nature machine intelligence

Review article

https://doi.org/10.1038/s42256-024-00843-5

Machine learning-aided generative molecular design

Received: 19 July 2023

Accepted: 24 April 2024

Published online: 18 June 2024

Check for updates

Yuanqi Du[1,10], Arian R. Jamasb[2,3,9,10], Jeff Guo[4,5,10], Tianfan Fu[6], Charles Harris[3], Yingheng Wang[1], Chenru Duan[7], Pietro Liò[3], Philippe Schwaller[4,5] & Tom L. Blundell[8]

Machine learning has provided a means to accelerate early-stage drug discovery by combining molecule generation and filtering steps in a single architecture that leverages the experience and design preferences of medicinal chemists. However, designing machine learning models that can achieve this on the fly to the satisfaction of medicinal chemists remains a challenge owing to the enormous search space. Researchers have addressed de novo design of molecules by decomposing the problem into a series of tasks determined by design criteria. Here we provide a comprehensive overview of the current state of the art in molecular design using machine learning models as well as important design decisions, such as the choice of molecular representations, generative methods and optimization strategies. Subsequently, we present a collection of practical applications in which the reviewed methodologies have been experimentally validated, encompassing both academic and industrial efforts. Finally, we draw attention to the theoretical, computational and empirical challenges in deploying generative machine learning and highlight future opportunities to better align such approaches to achieve realistic drug discovery end points.

Drug discovery and development is an iterative process of optimizing molecules to satisfy a set of specific properties, such as solubility[1], toxicity[2], pharmacokinetics[3] and other desirable therapeutic effects. Although many therapeutic modalities, such as (macrocyclic) peptides, biologics and oligonucleotide therapies, have shown promise and efficacy in the clinic, small molecules have received the most attention from the machine learning (ML) community, particularly for druggable targets. Owing to the vast search space of traditional drug-like small-molecule compounds, estimated to be between 10^{23} and 10^{60} (ref. 4), and the discontinuous nature of optimization functions[5], development is difficult, expensive and prone to failure. The number of novel molecular entities approved per dollar spent on industrial research and development activities has decreased exponentially over the past 70 years[6] and the average development cost has risen to more than US$2 billion over a 10–15-year timeline[7]. However, a recent uptick over the past decade is thought to be largely driven by the accumulation of better collection and use of high-quality decision-making information and experimental data in the drug discovery process[8], highlighting growing opportunities for ML in the research and development life cycle.

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA. [2]Department of Biochemistry, University of Cambridge, Cambridge, UK. [3]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. [4]Laboratory of Artificial Chemical Intelligence (LIAC), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [5]National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [6]Rensselaer Polytechnic Institute, Troy, NY, USA. [7]Microsoft Quantum, Redmond, WA, USA. [8]Heart and Lung Research Institute, University of Cambridge, Cambridge, UK. [9]Present address: Prescient Design, Genentech, Basel, Switzerland. [10]These authors contributed equally: Yuanqi Du, Arian R. Jamasb, Jeff Guo. e-mail: philippe.schwaller@epfl.ch; njm44@cam.ac.uk

https://doi.org/10.1038/s42256-024-00843-5

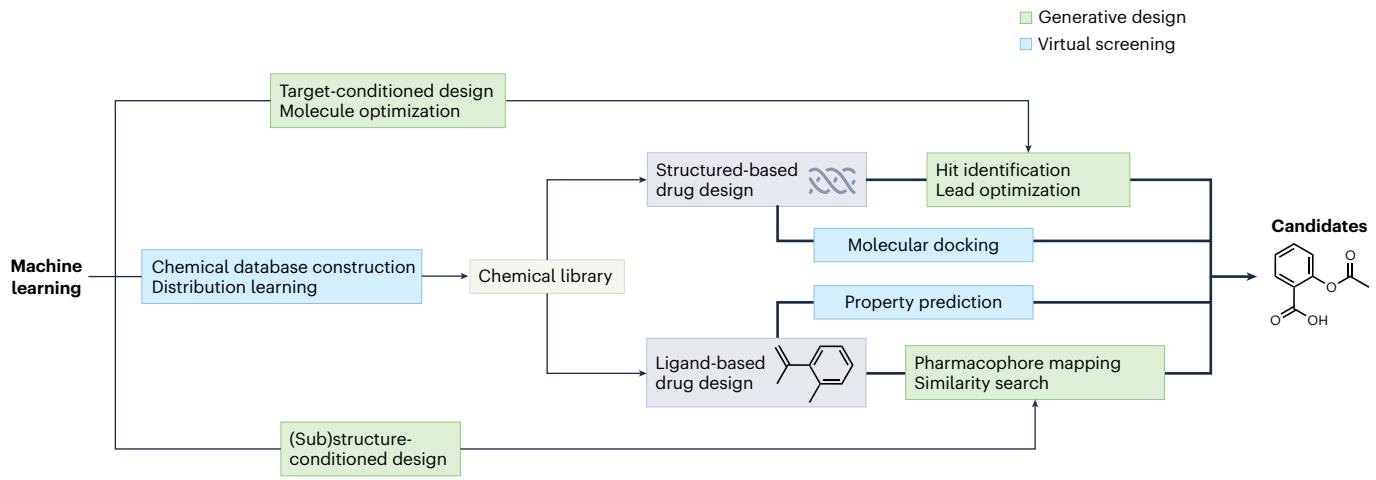

Fig. 1 | Generative ML-assisted molecular design pipeline. ML serves as a powerful accelerator for drug discovery and development, particularly in areas such as virtual screening and generative design. It notably streamlines several critical processes, including target identification, target-to-hit transformation, hit-to-lead conversion and lead optimization. By enhancing the efficiency of these steps, ML greatly reduces both the time and costs associated with traditional drug discovery methods, thereby bringing substantial benefits to the pharmaceutical industry.

Medicinal chemists face the challenge of designing potent molecules with optimal biological activity constrained by the intractability of exhaustive experimental validation. Often, traditional (computational) screening strategies are used. Ligand-based drug design (LBDD) leverages information from previously studied ligands to develop quantitative structure–activity relationship methods. These methods correlate molecular structure with activity, enabling scoring of novel molecules[9,10]. Given a target structure, structure-based drug design (SBDD) methods are effective in designing ligands with high complementarity, affinity and specificity for a given target pocket[11,12]. SBDD campaigns typically begin by docking molecular libraries against a target; the readouts can be used to directly discover potential 'hits' or iteratively optimized into 'lead' compounds[12]. However, these approaches can be limited by reliance on biased human knowledge[13] with limited exploration of chemical space[4].

Despite the challenges, increased availability of computational resources has enhanced the capabilities of traditional LBDD and SBDD approaches through ultralarge-scale virtual screening libraries (10^{8-10}; refs. 14–16). With increasingly larger make-on-demand libraries, these methods have immediate applicability and provide an attractive method to accelerate early-stage drug discovery. Chemogenomics offers a complementary approach by screening targeted libraries of molecules against panels of related targets. Such screens can identify and validate targets, and provide greater insight into the mechanism of action of potent compounds, resulting in hypotheses that can be used to constrain the search space. Alternatively and concurrently, ML methods have demonstrated potential in mitigating the bias challenge by augmenting human-derived knowledge with data-driven insights[13]. ML techniques enable the direct acquisition of molecular descriptors from data through representation learning, enabling, for example, improved molecular property prediction accuracy. Learned representations can subsequently be employed in scoring functions for screening molecular databases. Explicitly optimizing learned representations using ML is a complementary approach to virtual screening and enables directed exploration of chemical space, offering the potential to accelerate the identification of suitable candidates.

Generative modelling is a subfield of ML that focuses on developing algorithms capable of generating new data samples that resemble the data distribution from a given training dataset. These models strive to learn the underlying structure, patterns or probability distributions of the data, enabling the creation of novel compounds. Generative models have garnered substantial attention due to their versatility and wide-ranging applications, with notable successes in diffusion models for image synthesis and autoregressive language models for text generation[17,18]. Similarly, generative modelling has received substantial attention in drug discovery, where new molecular structures can be generated based on the properties of collections of known compounds.

Generative modelling has the potential to overcome the limitations of screening ultralarge chemical libraries, which are often biased owing to our existing knowledge of chemical space (generative molecular design pipeline in Fig. 1). These models can expand the range of discoverable novel molecular entities by generating novel compounds. Conditional generative models augur the design of molecules with specific desired properties, such as high target-specific activity. As a complementary approach to virtual screening, generative modelling has successfully identified experimentally validated inhibitor compounds outside existing chemical libraries[19,20].

Generative molecular design has experienced remarkable advancements in recent years, resulting in an extensive and rapidly expanding body of literature (Table 1 includes a glossary of technical terminologies used in this paper). Here we offer a comprehensive examination of ML methodologies applied to generative molecular design, elucidating their motivations and practical applications within pharmacology across diverse task settings ('Generative molecular design tasks' section). We categorize models by design criteria and highlight trade-offs ('Generative molecular design methodologies' section). We emphasize the importance of problem formulation and critical assessment to transform ML advancements into real-world drug discovery toolkits ('Evaluating designs of generative ML' section). Finally, we discuss the theoretical, computational and empirical challenges that the community is currently facing, as well as the opportunities to efficient and effective drug discovery assisted by ML ('Future directions' section).

Generative molecular design tasks

Generative molecular design can be classified into two main paradigms: distribution learning and goal-oriented generation (Fig. 2), where goal-oriented generation can be further broken down into conditional generation and molecule optimization. The suitability of each approach depends on the specific task and data involved.

