JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
Philipp Renz
01126686

Submitted at
Institute for Machine Learning

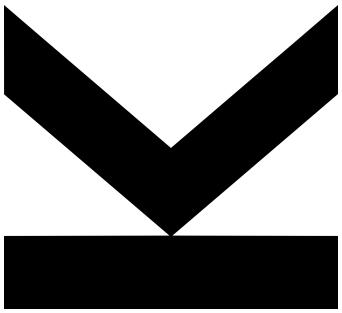
Thesis Supervisor / First Evaluator
Univ.-Prof. Mag. Dr.
Günter Klambauer

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

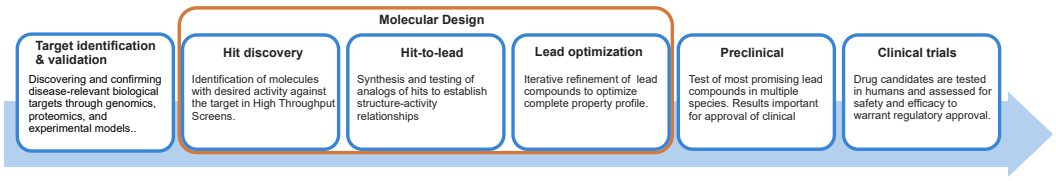


Figure 1.1: The drug discovery pipeline starts with the identification of a biological target. Once a target is identified, readily available molecules are screened for their activity against the target in high-throughput screening. Promising hits are then modified and optimized to lead compounds. These 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. Molecular design (orange box) is highly amenable to machine learning approaches and is the primary 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 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**).1 This translates to 63%, 19.5%, 11.3% and 9.6% of projects that make it to the respective stages.
- **Regulatory Approval:** Upon successful completion of clinical trials, the drug is submitted for regulatory approval. Agencies such as the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA) review the comprehensive data package, including preclinical and clinical trial results.

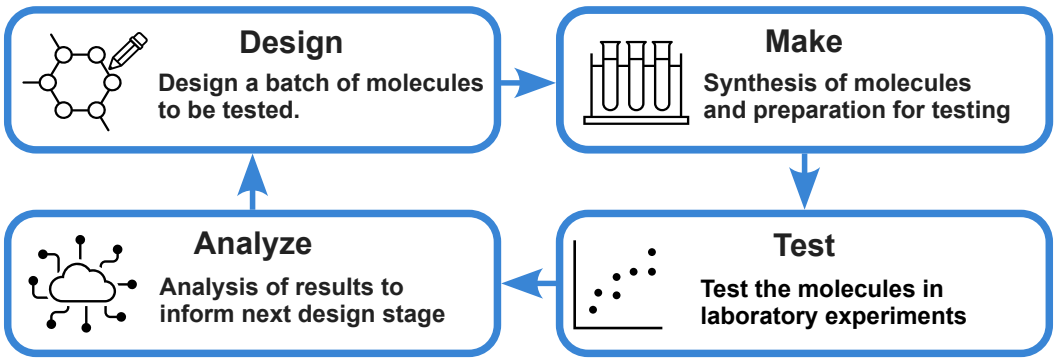


Figure 1.2: The Design-Make-Test-Analyze cycle is a key concept in drug discovery. The cycle consists of four stages: Design, Make, Test, and Analyze. Generative models can be used to design promising molecules to be tested. Computer-aided synthesis planning tools can be employed to make sure the molecules can be synthesized in the Make stage. The In the Analyze stage the experimental results can be used to update the property prediction models underlying the Design stage.

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.

1.1.3 Machine learning in drug discovery

Computer-aided drug design (CADD) has long been an integral part of the pharmaceutical research and development process. It encompasses a range of computational methods aimed at supporting and enhancing drug discovery. While traditional CADD approaches have proven valuable, the integration of machine learning (ML) has significantly expanded their capabilities.

