Generative Models in Drug Discovery: Advancing Assessments, Metrics and Retrosynthesis Prediction

Submitted by
Philipp Renz
01126686

Submitted at
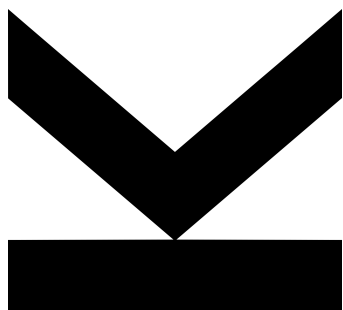**Institute for Machine Learning**

Thesis Supervisor / First Evaluator
Univ.-Prof. Mag. Dr. **Günter Klambauer**

Co-Supervisor
Univ.-Prof. Dr. **Sepp Hochreiter**

Second Evaluator
Name

May 2024

Doctoral Thesis

to obtain the academic degree of

Doktor der Naturwissenschaften

in the Doctoral Program

Naturwissenschaften

Abstract

In recent years the use of generative models in drug discovery has seen a surge, as novel deep learning architectures have shown great flexibility in generating molecular structures. However, the evaluation of generative models is challenging and existing benchmarks are often criticized for not reflecting the practical utility of the models. In this thesis, we propose new evaluation metrics and benchmarks for generative models in drug discovery. Another focus of this work is the application of generative models to retrosynthesis prediction, a crucial task in computer-aided synthesis planning (CASP).

The first part of this thesis focuses on observed failure modes in the evaluation of generative models for de novo molecular design. In particular we show that commonly used metrics used to evaluate distribution-learning are not sufficient to differentiate complex models from trivial baseline generators. Secondly, we show how generative models applied to molecular optimization can overfit to machine learning-based scoring functions, leading to biased evaluations.

The second part introduces a diversity-based benchmark for goal-directed molecule generators. Diverse, high-scoring compounds are crucial in drug discovery, as many candidates may fail in later stages. Previous studies on diverse molecule optimization have been limited by inadequate diversity measures, non-standardized compute budgets, and lack of model adaptation to diverse optimization settings. Our benchmark addresses these shortcomings, providing a standardized framework for evaluating diverse, goal-directed molecule generators and enabling fair model comparisons.

The third part of this thesis focuses on retrosynthesis prediction a crucial task in computer-aided synthesis planning (CASP). We propose a novel template-based retrosynthesis prediction model based on Modern Hopfield Networks. Our model takes both the target molecule and the reaction templates as input, which allows it to generalize over reaction templates, which improves performance, particularly on rare templates. Our model achieves state-of-the-art performance on the USPTO-50k dataset. while maintaining a significantly lower computational cost compared to existing methods.

Through our work, we provide insights into the capabilities and limitations of current generative models for molecules while proposing novel evaluation strategies. Additionally, our contributions in retrosynthesis prediction enable more accurate computer-aided synthesis planning. Collectively, these advances have the potential to accelerate the drug discovery pipeline and facilitate the development of novel pharmaceutical treatments.

Acknowledgement

Contents

List of Acronyms

CASP computer-aided synthesis planning

DMTA Design-Make-Test-Analyze

GAN generative adversarial network

NLL negative log-likelihood

QSPR quantitative structure-property relationship

VS virtual screening

Introduction

1.1 Small molecule drug design

The discovery of novel drugs has significantly contributed to the improvement of human health and well-being. There is a continuous demand for new drugs to expand the range of treatable diseases, improve the efficacy of existing treatments, and respond to the emergence of new health challenges.

Small molecule drugs constitute the majority of medicines in use, accounting for approximately 90% of global sales (Makurvet, 2021). These molecules, typically defined as having a molecular weight of less than 900 Da (**todo**), offer several advantages. They are generally stable, do not require specialized storage conditions, and can be conveniently administered orally. Moreover, they are relatively inexpensive to produce and can be easily synthesized in large quantities (Southey et al., 2023).

For a small molecule to be considered a viable drug candidate, it must fulfill a range of properties (Southey et al., 2023):

- **On-target activity:** The molecule must be active against the desired target to exhibit the intended therapeutic effect. At the molecular level, this means binding to the target and modulating its activity in the desired manner.

- **Specificity:** The molecule should demonstrate high specificity, selectively interacting with the intended target while minimizing undesirable off-target interactions. Such selectivity is crucial to prevent adverse side effects and maintain the drug's safety and efficacy profile.

- **Toxicity:** The absence of toxic effects is essential, as the molecule must be well-tolerated and free from potential harmful side effects. Toxicity can arise from various factors, including off-target interactions, metabolic byproducts, or allergic reactions.

- **Pharmacokinetics:** The molecule must possess favorable pharmacokinetic properties, encompassing adsorption, distribution, metabolism, and excretion (ADME). These properties determine how the molecule is absorbed into the body, distributed throughout, metabolized, and ultimately excreted. They are
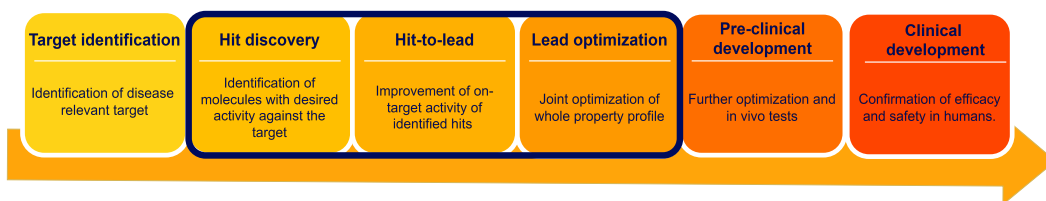
Figure 1.1: The drug discovery pipeline starts with the identification of a biological target. Once a target is identified, readily available molecules are screened for their activity against the target in high-throughput screening. Promising hits are then modified and optimized to lead compounds. These lead compounds are then further optimized and tested in preclinical. Finally, the most promising candidates are tested in clinical trials and eventually approved by regulatory agencies. The stages in the blue box are highly amenable to machine learning and computational methods and are the focus of this thesis.

crucial for ensuring the molecule reaches its target effectively and is processed safely by the body.

- **Synthesizability:** The molecule must be synthesizable in a cost-effective manner to be practically viable for large-scale production.

In addition to these properties, the molecule must be novel and not infringe on existing patents. While this is not inherently necessary for a drug's efficacy, it represents a significant practical consideration in the pharmaceutical industry.

The primary challenge in drug discovery lies in identifying a molecule that satisfies all these criteria simultaneously. The development of a new drug is a complex and expensive process, which can take 12–15 years and cost estimates range between $1.8-2.5 billion (Paul et al., 2010; DiMasi et al., 2016).

1.1.1 The drug discovery pipeline

The drug discovery process is usually divided into several stages (**todo**) depicted in Figure 1.1 and described below.

- **Target Identification and Validation:** The drug discovery process begins with the identification of a biological target, which is usually a molecule, protein, or gene involved in a disease pathway. Understanding the target's role in the disease is crucial for developing therapeutic interventions.

- **Hit Discovery:** This stage aims at identifying "hits", which are molecules that exhibit activity against the target. High-throughput screening (HTS) is a common approach used to test large libraries of molecules against the

target in a rapid and automated manner. Computational methods such as virtual screening can also be employed to increase the hit-rate of wet-lab experiments.

- **Hit-to-lead and lead optimization:** Promising hits are then refined and optimized to produce lead compounds. This stage focuses on improving the activity, selectivity, and pharmacological properties of the molecules. The process often follows a DMTA (Design-Make-Test-Analyze) cycle, where compounds are designed, synthesized, and tested iteratively in vitro. The goal is to find a lead compounds with a desirable balance of potency, selectivity, and drug-like properties.

- **Preclinical Development:** The most promising lead compounds are then tested in preclinical studies, which are typically conducted in animal models. These studies assess the safety, efficacy, pharmacokinetics, and toxicology of the drug candidate in vivo. The data generated during this stage are critical for determining whether the candidate is suitable for clinical trials in humans.

- **Clinical Trials:** Drug candidates that pass preclinical development proceed to clinical trials, which are conducted in humans and are typically divided into three phases:

  - **Phase I:** This phase focuses on assessing the safety, tolerability, and pharmacokinetics of the drug in a small group of healthy volunteers or patients.

