JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
Philipp Renz
01126686

Submitted at
Institute for Machine Learning

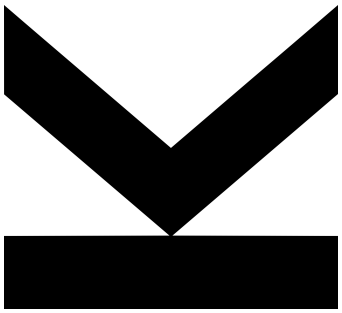
Thesis Supervisor / First Evaluator
Univ.-Prof. Mag. Dr.
Günter Klambauer

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

Second Evaluator
Name

May 2024

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



Doctoral Thesis

to obtain the academic degree of

Doktor der Naturwissenschaften

in the Doctoral Program

Naturwissenschaften

JOHANNES KEPLER UNIVERSITY LINZ
Altenbergerstraße 69
4040 Linz, Austria
www.jku.at
DVR 0093696

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

I would like to thank my supervisor, Prof. Dr. Günter Klambauer for his guidance and support throughout the course of this thesis. I am also grateful to Sepp Hochreiter without whom this work would not have been possible.

I would like to thank my colleagues at the Institute of Machine Learning for many hours of fruitful discussions and exchange. It was a pleasure being around you. Especially I would like to thank my co-author Philipp S. It was such a pleasure collaborating with you, and I'm grateful to have had such a great colleague to work with. Big thanks also go out to Vihang, Theresa and my favourite non-co-author Johannes. A big thank you also goes to Birgit and Jenny who do an awesome job of keeping the institute running, and keeping us all sane. Herbert and the IT team also deserve a big thank you for keeping our GPUs up and running.

A special thank you goes to my family for their unconditional support. Thank you Alfred, Eveline, Sarah, and Wolfgang, Ronja, Raphi for always being there for me.

Finally, I would like to all my friends for their support and for just being there. This includes the kayakers, acro people, and the legendary PL crew (including the ones who would not call it that.). You're al-rye-ght. Life would not as rich without you. Special thanks go out to Alex and Alina. I'm lucky to have you as friends.

Last but not least, I would like to thank my partner, Jordan. Although you only joined me for the last part of this journey, you have been a constant source of support and love. I'm grateful to have you in my life.

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

1 Introduction

1.1 Small molecule drug design

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

Small molecule drugs are the major kind of medicines in use, constituting as much as 90% of global sales (Makurvet, 2021). Small molecules are usually defined as molecules with a molecular weight of less than 900 Da. These molecules often are orally available, have good pharmacokinetic properties and can be synthesized in a cost-effective manner (Procas, 2019).

For a small molecule to be a viable drug candidate it needs to fulfill a whole range of properties (Procas, 2019):

- **On-target activity:** The molecule needs to be active against the desired target in order for it to show the desired therapeutic effect. On a molecular level this means that the molecule needs to bind to the target and modulate its activity in the desired way.
- **Pharmacokinetics:** The molecule must have favourable pharmacokinetic properties such as adsorption, distribution, metabolism and excretion (ADME). ADME determine how the molecule is absorbed into the body, how it is distributed in the body, how it is metabolized and how it is excreted from the body. These properties are crucial for the molecule to reach the target in the body and to be metabolized in a safe manner and to finally be excreted from the body.
- **Toxicity:** The absence of toxic effects is crucial, as the molecule must be well-tolerated and devoid of any potential harmful side effects. Toxicity can be caused by a range of factors, including off-target interactions, metabolic byproducts or allergies.

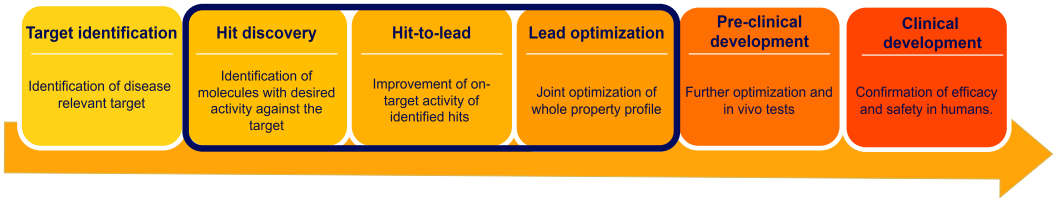


Figure 1.1: The drug discovery pipeline starts with the identification of a biological target. Once a target is identified, readily available molecules are screened for their activity against the target in high-throughput screening. Promising hits are then modified and optimized to lead compounds. These 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.