Machine learning has become an important tool in modern CADD, offering new approaches for predicting molecular properties and activities. ML models, trained on datasets molecular structures and their associated properties, can screen chemical libraries to identify potential drug candidates, reducing the time and resources required for experimental testing

Recent advances in deep learning have led to a surge in interested in generative models, introducing new possibilities in drug discovery. These models expand the application of ML beyond property prediction to the creation of novel molecular structures. Two key applications of generative models in drug discovery are:

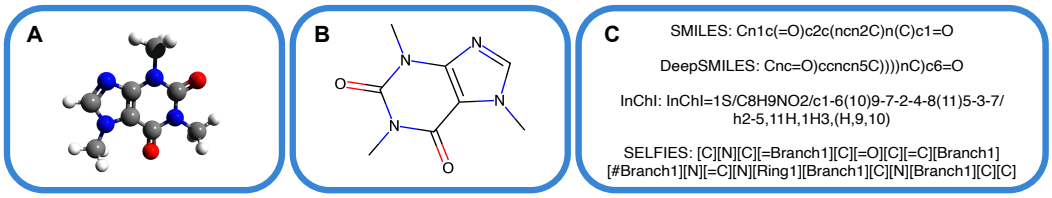


Figure 1.3: Different ways to represent a caffeine molecule **A:** The 3D structure of a molecule is given by the positions of its atoms in space. This structure is not necessarily fixed as some bonds can rotate and bonds can vibrate (Image source: (**EnglishCaffeine3D2010**)). **B:** The graph representation of the same molecule. **C:** Smiles, DeepSmiles, SELFIES and InChI are line notations that linearize the molecules graph representation.

- **De Novo Drug Design:** Generative models can create new molecular structures with specified property profiles, potentially expanding the chemical space explored in drug discovery. This approach allows for the generation of diverse compounds that may complement traditional design methods.
- **Computer-Aided Synthesis Planning:** Generative models are being applied to propose synthetic routes for target molecules, addressing a challenge in drug development. By suggesting potential synthesis pathways, these models aim to support the transition from in silico design to experimental realization.

The integration of generative models into the drug discovery pipeline represents a new approach, offering additional tools for molecular design and synthesis planning. As these technologies continue to develop, they may contribute to enhancing various aspects of the drug discovery process.

1.2 Generative models in drug discovery

1.2.1 Molecular representations

Molecules, though fundamentally complex quantum mechanical entities, can be represented through various simplified models for practical purposes. The most common representation depicts molecules as graphs, where atoms are nodes and chemical bonds are edges. Figure **??**b shows a graph representation of caffeine. This graph structure captures the molecule's connectivity, which defines its "identity". Additional properties such as atom type or charge are incorporated as features of the nodes and edges. While this representation doesn't capture the full quantum

complexity, it provides a stable and practical framework for understanding and working with molecular structures in many scientific and computational contexts.

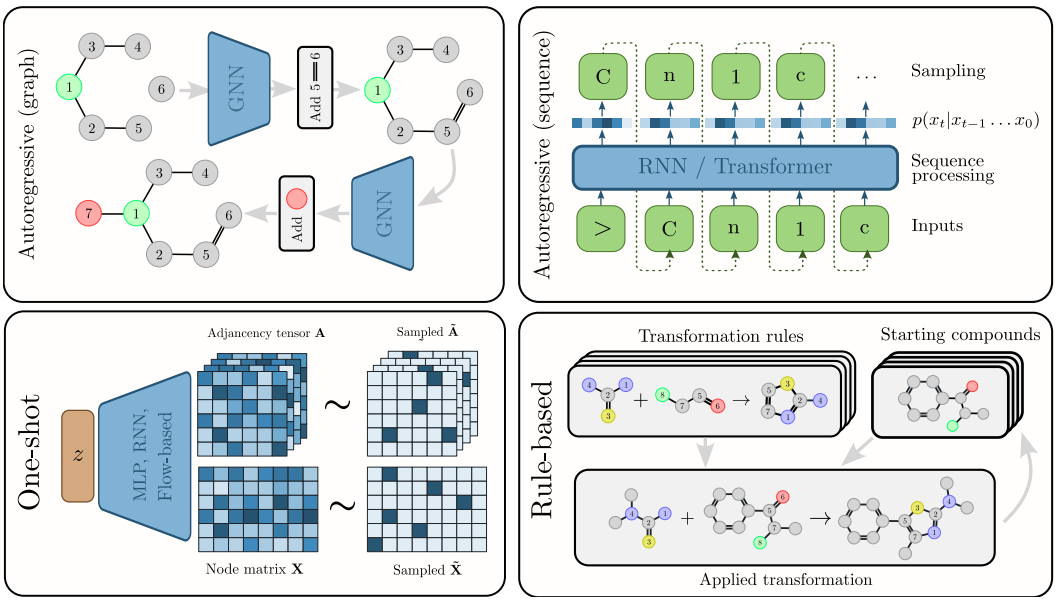
Molecular graphs can be linearized into one-dimensional character sequences, known as line notations. Figure **??**c shows examples of various line notations. SMILES (Simplified Molecular Input Line Entry System) (**weiningerSMILESChemicalLanguage1**) is a widely used line notation that represents molecules as strings of characters. SMILES strings encode the molecular graph in a human-readable format, making them convenient for storage and processing. 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. InChI (**hellerInChIIUPACInternational2015**) is less human-readable and less used in machine learning contexts, but provides a non-proprietory representation of molecules, with strict uniqueness and canonicalization rules.

