JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
Philipp Renz
01126686

Submitted at
Institute for Machine Learning

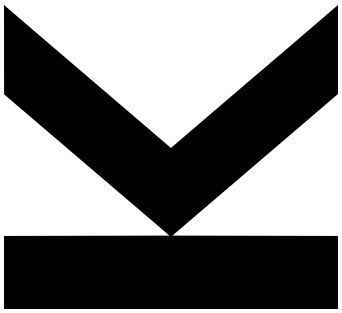
Thesis Supervisor / First Evaluator
Univ.-Prof. Mag. Dr.
Günter Klambauer

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

1 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 (**makurvetBiologicsVsSmall2021**). 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 (**southeyIntroductionSmallMolecule2023**).

For a small molecule to be considered a viable drug candidate, it must fulfill a range of properties (**southeyIntroductionSmallMolecule2023**):

- **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

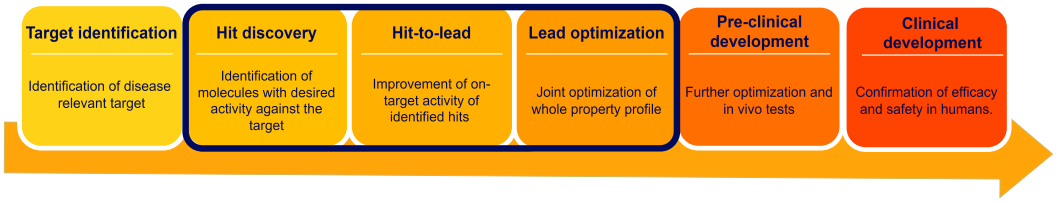


Figure 1.1: The drug discovery pipeline starts with the identification of a biological target. Once a target is identified, readily available molecules are screened for their activity against the target in high-throughput screening. Promising hits are then modified and optimized to lead compounds. These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

body, distributed throughout, metabolized, and ultimately excreted. They are 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 (**paulHowImproveProductivity2010**; **dimasiInnovationPharmaceuticalIndustry**

1.1.1 The drug discovery pipeline

The drug discovery process is usually divided into several stages (**todo**) depicted in **??** 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 (**mullardParsingClinicalSuccess2016**). This translates to 63%, 19.5%, 11.3% and 9.6% of projects that make it to the respective stages.

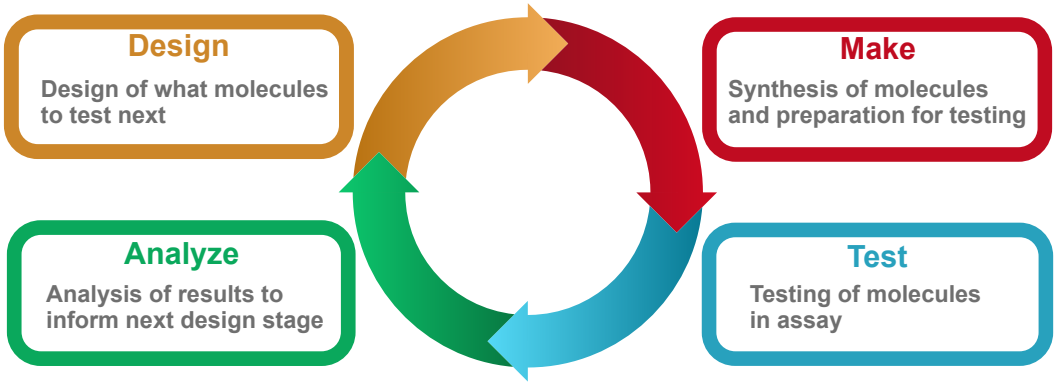


Figure 1.2: The DMTA cycle

- **Regulatory Approval:** Upon successful completion of clinical trials, the drug is submitted for regulatory approval. Agencies such as the U.S. Food and 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 **??**) 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*-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.

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 molecular properties. These predictions can be used in the design phase, to focus on 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. These models can be used to solve tasks which require molecular structures as outputs. In this thesis we focus on two critical 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 the production of a target molecule.

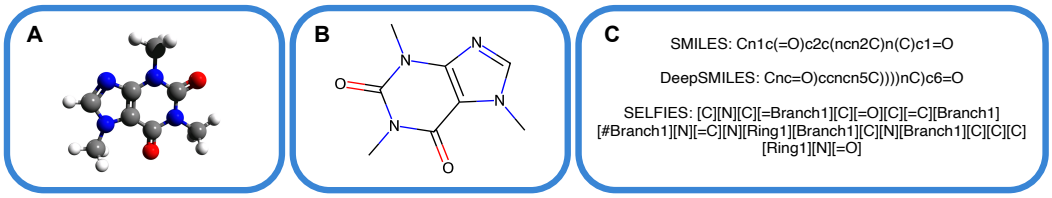


Figure 1.3: Different ways to represent a caffeine molecule **A:** The 3D structure of a molecule is given by the positions of its atoms in space. This structure is not necessarily fixed as some bonds can rotate and bonds can vibrate (Image source: (**EnglishCaffeine3D2010**)). **B:** The graph representation 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.