- **Specificity:** The molecule should exhibit high specificity, selectively interacting with the intended target while minimizing undesirable off-target interactions. Off-target binding can lead to adverse side effects and potentially compromise the drug's safety and efficacy profile.
- **Synthesizability:** The molecule must be synthesizable in a cost-effective manner to be practically useful.
- **Patentability:** The molecule must be novel and not infringe on any existing patents. While in general this is not needed for a drug to work, this constitutes a significant issue in practice.

The main challenge in drug discovery is to find a molecule that fulfills all these mentioned properties. The development of a new drug is a complex and expensive process, which can take up to 10–15 years and costs up to 3 billion USD (Procas, 2019).

1.1.1 The drug discovery pipeline

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

- **Target identification:** The drug discovery process starts with the identification of a biological target, which is a molecule or a protein that is involved in a disease process.
- **Hit discovery:** In the hit discovery stage molecules are screened for their activity against the target in high-throughput screening (HTS). These lab

experiments, often referred to as assays, are used to measure the activity of the molecules against the target in vitro. This stage results in a set of so called hits, which are molecules that show activity against the target.

- **Hit-to-lead:** Promising hits are then modified and optimized to lead compounds. In this stage, the optimization is primarily focused on improving the activity of the molecule against the target. This is usually done in a DMTA (Design-Make-Test-Analyze) cycle, where the molecule is designed, synthesized, tested in vitro. The results are then analyzed and the cycle continues until a satisfactory lead compound is found.
- **Lead optimization:** The lead compounds are then further optimized to improve their properties, such as pharmacokinetics, toxicity or specificity. This is usually done in a DMTA cycle as in the hit-to-lead stage and also involves the synthesis and testing of the molecules in vitro.
- **Preclinical development:** The most promising candidates are then tested in preclinical studies. These studies are usually done in animals and are used to assess the safety and efficacy of the drug candidate in vivo.
- **Clinical trials:** Finally, the candidates that pass the preclinical studies are tested in humans in clinical trials. These are usually divided into three phases, where the safety and efficacy of the drug are tested in increasing numbers of patients. Phase I trials are mainly focused on the safety of the drug, Phase II trials are focused on the efficacy of the drug and Phase III trials are focused on the safety and efficacy of the drug in a larger population.
- **Regulatory approval:** The final stage is the regulatory approval, where the drug is approved by regulatory agencies such as the FDA in the US or the EMA in Europe.

The general strategy of this stagewise approach is to reduce the uncertainty about the usefulness of a molecule at each stage. The earlier stages are usually cheaper and faster, but have higher uncertainty about the clinical success of the molecule. The later stages are more expensive and slower, but provide better information about the success chances of a molecule (Procas, 2019).

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). This translates to 63%, 19.5%, 11.3% and 9.6% of projects that make it to the respective stages.

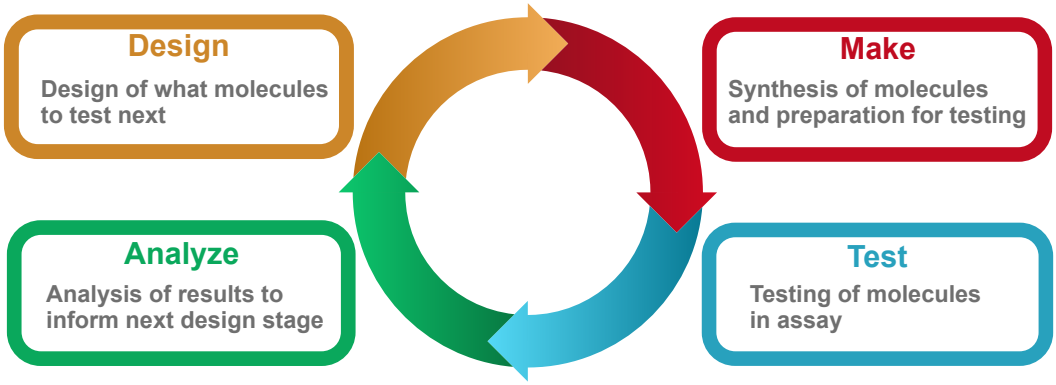


Figure 1.2: The DMTA cycle

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. This stage often relies on computational methods to predict the properties of the molecules.
- **Make:** The designed molecules are then synthesized 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.
- **Analyze:** The results of the experiments are then analyzed. involves the 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.