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

There are several approaches to constructing up molecular graphs. These mainly depend on the type of model used and the representation of the molecule. Some of the most common approaches are shown in Figure **??** and described below.

Sequence-based autoregressive models constitute one of the most popular approaches for generating molecules. This approach makes use of a linearized representation of the molecule, such as a SMILES string. Then the model generates



the molecule by sampling one token at a time, conditioned on the previously sampled tokens, similarly to how a language model generates text. Early work by (**seglerGeneratingFocusedMolecule2018**) and (**gomez-bombarelliAutomaticChemicalDesign2**) relied on recurrent neural networks (RNNs) to generate SMILES strings. This approach has since been popular and there has been work on string-based representations more suitable to generation (**oboyleDeepSMILESAdaptationSMILES2018**; **krennSelfReferencingEmbeddedStrings2020**; **noutahiGottaBeSAFE2023**), parsing the molecules into specialized data structures (**kusnerGrammarVariationalAutoencoder2017**; **jinJunctionTreeVariational2018**) and using other deep learning architectures such as transformers (**vaswaniAttentionAllYou2017**; **noutahiGottaBeSAFE2023**; **schwallerMolecular**; **bagalMolGPTMolecularGeneration2022**; **mazuzMoleculeGenerationUsing2023**).

Graph-based autoregressive models work similarly to its sequence-based counterpart, but instead of relying on a linearization of the molecule, they work directly on the graph representation of the molecule. The model generates the molecular graph by iteratively adding nodes and edges to the graph. The model can be trained in a similar manner to the string-based models, by predicting the next node or edge given the current state of the graph. However, the specification of possible actions is more complex than in the 1D case as there is no natural ordering of the nodes and edges in the graph (**liuConstrainedGraphVariational2018**; **liLearningDeepGenerative2018**; **youGraphConvolutionalPolicy2019**; **cohen-karlikOvercomin**

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

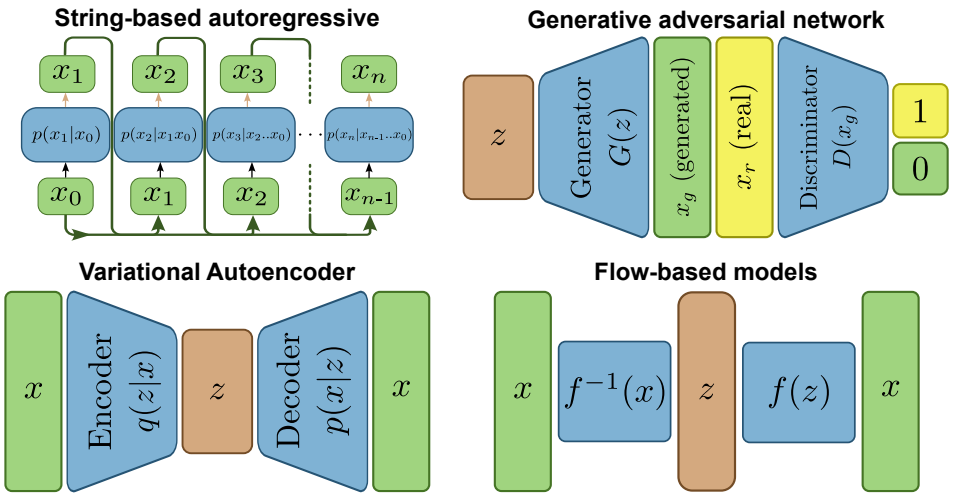


Figure 1.4: Different types of distribution-learning models. All model types try to fit the data distribution $p(x)$, but differ in the way they achieve this goal. While autoregressive models and generative flows the exact likelihood of the data can be calculated, and optimized, VAEs rely on a variational approximation of the likelihood, and generative adversarial networks indirectly fit the data distribution using a game-theoretic approach.

first generating a continous version of the molecule and then discretizing it to a valid molecule (**decaoMolGANImplicitGenerative2018**; **madhawaGraphNVPInvertibleFlow2019**).

