331. [CrossRef]
14. 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*, e1012229. [CrossRef] [PubMed]
15. Chen, S.; Zhang, O.; Jiang, C.; Zhao, H.; Zhang, X.; Chen, M.; Liu, Y.; Su, Q.; Wu, Z.; Wang, X.; et al. Deep lead optimization enveloped in protein pocket and its application in designing potent and selective ligands targeting LTK protein. *Nat. Mach. Intell.* **2025**, *7*, 448–458. [CrossRef]
16. Chan, H.S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing drug discovery via artificial intelligence. *Trends Pharmacol. Sci.* **2019**, *40*, 592–604. [CrossRef] [PubMed]
17. Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov, M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040. [CrossRef]

18. Skinnider, M.A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.* **2024**, *6*, 437–448. [CrossRef]
19. Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287. [CrossRef]
20. Liu, S.; Sui, J.; Luo, B.; Zhang, J.; Xiang, X.; Yang, T.; Luo, Y.; Liu, J. Discovery of 5-(Piperidin-4-yl)-1, 2, 4-oxadiazole derivatives as a new class of human caseinolytic protease p agonists for the treatment of hepatocellular carcinoma. *J. Med. Chem.* **2024**, *67*, 10622–10642. [CrossRef]
21. Koziarski, M.; Rekesh, A.; Shevchuk, D.; van der Sloot, A.; Gaiński, P.; Bengio, Y.; Liu, C.; Tyers, M.; Batey, R. Rgfn: Synthesizable molecular generation using gflownets. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 46908–46955.
22. Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. Learning multimodal graph-to-graph translation for molecular optimization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
23. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 1–13. [CrossRef] [PubMed]
24. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef] [PubMed]
25. Ji, C.; Zheng, Y.; Wang, R.; Cai, Y.; Wu, H. Graph polish: A novel graph generation paradigm for molecular optimization. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, *34*, 2323–2337. [CrossRef]
26. Langevin, M.; Minoux, H.; Levesque, M.; Bianciotto, M. Scaffold-constrained molecular generation. *J. Chem. Inf. Model.* **2020**, *60*, 5637–5646. [CrossRef] [PubMed]
27. 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. [CrossRef]
28. 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. [CrossRef]
29. 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. [CrossRef]
30. Lee, Y.; Choi, K.; Kim, C. Docking-based Multi-objective Molecular optimization Pipeline using Structure-constrained Genetic Algorithm. In Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 6–8 December 2022; pp. 3438–3445.
31. Nigam, A.; Pollice, R.; Krenn, M.; dos Passos Gomes, G.; 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. [CrossRef]
32. 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*, 1–14. [CrossRef]
33. Verhellen, J. Graph-based molecular Pareto optimisation. *Chem. Sci.* **2022**, *13*, 7526–7535. [CrossRef]
34. 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*, 6412–6422.
35. Zhou, Z.; Kearnes, S.; Li, L.; Zare, R.N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **2019**, *9*, 1–10. [CrossRef]
36. Shin, B.; Park, S.; Bak, J.; Ho, J.C. Controlled molecule generator for optimizing multiple chemical properties. In Proceedings of the Conference on Health, Inference, and Learning, Online, 8–9 April 2021; pp. 146–153.
37. Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchoł, M. Mol-CycleGAN: A generative model for molecular optimization. *J. Cheminform.* **2020**, *12*, 1–18.
38. Barshatski, G.; Radinsky, K. Unpaired generative molecule-to-molecule translation for lead optimization. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; pp. 2554–2564.
39. Barshatski, G.; Nordon, G.; Radinsky, K. Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Queensland, Australia, 1–5 November 2021; pp. 3727–3736.
40. Yang, X.; Fu, L.; Deng, Y.; Liu, Y.; Cao, D.; Zeng, X. GPMO: Gradient perturbation-based contrastive learning for molecule optimization. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 4940–4948.
41. Yu, J.; Xu, T.; Rong, Y.; Huang, J.; He, R. Structure-aware conditional variational auto-encoder for constrained molecule optimization. *Pattern Recognit.* **2022**, *126*, 108581. [CrossRef]
42. Chen, Z.; Min, M.R.; Parthasarathy, S.; Ning, X. A deep generative model for molecule optimization via one fragment modification. *Nat. Mach. Intell.* **2021**, *3*, 1040–1049. [CrossRef] [PubMed]

43. Kaminsky, N.; Singer, U.; Radinsky, K. CFOM: Lead optimization for drug discovery with limited data. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 1056–1066.