  - **Phase II:** In this phase, the efficacy of the drug is tested in a larger group of patients with the target disease. Safety and dosage optimization are also evaluated.

  - **Phase III:** This phase involves large-scale testing of the drug's safety and efficacy in a diverse patient population. It provides the critical data needed for regulatory approval.

  The success rates of clinical trials are low, with only about 10% of drugs that enter clinical trials eventually being approved by regulatory agencies More specifically, the success rates in Phase I/II/III and the final regulatory approval are 63%, 31%, 58% and 85% respectively (Mullard, 2016).[1] This translates to 63%, 19.5%, 11.3% and 9.6% of projects that make it to the respective stages.

- **Regulatory Approval:** Upon successful completion of clinical trials, the drug is submitted for regulatory approval. Agencies such as the U.S. Food and
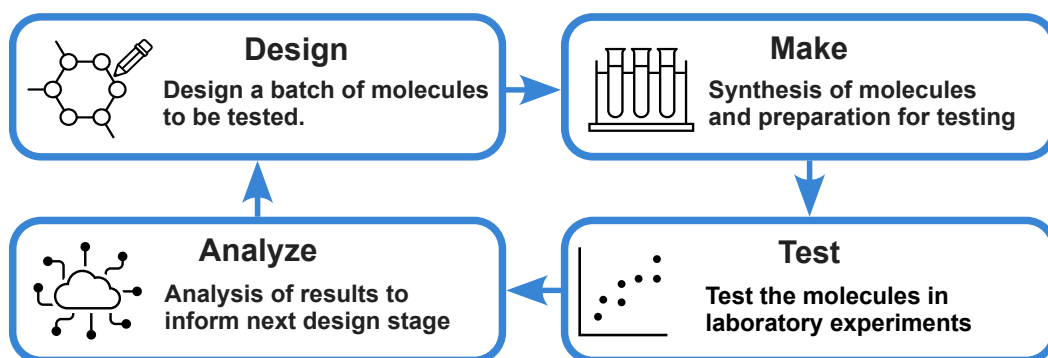
Figure 1.2: The DMTA cycle

Drug Administration (FDA) or the European Medicines Agency (EMA) review the comprehensive data package, including preclinical and clinical trial results. If the drug is deemed safe and effective, it receives approval for marketing and distribution.

- **Post-Market Surveillance:** After regulatory approval, the drug enters the market, but the process doesn't end here. Post-market surveillance, or Phase IV studies, are conducted to monitor the long-term safety and efficacy of the drug in the general population. This stage can reveal rare side effects or long-term risks that were not apparent during clinical trials, and it may lead to further modifications, warnings, or even withdrawal of the drug from the market.

The general strategy of this process is to start with a large number of molecules and then systematically reduce the number to a few candidates that are then tested in clinical trials. The early stages have lower per molecule costs but provide less information about the success chances of a molecule. The later stages are more expensive, but in the end provide accurate information whether a molecule is safe and effective in humans.

1.1.2 The Design-Make-Test-Analyze cycle

The hit discovery, hit-to-lead and lead optimization stages (blue box in Figure 1.1) usually operate in an iterative manner, resulting in a cycle of choosing molecules to be tested, synthesizing them, testing them in laboratory experiments and analyzing the results to guide the selection of the next molecule to be tested. This cycle is usually referred to as the *Design-Make-Test-Analyze* (*DMTA*)-cycle:

- **Design:** Under consideration of previous experimental results, the molecules to be tested are designed. The design generally aims to optimize the desired properties of the molecule, but also aims to maximize the information gained from the experiment.

- **Make:** The designed molecules are then synthesized and prepared for testing in the laboratory. This step requires a synthesis plan that outlines the steps needed to synthesize the molecule.

- **Test:** The synthesized molecules are then tested in laboratory experiments to measure the properties of interest. This can range from the activity of the molecule against a target, to its pharmacokinetic properties, to its toxicity and others.

- **Analyze:** The results of the experiments are analyzed. The obtained insights can then be used to guide the design of the next molecules to be tested. evaluation of the performance of the prediction models used in the design phase. The results of the analysis are then used to guide the design of the next molecule to be tested.

1.1.3 Computer-aided drug design

Computational methods are widely used throughout the DMTA cycle and beyond. One of the most important applications of computational methods in drug discovery is the prediction of molecular properties. These predictions can be used in the design phase, to find molecules that are likely to have the desired properties, reducing the number of molecules that need to be synthesized and tested. The measurements obtained in the test phase can be used to update the prediction models in the analyze phase, improving the quality of the predictions for the next iteration of the cycle.

Recent years have seen a surge in the use of generative models in drug discovery which was triggered by advances in deep learning, model architectures and increased compute capabilities. Generative models for molecules expand the use of machine learning models to tasks which require molecular structures as solutions. In this thesis we will focus on two core applications: Firstly, de novo drug design, which aims generate molecules with desired property profiles, and secondly, computer-aided synthesis planning, which seeks to generate synthetic routes for target molecules addressing the crucial "Make" step of the DMTA-cycle.
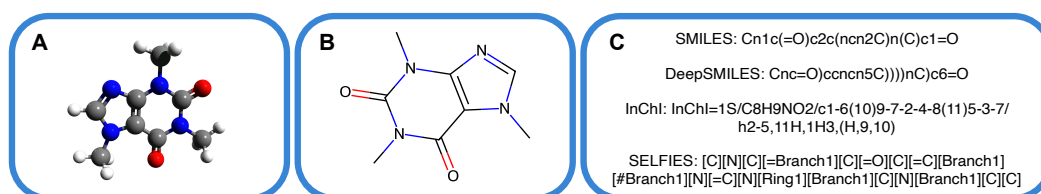
Figure 1.3: Different ways to represent a caffeine molecule **A:** The 3D structure of a molecule is given by the positions of its atoms in space. This structure is not necessarily fixed as some bonds can rotate and bonds can vibrate (Image source: (*English: Caffeine 3D Structure* 2010)). **B:** The graph representation of a molecule is given by the connectivity of its atoms. The nodes represent atoms and the edges represent bonds between atoms. As the bonds of a stable molecule are fixed this representation usually determines the identity/type of a molecule. **C:** The graph representation of a molecule can be linearized into a 1D sequence of characters. DeepSmiles and SELFIES are alternatives to the SMILES notation that simplify the syntax and make it easier to generate syntactically valid molecules. InChI is less popular in generative modelling.

1.2 Generative models in drug discovery

1.2.1 Molecular representations

Molecules, though fundamentally complex quantum mechanical entities, can be represented through various simplified models for practical purposes. The most common representation depicts molecules as graphs, where atoms are nodes and chemical bonds are edges. Figure 1.3b shows a graph representation of caffeine. This graph structure captures the molecule's connectivity, which defines its "identity". Additional properties such as atom type, charge, or chirality are incorporated as features of the nodes and edges. While this representation doesn't capture the full quantum complexity, it provides a stable and practical framework for understanding and working with molecular structures in many scientific and computational contexts.

Molecular graphs can be linearized into one-dimensional character sequences, known as line notations. Examples include InChI (Heller et al., 2015) and SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988). SMILES strings have proven particularly valuable for generative models, as they are easily processed by sequence-based models like recurrent neural networks (RNNs) and Transformers (Vaswani et al., 2017). Several extensions to SMILES have been proposed to make them more amenable for use in machine learning models. DeepSmiles (O'Boyle et al., 2018) attempted to make it easier to generate syntactically valid molecules, by changing the notation of branches and ring closures. SELFIES (Krenn

et al., 2022) provide a representation of molecules in which any sequence of tokens parses into a valid molecule. SAFE (Noutahi et al., 2023) provides a representation of molecules in which the substructures are represented by contiguous regions of a SMILES string.

Molecules can be represented in various complex forms beyond simple graphs and strings. Three-dimensional structures provide a spatial description of a molecule, detailing atomic positions in 3D space along with information about atom types and bonds. The most comprehensive representation is the quantum mechanical wavefunction, which captures the full complexity of molecular behavior. While these more sophisticated representations are valuable for modeling a wide range of molecular properties and interactions, but are not covered in the rest of this thesis.