Distribution learning

Given a collection of molecules, distribution learning seeks to model their probability distribution to describe the underlying data

https://doi.org/10.1038/s42256-024-00843-5

Table 1 | Definitions of technical terms covered in this paper (in alphabetical order)

Term	Description
Autoregressive models	Autoregressive models are a type of deep generative model that decompose the molecule generation process as a sequence of actions by iteratively adding atoms and/or bonds one by one, which loosely mimics the construction of the molecular structure by reaction synthesis.
Bayesian optimization	Bayesian optimization is a method for optimizing black-box functions that are not analytically available and expensive to evaluate. It explores the unknown areas sequentially with an acquisition function to determine the next point to query.
Conditional generation	Conditional generation refers to generating data given specific conditions such as context, property value and so on using a conditional distribution.
Convolutional neural network	CNNs are a type of neural network that process a grid-like data structure by applying a filter to extract features from its adjacent units (neighbours), focusing on the local pattern over one unit and its neighbours.
Diffusion models	Diffusion models (also known as score-based generative models) are inspired by the diffusion process of heat where it leverages a reverse-time diffusion process to transform from the equilibrium distribution (that is, Gaussian) to the complex data distribution. In the context of chemistry, diffusion models can be used in low-energy conformations of molecules, docking poses and so on.
Energy-based models	Energy-based models (EBMs) can be considered as learning an energy function over the chemical space, which assigns an energy value to each molecule; the model learns to sample the lower-energy molecule that is more stable to exist in the real world.
Equivariance	Equivariance is a property for a function that the function commutes with certain transformations (actions of a group).
Fine-tuning	Fine-tuning is a learning diagram that takes pre-trained models on similar tasks and tunes the parameters on a new task (often with a small dataset).
Flow network	Flow network is a directed graph where each edge has a non-negative capacity and a flow.
Fragment-based drug design	Fragment-based drug design is a method that uses fragments as basic building blocks.
Fragment linking	Fragment linking refers to combining fragments to form a larger compound.
Genetic algorithm	Genetic algorithm is an evolutionary algorithm that includes several operations inspired by biological evolution such as mutation, crossover, selection, reproduction and so on.
Generative adversarial networks	GANs resemble the process of a chemist proposing new molecule candidates and another chemist providing critical feedback on the proposed molecules. Through time, both chemists are trained better to propose better molecules and distinguish bad molecules.
Gradient-based optimization	Gradient-based optimization refers to optimization with a gradient descent/ascent algorithm, which defines directions by the gradient of the function at a certain point.
Graph neural network	GNNs are a type of neural network that take the graph representation of molecules (that is, atoms as nodes and bonds as edges) and model the interaction between neighbouring atoms, which learn a global representation by aggregating node features of the individual atoms.
Invariance	Invariance is a property where one object remains unchanged after certain transformations.
Latent variable models	Latent variable models (LVMs) are a type of statistical model that relate observational variables to latent variables. Many deep generative models can be considered latent variable models, such as VAEs, normalizing flows and autoregressive models.
Markov chain Monte Carlo	MCMC is a sampling method that samples from the target distribution via constructing a Markov chain and lets the equilibrium of the Markov chain be the target distribution.
Monte Carlo tree search	MCTS is a heuristic search method that estimates the expected improvement of all the branches via Monte Carlo sampling and prioritizes the promising branches to search.
Normalizing flows	Normalizing flows are a type of deep generative model that learn a series of transformations of the complex data distribution into a simple distribution (for example, Gaussian). Continuous normalizing flows (CNFs) are the continuous perspective of discrete sets of transformations that relaxes the restrictions on the neural architectures. Flow matching is an alternative training scheme to maximum likelihood that greatly improves the training efficiency of CNFs. In the context of chemistry, they help transform complex molecular distribution to simpler ones and the reverse process generates new molecules.
Optimal transport	Optimal transport measures the optimal cost of transportation from one distribution to another.
Recurrent neural network	Recurrent neural networks (RNNs) are a type of neural network that process a sequence of data; later input is dependent on the former input. The chemical metaphor is a sequence of reactions where later reactions may be affected by the previous reactions.
Reinforcement learning	Reinforcement learning is an ML method that involves an agent in an environment learning to take actions to maximize the reward signal. Reinforcement learning is typically formulated as a Markov decision process, which consists of states, actions, an environment and rewards: the agent receives a reward based on its actions and the environment, and the reward signal determines the next action and state.
Scaffold hopping	Scaffold hopping refers to changing to a similar scaffold while maintaining the desired properties of the original compound.
Scaffold elaboration	Scaffold elaboration refers to thoroughly searching within the chemical space of the scaffold.
Structure-based drug design	The goal of SBDD is to design a drug molecule that could bind tightly to a target structure (for example, target protein).
Symmetry	Symmetry refers to an object that is invariant with certain transformations such as translation, rotation, reflection and so on.
Transformer	Transformers comprise a type of neural network that model a molecule by learning all pairwise relationships over all atoms. They can better handle long-range interaction and are more suitable for scaling up the number of parameters as foundation models as in the case of GPT.
Variational autoencoders	Variational autoencoders (VAEs) learn an encoder to map the molecule into a continuous low-dimensional embedding space and a decoder to reconstruct the molecule back to its original space. Chemically, it emulates a chemist that extracts essential descriptors of the molecule and inverse designs the molecule based on the extracted descriptors.
Virtual screening	Virtual screening is a computational technique that exhaustively searches over large-scale molecule libraries/chemical space to identify drug molecules with desirable pharmaceutical properties.

https://doi.org/10.1038/s42256-024-00843-5

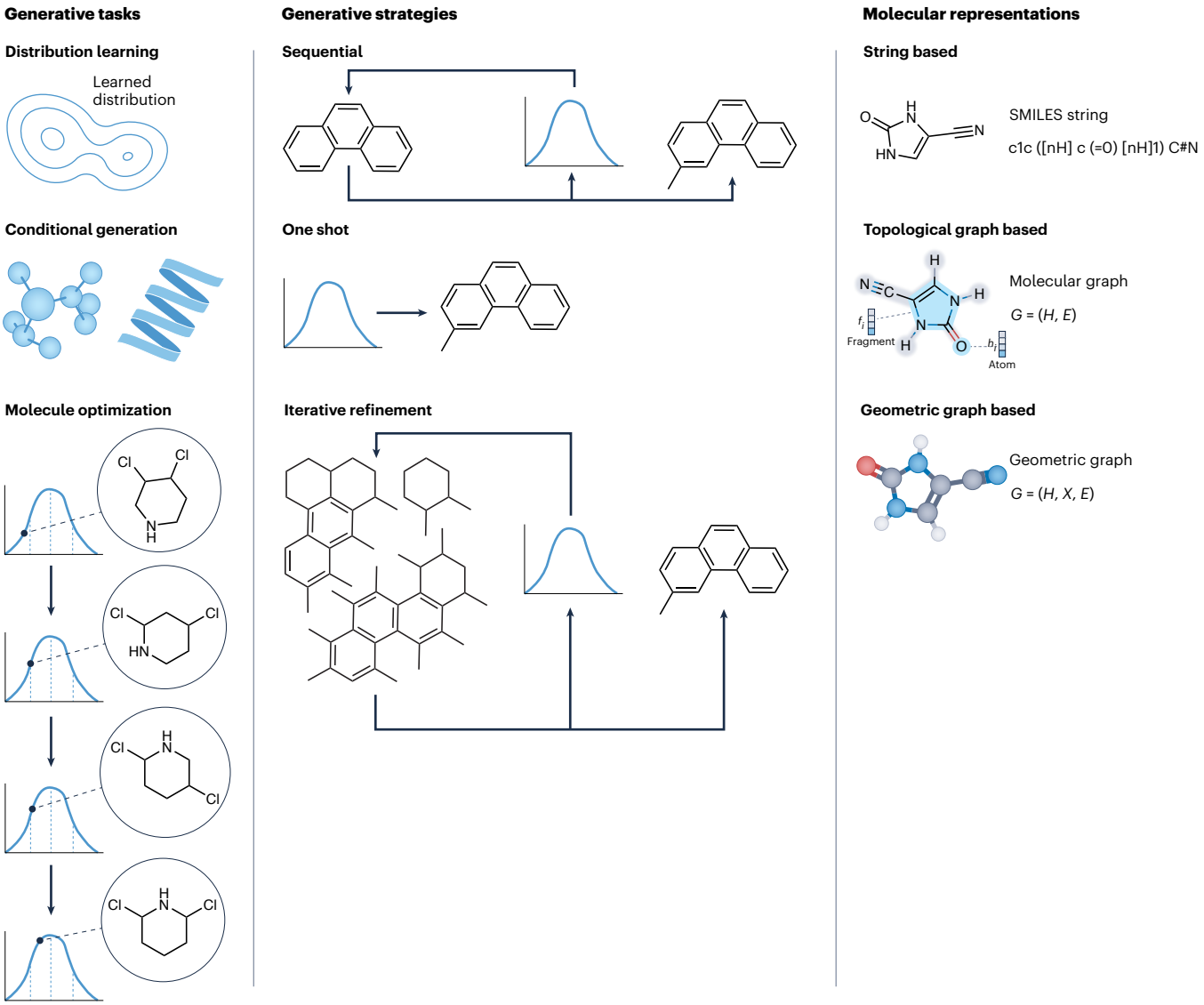

Fig. 2 | Illustrations for generative tasks, generative strategies and molecular representations. Left: generative tasks. Primarily, these encompass three fundamental components: distribution learning, conditional generation and molecule optimization. These aspects represent the core mechanics of the generative process in ML applications within drug discovery. Middle: generative strategies. These strategies typically fall into three categories: sequential or autoregressive generation, one-shot generation and iterative refinement. These paradigms dictate the approach for the generation process, each with unique benefits and applications. Right: molecular representations. There are three mainstream molecular representations used in this field. A popular string-based representation is SMILES, which provides a textual representation of the molecule structure. Topological graph-based representations utilize molecular graphs, a more natural representation of molecules. Finally, geometric graph-based representations take the form of geometric graphs, incorporating spatial data into the representation. Each representation offers different advantages and utilized according to the specific requirements of the task. $G = (H, X, E)$ represents a molecular graph with a set of vertices (H), a set of edges (E), and a set of Cartesian coordinates (X) if available. f_i and h_i denote the feature vectors for fragments and atoms, respectively.

generation process. With this distribution, novel molecules can be sampled that mimic the set of training molecules.

Applications. Distribution learning has two primary applications in real-world drug discovery. First, it can be used to construct targeted virtual screening libraries by sampling the chemical space based on a learned distribution of molecules. Second, although distribution learning does not directly enable generation of molecules with desirable properties, it can generate molecules that resemble the training set by sampling from the learned distribution. This can assist the design of 'me-too' compounds, which are structurally similar to known drugs and can be optimized to improve their efficacy or reduce their side effects. However, design criteria in medicinal chemistry can impose constraints on the suitability of molecules; while sampling from learned distributions can be guided towards suitable candidates, this procedure falls into conditional generation and molecule optimization.

Conditional generation

In conditional generation, the objective is to generate molecules that satisfy or possess specific attributes or properties, rather than simple random sampling from a distribution of molecules. Broadly, conditioning can be categorized into four types: property conditioned, molecular (sub) structure conditioned, target conditioned and phenotype conditioned.