Rule-based models generate molecules by applying a set of pre-defined graph transformation rules to combine molecular fragments. The BRICS (**degenArtCompilingUsing2008**) 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 in a dataset. This approach enables the model to grasp both the syntax and semantics of the molecules in the data.

These models can be trained on large chemical libraries of stable molecules, PubChem (**kimPubChemSubstanceCompound2016**), ChEMBL (**bentoChEMBLBioactivityDatabas**) or ZINC (**irwinZINCFreeTool2012**). Using this process the resulting models can learn what reasonable molecules look like in the context of drug discovery. This makes them useful for their two main purposes: they can expand virtual libraries and, more crucially, act as a foundation for other applications such as goal-directed generation, which we will explore in the subsequent section.

In recent years there has been a surge in interest distribution-learning models based on deep neural networks. Many architectures and training strategies originally 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 the use continuous optimization algorithms in latent space. VAEs belong in the class of approximate density models, as the likelihood of a given molecule can be calculated approximately via monte carlo sampling. VAEs have been a popular choice for generating molecules (**gomez-bombarelliAutomaticChemicalDesign2**
kusnerGrammarVariationalAutoencoder2017; **simonovskyGraphVAEGenerationSmall2018**;
samantaNeVAEDeepGenerative2018; **jinJunctionTreeVariational2018**; **daiSyntaxDirectedVari**
liuConstrainedGraphVariational2018).

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 $f : z \to x$. The likelihood of the training data can then be directly calculated and optimized via the change of variables formula:

$$p_x(x) = p_z(f^{-1}(x)) \left| \det \left[\frac{\partial f^{-1}(v)}{\partial v} \Big|_{v=x} \right] \right|. \tag{1.3}$$

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 (**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 is trained to distinguish between real and generated data, while the generator is trained to generate samples that fool the discriminator. *generative adversarial network*s are implicit density methods they sample from the model distribution, but do not provide a likelihood for a given sample. This approach has been combined with different generation strategies in the context of drug discovery (**decaoMolGANImplicitGenerative2018**; **kadurinDruGANAdvancedGenerative2017**;
guimaraesObjectiveReinforcedGenerativeAdversarial2017; **mendez-lucioNovoGenerationHi**
tangMolecularGenerativeAdversarial2024)

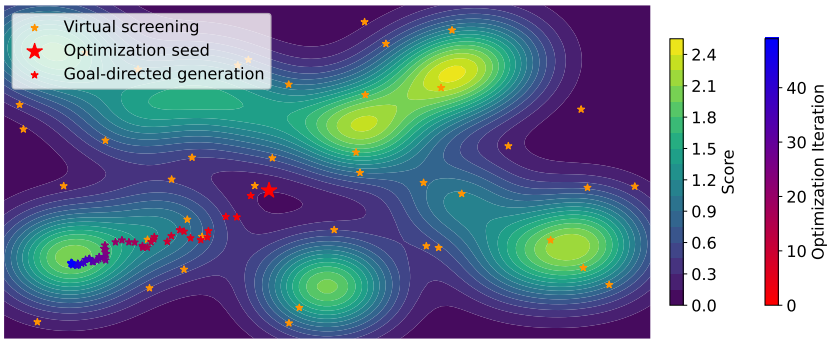


Figure 1.5: Illustration of the difference between goal-directed molecule generation and virtual screening in a 2D chemical space, where each point represents a molecule. The background color represents the molecules' scores. Goal-directed generation works akin to a numerical optimization algorithm and efficiently finds high-scoring molecules, shown in the transition from red to blue stars. In contrast VS amounts to a random search in chemical space, which is less efficient and is likely to miss high-scoring regions of chemical space.

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 has led to success in a variety of applications, such as the design of new drugs, materials or catalysts (**todo**).

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

- **Hill-climbing** (**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 achieved through 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** methods build 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**; **jensenGraphbasedGeneticAlgorithm2**
- **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 Challenges in Evaluating Generative Models in de Novo Design

1.2.5.1 Evaluation of distribution-learning models

The most basic and commonly used checks to assess the quality of the generated compounds are the validity, uniqueness and novelty of the generated molecules. A molecule is valid if it obeys chemical valence rules, which is usually checked using chemoinformatics toolkits such as RDKit (**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**), but is usually determined using a limited number of properties. The Frechet ChemNet Distance (FCD) (**preuerFrechetChemNetDistance2018**) provides a more comprehensive check of the distribution fit. 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 has been shown to be sensitive to distributional differences in many different molecular properties.