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 **??**b 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 (**hellerInChIIUPACInternational2015**) and SMILES (Simplified Molecular Input Line Entry System) (**weiningerSMILESChemicalLangu** 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 (**vaswaniAttentionAllYou2017**). Several extensions to SMILES have been proposed to make them more amenable for use in machine learning models. DeepSmiles (**oboyleDeepSMILESAdaptationSMILES2018**) attempted to make it easier to generate syntactically valid molecules, by changing the notation

of branches and ring closures. SELFIES (**krennSELFIESFutureMolecular2022**) provide a representation of molecules in which any sequence of tokens parses into a valid molecule. SAFE (**noutahiGottaBeSAFE2023**) 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 (**seglerGeneratingFocusedMolecule2018**) and (**gomez-bombarelliAutomaticChemicalDesign2018**) 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 (**oboyleDeepSMILESAdaptationSMILES2018**; **krennSelfReferencingEmbeddedStrings2020**), parsing the molecules into specialized data structures (**kusnerGrammarVariationalAutoencoder2**; **jinJunctionTreeVariational2018**) and using other architectures such as transformers (**vaswaniAttentionAllYou2017**; **noutahiGottaBeSAFE2023**; **schwallerMoleculartransforme**; **bagalMolGPTMolecularGeneration2022**; **mazuzMoleculeGenerationUsing2023**).

Graph-based autoregressive models generate molecules in graph-based representations. In this case the model generates the molecular graph by iteratively adding nodes6 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-karlikOvercomingOrderAutoregressive2024**; **youGraphConvolutionalPolicy2019**).

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 (**decaoMolGANImplicitGenerative2018**; **madhawaGraphNVPInvertibleFlow2019**; **kadurinCornucopiaMeaningfulLeads2016**).

Rule-based models generate molecules by applying a set of pre-defined graph transformation rules to combine molecular fragments. The BRICS (**degenArtCompilingUsing200**
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 (**hartenfellerDOGSReactionDrivenNovo2012**) 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. **jensenGraphbasedGeneticAlgorithm2019** 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* 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 (**gomez-bombarelliAutomaticChemicalDesign2018**; **seglerGeneratin**
olivecronaMolecularDenovoDesign2017; **guoAugmentedMemoryCapitalizing2023**; **thomasAugmentedHillClimbIncreases2022**; **jaquesSequenceTutorConservative2016**; **cohen-karlikOvercomingOrderAutoregressive2024**)

Variational autoencoders (VAEs) (**kingmaAutoEncodingVariationalBayes2013**) 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.

Generative flows (**rezendeVariationalInferenceNormalizing2016**) 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 $G : z \to x$. The likelihood of the training data can then be directly calculated and optimized via the change of variables formula:

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

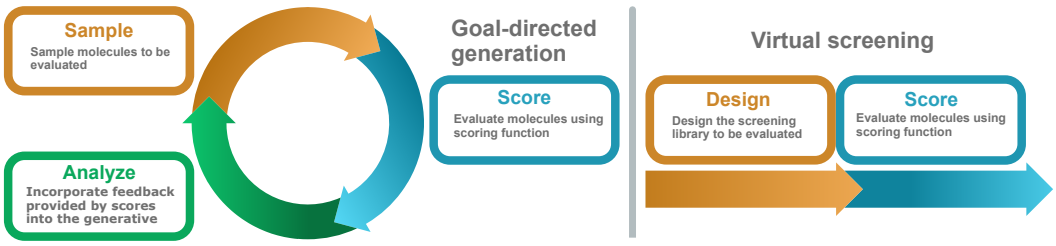


Figure 1.4: Comparison of goal-directed generative models and virtual screening. Goal-directed generation proceeds in a loop where already scored molecules inform what molecules to test next. Virtual screening proceeds in a linear fashion, where the molecules to be tested are determined beforehand.

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 (**madhawaGraphNVPInvertibleFlow2019**). 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) (**goodfellowGenerativeAdversarialNetworks2014**) 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 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 (**schneiderNovoMolecularDesign2013**) is a computational approach for automatically designing molecules with desired property profiles. Goal-directed generation expands upon *virtual screening*, a method in which a library of molecules is ranked according to the output of a *quantitative structure-property relationship* model. **waltersVirtualChemicalLibraries2019** estimates that approximately 10^{13} molecules can be routinely tested in a *virtual screening* experiment. While this number can vary significantly depending on the computational cost of running the *quantitative structure-property relationship* model, it is dwarfed by the size of drug-like chemical space, which is estimated to contain between 10^{30} and 10^{60} molecules (**waltersVirtualChemicalLibraries2019**; **ruddigkeitEnumeration166Billion2012**). Consequently, *virtual screening* 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 *virtual screening* by focusing the search on the most relevant parts of chemical space. In contrast to the random search approach taken by *virtual screening*, 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 *quantitative structure-property relationship* 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 (**eltonDeepLearningMolecular2019**; **sanchez-lengelingInverseMolecularDesign2018**; **duMachineLearningaidedGenerative2024**). 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. (**schneiderComputerbasedNovoDesign2005**; **schneiderNovoMolecularDesign2013**). 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 (**todo**).