- Property-conditioned generation refers to generating molecules with specific properties, such as binding affinity, synthetic

accessibility, ADME/T (absorption, distribution, metabolism, excretion and toxicity) profiles, other characterizations of biological activity, or development liability. Property-conditioning typically requires property-labelled molecules to access the conditional distribution of molecules. Chemogenomic data can be leveraged by training drug–target interaction predictors, which can be employed in conditional generation through classifier guidance.
- Molecular (sub)structure-conditioned generation generates molecules with specific structural constraints, such as designing partial structures, scaffold hopping, linker design, redesigning whole structures (lead optimization) or whole-molecule conditioning (conformer generation). Conditioning can be used to penalize the generation of molecules containing undesirable substructures, such as structural alerts.
- Target-conditioned generation aims to produce molecules with high binding affinities to specific disease-associated biomolecular targets[21]. Although property-conditioned generation can condition on scalar-valued binding affinity measurements to specific targets, it differs substantially from target-conditioned generation, which leverages explicit access to the target structure, facilitating molecular design through incorporating direct modelling of target–ligand interactions.
- Phenotype-conditioned generation involves learning phenotypic fingerprints from cell-based microscopy or other bioassay readouts such as transcriptomic data to provide conditioning signals to guide generation towards a direct biological outcome[22]. These data can be used learn joint embedding spaces (for example, via contrastive learning), which can be explored with generative models[23,24].

Applications. Conditional generation further equips distribution learning to handle conditional inputs. This constrains chemical space such that generated outputs are biased towards molecules satisfying certain design criteria.

- Conditional generation has direct applications to traditional problems such as linker design, lead optimization and scaffold hopping[25–28].
- A less considered application of conditional generation is patent circumvention such that designs are sufficiently distinct to existing chemical matter to bypass intellectual property liabilities[19].
- There is little work on conditional generation of highly specific molecules given a panel of targets, despite the common use of ligand–target activity matrices[29]. One study has explored the negative design of a kinase inhibitor using an evolutionary algorithm and a diffusion model[30].

Molecule optimization

Although conditional generation provides an efficient way to discover molecules with conditions as constraints, real-world drug discovery campaigns often involve optimizing properties with expensive and, typically, non-differentiable oracle functions. This is further complicated by optimization goals typically being multi-objective.

Applications. Molecule optimization has a crucial role in drug discovery by fine-tuning drug candidates' properties to improve their safety, efficacy and pharmacokinetic profile. This involves introducing small modifications to the molecular structure of a candidate to optimize drug-likeness properties, such as solubility, bioavailability and target affinity to enhance the therapeutic potential of candidates and increase downstream likelihood of success in clinical end points.

Generative molecular design methodologies

Several generative design choices need to be made when devising ML methods. We address four key design choices that largely affect the performance of generative methods: (1) representation types and the associated neural architectures, (2) generative methods, (3) generation strategies and (4) molecule optimization strategies. We stratify the representative literature along these dimensions in Table 2.

Molecular representations

Given data, the first step in developing a neural architecture for molecule generation is to determine the machine-readable input and output representations of molecular structures. Input representations enable the effective infusion of appropriate inductive biases into modelling. Output representations determine the optimization landscape search space of molecules. Moreover, representation types determine the suitability of families of generative methods, for example, discrete search algorithms can only be applied to combinatorial representations such as graphs and strings. While various input representations have been studied (this section), characterization of the trade-offs of the representation types and the neural architecture encoding them remains unclear. It is noted that conversions between representations are not necessarily bijective for molecules, for example, density maps and fingerprints cannot identify unique molecules and further techniques need to be developed to address the non-trivial mapping problem. Common molecular representations include strings, two-dimensional topological graphs and three-dimensional (3D) geometric graphs (Fig. 2). Further descriptions are provided in Supplementary Information section 1.

- String-based molecular structures are often encoded as strings such as simplified molecular-input line-entry system (SMILES)[31] or self-referencing embedded strings (SELFIES)[32]. SMILES represents molecules with syntactic rules, but strings can be invalid; SELFIES improves validity by refining these rules. Molecular strings are commonly encoded as sequential data by recurrent networks and transformer models (Table 1).
- Topological and geometric graph-based atoms and bonds are often represented as nodes and edges in topological graphs. Graph neural networks (GNNs) (Table 1) are commonly used to model graph-structured molecular data, updating node and edge features based on adjacent nodes. Where 3D information is available and relevant, geometric GNNs are often used to capture application-dependent symmetries in 3D space, such as translation and rotation invariance or equivariance (Table 1).

Representation granularity is another consideration for the design of generative models. Typically, approaches leverage either atoms or molecular fragments as the fundamental compositional units during generation. Fragment-based representations coarsen molecular structures into larger units containing groups of atoms carrying hierarchical information, such as functional group identity, thereby aligning with traditional fragment-based or pharmacophoric drug design approaches.

Generative methods

Deep generative models comprise a class of methods that estimate the probability distribution of the data and draw samples from the learned distribution (also called distribution learning).

- Variational autoencoders (VAEs)[33] (Table 1). Owing to their flexibility and the balance between efficiency and accuracy, VAEs have been a popular deep generative model for molecule generation since early developments in the field. CVAE[34] and GraphVAE[35] develop VAE-based methods employing SMILES and molecular graph representations, respectively.
- Generative adversarial networks (GANs)[36] (Table 1). Several GAN-based methods for molecule generation have been developed. Among them, ORGAN[37] and MolGAN[38] are representative works for molecule generation using SMILES and graph representations, respectively.

https://doi.org/10.1038/s42256-024-00843-5

Table 2 | Comparison of representative methods

Name	Method	Input	Output	Fragment	Sequential	Validity	Distribution	Property	(Sub) structure	Target	Goal oriented
CVAE[34]	VAE	SMILES	SMILES	✗	One shot	✗	✓	Single	✗	✗	BO
GraphVAE[35]	VAE	Graph	Graph	✗	One shot	✗	✓	✗	✗	✗	Conditional
JT-VAE[63]	VAE	Graph	Graph	✓	Sequential	✓[a]	✓	Single	✗	✗	BO
DVAE[122]	VAE	Graph	Graph	✗	One shot	✓[a]	✓	Single	✗	✗	Conditional
RationaleRL[123]	VAE	Graph	Graph	✓	Sequential	✗	✓	Multi	✗	✗	RL
ChemSpacE[66]	LVM	Graph	Graph	✓	One shot	✓[a]	✗	Multi	✗	✗	Latent space traversal
QMO[124]	LVM	SMILES	SMILES	✗	One shot	✗	✓	Multi	✗	✗	Optimization
GraphNVP[125]	NF	Graph	Graph	✗	One shot	✗	✓	Single	✗	✗	Interpolation
AAE[126]	GAN + AE	Fingerprint	Fingerprint	✗	One shot	✗	✓	Single	✗	✗	Conditional
ORGAN[37]	GAN	SMILES	SMILES	✗	One shot	✗	✓	Multi	✗	✗	RL
MolGAN[38]	GAN	Graph	Graph	✗	One shot	✗	✓	Multi	✗	✗	RL
MolecularRNN[44]	AR	Graph	Graph	✗	Sequential	✓[a]	✓	Single	✗	✗	RL
ChemTS[56]	AR	SMILES	SMILES	✗	Sequential	✗	✓	Single	✗	✗	MCTS
REINVENT[57]	AR	SMILES	SMILES	✗	Sequential	✗	✓	Multi	✗	✗	RL
DeLinker[127]	VAE	Graph	Graph	✗	One shot	✓[a]	✓	✗	Fragment	Protein	Conditional
GraphAF[40]	AR + NF	Graph	Graph	✗	Sequential	✓[a]	✓	Single	✗	✗	RL
GraphEBM[128]	EBM	Graph	Graph	✗	One shot	✓[a]	✓	Multi	✗	✗	EBM
DiGress[129]	Diffusion	Graph	Graph	✗	Iterative	✗	✓	Multi	✗	✗	Conditional
GCPN[59]	RL	Graph	Graph	✗	Sequential	✓[a]	✗	Single	✗	✗	RL
GB-GA[55]	GA	Graph	Graph	✗	One shot	✗	✗	Single	✗	✗	MCTS
GA+D[130]	GA	SELFIES	SELFIES	✗	One shot	✓	✗	Single	✗	✗	GA
SynNet[96]	GA	Fingerprint	Fingerprint	✓	One shot	✗	✗	Single	✗	✗	GA
AutoGrow 4.0[131]	GA	Graph	Graph	✓	One shot	✓	✗	Single	Lead molecule	Protein	GA
Reinforced GA[60]	GA + RL	Graph	Graph	✓	One shot	✓	✗	Single	✗	Protein	RL + GA
MIMOSA[61]	MCMC	Graph	Graph	✓	Iterative	✓	✗	Multi	✗	✗	MCMC
GFlowNet[53]	Flow network	Graph	Graph	✓	Sequential	✓[a]	✗	Single	✗	✗	Flow network
DST[62]	None	Graph	Graph	✓	Iterative	✓	✗	Multi	✗	✗	Optimization
GraphDG[132]	VAE	Graph	Geometry	✗	One shot	NA	✓	✗	Molecule	✗	Conditional
GeoMol[133]	OT	Graph	Geometry	✓	Sequential	NA	✗	✗	Molecule	✗	Conditional
E-FM[134]	FM	Graph	Geometry	✗	Iterative	NA	✗	✗	Molecule	✗	Conditional
EquiBind[135]	OT	Graph	Geometry	✗	One shot	NA	✗	✗	Molecule	Protein	Conditional
Torsional Diff[136]	Diffusion	Graph	Geometry	✗	Iterative	NA	✓	✗	Molecule	✗	Conditional
DiffDock[137]	Diffusion	Graph	Geometry	✗	Iterative	NA	✓	✗	Molecule	Protein	Conditional
liGAN[138]	VAE	Density	Density	✗	One shot	✗	✓	✗	✗	Protein	Conditional
G-SchNet[45]	AR	Geometry	Geometry	✗	Sequential	✗	✓	Single	✗	✗	Fine-tuning
Drótar et al.[139]	AR + VAE	Geometry	Geometry	✗	Sequential	✓	✗	✗	✗	Protein	Conditional
3D-SBDD[79]	AR	Geometry	Geometry	✗	Sequential	✗	✓	✗	Fragment	Protein	Conditional
3D-Scaffold[140]	AR	Geometry	Geometry	✗	Sequential	✗	✓	✗	Scaffold	✗	Conditional
GraphBP[141]	AR + NF	Geometry	Geometry	✗	Sequential	✗	✓	✗	✗	Protein	Conditional
E-NF[142]	CNF	Geometry	Geometry	✗	Iterative	✗	✓	✗	✗	✗	None
EDM[47]	Diffusion	Geometry	Geometry	✗	Iterative	✗	✓	Single	✗	✗	Conditional
EquiFM[42]	FM	Geometry	Geometry	✗	Iterative	✗	✓	Singe	✗	✗	Conditional
DiffSBDD[48]	Diffusion	Geometry	Geometry	✗	Iterative	✗	✓	Multi	All substructures	Protein	Conditional

[a]Validity is guaranteed by additional post-processing validity checks or correction steps. The table is separated by whether methods involve geometric representations. The ticks (✓) and crosses (✗) indicate whether a particular method possesses a certain property or capability. A tick means that the method has that property or capability, whereas a cross means that it does not. BO, Bayesian optimization; FM, flow matching; GA, genetic algorithm; NF, normalizing flow; RL, reinforcement learning; OT, optimal transport.