44. Wu, K.; Xia, Y.; Deng, P.; Liu, R.; Zhang, Y.; Guo, H.; Cui, Y.; Pei, Q.; Wu, L.; Xie, S.; et al. TamGen: Drug design with target-aware molecule generation through a chemical language model. *Nat. Commun.* **2024**, *15*, 9360. [CrossRef]
45. 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. [CrossRef]
46. Fu, T.; Gao, W.; Xiao, C.; Yasonik, J.; Coley, C.W.; Sun, J. Differentiable scaffolding tree for molecular optimization. *arXiv* **2022**, arXiv:2109.10469.
47. Eckmann, P.; Sun, K.; Zhao, B.; Feng, M.; Gilson, M.K.; Yu, R. LIMO: Latent Inceptionism for Targeted Molecule Generation. *arXiv* **2022**, arXiv:2206.09010.
48. Niu, Y.; Gao, Z.; Xu, T.; Liu, Y.; Bian, Y.; Rong, Y.; Huang, J.; Li, J. InversionGNN: A Dual Path Network for Multi-Property Molecular Optimization. *arXiv* **2025**, arXiv:2503.01488.
49. Xia, X.; Liu, Y.; Zheng, C.; Zhang, Y.; Wu, Q.; Gao, X.; Zeng, X.; Su, Y. Evolutionary Multiobjective Molecule Optimization in an Implicit Chemical Space. *J. Chem. Inf. Model.* **2024**, *64*, 5161–5174. [CrossRef] [PubMed]
50. Zhou, X.; Cheng, X.; Yang, Y.; Bao, Y.; Wang, L.; Gu, Q. Controllable and decomposed diffusion models for structure-based molecular optimization. *arXiv* **2024**, arXiv:2403.13829.
51. Morehead, A.; Cheng, J. Geometry-complete diffusion for 3D molecule generation and optimization. *Commun. Chem.* **2024**, *7*, 150. [CrossRef] [PubMed]
52. Wang, Z.; Nie, W.; Qiao, Z.; Xiao, C.; Baraniuk, R.; Anandkumar, A. Retrieval-based Controllable Molecule Generation. *arXiv* **2023**, arXiv:2208.11126.
53. Abeer, A.N.; Urban, N.M.; Weil, M.R.; Alexander F.J.; Yoon, B.-J. Multi-objective latent space optimization of generative molecular design models. *Patterns* **2024**, *5*, 101042. [CrossRef]
54. Wu, Z.; Zhang, O.; Wang, X.; Fu, L.; Zhao, H.; Wang, J.; Du, H.; Jiang, D.; Deng, Y.; Cao, D.; et al. Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nat. Mach. Intell.* **2024**, *6*, 1359–1369. [CrossRef]
55. Ye, G.; Cai, X.; Lai, H.; Wang, X.; Huang, J.; Wang, L.; Liu, W.; Zeng, X. Drugassist: A large language model for molecule optimization. *Briefings Bioinform.* **2025**, *26*, bbae693. [CrossRef] [PubMed]
56. Moerland, T.M.; Broekens, J.; Plaat, A.; Jonker, C.M.; et al. Model-based reinforcement learning: A survey. *Found. Trends Mach. Learn.* **2023**, *16*, 1–118. [CrossRef]
57. Pogány, P.; Arad, N.; Genway, S.; Pickett, S.D. De novo molecule design by translating from reduced graphs to SMILES. *J. Chem. Inf. Model.* **2018**, *59*, 1136–1146. [CrossRef]
58. Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In Proceedings of the International conference on machine learning. PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2323–2332.
59. Lee, S.; Lee, D.B.; Hwang, S.J. Contrastive learning with adversarial perturbations for conditional text generation. *arXiv* **2021**, arXiv:2012.07280.