1.2.2 Generation strategies

Sequence-based autoregressive models constitute one of the most popular approaches for generating molecules. Auto-regressive modelling is based on the idea of generating a molecule by iteratively predicting the next element of the sequence given the preciding characters. The likelihood of sampling a sequence x is thus modelled by $p(x) = \prod_{i=0}^{n} p(x_i|x_{1:i-1})$. This approach has been widely used in natural language processing and has been adapted to the generation of molecules. Early work by (Segler et al., 2018a) and (Gómez-Bombarelli et al., 2018) relied on recurrent neural networks (RNNs) to generate SMILES strings. This approach has since been popular and there has been work on string-based representations more suitable to generation (O'Boyle et al., 2018; Krenn et al., 2020; Noutahi et al., 2023), parsing the molecules into specialized data structures (Kusner et al., 2017; Jin et al., 2018) and using other deep learning architectures such as transformers (Vaswani et al., 2017; Noutahi et al., 2023; Schwaller et al., 2019; Bagal et al., 2022; Mazuz et al., 2023).

Graph-based autoregressive models work similarly to its sequence-based counterpart, but instead of relying on a linearization of the molecule, they work directly on the graph representation of the molecule. The model generates the molecular graph by iteratively adding nodes and edges to the graph. The model can be trained in a similar manner to the string-based models, by predicting the next node or edge given the current state of the graph. However, the specification of possible actions is more complex than in the 1D case as there is no natural ordering of the nodes

and edges in the graph (Liu et al., 2018; Li et al., 2018; You et al., 2019; Cohen-Karlik et al., 2024).

One-shot methods are a class of models that generate molecules in one step, without the need for an iterative generation process. These models generate an adjacency matrix and node feature vector of a molecule in a single step. This is usually done by first generating a continous version of the molecule and then discretizing it to a valid molecule (De Cao et al., 2018; Madhawa et al., 2019; Kadurin et al., 2016).

Rule-based models generate molecules by applying a set of pre-defined graph transformation rules to combine molecular fragments. The BRICS (Degen et al., 2008) method provides a set of molecular fragments and rules how to meaningfully combine them. This enables the generation of new molecules by combining these fragments. DOGS (Hartenfeller et al., 2012) generates molecules by applying a set of chemical reaction rules to a set of starting molecules, which has the advantage of biasing generation towards synthesizable molecules. Jensen (2019) defines graph mutation and crossover operations to generate new molecules. These models allow the generation of molecules that are chemically valid, or resemble known "reasonable" molecules.

1.2.3 Distribution-learning

Distribution-learning is a fundamental application of generative models in drug design. Its objective is to create a model that accurately captures the distribution of molecules within a dataset. Formally, the model learns a distribution $q(x)$ that approximates the true distribution $p(x)$ of molecules. This approach enables the model to grasp both the syntax and semantics of the molecules in the training set. These models can be trained on large chemical libraries of stable molecules, PubChem (**pubchem**), ChEMBL (Bento et al., 2014) or (**zinc**). Using this process the resulting models can learn what reasonable molecules look like in the context of drug discovery. This makes them useful for their two main purposes: they can expand virtual libraries and, more crucially, act as a foundation for other applications such as goal-directed generation, which we will explore in the subsequent section.

In recent years there has been a surge in interest distribution-learning models based on deep neural networks. Many architectures and training strategies originally
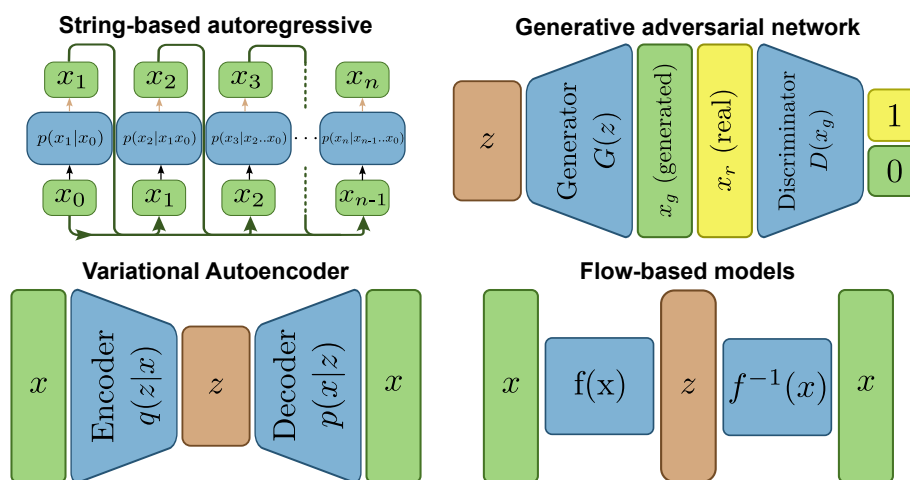
Figure 1.4: Illustration of how different distribution-learning methods try to generate samples that resemble the data in distribution. String-based autoregressive models sample the next token according to a probability distribution which is based on the previously sampled tokens. The likelihood of the samples can be optimized using the probability chain rule. The generator in generative adversarial networks is trained to generate samples, that a discriminator cannot distinguish from real samples. Variational autoencoders first transform an input sample into a latent space and then reconstruct the input from the latent representation. Flow-based models make use of an invertible mapping which allows to calculate and maximize the exact likelihood of the data.

proposed for text and image generation have been adapted and specialized to generate molecules. While all of them aim to approximate $p(x)$, they differ in the way they model the distribution and the choice of molecular representation.

Autoregressive models can be directly trained using a maximum likelihood approach by minimizing the cross entropy or *negative log-likelihood* (*NLL*) of the training data

$$\mathcal{L} = -\mathbb{E}_{x \sim p(x)} \log q(x) \approx -\frac{1}{N} \sum_{i=1}^{N} \log q(x_i), \qquad (1.1)$$

where $q(x)$ is the model distribution and $p(x)$ is the true distribution of the data. These models are explicit density models, as the likelihood for a given molecule can be calculated exactly. Autoregressive models form the backbone of many generative models in drug discovery (Gómez-Bombarelli et al., 2018; Segler et al., 2018a; Olivecrona et al., 2017; Guo et al., 2023; Thomas et al., 2022b; Jaques et al., 2016; Cohen-Karlik et al., 2024)

Variational autoencoders (VAEs) (Kingma et al., 2013) generate molecules by first sampling from a simple latent distribution $p(z)$, and then mapping the samples to molecular space via a probabilistic decoder network $p(x|z)$. To make training tractable a second network, the encoder network $q(z|x)$ is used to map the data to the latent space. The model is then trained to maximize the evidence lower bound (ELBO) of the data

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)||p(z)), \qquad (1.2)$$

where KL is the Kullback-Leibler divergence. This model has the advantage of providing a continuous latent space, which can be used to interpolate between molecules and allows the use continuous optimization algorithms in latent space. VAEs belong in the class of approximate density models, as the likelihood of a given molecule can be calculated approximately via monte carlo sampling. VAEs have been a popular choice for generating molecules (Gómez-Bombarelli et al., 2018; Kusner et al., 2017; Simonovsky et al., 2018; Samanta et al., 2018; Jin et al., 2018; Dai et al., 2018).

Generative flows (Rezende et al., 2016) are based on the idea of learning a bijective mapping between molecular space and a latent space. Generative flows transform from a simple distribution $p(z)$ in latent space to a distribution in chemical space, p(x), via a bijective mapping $G : z \rightarrow x$. The likelihood of the training data can then be directly calculated and optimized via the change of variables formula:

$$p(x) = p(z) \left| \det \frac{\partial G}{\partial z} \right|. \qquad (1.3)$$

Originally generative flows have been proposed for continuous data, but have been adapted to discrete data such as molecules by using a continuous relaxation of the molecule (Madhawa et al., 2019). These models also belong to the class of explicit density models, as the likelihood of a given molecule can be calculated exactly.