- Normalizing flows[39] (Table 1). Normalizing flows are desirable as they offer access to the exact log-likelihood of the distribution of molecules but impose additional restrictions on the model to be invertible. However, they have proven popular by showing strong practical performance on molecule generation[40]. It is noted that the discrete sets of invertible layers can be considered as a continuous transformation that relaxes the restrictions on the neural architectures. Flow matching[41] is an alternative optimization objective in addition to maximum likelihood for continuous normalizing flow that largely improves the performance and has been applied to molecular design[42].
- Autoregressive models[43] (Table 1). Although autoregressive models can be trained by directly maximizing the log-likelihood of the molecule distribution, they introduce an additional assumption about the order of the generation trace for molecules, typically requiring supervision through teacher forcing. Many autoregressive models have been developed to generate molecules with graph and point cloud representations[44,45].
- Diffusion models[46] (Table 1). EDM[47] devises a diffusion model equivariant to translation and rotation to generate molecules in 3D space. It has also been applied to target-conditioned design, linker design and molecular conformation generation[48–50]. It has also been connected with an evolutionary algorithm for molecule optimization and inpainting for substructure-conditioned generation[48].

Generation strategies

Generation strategies refer to the manner in which models output molecular structures, and can generally be grouped into one-shot, sequential or iterative refinement (Fig. 2).

- One-shot generation produces a complete molecular structure in a single forward pass of a model. This approach is often limited by difficulty in producing realistic and plausible molecular structures with high accuracy. Furthermore, one-shot generation often cannot accommodate explicit constraints, such as valence constraints[51], which are critical for ensuring the accuracy and validity of the generated structures.
- Sequential generation constructs the molecular structure through a sequence of steps, either in atoms or fragments[52]. It is often easy to inject valency constraints into sequential generation, which improves the quality of the generated molecules. However, the major limitation of sequential generation is that it requires the definition of an arbitrary ordering of the generation trace during training and results in slow inference. GFlowNets[53] overcomes this problem by forcing a reward function to be proportional to the probability of sampling the same molecule, regardless of the trajectory taken.
- Iterative refinement frames design by predicting a series of updates to manipulate predictions, side-stepping difficulties in one-shot methods, as exemplified by the successful application of recycling and the recurrent structure module to refine backbone frames in AlphaFold2[54]. This approach has inspired related strategies to molecule generation. Diffusion models[46] are a prevalent technique, generating new data through a sequence of denoising steps. So far, diffusion models have been applied to a variety of molecule generation problems, including conformer generation[50], SBDD[48] and linker design[49].

Optimization strategies

Combinatorial optimization. For combinatorial encodings of molecules, such as graphs or strings, optimization techniques from the combinatorial optimization domain can be directly applied.

- Genetic algorithm (Table 1) focuses on the interaction between candidates within a defined population with crossover and mutation operations to increase the diversity of the offspring[55]. Each new population is selected by an oracle function as the starting point of the next iteration.
- Monte Carlo tree search (MCTS) (Table 1) starts with a root node and iteratively updates and expands the search tree by the reward function over the simulated molecule in each step until a stopping criterion is reached[55,56].
- Reinforcement learning (Table 1) has been used to estimate expected rewards for search branches, suppressing the randomness of traditional heuristic search by adaptively prioritizing promising branches. Reinforcement learning methods have been developed to generate molecules using policy networks to predict a series of actions for the generation process[57–60].
- Markov chain Monte Carlo (MCMC) (Table 1) formulates the optimization problem as a sampling problem based on probabilistic formulation and constructs a Markov chain to traverse the chemical space[61].

Continuous optimization. Molecules can be either represented or encoded in continuous domains, that is, point clouds and geometric graphs embedding in Euclidean space, or deep generative models that encode discrete data into a continuous latent space.

- Gradient-based optimization (Table 1) requires a differentiable oracle function (often trained neural networks) and directly optimizes the molecule input[62] or encoded latent vectors[34].
- Bayesian optimization (Table 1) uses surrogate models (often Gaussian processes) to estimate the uncertainty of the optimization process. Bayesian optimization can be used to optimize expensive oracle functions and can be combined with active learning to prioritize experiments. Frequently, Bayesian optimization is combined with VAEs, enabling property optimization through latent space navigation[34,63–65].
- Latent space traversal (Table 1) directly leverages simple heuristics (for example, linear separability) about the learned latent space of deep generative models and often achieves a good trade-off between accuracy and efficiency[66].

Evaluating designs of generative ML

For practical applications, generated molecules must be qualitatively and quantitatively evaluated. Depending on the model deployment, careful consideration into how metrics are calculated is important. Often, individual property profiles of generated molecules are reported, for example, binding affinity, which fails to consider the fundamental multi-objective optimization challenge of drug discovery; for example, jointly optimizing binding affinity and ADME/T properties.

Although we discuss evaluation methodologies for ML-based models, it is important to acknowledge existing computational chemistry tools for molecular design. Software from Schrödinger[67] and OpenEye[68] provide industry-standard tooling, including docking[69,70] and molecular dynamics simulations. Comparisons of generative modelling should also include traditional methods, with appropriate resources made available to both approaches.

Computational evaluation

Standard metrics include validity, uniqueness and novelty (Table 3), which broadly assess the ability to generate valid molecules and extrapolate beyond the training data. Generative models can cover an expansive chemical space, despite being trained on comparatively small datasets[71]. Early benchmarks such as GuacaMol[72] and MOSES[73] highlight these metrics and propose a set of diverse benchmark tasks. However, these metrics by themselves provide only a baseline assessment, as the goal is to generate molecules satisfying a target objective.

https://doi.org/10.1038/s42256-024-00843-5

Physicochemical properties. Commonalities in requirements for drug-like molecules typically result in compounds satisfying empirical profiles of physicochemical properties as codified in Lipinski's rule of 5[74]. Often, measured properties include log*P*, solubility or aggregates thereof, such as the quantitative estimate of drug likeness[75]. It is noted that these are coarse descriptors and many current drugs would be deemed 'unsuitable'; blindly optimizing for these metrics is currently a common pitfall in the ML literature.

Three-dimensional structural design. Molecular docking is the most common method to assess protein–ligand complementarity, and can be explicitly optimized for by generative models[30,76]. It is noted that many docking algorithms reward large molecules on the basis of forming more interactions, which can be promiscuous binders[77]. A complementary metric is ligand efficiency, where the score is normalized per heavy atom[78]. Absent protein structural information, LBDD performs pharmacophore matching (Table 3) to mimic ligand–protein interactions of a known molecule. Several studies jointly generate molecules and their 3D pose within a binding site[48,49,79]. However, Harris et al.[80] demonstrated that many methods produce poses of dubious biophysical plausibility, often with steric clashes. Furthermore, the common practice of re-docking generated ligands masks issues with poses during generation[80].

Diversity. Diversity is an important and oft-overlooked metric. To maximize the likelihood that generated ideas are actionable, that is, deemed reasonable candidates after post-processing to propose for experimental validation, it is usually important to generate diverse solutions, or at least possess the capability to do so. An exception to encouraging diversity is lead optimization, where the goal may be to explore minor structural modifications on an intermediate molecule to fine-tune its properties.

Diversity-assessment metrics include scaffold diversity, that is, how many unique scaffolds are generated that satisfy the target objective. Another metric is pharmacophore diversity or functional group diversity. In addition, an internal diversity metric was introduced in MOSES and quantifies the fingerprint similarity within a set of generated molecules[73].

Synthetic accessibility. Molecules proposed by generative models typically fail to explicitly account for synthetic accessibility. Addressing synthesizability is an ongoing challenge, considering many proposed ideas may not have known synthetic routes and a chemist can only triage a fraction of proposed ideas. Consequently, this has become a core focus of research and proposed solutions can be broadly categorized into heuristic-based, computer-aided synthesis planning (CASP) and generative model-based approaches.

Many heuristic-based approaches (Table 3) count substructure frequency compared with known synthesized molecules[81–83]. Alternatively, ref. 84 tasked a neural network to learn a behaviour where products are more complex than their reactants. It is noted that these scores assess molecular complexity as a proxy for synthetic feasibility, and one should only expect to derive correlation from such metrics. Contrariwise, CASP tools can post-process generated ideas by suggesting potential retrosynthetic routes[85–87]. MCTS[88,89] recursively decomposes a target molecule into precursors and can explicitly encode synthetic feasibility through reaction templates. However, the generalizability of these methods deeply depends on the template set and abstraction level. A related paradigm involves training models on the output of CASP tools to accelerate inference[90,91].

Recent studies have proposed to incorporate synthetic constraints into generative models explicitly. Exemplary studies have trained models capable of predicting how to combine commercially available building blocks with reaction rules, while optimizing for target properties[92–97]. Similarly, ref. 98 biases generation with user-specified reaction rules to promote synthetic compatibility. Other studies perform de novo design using reaction-based expansion from a library of building blocks[99,100]. Again, performance depends on template quality and granularity in the encoded chemistry. While continued efforts in the research community will drive the field forward, synthesizability will probably remain an important consideration in post-processing pipelines for the foreseeable future.

Table 3 | Common metrics (but not exhaustive) to assess molecular generative models

Category	Metric	Definition
Standard	Validity	Percentage of viable molecules, that is, obey chemical laws.
	Uniqueness	Fraction of non-repeated molecules.
	Novelty	Fraction of molecules not in the training set.
Physicochemical	log*P*	Octanol–water partition coefficient that measures lipophilicity.
	logS	log of the solubility.
	tPSA	Topological polar surface area measuring the surface spanned by the polar atoms of a molecule.
	QED	Quantitative estimate of drug likeness.
	Lipinski	'Rule of 5' empirical guidelines: ≤5 hydrogen-bond donors, ≤10 hydrogen-bond acceptors, ≤500 molecular weight, ≤5 log*P*.
Structural	Molecular docking	Simulation predicting the orientation and binding affinity of a molecule against a target protein. The number of steric clashes and interaction fingerprints are important.
	Pharmacophore matching	Volume and functional groups overlap relative to a known ligand.
Diversity	Internal diversity	Fingerprint similarity within a set of generated molecules.
Synthesis	SA score	Heuristic measuring synthetic complexity by scoring substructures.
	SC score	Feed-forward neural network trained on Reaxys data to predict synthetic complexity.