Computational methods are widely used throughout the DMTA cycle. One of the most important applications of computational methods in drug discovery is the prediction of the properties of molecules. These properties can range from the activity of a molecule against a target, to its pharmacokinetic properties, to its

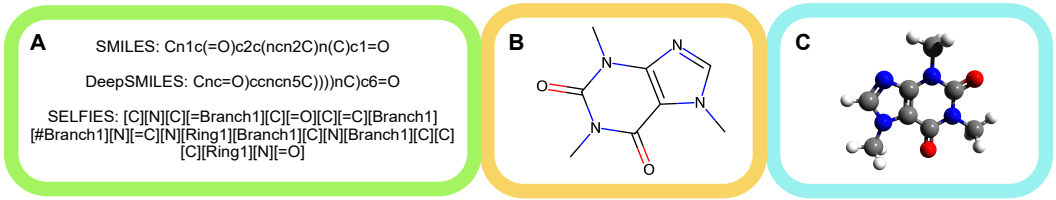


Figure 1.3: Different ways to represent molecules. The molecule shown is caffeine **A:** Different 1D representations of a molecule. SMILES is an established line notation for molecules. DeepSmiles enables easier generation of molecules by getting rid of pair brackets and ring numbers. SELFIES guarantees that a sequence of tokens parses into a valid molecule. **B:** 2D graph representation of a molecule. The nodes represent atoms and the edges represent bonds. **C:** 3D structure of a molecule. The atoms are positioned in 3D space. The positions of these atoms can change as some bonds are allow rotations. Source of the 3D structure: (*English: Caffeine 3D Structure* 2010).

toxicity. These *quantitative structure-property relationship* (*QSPR*) models are used to predict the properties of molecules in the design phase, to guide the selection of molecules to be tested.

Recent years have seen a surge of interest in generative models for drug discovery, expanding the capabilities of computer-aided drug design to tasks that produce molecular structures as outputs. This thesis focuses on two critical applications: de novo drug design, which aims to discover molecules with desired property profiles, and computer-aided synthesis planning, which seeks to determine viable synthesis routes for molecules of interest. By harnessing the power of generative models in these areas, we aim to accelerate the drug discovery process and pave the way for more efficient development of potential new therapies.

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. 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. Figure 1.3b shows a graph representation of caffeine. 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. Early work by (Segler et al., 2018) and (Gómez-Bombarelli et al., 2018) used recurrent neural networks (RNNs) to generate molecules in SMILES format. Auto-regressive modelling is based on the idea of generating a molecule by iteratively predicting the next characters of the SMILES string given the preciding characters. The likelihood is thus modelled by $p(x) = \prod_{i=0}^{n} p(x_i|x_{1:i-1})$. 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), parsing the molecules into specialized data structures (Kusner et al., 2017; Jin et al.,

2018) and using other 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 generate molecules in graph-based representations. In this case 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 (Cohen-Karlik et al., 2024; You et al., 2019).

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. As a self-supervised learning task, it allows models to leverage large datasets. The

resulting models serve 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 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

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

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., 2018; 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

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

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 to run continuous optimization algorithms in latent space. VAEs belong in the class of approximate density model, as the likelihood of a given molecule can be calculated only approximately via monte carlo sampling.

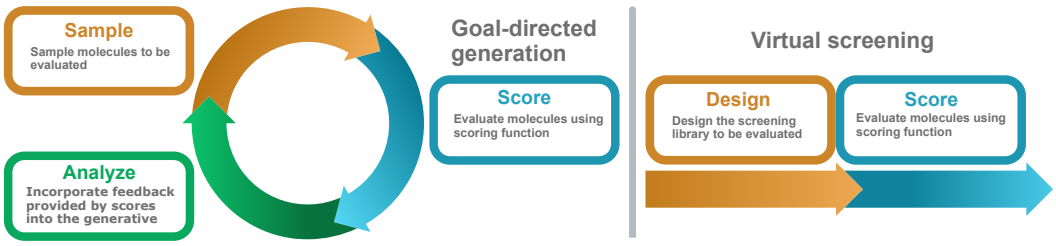


Figure 1.4: Comparison of goal-directed generative models and virtual screening. Goal-directed generation proceeds in a loop where already scored molecules inform what molecules to test next. Virtual screening proceeds in a linear fashion, where the molecules to be tested are determined beforehand.