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are latent space models that map a simple distribution in latent space to molecular space, but rely on a game-theoretic approach to training. A generator network is trained to generate data, which is then fed to a discriminator network. The two networks then engage in a minimax game, where the discriminator tries is trained to distinguish between real and generated data, while the generator is trained to fool the discriminator. *generative adversarial network*s (*GAN*s) are implicit density methods they sample from the model distribution, but do not provide a likelihood for a given sample.
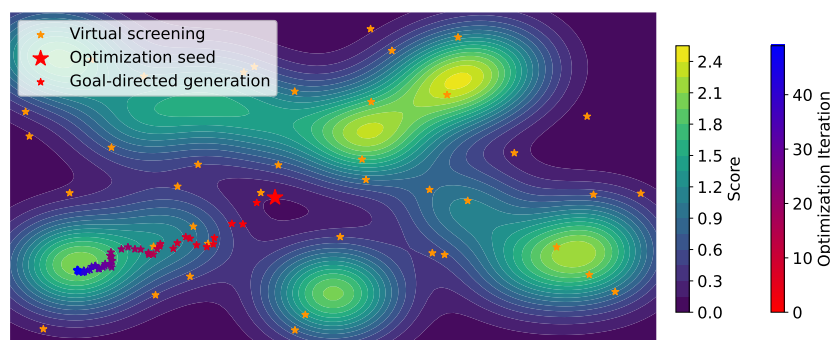
Figure 1.5: Illustration of the difference between goal-directed molecule generation and virtual screening in a 2D chemical space, where each point represents a molecule. The background color represents the molecules' scores. Goal-directed generation works akin to a numerical optimization algorithm and efficiently finds high-scoring molecules, shown in the transition from red to blue stars. In contrast VS amounts to a random search in chemical space, which is less efficient and is likely to miss high-scoring regions of chemical space.

This approach has been combined with different generation strategies in the context of drug discovery (De Cao et al., 2018; Kadurin et al., 2017; Guimaraes et al., 2017; Méndez-Lucio et al., 2018)

1.2.4 Goal-directed molecule generation

Goal-directed molecule generation (Schneider, 2013) is a computational approach for automatically designing molecules with desired property profiles. Goal-directed generation expands upon *virtual screening* (*VS*), a method in which a library of molecules is ranked according to the output of a *quantitative structure-property relationship* (*QSPR*) model. Walters (2019) estimates that approximately 10^{13} molecules can be routinely tested in a *VS* experiment. While this number can vary significantly depending on the computational cost of running the *QSPR* model, it is dwarfed by the size of drug-like chemical space, which is estimated to contain between 10^{30} and 10^{60} molecules (Walters, 2019; Ruddigkeit et al., 2012). Consequently, *VS* is limited to exploring only a small fraction of chemical space and cannot fully leverage the vast number of possible candidates that drug-like chemical space offers.

Goal-directed generators address this limitation of *VS* by focusing the search on the most relevant parts of chemical space. In contrast to the random search approach taken by *VS*, goal-directed generators act more like optimizers that are able to efficiently locate maxima. This is achieved by an iterative process in which a model generates a set of molecules, which are then scored by a *QSPR* model. These scores

are then used to update the model, shifting the sampling distribution to regions of chemical space with higher scores.

Recently, there has been a surge of deep learning-based goal-directed generators (Elton et al., 2019; Sanchez-Lengeling et al., 2018; Du et al., 2024). A multitude of different models have been proposed, which are based on a variety of neural network architectures, training strategies and molecular representations. These methods augment traditional rule-based generation approaches that have been combined with graph search and evolutionary algorithms. (Schneider et al., 2005; Schneider, 2013). The new wave of deep-learning methods has shown great promise in generating novel molecules with desired property profiles and has led to success in a variety of applications, such as the design of new drugs, materials or catalysts (**todo**).

Some of the most commonly used approaches to goal-directed molecular generation are:

- **Hill-climbing** (Segler et al., 2018a; Xie et al., 2021; Thomas et al., 2022b) is a simple optimization algorithm that relies on an underlying distribution-learning model. Molecules are sampled from the model's learned distribution and their scores are evaluated. The model is then retrained on the top-scoring molecules and the process is repeated.

- **Reinforcement learning** uses the molecule scores as a reward signal to update the model distribution. This is commonly via methods based on the REINFORCE algorithm (Williams, 1992) which allows to update the model distribution in a way that increases expected scores of the generated molecules (Olivecrona et al., 2017; Thomas et al., 2022b; You et al., 2019; Guo et al., 2023).

- **Genetic algorithms** in molecular generation operate by evolving an initial population of molecules through iterative cycles of mutation, crossover, and selection (Jensen, 2019; Nigam et al., 2021; Yoshikawa et al., 2018). Starting from an initial set of molecules, new molecules are generated by applying mutation and crossover operations. The molecules are then scored, and the best ones are selected for the next generation. This process is repeated for multiple generations, gradually optimizing the population towards desired molecular characteristics.

- **Tree search** builds a tree of possible molecules by recursively applying a set of transformation rules to some initial molecules. Using techniques such as Monte Carlo Tree Search, the tree is explored to find the most promising molecules (Yang et al., 2017; Jensen, 2019).

- **Continuous optimization** employ classical optimization algorithms in the continuous latent space of (variational) autoencoders (Gómez-Bombarelli et al., 2018; Kusner et al., 2017; Winter et al., 2019) or generative flows (Madhawa et al., 2019).

- **Generative Flow Networks** (Bengio et al., 2021) aim to generate molecules with probability proportional to their score. This method relies on an iterative generation process and models chemical space as a directed acyclic graph, with nodes being intermediate molecules and edges graph edits. The transition probabilities between nodes are given by a "flow" of probability mass from the root node to finished molecules, such that the probability of each finished molecule is proportional to its score. This has the advantage of being able to explore multiple modes of the scoring function.

1.2.5 Challenges in Evaluating Generative Models in de Novo Design

1.2.5.1 Evaluation of distribution-learning models

The most basic and commonly used checks to assess the quality of the generated compounds are the validity, uniqueness and novelty of the generated molecules. A molecule is valid if it obeys chemical valence rules, which is usually checked using chemoinformatics toolkits such as RDKit (Landrum, 2006). The uniqueness of a set of molecules measures the fraction of unique molecules in the set and can flag models that output many duplicates. The novelty a set of generated molecules is the fraction of molecules that are not in the training set and can, to a certain extent, detect whether a model overfits to the training set.

A variety approaches exist to assess how well a model can learn the distribution of the training set. Explicit/approximate density models allow principled evaluation using the negative log-likelihood on a hold-out test set. However, this is not applicable for implicit density models such as GANs. The KL-divergence between the distributions of scalar molecular properties (e.g. molecular weight, logP, ...) of the generated molecules and the training set is a commonly used metric, to evaluate the distribution fit (Brown et al., 2019). The Frechet ChemNet Distance (FCD) (Preuer et al., 2018) offers a more comprehensive evaluation of the distribution fit. Instead of comparing the distributions of few select property values, the FCD compares the distributions of the activations of a neural network trained to predict bioactivities. The Frechet distance between the distributions of the activations of

the generated molecules and a reference set is then calculated. This metric has been shown to be sensitive to distributional differences in many different molecular properties.

The Moses (Polykovskiy et al., 2020) and GuacaMol (Brown et al., 2019) benchmarks bundle these metrics into standardized distribution-learning benchmarks. While improving the evaluation of distribution-learning models, these benchmarks mainly rely on ad-hoc metrics and it is unclear whether they provide a comprehensive evaluation of distribution-learning models.

1.2.5.2 Goal-directed optimization of ML-based scoring functions

Scoring functions based on machine learning models are commonly used in goal-directed generation tasks (**todo**). The fact that such machine learning models are trained on limited amounts of experimental data, adds additional aspects to a proper model evaluation. In this setting there already are known molecules with high scores which are used to train the scoring function. The task thus becomes to find *novel* high-scoring molecules using the ML model's generalization capabilities. However, the machine learning models are often biased towards their training data, which might lead to a lack of novelty in the generated molecules. It is not clear whether this actually leads to a lack of novelty and how to quantify such biases.