Standard metrics assess the base behaviour of the generative model while the other categories represent properties to optimize for. In real-world applications, a combination of metrics (necessarily multi-objective) should be used to assess the generative performance.

Experimental validation of generated molecules

Generated molecules can only be unequivocally validated by wet-lab experimentation and is in stark contrast to existing studies that predominantly focus solely on the computational contribution. While generative models are far from without weaknesses, the disconnect between prediction and experiment is also attributed to the expertise required to perform such validation. Albeit much fewer, we summarize existing studies reporting experimental validation (Table 4) and highlight selected examples of generated molecules (Fig. 3).

Observations from the literature. Most studies reporting experimental validation utilize either RNNs and/or VAEs and operate on SMILES (Table 4). We make four key observations: first, SMILES, although capturing limited 3D information, are an efficient representation

https://doi.org/10.1038/s42256-024-00843-5

Table 4 | Experimentally validated small-molecule generative design case studies

Model	Input	Output	Design task	Target	Hit rate	Outcome	Publication year
Distribution learning							
LSTM RNN[146]	SMILES	SMILES	De novo	RXR	4/5 (80%)	nM agonist	2018
LSTM RNN[147]	SMILES	SMILES	De novo	RXR	2/4 (50%)	μM agonist	2018
GraphGMVAE[148]	Graph	SMILES	Scaffold hopping	JAK1	7/7 (100%)	nM inhibitor	2021
LSTM RNN[106]	SMILES	SMILES	De novo	LXR	17/25 (68%)	μM agonist	2021
LSTM RNN[149]	SMILES	SMILES	De novo	RORγ	3/3 (100%)	μM agonist	2021
LSTM RNN[150]	SMILES	SMILES	De novo	FLT-3	1/1 (100%)	μM inhibitor	2022
GGNN GNN[151]	Graph	Graph	Fragment linking	CDK8	9/43 (21%)	nM inhibitor	2022
GRU RNN[152]	SMILES	SMILES	De novo	Bacteria	0/1 (0%)[a]	μM inhibitor	2022
BiRNN encoder–decoder[153]	SMILES	SMILES	De novo	DDR1	2/2 (100%)	nM inhibitor	2021
GRU RNN[154]	SMILES	SMILES	Reaction-based de novo	MERTK	15/17 (88%)	μM inhibitor	2022
LSTM RNN[155]	SMILES	SMILES	De novo	PI3Kγ	3/18 (17%)	nM inhibitor	2023
Transformer[156]	SMILES	SMILES	Fragment linking	TBK1	1/1 (100%)	nM inhibitor	2023
VAE and transformer[157]	SMILES	SMILES	Fragment hopping/linking	CDK2	17/23 (74%)[c]	nM inhibitor (MC)[b]	2023
LSTM RNN[102]	SMILES	SMILES	De novo	Nurr1γ	2/6 (33%)	nM inhibitor	2023
Graph transformer-LSTM RNN[158]	Graph	SMILES	De novo	PPARγ	2/2 (100%)	μM agonist	2023
Goal oriented							
DNC[159]	SMILES	SMILES	De novo	Kinases	0[d]	μM inhibitor	2018
AAE (conditional)[160]	SMILES	SMILES	De novo	JAK3	1/1 (100%)	μM inhibitor	2018
VAE[19]	SMILES	SMILES	De novo	DDR1	4/6 (67%)	nM inhibitor[b]	2019
LSTM RNN[108]	SMILES	SMILES	De novo ligand based	DDR1	4/6 (67%)	nM inhibitor	2021
Stack-GRU RNN[161]	SMILES	SMILES	De novo	EGFR	4/15 (27%)	nM inhibitor	2022
LSTM RNN (conditional)[107]	SMILES	SMILES	De novo	RIPK1	4/8 (50%)	nM inhibitor[b]	2022
Chemistry42[20]	Mixed	Mixed	De novo structure based	CDK20	6/13 (46%)[a]	nM inhibitor	2023
Chemistry42[162]	Mixed	Mixed	De novo structure based	CDK8	1/1 (100%)	nM inhibitor[b]	2023
Chemistry42[105]	Mixed	Mixed	De novo structure based (R-group)	SIK2	6/6 (100%)	nM inhibitor	2023
VAE[163]	SMILES	SMILES	De novo structure based	KOR	2/5 (40%)	μM antagonist	2023
Chemistry42[164]	Mixed	Mixed	De novo structure based	PHD enzymes	1/1 (100%)	nM inhibitor[b]	2024
GRU RNN-transformer[165]	SMILES	SMILES	De novo activity model	NLRP3	0[e]	nM inhibitor[b]	2024
Transformer-VAE (conditional)[166]	Geometry-SMILES	SMILES	De novo	Tuberculosis ClpP	1/6 (17%)[a]	μM inhibitor	2024
QC-LSTM RNN-Chemistry42[167]	SMILES	SMILES	De novo structure based	KRAS	1/12 (8%)[a]	μM inhibitor	2024
Graph transformer[168]	Graph	Graph	De novo activity model	MGLL	1/3 (33%)[a]	μM inhibitor	2024
Chemistry42[169]	Mixed	Mixed	Fragment linking	Polθ	4/6 (67%)	μM inhibitor[b]	2024
Chemistry42[114,115]	Mixed	Mixed	De novo structure based	TNIK	Unknown[f]	nM inhibitor[b]	2024
Attention-convolution layers[170]	Substructure vector	SMILES	Scaffold based	Factor Xa	Unknown[g]	μM inhibitor	2024
Flow (conditional)[171]	Geometry	Geometry	De novo	HAT1 and YTHDC1	0/2 and 0/3 (0%)[a]	Both μM inhibitor[a]	2024

https://doi.org/10.1038/s42256-024-00843-5

Table 4 (continued) | Experimentally validated small-molecule generative design case studies

Model	Input	Output	Design task	Target	Hit rate	Outcome	Publication year
Activity model (MCTS)[97]	Variable	Variable	Reaction based	Bacteria	6/58 (10%)	μg inhibitor[b]	2024
Chemistry42[172]	Mixed	Mixed	De novo structure based	KIF18A	Unknown[h]	nM inhibitor[b]	2024
Diffusion (conditional)[173]	Geometry	Geometry	Lead optimization	CDK2	7/7 (100%)	nM inhibitor (MC)	2024

The table is separated by whether distribution learning or goal-oriented generation was used. Hit rate is defined as the percentage of actives in an in vitro assay (<10 μM potency) out of all reported synthesized designs. Hit rate is defined strictly based on generated designs and omits actives from manual domain-expert modifications. Denoted nanomolar (nM) potent if the most potent design (including generated and domain-expert designed derivatives based on generated designs) possessed half-maximal inhibitory concentration (IC_{50}) or half-maximal effective concentration (EC_{50}) values <10 nM. R-group indicates a core scaffold was fixed and only variable R-groups were generated. Chemistry42[104] contains over 40 generative models with varying input and output encompassing SMILES, fingerprints and graphs. To summarize this information, we have denoted the input and output of Chemistry42 as 'mixed'. 'Variable' denotes variable inputs/outputs were used depending on the model. [a]There were additional actives with a concrete potency measured >10 μM. [b]In vivo validation was also performed. [c]Some manual modifications were made to the generated molecules for synthetic ease. [d]An in-house library was screened to identify high-Tanimoto-similarity molecules to the generated set. Therefore, none of the generated molecules were directly experimentally validated. [e]12 generated molecules were selected for docking and analysis of the binding poses short-listed two scaffolds. Derivatives were designed based on these two scaffolds, resulting in a nM inhibitor. Therefore, none of the generated molecules were directly experimentally validated. [f]79 molecules were synthesized in total[114]. [g]8 commercially available generated molecules were purchased. The most potent affinity was reported. [h]110 molecules were synthesized in total[145]. AAE, adversarial autoencoder[144]; BiRNN, bidirectional recurrent neural network; DNC, differentiable neural computer[143]; QC, quantum computing; GGNN, gated graph neural network; GRU, gated recurrent unit neural network; LSTM, long short-term memory recurrent neural network; MC, macrocycles.

with encoded chemical knowledge. This makes SMILES-based models well suited for distribution learning and fine-tuning with small datasets[101,102]. Fine-tuning is the dominant method used in the distribution learning examples in Table 4. Second, many studies with experimental validation target kinases, which are prevalent in popular open-source datasets such as ChEMBL[103]. Third, most goal-oriented approaches use reinforcement learning (either alone or as a component) as the optimization algorithm, and encompass both LBDD and SBDD. These studies also extensively feature Chemistry42[104], which is a proprietary generative platform from Insilico Medicine containing over 40 models. Lastly, AlphaFold[54] predicted structures can be successfully used for generative SBDD[20,105].

Generated molecules have novel scaffolds. Figure 3 highlights selected examples of generated designs encompassing distribution learning and goal-oriented generation approaches. Reference[106] used a SMILES-based RNN to perform distribution learning to generate liver X receptor (LXR) agonists validated using automated on-chip synthesis, providing a glimpse of the integration of generative models with synthesis platforms. Compound 17 has micromolar potent activity with a sixfold selectivity for LXRα over LXRβ. Reference[102] used the same model in an extreme low-data regime (fine-tuning with a single positive example) to design a high nanomolar potent orphan nuclear receptor related 1 (NURR1) agonist. Using another RNN, ref. 107 demonstrated the combination of generative design and virtual screening by triaging a generated library of potential receptor-interacting protein kinase 1 (RIPK1) inhibitors with pharmacophore features and discovering a nanomolar potent inhibitor with in vivo tolerability. Goal-oriented generative approaches have also been experimentally validated. Reference [108] used REINVENT[57,109] and pharmacophore matching to design a novel nanomolar potent discoidin domain receptor1 (DDR1) inhibitor. In a seminal work, ref. 19, introduced the generative tensorial reinforcement learning (GENTRL) VAE model and successfully demonstrated design, synthesis and experimental validation of a DDR1 inhibitor within 21 days. However, the novelty of the most potent design is a point of contention as it has high structural similarity to potanitib, a known inhibitor. Notably, ref. 20 used a similar model to design a micromolar potent cyclin-dependent kinase 20 (CDK20) inhibitor in 30 days using an AlphaFold[54] predicted structure. A second round of generation identified a nanomolar potent inhibitor. Overall, both distribution learning and goal-oriented approaches have been experimentally validated. Generated molecules have novel scaffolds and demonstrate the possibility of generative design to accelerate discovery. Moreover, apart from generative approaches, in vitro validation has also been demonstrated by leveraging reaction-based expansion[110–112]. Finally, we highlight the growing prevalence of ML-aided commercial drug discovery campaigns from industry that have been exhaustively compiled in ref. 113. A notable example is from Insilico Medicine[114,115], which has progressed the first drug candidate from generative design to phase two clinical trials.