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:

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

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 as calculating the likelihood of a given molecule is not tractable.

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 *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 have been used in a variety of applications, such as the design of new drugs, materials or catalysts (Procas, 2019).

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

- **Hill-climbing** (Segler et al., 2018; 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** transform an initial population of molecules by applying mutations and crossovers. The molecules are then evaluated and the best ones are selected for the next generation. Applying this process iteratively leads to a population of high-scoring molecules (Jensen, 2019; Nigam et al., 2021; Yoshikawa et al., 2018)
- **Tree search** builds a tree of possible molecules by recursively applying a set of rules to to the current molecules. Molecules with higher scores are more likely to be expanded further (Yang et al., 2017; Jensen, 2019)
- **Continuous optimization** employ classical optimization algorithms in the continuous latent space of (variational) autoencoders (Winter et al., 2019; Gómez-Bombarelli et al., 2018; Kusner et al., 2017) 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 Evaluation challenges

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 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 of the generated molecules and the training set is a commonly used metric, to evaluate the distribution fit. The Frechet ChemNet Distance (Preuer et al., 2018) provides a more comprehensive evaluation of the distribution fit by comparing the distribution of the activations a neural network trained to predict the bioactivities. This bioactivity informed metric has been shown to be sensitive to distributional differences 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 (Procas, 2019). However, the fact that machine learning models are trained on limited amounts of experimental data, adds additional aspects to a proper model evaluation. In this setting there are already 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. It is not clear whether the high-scoring molecules generated by ML algorithms are sufficiently novel or if they tend to be heavily biased to the high-scoring compounds in the training set.

Another issue is that optimizing an ML model's output with respect to its input can lead to problems. It has been shown that samples generated in this way can wrongly be awarded high scores as shown in (Szegedy et al., 2014; Goodfellow et al., 2015). It is unclear whether this problem transfers to the context of goal-directed molecule generation. While the mentioned studies worked in the image domain where ground truth evaluation can easily be achieved using human vision in molecular optimization this problem is harder to quantify and detect.

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.

However, the concept of diversity is multifaceted and the importance of different aspects depends on the application. The internal diversity or average pairwise distance between generated molecules is a common metric to evaluate the diversity of compounds, but it has been shown to be a poor metric in the context of goal-directed generation (Waldman et al., 2000; Xie et al., 2021; Thomas et al., 2021).

Thomas et al. (2021) highlight that the internal diversity is not in line with chemical intuition in some descriptive cases and propose the sphere exclusion diversity (SEDiv) metric which measures a sets diversity by the number of diverse compounds selected using a sphere exclusion algorithm (Gobbi et al., 2003; Sayle, 2019) over the sets size. While this metric is more in line with chemical intuition, it is a relative metric and can lead to misleading results for sets of different sizes. For example, a single molecule has perfect diversity according to this metric, which is not in line with the intuitive understanding of diversity.

Recently, Xie et al. (2023) introduced the #Circles metric, which is identical to the SEDiv metric but skips the normalization by the number of molecules. This metric is more in line with the needs in goal-directed generation where one is interested in coverage of the chemical space rather than having sets with low redundancy. While the authors evaluate and compare a limited number of different goal-directed models using #Circles, a comprehensive comparison of models using this metric is still missing, leaving open the question of how well different models perform in the task of finding diverse high-scoring molecules.

1.2.5.4 Standardized Computational Resources

A frequently neglected aspect in evaluating goal-directed models is the absence of standardized computational resource allocation. 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 de novo design 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.

The computational cost of running a goal-directed model comprises two main components: molecule generation and scoring. In applications where scoring is more costly than molecule generation, a model's sample efficiency i.e. the number of scoring function evaluations needed to reach a certain performance level, mainly determines its performance. Recently, sample efficiency has gained increased attention, with (Gao et al., 2022) proposing a benchmark focused on this aspect. Other researchers have adapted to this approach (Thomas et al., 2022a; Thomas et al., 2022b; Guo et al., 2023).

The converse in which scoring is relatively cheap compared to molecule generation has received less attention in the literature. In this case a models performance is mainly determined by a combination of its generation speed and sample efficiency.

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 (Procas, 2019).

Template-based methods represent another approach to retrosynthesis prediction (Segler et al., 2017; Dai et al., 2020; Fortunato 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.