Another issue is that optimizing an ML model's output with respect to its input can lead to unexpected problems. Research has demonstrated that samples generated through this optimization process can incorrectly receive high scores from the model, as shown in (Szegedy et al., 2014; Goodfellow et al., 2015). This phenomenon occurs because discriminative models don't necessarily learn all characteristics of high-scoring samples to perform classification. Instead, they often rely on a limited set of correlations between the input sample and the target property. Consequently, the model may classify certain samples as high-scoring, even though they wouldn't actually score highly in reality. While this effect was readily detectable in the image domain, where human vision can easily provide ground truth evaluation, it's more challenging to identify in molecular optimization. It remains unclear whether this issue extends to the context of goal-directed molecule generation and how to quantitatively assess its impact in this field.

1.2.5.3   Diversity of generated molecules

The diversity of the generated molecules is an important aspect in the application of goal-directed generative models (Martin, 2001; Gorse, 2006). The used scoring functions are usually only imperfect and incomplete approximations of the desired properties. Given the expected failure of some of the candidates in later experiments, it is important to generate diverse sets of molecules. Diversity encourages uncorrelated outcomes in downstream experiments, which increases the chances of finding a successful candidate. In essence, a varied molecular portfolio serves as a hedge against the inherent uncertainties in the modeling and experimental processes.

The concept of diversity in molecular generation is complex, with its measurement depending on the specific problem at hand. While internal diversity (average pairwise distance between molecules) is commonly used, it has proven inadequate for goal-directed generation (Waldman et al., 2000; Xie et al., 2021; Thomas et al., 2021). Thomas et al. (2021) proposed the sphere exclusion diversity (SEDiv) metric, which aligns better with chemical intuition but can be misleading for differently sized sets. Xie et al. (2023) introduced the #Circles metric, a non-normalized version of SEDiv, which better addresses the needs in goal-directed generation. It does so by by focusing on chemical space coverage of the generated molecules which correlates with the probability of finding successful candidates.

While initial evaluations using #Circles have been conducted, most comparisons are limited by the fact that the models were not adapted to the diverse optimization setting. A comprehensive comparison of models specifically designed for diverse optimization is still missing, leaving open the question of how well different approaches perform in generating diverse, high-scoring molecules.

1.2.5.4   Standardized Computational Resources

A frequently neglected aspect in evaluating goal-directed models is the use of standardized computational resources. At its core, optimizing molecular properties is a search problem that—given unlimited resources—can be solved through exhaustive enumeration of the chemical space. Consequently, the primary challenge in goal-directed generation lies in identifying high-scoring molecules while minimizing resource consumption.

However, many studies compare different models without accounting for this crucial factor, potentially leading to biased comparisons. For instance, some algorithms might run for days or weeks, while others operate for mere minutes or hours. Recently, this issue has gained increased attention, after (Gao et al., 2022) proposed a benchmark that measures the sample efficiency of goal-directed generation algorithms. Other researchers have adapted to this approach (Thomas et al., 2022a; Thomas et al., 2022b; Guo et al., 2023).

Both of these aspects remain underexplored in the literature especially in the context of finding diverse high-scoring molecules.

1.2.6 Retrosynthesis prediction

Drug candidates, whether designed by generative models or other means, eventually need to be synthesized for testing and eventually for use in patients. However, finding a synthesis route for a given molecule can be a complex and time-consuming process. *Computer-aided synthesis planning* (*CASP*) methods help chemists to find synthesis routes, enabling synthesis of previously inaccessible molecules or making synthesis more efficient and cheaper.

This problem is often approached using a retrosynthesis approach (Corey et al., 1969; Corey, 1991), which recursively deconstructs the target molecule into simpler precursors until they match available starting materials. At each step, single-step retrosynthesis prediction models suggest sets of reactants that could theoretically combine to produce the current (intermediate) target molecule. The success of retrosynthesis planning hinges on highly accurate chemical reaction models, as these ensure that the proposed synthetic routes are feasible in laboratory conditions.

Early work in retrosynthesis prediction relied on carefully curated expert rules encoding possible reactions. Recently, machine learning models that learn the patterns of chemical reactions from examples stored in reaction databases have received increased attention (Coley et al., 2018). One line of work relies on sequence-to-sequence SMILES strings of reactants given that of the product, using models originally developed for machine translation (Schwaller et al., 2019; Nam et al., 2016; Schwaller et al., 2018; Karpov et al., 2019; Tetko et al., 2020). Another set of approaches exploit the fact that connectivity in a reaction is often preserved, and use graph neural networks to edit the connectivity of the target molecule in order to yield possible reactants (Sacha et al., 2020; Shi et al., 2020; Somnath et al., 2020; Yan et al., 2020).

Template-based methods represent another approach to retrosynthesis prediction (Segler et al., 2017; Segler et al., 2018b; Dai et al., 2020; Sun et al., 2020). These models first extract a set of graph transformation rules, or templates, from a large reaction database. These templates encode common reaction patterns. Given a target molecule ranks the templates based on their likelihood of producing a feasible reaction. Finally, the highest-ranked templates are applied to the target to yield sets of reactants.

While template-based methods have shown excellent performance in retrosynthesis prediction, they face challenges with rare templates. Template extraction often leads to many templates being represented by only a few training samples, resulting in a few-shot learning problem where models struggle to perform well on these uncommon templates. While some strategies have been proposed to alleviate this issue, such as data augmentation (Fortunato et al., 2020) and specialized architectures and training objectives (Dai et al., 2020), the problem remains a challenge in the field.

1.3 Aims and Objectives

1.3.1 Identifying Failure Modes in Generative Model Evaluation

In (Renz et al., 2019b) we investigate possible failure modes in the evaluation of distribution-learning and goal-directed generative models. We show that the distribution-learning benchmark proposed in GuacaMol (Brown et al., 2019) is not able to distinguish recently published generative models from simple baseline models. We show that most of the tested generative models do not outperform the simple baseline model, or only do so marginally. While this does not necessarily mean that the generative models are not useful, it calls for a more comprehensive evaluation of distribution-learning models, such as evaluations using the negative log-likelihood of the test set when applicable.

In the context of goal-directed optimization we introduce *control scores* that give information whether the optimization leads to the generation of molecules that are biased towards the training data and whether the optimization overfits to the used scoring function. The control scores are obtained by retraining the scoring function on a hold-out set of the training data or using a different random initialization.

We show that the generated molecules are biased towards the high-scoring molecules in the training set, which might lead to a lack of novelty in the generated molecules. We also show that the generative models are able to overfit to the scoring function's random initialization. This might lead to an overestimation of the models' performance and might lead to the generation of molecules that are not high-scoring in reality. The proposed control scores serve as a diagnostic tool to detect these issues. Section 2.1 reprints the corresponding publication.

1.3.2 Diversity-based comparison of goal-directed generators

In (Renz et al., 2024) we introduce a benchmark for diverse optimization that addresses the above-mentioned issues. In this benchmark, we evaluate the diversity of the generated molecules using a recently proposed diversity metric #Circles (Xie et al., 2023). We compare the performance of diverse optimization approaches under two different compute budgets, namely a fixed number of scoring function evaluations and a fixed time budget. The first setting is relevant for applications where the cost of evaluating the scoring function dominates the optimization process, while the second setting is relevant for scoring functions that are cheap to evaluate. Using this setup we test 14 goal-directed optimization methods and show how SMILES-based auto-regressive models dominate the benchmark. Section 2.2 reprints the corresponding publication.

1.3.3 Improving few-shot and zero-shot retrosynthesis prediction

In (Seidl et al., 2022) we propose a novel approach to template-based retrosynthesis prediction. We use a multimodal learning approach that learns to associate relevant templates to product molecules using a Modern Hopfield Network (Ramsauer et al., 2020). Our model can leverage structural information about the templates and can make use of similarities between them. This allows for improved generalization, especially for templates with few training samples and even for unseen templates. This model is several times faster than comparable methods and shows good predictive performance. Section 2.3 reprints the corresponding publication.