Future directions

In this paper, we have reviewed the current progress of ML-aided molecular design. We started by introducing the problem setting before delving into how design decisions around the granularity of molecular representation and architecture can affect downstream performance. In the latter part of the Review, we discussed the evaluation of generative models, specifically around the importance but also limitations of in silico metrics, and emphasizing that experimental validation should always be the end goal. Finally, we comprehensively report literature examples that have achieved experimental validation, encompassing both academic and industrial efforts. While progress in the field has demonstrated that generative molecular design can accelerate early-stage drug discovery, challenges in adopting such workflows to real-world discovery campaigns remain, partly owing to unrealistic problem formulations and evaluation protocols. Correspondingly, we close our paper with a discussion of current challenges and future opportunities to better align computational efforts to achieve realistic drug discovery end points.

Challenges

- Out-of-distribution generation. Known chemical matter occupies a small fraction of chemical space. While deep generative models can propose molecules outside the training distribution, care must be taken to ensure plausibility.
- Unrealistic problem formulation. Precise formulation of design tasks is crucial to develop models applicable to real-world drug discovery. Fundamental aspects, such as conformational dynamics, the role of water and entropic contributions, are frequently overlooked and unrealistic assumptions such as unlimited access to oracle calls are often made. The latter is encapsulated by the sample efficiency problem and recent studies make progress in efficient goal-oriented generation under a limited oracle budget[116–119].

https://doi.org/10.1038/s42256-024-00843-5

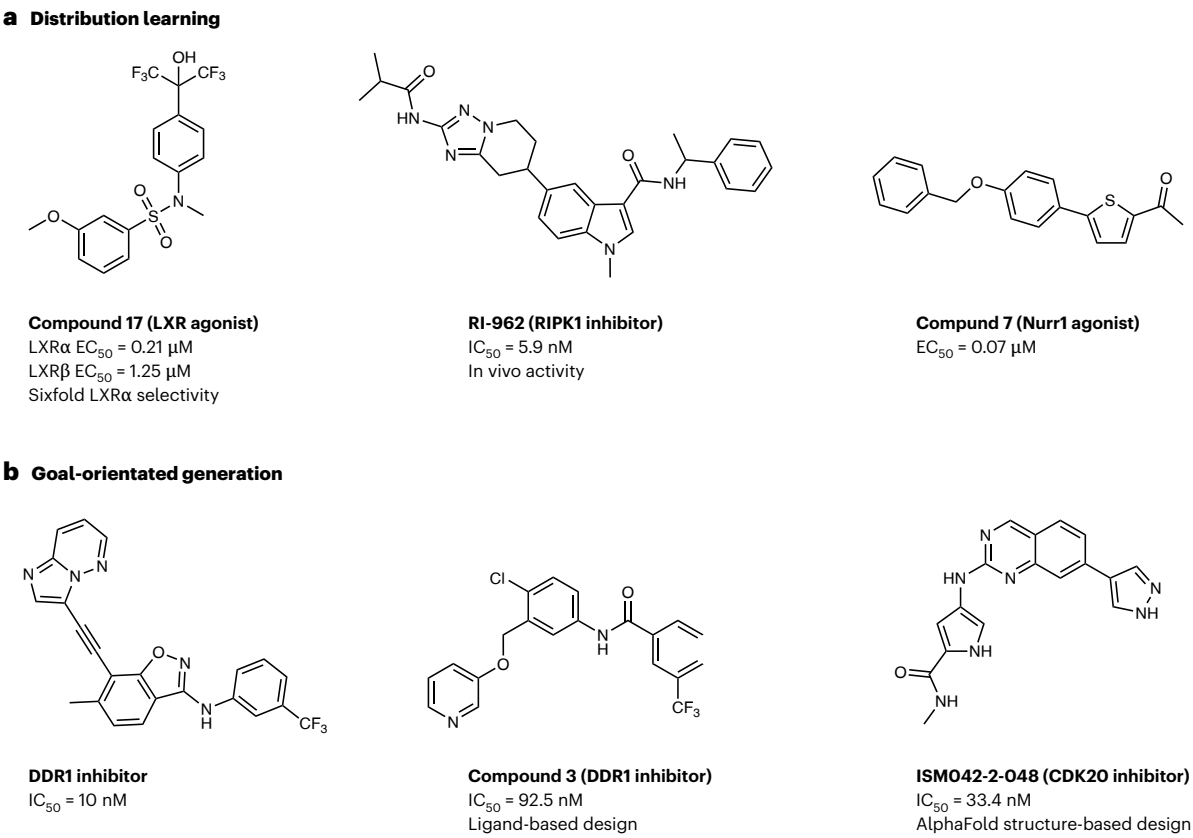

Fig. 3 | Selected examples of experimentally validated generative designs. a, Distribution learning approaches. **b**, Goal-oriented generation approaches.

- Low-fidelity oracles. Effectively scoring designs along dimensions relevant to drug discovery remains difficult and a bottleneck in deploying generative models in industrial settings. For example, high-throughput binding affinity prediction is typically inaccurate in both data-driven and physics-based workflows. While alternative, higher-accuracy oracles exist, their computational demands limit scalability. Moreover, inaccessibility of high-quality labelled data presents a barrier to developing oracles with both high accuracy and manageable compute during inference.
- Lack of unified evaluation protocols. The evaluation protocol used for assessing the quality of a drug candidate is closely linked to how we define what makes a good drug. Easy-to-compute physicochemical descriptors typically used by the ML community remain questionable and certainly paint an incomplete picture of performance. Rigorous comparisons between generative molecular design and virtual screening are also lacking.
- Lack of large-scale studies and benchmarks. Many ML methods have been developed but there are no fair benchmark results on many types of model in different key tasks. For example, only a small fraction of available data is used for training, limiting understanding of model scalability. Recent benchmarks are important contributions to standardizing computational evaluation protocols[80,120].
- Lack of interpretability. Interpretability is an important yet under-explored area for molecular generative models. For example, insights into how the generative or optimization process composes molecules could lead to chemical rules interpretable by medicinal chemists. This is especially important in the small-molecule regime where generative models are often used to submit ideas to medicinal chemists as synthesis barriers precludes testing of all generated designs.

Opportunities

- Applications beyond small-molecule design. The approaches discussed here may have broader applications for designing other richly structured materials, such as polysaccharides, proteins (in particular, antibodies), nucleic acids, crystal structures and polymers[121].
- Large language models show the potential to revolutionize molecular design with text-guided discovery and decision-making as agents, due to vast available training data, including the body of scientific literature. In addition, models tailored or fine-tuned on molecular structures present additional opportunities for researchers to draw on proven advances in natural language processing.
- Later phases in drug development. Molecular design/optimization occupies early phases in drug discovery. However, late-stage failures due to limited efficacy, poor ADME/T profiles and safety concerns are pain points in drug development pipelines. While limited, integrating clinical data into design pipelines is a promising direction to improve the downstream success rate.
- Focused model purpose. The drug discovery pipeline is the result of decades of experience and hard-learned lessons within pharmaceutical companies. ML researchers should aim to design not only pure de novo models, (particularly when lacking the capabilities to characterize them in depth) but also models that are focused on improving specific steps subject to realistic constraints in the multi-year process.
- Self-driving labs. Increasing demand for high-throughput experiments to provide feedback for ML-designed molecules place growing attention on self-driving labs to serve a critical role in expediting the design–make–test–analyse cycle.

References

1. Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput.* **44**, 1000–1005 (2004).
2. Gillette, J. R., Mitchell, J. R. & Brodie, B. B. Biochemical mechanisms of drug toxicity. *Annu. Rev. Pharmacol.* **14**, 271–288 (1974).
3. Gibaldi, M. & Perrier, D. *Pharmacokinetics* (CRC Press, 1982).
4. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
5. Stumpfe, D. & Bajorath, J. Exploring activity cliffs in medicinal chemistry: miniperspective. *J. Med. Chem.* **55**, 2932–2942 (2012).
6. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).
7. Berdigaliyev, N. & Aljofan, M. An overview of drug discovery and development. *Future Med. Chem.* **12**, 939–947 (2020).
8. Ringel, M. S., Scannell, J. W., Baedeker, M. & Schulze, U. Breaking Eroom's law. *Nat. Rev. Drug Discov.* **19**, 833–834 (2020).
9. Aparoy, P., Kumar Reddy, K. & Reddanna, P. Structure and ligand based drug design strategies in the development of novel 5-LOX inhibitors. *Curr. Med. Chem.* **19**, 3763–3778 (2012).
10. Baskin, I. & Varnek, A. in *Chemoinformatics Approaches to Virtual Screening* Ch. 1, 1–43 (Royal Society of Chemistry, 2008).
11. Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082 (1992).
12. Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **10**, 787–797 (2003).
13. Choung, O.-H., Vianello, R., Segler, M., Stiefl, N. & Jiménez-Luna, J. Extracting medicinal chemistry intuition via preference machine learning. *Nat. Commun.* **14**, 6651 (2023).
14. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
Ultralarge-scale virtual screening of a make-on-demand library identified hits with previously unknown chemical motifs that were experimentally validated.
15. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
16. Gorgulla, C. et al. Virtualflow 2.0—the next generation drug discovery platform enabling adaptive screens of 69 billion molecules. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.25.537981 (2023).
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10684–10695 (IEEE, 2022).
18. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
19. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
One of the first studies to experimentally validate ML-generated molecules and highlighted the potential for accelerated drug discovery.
20. Ren, F. et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem. Sci.* **14**, 1443–1452 (2023).
A study performing molecular docking using an AlphaFold-generated structure on ML-generated molecules with experimental validation.
21. Wu, C.-T. et al. COT: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinf. Adv.* **2**, vbac037 (2022).
22. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **11**, 10 (2020).
23. Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S. & Klambauer, G. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nat. Commun.* **14**, 7339 (2023).
24. Nguyen, C. Q., Pertusi, D. & Branson, K. M. Molecule-morphology contrastive pretraining for transferable molecular representation. Preprint at https://arxiv.org/abs/2305.09790 (2023).
25. Schaller, D. et al. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1468 (2020).
26. Imrie, F., Hadfield, T. E., Bradley, A. R. & Deane, C. M. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **12**, 14577–14589 (2021).
27. Guo, J. et al. Link-INVENT: generative linker design with reinforcement learning. *Digit. Discov.* **2**, 392–408 (2023).
28. Torge, J., Harris, C., Mathis, S. V. & Lio, P. DiffHopp: a graph diffusion model for novel drug design via scaffold hopping. Preprint at https://arxiv.org/abs/2308.07416 (2023).
29. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206 (2007).
30. Harris, C. et al. Flexible small-molecule design and optimization with equivariant diffusion models. In *ICLR 2023—Machine Learning for Drug Discovery Workshop* (OpenReview, 2023).
31. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
32. 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.* **1**, 045024 (2020).
33. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In *Proc. 2nd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (OpenReview, 2014).
34. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
One of the first studies to apply a variational autoencoder for molecular design and is a foundational work for many recently reported methods.
35. Simonovsky, M. & Komodakis, N. GraphVAE: towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning—ICANN 2018* (eds Kůrková, V. et al.) 412–422 (Springer, 2018).
36. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, 2672–2680 (2014).
37. 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. Preprint at https://arxiv.org/abs/1705.10843 (2017).
38. De Cao, N. & Kipf, T. MolGAN: an implicit generative model for small molecular graphs. Preprint at https://arxiv.org/abs/1805.11973 (2018).
39. Rezende, D. & Mohamed, S. Variational inference with normalizing flows. In *Proc. 32nd International Conference on Machine Learning* 1530–1538 (PMLR, 2015).
40. Shi, C. et al. GraphAF: a flow-based autoregressive model for molecular graph generation. In *Proc. 8th International Conference on Learning Representations* (OpenReview, 2020).
41. Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. In *Proc. 11th International Conference on Learning Representations* (OpenReview, 2023).