60. Grantham, K.; Mukaidaisi, M.; Ooi, H.K.; Ghaemi, M.S.; Tchagang, A.; Li, Y. Deep evolutionary learning for molecular design. *IEEE Comput. Intell. Mag.* **2022**, *17*, 14–28. [CrossRef]
61. Barreto Gomes, D.E.; Galentino, K.; Sisquellas, M.; Monari, L.; Bouysset, C.; Cecchini, M. ChemFlow—From 2D Chemical Libraries to Protein–Ligand Binding Free Energies. *J. Chem. Inf. Model.* **2023**, *63*, 407–411. [CrossRef] [PubMed]
62. Cheng, A.H.; Cai, A.; Miret, S.; Malkomes, G.; Phielipp, M.; Aspuru-Guzik, A. Group SELFIES: A robust fragment-based molecular string representation. *Digit. Discov.* **2023**, *2*, 748–758. [CrossRef]
63. Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular design in drug discovery: A comprehensive review of deep generative models. *Briefings Bioinform.* **2021**, *22*, bbab344. [CrossRef] [PubMed]
64. Zeng, X.; Xiang, H.; Yu, L.; Wang, J.; Li, K.; Nussinov, R.; Cheng, F. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat. Mach. Intell.* **2022**, *4*, 1004–1016 . [CrossRef]
65. 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. [CrossRef]
66. Kuzminykh, D.; Polykovskiy, D.; Kadurin, A.; Zhebrak, A.; Baskov, I.; Nikolenko, S.; Shayakhmetov, R.; Zhavoronkov, A. 3D molecular representations based on the wave transform for convolutional neural networks. *Mol. Pharm.* **2018**, *15*, 4378–4385. [CrossRef]
67. Kusner, M.J.; Paige, B.; Hernández-Lobato, J.M. Grammar variational autoencoder. In Proceedings of the International conference on machine learning. PMLR, Sydney, Australia, 6–11 August 2017; pp. 1945–1954.
68. Barthel, S.; Krenn, M.; Ai, Q.; Carson, N.; Frei, A.; Frey, N.C.; Friederich, P.; Gaudin, T.; Gayle, A.A.; Jablonka, K.M.; et al. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3*, 100588.

69. Samanta, B.; De, A.; Jana, G.; Gómez, V.; Chattaraj, P.K.; Ganguly, N.; Gomez-Rodriguez, M. Nevae: A deep generative model for molecular graphs. *J. Mach. Learn. Res.* **2020**, *21*, 1–33. [CrossRef]
70. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. [CrossRef]
71. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Front. Pharmacol.* **2020**, *11*, 565644. [CrossRef] [PubMed]
72. Sterling, T.; Irwin, J.J. ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef] [PubMed]
73. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
74. 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. [CrossRef] [PubMed]
75. Blum, L.C.; Reymond, J.L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733. [CrossRef]
76. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [CrossRef]
77. Chen, C.; Yaari, Z.; Apfelbaum, E.; Grodzinski, P.; Shamay, Y.; Heller, D.A. Merging Data Curation and Machine Learning to Improve Nanomedicines. *Adv. Drug Deliv. Rev.* **2022**, *183*, 114172. [CrossRef]
78. Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J.P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.* **2015**, *29*, 885–896. [CrossRef]
79. Bowers, E.C.; Hassanin, A.A.; Ramos, K.S. In vitro models of exosome biology and toxicology: New frontiers in biomedical research. *Toxicol. Vitr.* **2020**, *64*, 104462. [CrossRef]
80. Du, H.; Jiang, D.; Zhang, O.; Wu, Z.; Gao, J.; Zhang, X.; Wang, X.; Deng, Y.; Kang, Y.; Li, D.; et al. A flexible data-free framework for structure-based de novo drug design with reinforcement learning. *Chem. Sci.* **2023**, *14*, 12166–12181. [CrossRef] [PubMed]
81. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
82. Wang, J.; Zheng, S.; Chen, J.; Yang, Y. Meta learning for low-resource molecular optimization. *J. Chem. Inf. Model.* **2021**, *61*, 1627–1636. [CrossRef]
83. Ma, R.; Li, Y.; Li, C.; Wan, F.; Hu, H.; Xu, W.; Zeng, J. Secure multiparty computation for privacy-preserving drug discovery. *Bioinformatics* **2020**, *36*, 2872–2880. [CrossRef] [PubMed]
84. Chen, S.; Xue, D.; Chuai, G.; Yang, Q.; Liu, Q. FL-QSAR: A federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics* **2021**, *36*, 5492–5498. [CrossRef] [PubMed]
85. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318.
86. Shahverdi, A.; Trieu, N.; Weng, C.; Youmans, W. Privacy-Preserving Prescription Drug Management Using Fully Homomorphic Encryption. In *Protecting Privacy through Homomorphic Encryption*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 169–176.
87. Peng, M.; Zhang, Q.; Xing, X.; Gui, T.; Huang, X.; Jiang, Y.G.; Ding, K.; Chen, Z. Trainable undersampling for class-imbalance learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4707–4714.
88. Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; Dokania, P. Calibrating deep neural networks using focal loss. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15288–15299.
89. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*, 5281.
90. Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C.W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv* **2021**, arXiv:2102.09548.
91. Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. admetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* **2019**, *35*, 1067–1069. [CrossRef]
92. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef]
93. Daina, A.; Michielin, O.; Zoete, V. iLOGP: A simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using the GB/SA approach. *J. Chem. Inf. Model.* **2014**, *54*, 3284–3301. [CrossRef]
94. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 1–11. [CrossRef] [PubMed]

95. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 1–14. [CrossRef]
96. Jin, W.; Barzilay, R.; Jaakkola, T. Multi-objective molecule generation using interpretable substructures. In Proceedings of the International Conference on Machine Learning. PMLR, Online, 13–18 July 2020; pp. 4849–4859.
97. Nigam, A.; Pollice, R.; Tom, G.; Jorner, K.; Willes, J.; Thiede, L.A.; Kundaje, A.; Aspuru-Guzik, A. Tartarus: A benchmarking platform for realistic and practical inverse molecular design. *arXiv* **2022**, arXiv:2209.12487.
98. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
99. Vogt, M.; Yonchev, D.; Bajorath, J. Computational method to evaluate progress in lead optimization. *J. Med. Chem.* **2018**, *61*, 10895–10900. [CrossRef] [PubMed]
100. Renz, P.; Van Rompaey, D.; Wegner, J.K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today Technol.* **2019**, *32*, 55–63. [CrossRef] [PubMed]
101. 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.
102. Shin, D.H.; Son, Y.H.; Lee, D.J.; Han, J.W.; Kam, T.E. Dynamic many-objective molecular optimization: Unfolding complexity with objective decomposition and progressive optimization. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Jeju, Republic of Korea, 3–9 August 2024; pp. 6026–6034.
103. Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.C.; et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **2024**, *15*, 2657. [CrossRef]
104. Liang, J.; Ban, X.; Yu, K.; Qu, B.; Qiao, K.; Yue, C.; Chen, K.; Tan, K.C. A survey on evolutionary constrained multi-objective optimization. *IEEE Trans. Evol. Comput.* **2022**, *27*, 201–221. [CrossRef]
105. Post, K.L.; Belmadani, M.; Ganguly, P.; Meili, F.; Dingwall, R.; McDiarmid, T.A.; Meyers, W.M.; Herrington, C.; Young, B.P.; Callaghan, D.B.; et al. Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nat. Commun.* **2020**, *11*, 2073. [CrossRef]
106. Luo, S.; Chen, T.; Xu, Y.; Zheng, S.; Liu, T.Y.; Wang, L.; He, D. One transformer can understand both 2d & 3d molecular data. *arXiv* **2022**, arXiv:2210.01765.
107. Konstantinidou, M.; Visser, E.J.; Vandenboorn, E.; Chen, S.; Jaishankar, P.; Overmans, M.; Dutta, S.; Neitz, R.J.; Renslo, A.R.; Ottmann, C.; et al. Structure-based optimization of covalent, small-molecule stabilizers of the 14-3-3σ/ERα protein–protein interaction from nonselective fragments. *J. Am. Chem. Soc.* **2023**, *145*, 20328–20343. [CrossRef] [PubMed]
108. Son, Y.H.; Shin, D.H.; Lee, D.J.; Kam, T.E. Molecular Optimization with Mamba-Based GFlowNet. In Proceedings of the 2025 International Conference on Electronics, Information, and Communication (ICEIC), Osaka, Japan, 19–22 January 2025; pp. 1–4.
109. Gusev, F.; Gutkin, E.; Kurnikova, M.G.; Isayev, O. Active learning guided drug design lead optimization based on relative binding free energy modeling. *J. Chem. Inf. Model.* **2022**, *63*, 583-594 [CrossRef] [PubMed]
110. Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer learning for drug discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694. [CrossRef]
111. Zhu, J.; Liu, Y.; Zhang, Y.; Chen, Z.; She, K.; Tong, R. DAEM: Deep attributed embedding based multi-task learning for predicting adverse drug–drug interaction. *Expert Syst. Appl.* **2023**, *215*, 119312. [CrossRef]
112. Yu, J.; Zheng, Y.; Koh, H.Y.; Pan, S.; Wang, T.; Wang, H. Collaborative Expert LLMs Guided Multi-Objective Molecular Optimization. *arXiv* **2025**, arXiv:2503.03503.