1.4 List of publications

This thesis comprises the work published in the following papers:

- P. Renz et al. (2019b). "On Failure Modes in Molecule Generation and Optimization". In: *Drug Discovery Today: Technologies*. Artificial Intelligence 32–33, pp. 55–63. DOI: 10.1016/j.ddtec.2020.09.003

- P. Renz et al. (2024). "Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed Generators". In: *J. Chem. Inf. Model.* DOI: 10.1021/acs.jcim.4c00519

- P. Seidl et al. (2022). "Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks". In: *J. Chem. Inf. Model.* 62.9, pp. 2111–2120. DOI: 10.1021/acs.jcim.1c01065

Other Publications Besides the papers listed above, I have also contributed to the following publications:

- K. Preuer et al. (2018). "Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery". In: *J. Chem. Inf. Model.* 58.9, pp. 1736–1741. DOI: 10.1021/acs.jcim.8b00234

- P. Renz et al. (2019a). "Uncertainty Estimation Methods to Support Decision-Making in Early Phases of Drug Discovery". In: *NeurIPS-2019 Workshop on Safety and Robustness in Decision Making*

- M. Hofmarcher et al. (2020). *Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks.* DOI: 10.48550/arXiv.2004.00979. arXiv: 2004.00979 [cs, q-bio, stat]. Pre-published

- P. Renz et al. (2023). "Low-Count Time Series Anomaly Detection". In: *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. DOI: 10.1109/MLSP55844.2023.10285979

Publications 2

This chapter presents publications as originally published, reprinted with permission from the corresponding publishers. The copyright of the original publications is held by the respective copyright holders. In order to fit the paper dimension, reprinted publications may be scaled in size and/or cropped.

2.1 On Failure Modes in Molecule Generation and Optimization

This publication is reprinted under a CC BY-NC-ND license.

2.2 Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed Generators

This publication is reprinted under a CC BY 4.0 license.

2.3 Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks

This publication is reprinted under a CC BY 4.0 license.

Conclusion and Outlook 3

The work in this thesis has focused on advancing the application of generative models in drug discovery, concentrating on two main aspects: Firstly, we identified limitations in the evaluation of generative models for de novo molecular design, and proposed ways to make evaluation more informative and relevant to practical applications. Secondly we introduced a novel template-based model for retrosynthesis prediction that matches or exceeds the performance of existing methods, performing particularly well on rare reaction templates.

In the first part of this thesis, we showed how established ways of evaluating distribution-learning models cannot differentiate complex models from trivial baseline generators. We also showed how goal-directed generative models used to optimize machine learning-based scoring functions, can overfit to the scoring function and exhibit biases towards already known high scoring molecules contained in the training data.

The second part of this thesis introduced a diversity-based benchmark for goal-directed molecule generators. This benchmark addresses the shortcomings of previous benchmarks by adressing the issues of inadequate diversity measures, non-standardized compute budgets, and lack of model adaptation to the diverse optimization setting. We used this benchmark to evaluate a range of generative models comparing them in a meaningful way.

The last part of this thesis introduced a novel template-based model for retrosynthesis prediction based on Modern Hopfield Networks. This model leverages a multi-modal approach that combines reaction templates and target molecules. Our model is able to generalize over reaction templates and performs particularly well on rare templates. We showed that our model matches or exceeds the performance.

In conclusion, our work provides insights into the capabilities and limitations of current generative models for molecules and proposes novel evaluation strategies. Additionally, our contributions in retrosynthesis prediction enable more accurate computer-aided synthesis planning. We hope that our work will help to accelerate

the drug discovery pipeline and facilitate the development of novel pharmaceutical treatments.

Bibliography

Bagal, V., Aggarwal, R., Vinod, P. K., and Priyakumar, U. D. (2022). "MolGPT: Molecular Generation Using a Transformer-Decoder Model". In: *J. Chem. Inf. Model.* 62.9, pp. 2064–2076. DOI: 10.1021/acs.jcim.1c00600 (cit. on p. 8).

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. (2021). *Flow Network Based Generative Models for Non-Iterative Diverse Candidate Generation*. DOI: 10.48550/arXiv.2106.04399. arXiv: 2106.04399 [cs]. Pre-published (cit. on p. 14).

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. P. (2014). "The ChEMBL Bioactivity Database: An Update". In: *Nucleic Acids Res* 42.D1, pp. D1083–D1090. DOI: 10.1093/nar/gkt1031 (cit. on p. 9).

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. (2019). "GuacaMol: Benchmarking Models for de Novo Molecular Design". In: *J. Chem. Inf. Model.* 59.3, pp. 1096–1108. DOI: 10.1021/acs.jcim.8b00839 (cit. on pp. 14, 15, 18).

Cohen-Karlik, E., Rozenberg, E., and Freedman, D. (2024). "Overcoming Order in Autoregressive Graph Generation for Molecule Generation". In: *Transactions on Machine Learning Research* (cit. on pp. 9, 10).

Coley, C. W., Green, W. H., and Jensen, K. F. (2018). "Machine Learning in Computer-Aided Synthesis Planning". In: *Acc. Chem. Res.* 51.5, pp. 1281–1289. DOI: 10.1021/acs.accounts.8b00087 (cit. on p. 17).

Corey, E. J. and Wipke, W. T. (1969). "Computer-Assisted Design of Complex Organic Syntheses". In: *Science* 166.3902, pp. 178–192. JSTOR: 1727162 (cit. on p. 17).

Corey, E. J. (1991). "The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture)". In: *Angewandte Chemie International Edition in English* 30.5, pp. 455–465. DOI: 10.1002/anie.199104553 (cit. on p. 17).

Dai, H., Li, C., Coley, C. W., Dai, B., and Song, L. (2020). "Retrosynthesis Prediction with Conditional Graph Logic Network". arXiv: 2001.01408 [cs, stat] (cit. on p. 18).

Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. (2018). "Syntax-Directed Variational Autoencoder for Structured Data". arXiv: 1802.08786 [cs] (cit. on p. 11).

De Cao, N. and Kipf, T. (2018). "MolGAN: An Implicit Generative Model for Small Molecular Graphs". arXiv: 1805.11973 [cs, stat] (cit. on pp. 9, 12).

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008). "On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces". In: *ChemMedChem* 3.10, pp. 1503–1507. DOI: 10.1002/cmdc.200800178 (cit. on p. 9).

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs". In: *Journal of Health Economics* 47, pp. 20–33. DOI: 10.1016/j.jhealeco.2016.01.012 (cit. on p. 3).

Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., and Blundell, T. L. (2024). "Machine Learning-Aided Generative Molecular Design". In: *Nat Mach Intell* 6.6, pp. 589–604. DOI: 10.1038/s42256-024-00843-5 (cit. on p. 13).

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). "Deep Learning for Molecular Design—a Review of the State of the Art". In: *Mol. Syst. Des. Eng.* 4.4, pp. 828–849. DOI: 10.1039/C9ME00039A (cit. on p. 13).

English: Caffeine 3D Structure (2010). URL: https://commons.wikimedia.org/wiki/File:Caffeine_3d_structure.png (visited on 06/09/2024) (cit. on p. 7).

Fortunato, M. E., Coley, C. W., Barnes, B. C., and Jensen, K. F. (2020). "Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning". In: *J. Chem. Inf. Model.* 60.7 (7), pp. 3398–3407. DOI: 10.1021/acs.jcim.0c00403 (cit. on p. 18).

Gao, W., Fu, T., Sun, J., and Coley, C. W. (2022). *Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization*. DOI: 10.48550/arXiv.2206.12411. arXiv: 2206.12411 [cs, q-bio]. Pre-published (cit. on p. 17).

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". In: *ACS Cent. Sci.* 4.2, pp. 268–276. DOI: 10.1021/acscentsci.7b00572 (cit. on pp. 8, 10, 11, 14).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). "Generative Adversarial Networks". arXiv: 1406.2661 [cs, stat] (cit. on p. 11).

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. DOI: 10.48550/arXiv.1412.6572. arXiv: 1412.6572 [cs, stat]. Pre-published (cit. on p. 15).

Gorse, A.-D. (2006). "Diversity in Medicinal Chemistry Space". In: *Current Topics in Medicinal Chemistry* 6.1, pp. 3–18 (cit. on p. 16).

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. (2017). "Objective-Reinforced Generative Adversarial Networks (OR-GAN) for Sequence Generation Models". arXiv: 1705.10843 [cs, stat] (cit. on p. 12).