42. Song, Y. et al. Equivariant flow matching with hybrid probability transport for 3D molecule generation. *Adv. Neural Inf. Process. Syst.* **36**, 549–568 (2023).
43. Van Oord, A., Kalchbrenner, N. & Kavukcuoglu, K. Pixel recurrent neural networks. In *Proc. 33rd International Conference on Machine Learning* 1747–1756 (PMLR, 2016).
44. Popova, M., Shvets, M., Oliva, J. & Isayev, O. MolecularRNN: generating realistic molecular graphs with optimized properties. Preprint at https://arxiv.org/abs/1905.13372 (2019).
45. Gebauer, N., Gastegger, M. & Schütt, K. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. *Adv. Neural Inf. Process. Syst.* **32**, 7566–7578 (2019).
46. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
47. Hoogeboom, E., Satorras, V. G., Vignac, C. & Welling, M. Equivariant diffusion for molecule generation in 3D. In *Proc. 39th International Conference on Machine Learning* 8867–8887 (PMLR, 2022).
48. Schneuing, A. et al. Structure-based drug design with equivariant diffusion models. In *NeurIPS 2022 Machine Learning for Structural Biology* (OpenReview, 2022).
One of the first studies to leverage the flexibility of diffusion models to achieve a variety of types of conditional generation and molecule optimization.
49. Igashov, I. et al. Equivariant 3D-conditional diffusion models for molecular linker design. *Nat. Mach. Intell.* **6**, 417–427 (2024).
50. Xu, M. et al. GeoDiff: a geometric diffusion model for molecular conformation generation. In *Proc. 9th International Conference on Learning Representations* (OpenReview, 2021).
51. Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. Constrained graph variational autoencoders for molecule design. *Adv. Neural Inf. Process. Syst.* **31**, 7795–7804 (2018).
52. Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
53. Bengio, E., Jain, M., Korablyov, M., Precup, D. & Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Adv. Neural Inf. Process. Syst.* **34**, 27381–27394 (2021).
54. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
55. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572 (2019).
56. Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. & Tsuda, K. ChemTS: an efficient Python library for de novo molecular generation. *Sci. Technol. Adv. Mat.* **18**, 972–976 (2017).
57. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **9**, 48 (2017).
One of the first studies applying reinforcement learning to molecular design and is the first version of 'REINVENT', an industrially used method, which is still under active development.
58. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
59. 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.* **31**, 6410–6421 (2018).
60. Fu, T., Gao, W., Coley, C. & Sun, J. Reinforced genetic algorithm for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **35**, 12325–12338 (2022).
61. Fu, T., Xiao, C., Li, X., Glass, L. M. & Sun, J. MIMOSA: multi-constraint molecule sampling for molecule optimization. In *Proc. 35th AAAI Conference on Artificial Intelligence* 125–133 (AAAI, 2021).
62. Fu, T. et al. Differentiable scaffolding tree for molecular optimization. In *Proc. 10th International Conference on Learning Representations* (OpenReview, 2022).
63. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *Proc. 35th International Conference on Machine Learning, Proc. Machine Learning Research* Vol. 80 (eds Dy, J. & Krause, A.) 2323–2332 (PMLR, 2018).
64. Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **11**, 577–586 (2020).
65. Gao, W., Fu, T., Sun, J. & Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Adv. Neural Inf. Process. Syst.* **35**, 21342–21357 (2022).
One of the benchmarks for molecule optimization evaluating more than 25 methods on 20 commonly used oracle functions.
66. Du, Y. et al. ChemSpacE: interpretable and interactive chemical space exploration. In *Transactions on Machine Learning Research* (OpenReview, 2023).
67. Schrödinger release 2024-1 (Schrödinger, 2024).
68. OpenEye: Applications (Cadence Molecular Sciences, 2023); https://www.eyesopen.com/applications
69. Friesner, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
70. OpenEye: OEDocking (Cadence Molecular Sciences, 2023); https://www.eyesopen.com/oedocking
71. Arús-Pous, J. et al. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **11**, 71 (2019).
72. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
One of the first proposed benchmarks for ML-based molecular design and many proposed tasks are still used in newer benchmarks.
73. Polykovskiy, D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* **11**, 565644 (2020).
74. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **23**, 3–25 (1997).
75. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
76. Guo, J. et al. DockStream: a docking wrapper to enhance de novo molecular design. *J. Cheminf.* **13**, 1–21 (2021).
77. Arnott, J. A. & Planey, S. L. The influence of lipophilicity in drug discovery and design. *Expert Opin. Drug Discov.* **7**, 863–875 (2012).
78. Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C. & Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* **13**, 105–121 (2014).
79. Luo, S., Guan, J., Ma, J. & Peng, J. A 3D generative model for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **34**, 6229–6239 (2021).
80. Harris, C. et al. Benchmarking generated poses: how rational is structure-based drug design with generative models? Preprint at https://arxiv.org/abs/2308.07413 (2023).
81. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **1**, 8 (2009).
82. Fukunishi, Y., Kurosawa, T., Mikami, Y. & Nakamura, H. Prediction of synthetic accessibility based on commercially available compound databases. *J. Chem. Inf. Model.* **54**, 3259–3267 (2014).

83. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminf.* **12**, 35 (2020).
84. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
85. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
86. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
87. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
88. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
Seminal work applying deep learning for retrosynthesis, which can be used to filter ML-generated molecules for synthesizability.
89. Genheden, S. et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **12**, 70 (2020).
90. 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.* **12**, 3339–3349 (2021).
91. Liu, C.-H. et al. RetroGNN: fast estimation of synthesizability for virtual screening and de novo design by learning from slow retrosynthesis software. *J. Chem. Inf. Model.* **62**, 2293–2300 (2022).
92. 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.* **32**, 79377949 (2019).
93. Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S. & Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis dags. *Adv. Neural Inf. Process. Syst.* **33**, (2020).
94. Horwood, J. & Noutahi, E. Molecular design in synthetically accessible chemical space via deep reinforcement learning. *ACS Omega* **5**, 32984–32994 (2020).
95. Gottipati, S. K. et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *Proc. 37th International Conference on Machine Learning* (eds Daumé H. & Singh, A.) 3668–3679 (PMLR, 2020).
96. Gao, W., Mercado, R. & Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. In *Proc. 10th International Conference on Learning Representations* (OpenReview, 2022).
97. Swanson, K. et al. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* **6**, 338–353 (2024).
98. Fialková, V. et al. LibINVENT: reaction-based generative scaffold decoration for in silico library design. *J. Chem. Inf. Model.* **62**, 2046–2063 (2021).
99. Hartenfeller, M. et al. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* **8**, e1002380 (2012).
100. Ghiandoni, G. M. et al. RENATE: a pseudo-retrosynthetic tool for synthetically accessible de novo design. *Mol. Inform.* **41**, 2100207 (2022).
101. Flam-Shepherd, D., Zhu, K. & Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nat. Commun.* **13**, 3293 (2022).
102. Ballarotto, M. et al. De novo design of Nurr1 agonists via fragment-augmented generative deep learning in low-data regime. *J. Med. Chem.* **66**, 8170–8177 (2023).
103. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
104. Ivanenkov, Y. A. et al. Chemistry42: an AI-driven platform for molecular design and optimization. *J. Chem. Info. Model.* **63**, 695–701 (2023).
105. Zhu, W. et al. Discovery of novel and selective SIK2 inhibitors by the application of AlphaFold structures and generative models. *Bioorg. Med. Chem.* **91**, 117414 (2023).
106. Grisoni, F. et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **7**, eabg3338 (2021).
107. Li, Y. et al. Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor. *Nat. Commun.* **13**, 6891 (2022).
A study showing the complementarity of virtual screening and generative molecular design with experimental validation.
108. Yoshimori, A. et al. Design and synthesis of DDR1 inhibitors with a desired pharmacophore using deep generative models. *ChemMedChem* **16**, 955–958 (2021).
109. Blaschke, T. et al. REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **60**, 5918–5922 (2020).
110. Merk, D., Grisoni, F., Friedrich, L., Gelzinyte, E. & Schneider, G. Scaffold hopping from synthetic RXR modulators by virtual screening and de novo design. *MedChemComm* **9**, 1289–1292 (2018).
111. Merk, D., Grisoni, F., Friedrich, L., Gelzinyte, E. & Schneider, G. Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J. Med. Chem.* **61**, 5442–5447 (2018).
112. Gillet, V. et al. Synthetically accessible de novo design using reaction vectors: application to PARP1 inhibitors. *Mol. Inform.* **43**, e202300183 (2024).
113. Pun, F. W., Ozerov, I. V. & Zhavoronkov, A. AI-powered therapeutic target discovery. *Trends Pharmacol. Sci.* **44**, 561–572 (2023).
114. First generative AI drug begins phase II trials with patients. *Insilico Medicine* https://insilico.com/blog/first_phase2 (2023).
115. Ren, F. et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* (2024).
116. Guo, J. et al. Improving de novo molecular design with curriculum learning. *Nat. Mach. Intell.* **4**, 555–563 (2022).
117. Guo, J. & Schwaller, P. Augmented memory: capitalizing on experience replay to accelerate de novo molecular design. Preprint at https://doi.org/10.48550/arXiv.2305.16160 (2024).
118. Guo, J. & Schwaller, P. Beam enumeration: probabilistic explainability for sample efficient self-conditioned molecular design. In *Proc. 12th International Conference on Learning Representations* (OpenReview, 2024).
119. Dodds, M. et al. Sample efficient reinforcement learning with active learning for molecular design. *Chem. Sci.* **15**, 4146–4160 (2024).
120. Buttenschoen, M., Morris, G. M. & Deane, C. M. Posebusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).
121. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
122. Du, Y., Guo, X., Wang, Y., Shehu, A. & Zhao, L. Small molecule generation via disentangled representation learning. *Bioinformatics* **38**, 3200–3208 (2022).
123. Jin, W., Barzilay, R. & Jaakkola, T. Multi-objective molecule generation using interpretable substructures. In *Proc. 37th International Conference on Machine Learning* 4849–4859 (PMLR, 2020).