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

- **Hill-climbing** (**seglerGeneratingFocusedMolecule2018**; **xieMARSMarkovMolecular2021**; **thomasAugmentedHillClimbIncreases2022**) 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 (**williamsSimpleStatisticalGradientfollowing1992**) which allows to update the model distribution in a way that increases expected scores of the generated molecules (**olivecronaMolecularDenovoDesign2017**; **thomasAugmentedHillClimbIncreases2022**; **youGraphConvolutionalPolicy2019**; **guoAugmentedMemoryCapitalizing2023**).

- **Genetic algorithms** in molecular generation operate by evolving an initial population of molecules through iterative cycles of mutation, crossover, and selection (**jensenGraphbasedGeneticAlgorithm2019**; **nigamGenerativeModelsSuperfast**; **yoshikawaPopulationbasedNovoMolecule2018**). 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 (**yangChemTSEfficientPython2017**; **jensenGraphbasedGeneticAlgorithm2019**

- **Continuous optimization** employ classical optimization algorithms in the continuous latent space of (variational) autoencoders (**gomez-bombarelliAutomaticChemi**; **kusnerGrammarVariationalAutoencoder2017**; **winterEfficientMultiobjectiveMolecular2**) or generative flows (**madhawaGraphNVPInvertibleFlow2019**).

- **Generative Flow Networks** (**bengioFlowNetworkBased2021**) 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 (**landrumRDKitOpensourceCheminformatics2006**). 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 (**brownGuacaMolBenchmarkingModels2019**). The Frechet ChemNet Distance (FCD) (**preuerFrechetChemNetDistance2018**) 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 (**polykovskiyMolecularSetsMOSES2020**) and GuacaMol (**brownGuacaMolBenchma**
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**). 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 (**szegedyIntriguingPropertiesNeural2014**; **goodfellowExplainingHarnessingAdversarial2015**). While this effect was easy to

detect in the image domain where ground truth evaluation can easily be achieved using human vision, it is harder to detect in molecular optimization. It is unclear whether this problem transfers to the context of goal-directed molecule generation and how to quantify it.

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 (**martinDiverseViewpointsComputational2001**; **gorseDiversityMedicinalChemistry2006**). 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 (**waldmanNovelAlgorithmsOptimization2000**; **xieMARSMarkovMolecular2021**; **thomasComparisonStructureLigandbased2021**).

thomasComparisonStructureLigandbased2021 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 (**gobbiDISEDirectedSphere2003**; **sayle2DSimilarityDiversity2019**) 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, **xieHowMuchSpace2023** 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 (**gaoSampleEfficiencyMatters2022**) proposing a benchmark focused on this aspect. Other researchers have adapted to this approach (**thomasReevaluatingSampleEfficiency2022**; **thomasAugmentedHillClimbIncreases2022**; **guoAugmentedMemoryCapitalizing2023**).

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* 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 (**coreyComputerAssistedDesig**
coreyLogicChemicalSynthesis1991a), 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 (**coleyMachineLearningComputerAided2018**). 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 (**schwallerMolecularTransformerModel2019**; **namLinkingNeuralMachine2016**; **schwallerFour**
karpovTransformerModelRetrosynthesis2019; **tetkoStateoftheartAugmentedNLP2020**). 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 (**todo**).

Template-based methods represent another approach to retrosynthesis prediction (**seglerNeuralSymbolicMachineLearning2017**; **daiRetrosynthesisPredictionConditional2020**; **fortunatoDataAugmentationPretraining2020**). 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 (**fortunatoDataAugmentationPretraining2020**) and specialized architectures and training objectives (**daiRetrosynthesisPredictionConditional2020**), the problem remains a challenge in the field.

1.3 Aims and Objectives

1.3.1 Identifying Failure Modes in Generative Model Evaluation

In (**renzFailureModesMolecule2019**) 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 (**brownGuacaMolBenchmarkin** 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. **??** reprints the corresponding publication.

1.3.2 Diversity-based comparison of goal-directed generators

In (**renzDiverseHitsNovo2024**) 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 (**xieHowMuchSpace2023**). 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. **??** reprints the corresponding publication.

1.3.3 Improving few-shot and zero-shot retrosynthesis prediction

In (**seidlImprovingFewZeroShot2022**) 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 (**ramsauerHopfieldNetworksAll2020**). 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. **??** reprints the corresponding publication.

1.4 List of publications

This thesis comprises the work published in the following papers:

- **renzFailureModesMolecule2019**
- **renzDiverseHitsNovo2024**
- **seidlImprovingFewZeroShot2022**

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

- **preuerFrechetChemNetDistance2018**
- **renzUncertaintyEstimationMethods2019**
- **hofmarcherLargescaleLigandbasedVirtual2020**
- **renzLowCountTimeSeries2023**

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

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 scoring functions can be biased towards the scoring function, leading to overfitting and biases to 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.