Guo, J. and Schwaller, P. (2023). *Augmented Memory: Capitalizing on Experience Replay to Accelerate De Novo Molecular Design*. arXiv: `2305.16160 [cs, q-bio]`. URL: `http://arxiv.org/abs/2305.16160` (visited on 09/11/2023). Pre-published (cit. on pp. 10, 13, 17).

Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). "DOGS: Reaction-Driven de Novo Design of Bioactive Compounds". In: *PLOS Computational Biology* 8.2, e1002380. DOI: `10.1371/journal.pcbi.1002380` (cit. on p. 9).

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). "InChI, the IUPAC International Chemical Identifier". In: *Journal of Cheminformatics* 7.1, p. 23. DOI: `10.1186/s13321-015-0068-4` (cit. on p. 7).

Hofmarcher, M., Mayr, A., Rumetshofer, E., Ruch, P., Renz, P., Schimunek, J., Seidl, P., Vall, A., Widrich, M., Hochreiter, S., and Klambauer, G. (2020). *Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks*. DOI: `10.48550/arXiv.2004.00979`. arXiv: `2004.00979 [cs, q-bio, stat]`. Pre-published (cit. on p. 20).

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E., and Eck, D. (2016). "Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control". arXiv: `1611.02796 [cs]` (cit. on p. 10).

Jensen, J. H. (2019). "A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space". In: *Chem. Sci.* 10.12, pp. 3567–3572. DOI: `10.1039/C8SC05372C` (cit. on pp. 9, 13).

Jin, W., Barzilay, R., and Jaakkola, T. (2018). "Junction Tree Variational Autoencoder for Molecular Graph Generation". arXiv: `1802.04364 [cs, stat]` (cit. on pp. 8, 11).

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., and Zhavoronkov, A. (2016). "The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology". In: *Oncotarget* 8.7, pp. 10883–10890. DOI: `10.18632/oncotarget.14073`. pmid: `28029644` (cit. on p. 9).

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). "druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico". In: *Mol. Pharmaceutics* 14.9, pp. 3098–3104. DOI: `10.1021/acs.molpharmaceut.7b00346` (cit. on p. 12).

Karpov, P., Godin, G., and Tetko, I. V. (2019). "A Transformer Model for Retrosynthesis". In: *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*. Ed. by I. V. Tetko, V. Kůrková, P. Karpov, and F. Theis. Lec-

ture Notes in Computer Science. Springer International Publishing, pp. 817–830 (cit. on p. 17).

Kingma, D. P. and Welling, M. (2013). *Auto-Encoding Variational Bayes*. DOI: 10. 48550/arXiv.1312.6114. arXiv: 1312.6114 [cs, stat]. Pre-published (cit. on p. 11).

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., Lameiro, R. F., Lemm, D., Lo, A., Moosavi, S. M., Nápoles-Duarte, J. M., Nigam, A., Pollice, R., Rajan, K., Schatzschneider, U., Schwaller, P., Skreta, M., Smit, B., Strieth-Kalthoff, F., Sun, C., Tom, G., Falk Von Rudorff, G., Wang, A., White, A. D., Young, A., Yu, R., and Aspuru-Guzik, A. (2022). "SELFIES and the Future of Molecular String Representations". In: *Patterns* 3.10, p. 100588. DOI: 10.1016/j.patter.2022.100588 (cit. on p. 7).

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). "Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation". arXiv: 1905.13741 [physics, physics:quant-ph, stat] (cit. on p. 8).

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). "Grammar Variational Autoencoder". arXiv: 1703.01925 [stat] (cit. on pp. 8, 11, 14).

Landrum, G. (2006). *RDKit: Open-source Cheminformatics*. URL: http://www.rdkit.org (cit. on p. 14).

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). "Learning Deep Generative Models of Graphs". arXiv: 1803.03324 [cs, stat] (cit. on p. 9).

Liu, Q., Allamanis, M., Brockschmidt, M., and Gaunt, A. (2018). "Constrained Graph Variational Autoencoders for Molecule Design". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 7795–7804 (cit. on p. 9).

Madhawa, K., Ishiguro, K., Nakago, K., and Abe, M. (2019). "GraphNVP: An Invertible Flow Model for Generating Molecular Graphs". arXiv: 1905.11600 [cs, stat] (cit. on pp. 9, 11, 14).

Makurvet, F. D. (2021). "Biologics vs. Small Molecules: Drug Costs and Patient Access". In: *Medicine in Drug Discovery* 9, p. 100075. DOI: 10.1016/j.medidd.2020.100075 (cit. on p. 2).

Martin, Y. C. (2001). "Diverse Viewpoints on Computational Aspects of Molecular Diversity". In: *J. Comb. Chem.* 3.3, pp. 231–250. DOI: 10.1021/cc000073e (cit. on p. 16).

Mazuz, E., Shtar, G., Shapira, B., and Rokach, L. (2023). "Molecule Generation Using Transformers and Policy Gradient Reinforcement Learning". In: *Sci Rep* 13.1, p. 8799. DOI: 10.1038/s41598-023-35648-w (cit. on p. 8).

Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. (2018). "De Novo Generation of Hit-like Molecules from Gene Expression Signatures Using Artificial Intelligence". In: DOI: 10.26434/chemrxiv.7294388.v1 (cit. on p. 12).

Mullard, A. (2016). "Parsing Clinical Success Rates". In: *Nature Reviews Drug Discovery* 15.7, pp. 447–447. DOI: 10.1038/nrd.2016.136 (cit. on p. 4).

Nam, J. and Kim, J. (2016). "Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions". arXiv: 1612.09529 [cs] (cit. on p. 17).

Nigam, A., Pollice, R., Krenn, M., Gomes, G. d. P., and Aspuru-Guzik, A. (2021). "Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES". In: *Chem. Sci.* 12.20, pp. 7079–7090. DOI: 10.1039/D1SC00231G (cit. on p. 13).

Noutahi, E., Gabellini, C., Craig, M., Lim, J. S. C., and Tossou, P. (2023). *Gotta Be SAFE: A New Framework for Molecular Design*. DOI: 10.48550/arXiv.2310.10773. arXiv: 2310.10773 [cs, q-bio]. Pre-published (cit. on p. 8).

O'Boyle, N. and Dalke, A. (2018). "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures". In: DOI: 10.26434/chemrxiv.7097960.v1 (cit. on pp. 7, 8).

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). "Molecular De-Novo Design through Deep Reinforcement Learning". In: *Journal of Cheminformatics* 9.1, p. 48. DOI: 10.1186/s13321-017-0235-x (cit. on pp. 10, 13).

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. (2010). "How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge". In: *Nat Rev Drug Discov* 9.3, pp. 203–214. DOI: 10.1038/nrd3078 (cit. on p. 3).

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., and Zhavoronkov, A. (2020). "Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models". In: *Frontiers in Pharmacology* 11 (cit. on p. 15).

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. (2018). "Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery". In: *J. Chem. Inf. Model.* 58.9, pp. 1736–1741. DOI: 10.1021/acs.jcim.8b00234 (cit. on pp. 14, 20).

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. (2020). "Hopfield Networks Is All You Need". arXiv: 2008.02217 [cs, stat] (cit. on p. 19).

Renz, P., Cutajar, K., Twomey, N., Cheung, G. K. C., and Xie, H. (2023). "Low-Count Time Series Anomaly Detection". In: *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*. 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. DOI: `10.1109/MLSP55844.2023.10285979` (cit. on p. 20).

Renz, P., Hochreiter, S., and Klambauer, G. (2019a). "Uncertainty Estimation Methods to Support Decision-Making in Early Phases of Drug Discovery". In: *NeurIPS-2019 Workshop on Safety and Robustness in Decision Making* (cit. on p. 20).

Renz, P., Luukkonen, S., and Klambauer, G. (2024). "Diverse Hits in De Novo Molecule Design: Diversity-Based Comparison of Goal-Directed Generators". In: *J. Chem. Inf. Model.* DOI: `10.1021/acs.jcim.4c00519` (cit. on pp. 19, 20).