124. Hoffman, S. C., Chenthamarakshan, V., Wadhawan, K., Chen, P.-Y. & Das, P. Optimizing molecules using efficient queries from property evaluations. *Nat. Mach. Intell.* **4**, 21–31 (2021).
125. Madhawa, K., Ishiguro, K., Nakago, K. & Abe, M. GraphNVP: an invertible flow model for generating molecular graphs. Preprint at https://arxiv.org/abs/1905.11600 (2019).
126. Kadurin, A. et al. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **8**, 10883 (2017).
127. Imrie, F., Bradley, A. R., van der Schaar, M. & Deane, C. M. Deep generative models for 3D linker design. *J. Chem. Inf. Model.* **60**, 1983–1995 (2020).
128. Liu, M., Yan, K., Oztekin, B. & Ji, S. GraphEBM: molecular graph generation with energy-based models. Preprint at https://arxiv.org/abs/2102.00546 (2021).
129. Vignac, C. et al. DiGress: discrete denoising diffusion for graph generation. In *Proc. 11th International Conference on Learning Representations* (OpenReview, 2023).
130. Nigam, A., Friederich, P., Krenn, M. & Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. In *Proc. 8th International Conference on Learning Representations* (OpenReview, 2020).
131. Spiegel, J. O. & Durrant, J. D. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminf.* **12**, 25 (2020).
One of the representative works that leverage genetic algorithms for molecular design.
132. Simm, G. & Hernandez-Lobato, J. M. A generative model for molecular distance geometry. In *Proc. 37th International Conference on Machine Learning*, *Proc. Machine Learning Research* Vol. 119 (eds Daumé H. & Singh, A.) 8949–8958 (PMLR, 2020).
133. Ganea, O. et al. GeoMol: torsional geometric generation of molecular 3D conformer ensembles. *Adv. Neural Inf. Process. Syst.* **34**, 13757–13769 (2021).
134. Klein, L., Krämer, A. & Noe, F. Equivariant flow matching. *Adv. Neural Inf. Process. Syst.* **36**, 59886–59910 (2023).
135. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. & Jaakkola, T. EquiBind: geometric deep learning for drug binding structure prediction. In *Proc. 39th International Conference on Machine Learning*, *Proc. Machine Learning Research* Vol. 162 (eds Chaudhuri, K. et al.) 20503–20521 (PMLR, 2022).
136. Jing, B., Corso, G., Chang, J., Barzilay, R. & Jaakkola, T. Torsional diffusion for molecular conformer generation. *Adv. Neural Inf. Process. Syst.* **35**, 24240–24253 (2022).
137. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. S. DiffDock: diffusion steps, twists, and turns for molecular docking. In *Proc. 11th International Conference on Learning Representations* (OpenReview, 2023).
138. Ragoza, M., Masuda, T. & Koes, D. R. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **13**, 2701–2713 (2022).
139. Drotár, P., Jamasb, A. R., Day, B., Cangea, C. & Liò, P. Structure-aware generation of drug-like molecules. Preprint at https://arxiv.org/abs/2111.04107 (2021).
140. Joshi, R. P. et al. 3D-scaffold: a deep learning framework to generate 3D coordinates of drug-like molecules with desired scaffolds. *J. Phys. Chem. B* **125**, 12166–12176 (2021).
141. Liu, M., Luo, Y., Uchino, K., Maruhashi, K. & Ji, S. Generating 3D molecules for target protein binding. In *Proc. 39th International Conference on Machine Learning*, *Proc. Machine Learning Research*, Vol. 162 (eds Chaudhuri, K. et al) 13912–13924 (PMLR, 2022).
142. Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I. & Welling, M. E(n) equivariant normalizing flows. *Adv. Neural Inf. Process. Syst.* **34**, 4181–4192 (2021).
143. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
144. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. Adversarial autoencoders. Preprint at https://arxiv.org/abs/1511.05644 (2015).
145. Philippidis, A. Insilico joins scramble to treat solid tumors by targeting KIF18A; https://www.genengnews.com/topics/artificial-intelligence/insilico-joins-scramble-to-treat-solid-tumors-by-targeting-kif18a/ (2024).
146. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).
147. 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.* **1**, 68 (2018).
148. Yu, Y. et al. A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors. *ACS Omega* **6**, 22945–22954 (2021).
149. 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.* **60**, 19477–19482 (2021).
150. Jang, S. H. et al. PCW-A1001, AI-assisted de novo design approach to design a selective inhibitor for FLT-3 (D835Y) in acute myeloid leukemia. *Front. Mol. Biosci.* **9**, 1072028 (2022).
151. Eguida, M., Schmitt-Valencia, C., Hibert, M., Villa, P. & Rognan, D. Target-focused library design by pocket-applied computer vision and fragment deep generative linking. *J. Med. Chem.* **65**, 13771–13783 (2022).
152. Chen, N. et al. Recurrent neural network (RNN) model accelerates the development of antibacterial metronidazole derivatives. *RSC Adv.* **12**, 22893–22901 (2022).
153. Tan, X. et al. Discovery of pyrazolo [3,4-d] pyridazinone derivatives as selective DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation. *J. Med. Chem.* **65**, 103–119 (2021).
154. Hua, Y. et al. Effective reaction-based de novo strategy for kinase targets: a case study on MERTK inhibitors. *J. Chem. Inf. Model.* **62**, 1654–1668 (2022).
155. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 114 (2023).
156. Song, S. et al. Application of deep generative model for design of pyrrolo [2,3-d] pyrimidine derivatives as new selective tank binding kinase 1 (TBK1) inhibitors. *Eur. J. Med. Chem.* **247**, 115034 (2023).
157. Yu, Y. et al. Accelerated discovery of macrocyclic CDK2 inhibitor QR-6401 by generative models and structure-based drug design. *ACS Med. Chem. Lett.* **14**, 297–304 (2023).
158. Atz, K. et al. Prospective de novo drug design with deep interactome learning. *Nat. Commun.* **15**, 3408 (2024).
159. Putin, E. et al. Adversarial threshold neural computer for molecular de novo design. *Mol. Pharmaceutics* **15**, 4386–4397 (2018).
160. Polykovskiy, D. et al. Entangled conditional adversarial autoencoder for de novo drug discovery. *Mol. Pharm.* **15**, 4398–4405 (2018).
161. Korshunova, M. et al. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Commun. Chem.* **5**, 129 (2022).
162. Li, Y. et al. Discovery of potent, selective, and orally bioavailable small-molecule inhibitors of CDK8 for the treatment of cancer. *J. Med. Chem.* **66**, 5439–5452 (2023).

163. Salas-Estrada, L. et al. De novo design of κ-opioid receptor antagonists using a generative deep-learning framework. *J. Chem. Inf. Model.* **63**, 5056–5065 (2023).
164. Xu, J. et al. Discovery of novel and potent prolyl hydroxylase domain-containing protein (PHD) inhibitors for the treatment of anemia. *J. Med. Chem.* **67**, 1393–1405 (2024).
165. Bo, W. et al. Local scaffold diversity-contributed generator for discovering potential NLRP3 inhibitors. *J. Chem. Inf. Model.* **64**, 737–748 (2024).
166. Xia, Y. et al. Target-aware molecule generation for drug design using a chemical language model. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.08.574635 (2024).
167. Vakili, M. G. et al. Quantum computing-enhanced algorithm unveils novel inhibitors for KRAS. Preprint at https://arxiv.org/abs/2402.08210 (2024).
168. Hassen, A. K. et al. Generate what you can make: achieving in-house synthesizability with readily available resources in de novo drug design. Preprint at *chemRxiv* https://doi.org/10.26434/chemrxiv-2024-wtjt6 (2024).
169. Wang, Y. et al. Discovery of 3-hydroxymethyl-azetidine derivatives as potent polymerase theta inhibitors. *Bioorg. Med. Chem.* **103**, 117662 (2024).
170. Zhao, Y. et al. Accelerating factor Xa inhibitor discovery with a de novo drug design pipeline. *Chin. J. Chem. Eng.* (2024).
171. Jiang, Y. et al. Pocketflow is a data-and-knowledge-driven structure-based molecular generative model. *Nat. Mach. Intell.* **6**, 326–337 (2024).
172. Zhang, J. et al. ISM9682A, a novel and potent KIF18A inhibitor, shows robust antitumor effects against chromosomally unstable cancers. *Cancer Res.* **84**, 5727–5727 (2024).
173. Huang, L. et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **15**, 2657 (2024).

Acknowledgements

J.G. and P.S. acknowledge support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. A.R.J. is funded by a Biotechnology and Biological Sciences Research Council (BBSRC) DTP studentship (BB/M011194/1). Y.W. acknowledges the support of Cornell Presidential Life Science Fellowship. We are grateful to K. Atz and A. Mueller for helpful feedback and discussion.

Author contributions

Y.D., A.R.J. and J.G. led this work under the supervision of P.S. and T.L.B. and contributed equally. T.F. and C.H. also contributed equally. All authors contributed ideas and discussions to writing, reviewing and editing of the paper before submission.

Competing interests

A.R.J. declares a potential financial conflict of interest due to his role as a machine learning scientist at Prescient Design, Genentech. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00843-5.

Correspondence and requests for materials should be addressed to Philippe Schwaller or Tom L. Blundell.

Peer review information *Nature Machine Intelligence* thanks J. B. Brown and Ola Engkvist for their contribution to the peer review of this work.