For goal-directed models we study to which extent introduce *control scores* that give information whether the optimization overfits to artifacts of the scoring functions, or the training data. We train additional scoring functions, using either a different random initialization or training it on a hold-out subset of the the available training data. Using this approach, we show that the generated samples are biased towards the training data and show biases towards the scoring function's random initialization.

This shows that the reported performance of these models is an overestimation, and that generative models overfit to the scoring function's random initialization and to high-scoring training samples. This shows that the reported performance of these models is an overestimation, and that our control scores can be used to obtain a more meaningful evaluation of goal-directed molecule generators. 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]. URL: http://arxiv.org/abs/2004.00979 (visited on 09/16/2022). preprint

- 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

2 Publications

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

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

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

3 Conclusion and Outlook

In this thesis, we have made contributions to applications of machine learning in drug discovery. Our focus lay on advancing the evaluation of generative models for molecules and on improving single-step retrosynthesis prediction.

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. 12).

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. 13, 16).

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. 8, 9).

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. 16).

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. 15).

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. 15).

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. 16).

De Cao, N. and Kipf, T. (2018). "MolGAN: An Implicit Generative Model for Small Molecular Graphs". arXiv: 1805.11973 [cs, stat] (cit. on p. 8).

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. 8).

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. 11).

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. 11).

English: Caffeine 3D Structure (2010). URL: https://commons.wikimedia.org/wiki/File:Caffeine_3d_structure.png (visited on 06/09/2024) (cit. on p. 6).

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. 16).

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. 15).

Gobbi, A. and Lee, M.-L. (2003). "DISE: Directed Sphere Exclusion". In: *J. Chem. Inf. Comput. Sci.* 43.1, pp. 317–323. DOI: 10.1021/ci025554v (cit. on p. 14).

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. 7, 9, 12).

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. 10).

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. 13).

Gorse, A.-D. (2006). "Diversity in Medicinal Chemistry Space". In: *Current Topics in Medicinal Chemistry* 6.1, pp. 3–18 (cit. on p. 14).

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. 9, 12, 15).

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. 8).

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]. URL: http://arxiv.org/abs/2004.00979 (visited on 09/16/2022). preprint (cit. on p. 18).

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. 9).

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. 8, 12).

Jin, W., Barzilay, R., and Jaakkola, T. (2018). "Junction Tree Variational Autoencoder for Molecular Graph Generation". arXiv: 1802.04364 [cs, stat] (cit. on p. 7).

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. 8).

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. Lecture Notes in Computer Science. Springer International Publishing, pp. 817–830 (cit. on p. 16).

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. 9).

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. 7).

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). "Grammar Variational Autoencoder". arXiv: 1703.01925 [stat] (cit. on pp. 7, 12).

Landrum, G. (2006). *RDKit: Open-source Cheminformatics*. URL: http://www.rdkit.org (cit. on p. 12).

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. 8, 10, 12).

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. 14).

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).

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. 16).

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. 12).

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 pp. 7, 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 p. 7).

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. 9, 12).

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. 13).

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. 13, 18).

Procas, t. (2019). "TODO: Why did I not do this already?" In: *Journal of Unfinished Things* (cit. on pp. 2–4, 11, 13, 16).

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. 17).

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. 19).

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. 18).

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. 17, 18).

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. 16, 18).

Rezende, D. J. and Mohamed, S. (2016). "Variational Inference with Normalizing Flows". arXiv: 1505.05770 [cs, stat] (cit. on p. 10).

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. 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. 11).

Sayle, R. (2019). *2D Similarity, Diversity and Clustering in RDKit*. URL: https://www.nextmovesoftware.com/talks/Sayle_2DSimilarityDiversityAndClusteringInRdkit_RDKITUGM_201909.pdf (visited on 11/17/2023) (cit. on p. 14).

Schneider, G. (2013). *De Novo Molecular Design*. John Wiley & Sons, Ltd. DOI: 10.1002/9783527677016 (cit. on pp. 10, 11).

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. 11).

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. 16).

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, 16).

Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). "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. 7, 9, 11).

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. 16).

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. 17, 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. 13).

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. 16).

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. 15).

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. 9, 11, 12, 15).

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. 14).

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. 14).

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. 11).

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. 11).

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. 12).

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. 11, 14).

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. 14, 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. 12).

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. 12).

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. 8, 12).