The MOSES (**polykovskiyMolecularSetsMOSES2020**) and GuacaMol (**brownGuacaMolBenchn** benchmarks provide standardized frameworks for evaluating distribution-learning models in molecular generation. While these benchmarks represent progress in assessment methodology, questions remain about their comprehensiveness and ability to fully capture the complexities of molecular generation tasks in drug discovery contexts.

1.2.5.2 Goal-directed optimization of ML-based scoring functions

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

Another issue is that optimizing an ML model's output with respect to its input can lead to unexpected problems. Research has demonstrated that samples generated through this optimization process can incorrectly receive high scores from the model, as shown in (**szegedyIntriguingPropertiesNeural2014**; **goodfellowExplainingHarnessingAdversarial2015**). This phenomenon occurs because discriminative models don't necessarily learn all characteristics of high-scoring samples to perform classification. Instead, they often rely on a limited set of correlations between the input sample and the target property.

Consequently, the model may classify certain samples as high-scoring, even though they wouldn't actually score highly in reality. While this effect was readily detectable in the image domain, where human vision can easily provide ground truth evaluation, it's more challenging to identify in molecular optimization. It

remains unclear whether this issue extends to the context of goal-directed molecule generation and how to quantitatively assess its impact in this field.

1.2.5.3 Diversity of generated molecules

The diversity of the generated molecules is an important aspect in the application of goal-directed generative models (**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. In essence, a varied molecular portfolio serves as a hedge against the inherent uncertainties in the modeling and experimental processes.

The concept of diversity in molecular generation is complex, with its measurement depending on the specific problem at hand. While internal diversity (average pairwise distance between molecules) is commonly used, it has proven inadequate for goal-directed generation (**waldmanNovelAlgorithmsOptimization2000**; **xieMARSMarkovMolecular2021**; **thomasComparisonStructureLigandbased2021**). **thomasComparisonStructureLigandbased2021** proposed the sphere exclusion diversity (SEDiv) metric, which aligns better with chemical intuition but can be misleading for differently sized sets. **xieHowMuchSpace2023** introduced the #Circles metric, a non-normalized version of SEDiv, which better addresses the needs in goal-directed generation. It does so by by focusing on chemical space coverage of the generated molecules which correlates with the probability of finding successful candidates.

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

1.2.5.4 Standardized Computational Resources

A frequently neglected aspect in evaluating goal-directed models is the use of standardized computational resources. At its core, optimizing molecular properties is a search problem that—given unlimited resources—can be solved through exhaustive

enumeration of the chemical space. Consequently, the primary challenge in goal-directed generation lies in identifying high-scoring molecules while minimizing resource consumption.

However, many studies compare different models without accounting for this crucial factor, potentially leading to biased comparisons. For instance, some algorithms might run for days or weeks, while others operate for mere minutes or hours. Recently, this issue has gained increased attention, after (**gaoSampleEfficiencyMatters2022**) proposed a benchmark that measures the sample efficiency of goal-directed generation algorithms. Other researchers have adapted to this approach (**thomasReevaluatingSampleEf**
thomasAugmentedHillClimbIncreases2022; **guoAugmentedMemoryCapitalizing2023**).

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 (**sachaMoleculeEditGraph2020**; **shiGraphGraphsFramework2020**; **somnathLearningGraphModels2020**; **yanRetroXpertDecom**

Template-based methods represent another approach to retrosynthesis prediction (**seglerNeuralSymbolicMachineLearning2017**; **seglerPlanningChemicalSyntheses2018**; **daiRetrosynthesisPredictionConditional2020**; **sunEnergybasedViewRetrosynthesis2020**). 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.

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

We show that the generated molecules are biased towards the high-scoring molecules in the training set, which might lead to a lack of novelty in the generated molecules. We also show that the generative models are able to overfit to the scoring function's random initialization. This might lead to an overestimation of the models' performance and might lead to the generation of molecules that are not high-scoring in reality. The proposed control scores serve as a diagnostic tool to detect these issues. **??** 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 machine learning-based scoring functions, can overfit to the scoring function and exhibit biases towards already known high scoring molecules contained in the training data.

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

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

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

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