Renz, P., Van Rompaey, D., Wegner, J. K., Hochreiter, S., and Klambauer, G. (2019b). "On Failure Modes in Molecule Generation and Optimization". In: *Drug Discovery Today: Technologies*. Artificial Intelligence 32–33, pp. 55–63. DOI: `10.1016/j.ddtec.2020.09.003` (cit. on pp. 18, 20).

Rezende, D. J. and Mohamed, S. (2016). "Variational Inference with Normalizing Flows". arXiv: `1505.05770 [cs, stat]` (cit. on p. 11).

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17". In: *J. Chem. Inf. Model.* 52.11, pp. 2864–2875. DOI: `10.1021/ci300415d` (cit. on p. 12).

Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P., and Jastrzębski, S. (2020). "Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits". arXiv: `2006.15426 [physics, stat]` (cit. on p. 17).

Samanta, B., De, A., Jana, G., Chattaraj, P. K., Ganguly, N., and Gomez-Rodriguez, M. (2018). "NeVAE: A Deep Generative Model for Molecular Graphs". arXiv: `1802.05283 [physics, stat]` (cit. on p. 11).

Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018). "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering". In: *Science* 361.6400, pp. 360–365. DOI: `10.1126/science.aat2663`. pmid: `30049875` (cit. on p. 13).

Schneider, G. (2013). *De Novo Molecular Design*. John Wiley & Sons, Ltd. DOI: `10.1002/9783527677016` (cit. on pp. 12, 13).

Schneider, G. and Fechner, U. (2005). "Computer-Based de Novo Design of Drug-like Molecules". In: *Nat Rev Drug Discov* 4.8 (8), pp. 649–663. DOI: `10.1038/nrd1799` (cit. on p. 13).

Schwaller, P., Gaudin, T., Lányi, D., Bekas, C., and Laino, T. (2018). ""Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using

Neural Sequence-to-Sequence Models". In: *Chemical Science* 9.28 (28), pp. 6091–6098. DOI: 10.1039/C8SC02339E (cit. on p. 17).

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019). "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction". In: *ACS Cent. Sci.* 5.9, pp. 1572–1583. DOI: 10.1021/acscentsci.9b00576 (cit. on pp. 8, 17).

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018a). "Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks". In: *ACS Cent. Sci.* 4.1, pp. 120–131. DOI: 10.1021/acscentsci.7b00512 (cit. on pp. 8, 10, 13).

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018b). "Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI". In: *Nature* 555.7698 (7698), pp. 604–610. DOI: 10.1038/nature25978. arXiv: 1708.04202 (cit. on p. 18).

Segler, M. H. S. and Waller, M. P. (2017). "Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction". In: *Chemistry* 23.25 (25), pp. 5966–5971. DOI: 10.1002/chem.201605499 (cit. on p. 18).

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Segler, M., Hochreiter, S., and Klambauer, G. (2022). "Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks". In: *J. Chem. Inf. Model.* 62.9, pp. 2111–2120. DOI: 10.1021/acs.jcim.1c01065 (cit. on pp. 19, 20).

Shi, C., Xu, M., Guo, H., Zhang, M., and Tang, J. (2020). "A Graph to Graphs Framework for Retrosynthesis Prediction". arXiv: 2003.12725 (cit. on p. 17).

Simonovsky, M. and Komodakis, N. (2018). "GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders". arXiv: 1802.03480 [cs] (cit. on p. 11).

Somnath, V. R., Bunne, C., Coley, C. W., Krause, A., and Barzilay, R. (2020). "Learning Graph Models for Template-Free Retrosynthesis". arXiv: 2006.07038 [cs, stat] (cit. on p. 17).

Southey, M. W. Y. and Brunavs, M. (2023). "Introduction to Small Molecule Drug Discovery and Preclinical Development". In: *Front. Drug Discov.* 3. DOI: 10.3389/fddsv.2023.1314077 (cit. on p. 2).

Sun, R., Dai, H., Li, L., Kearnes, S., and Dai, B. (2020). "Energy-Based View of Retrosynthesis". arXiv: 2007.13437 (cit. on p. 18).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). "Intriguing Properties of Neural Networks". arXiv: 1312.6199 [cs] (cit. on p. 15).

Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). "State-of-the-Art Augmented NLP Transformer Models for Direct and Single-Step Retrosynthesis".

In: *Nature Communications* 11.1 (1), p. 5575. DOI: 10.1038/s41467-020-19266-y. arXiv: 2003.02804 (cit. on p. 17).

Thomas, M., O'Boyle, N. M., Bender, A., and De Graaf, C. (2022a). *Re-Evaluating Sample Efficiency in de Novo Molecule Generation*. arXiv: 2212.01385 [cs, q-bio]. URL: http://arxiv.org/abs/2212.01385 (visited on 09/04/2023). Pre-published (cit. on p. 17).

Thomas, M., O'Boyle, N. M., Bender, A., and de Graaf, C. (2022b). "Augmented Hill-Climb Increases Reinforcement Learning Efficiency for Language-Based de Novo Molecule Generation". In: *Journal of Cheminformatics* 14.1, p. 68. DOI: 10.1186/s13321-022-00646-z (cit. on pp. 10, 13, 17).

Thomas, M., Smith, R. T., O'Boyle, N. M., de Graaf, C., and Bender, A. (2021). "Comparison of Structure- and Ligand-Based Scoring Functions for Deep Generative Models: A GPCR Case Study". In: *Journal of Cheminformatics* 13.1, p. 39. DOI: 10.1186/s13321-021-00516-0 (cit. on p. 16).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention Is All You Need". arXiv: 1706.03762 [cs] (cit. on pp. 7, 8).

Waldman, M., Li, H., and Hassan, M. (2000). "Novel Algorithms for the Optimization of Molecular Diversity of Combinatorial libraries11Color Plates for This Article Are on Pages 533–536." In: *Journal of Molecular Graphics and Modelling* 18.4, pp. 412–426. DOI: 10.1016/S1093-3263(00)00071-1 (cit. on p. 16).

Walters, W. P. (2019). "Virtual Chemical Libraries". In: *J. Med. Chem.* 62.3, pp. 1116–1124. DOI: 10.1021/acs.jmedchem.8b01048 (cit. on p. 12).

Weininger, D. (1988). "SMILES, a Chemical Language and Information System". In: *J. Chem. Inf. Comput. Sci.* 28.1, pp. 31–36. DOI: 10.1021/ci00057a005 (cit. on p. 7).

Williams, R. J. (1992). "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Mach Learn* 8.3-4, pp. 229–256. DOI: 10.1007/BF00992696 (cit. on p. 13).

Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., and Clevert, D.-A. (2019). "Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space". In: *Chem. Sci.* 10.34, pp. 8016–8024. DOI: 10.1039/C9SC01928F (cit. on p. 14).

Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., and Li, L. (2021). "MARS: Markov Molecular Sampling for Multi-objective Drug Discovery". arXiv: 2103.10432 [cs, q-bio] (cit. on pp. 13, 16).

Xie, Y., Xu, Z., Ma, J., and Mei, Q. (2023). "How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules". In: ICLR (cit. on pp. 16, 19).

Yan, C., Ding, Q., Zhao, P., Zheng, S., Yang, J., Yu, Y., and Huang, J. (2020). "RetroX-pert: Decompose Retrosynthesis Prediction like a Chemist". arXiv: 2011.02893 [cs, q-bio] (cit. on p. 17).

Yang, X., Zhang, J., Yoshizoe, K., Terayama, K., and Tsuda, K. (2017). "ChemTS: An Efficient Python Library for de Novo Molecular Generation". In: *Science and Technology of Advanced Materials* 18.1, pp. 972–976. DOI: 10.1080/14686996.2017.1401424. pmid: 29435094 (cit. on p. 13).

Yoshikawa, N., Terayama, K., Honma, T., Oono, K., and Tsuda, K. (2018). "Population-Based de Novo Molecule Generation, Using Grammatical Evolution". arXiv: 1804.02134 [physics, q-bio] (cit. on p. 13).

You, J., Liu, B., Ying, R., Pande, V., and Leskovec, J. (2019). *Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation*. DOI: 10.48550/arXiv.1806.02473. arXiv: 1806.02473 [cs, stat]. Pre-published (cit. on pp. 9, 13).