Fake it until you make it? Generative De novo Design and Virtual Screening of Synthesizable Molecules[1]

Megan Stanley, Marwin Segler[2]
Microsoft Research AI4Science

Abstract

Computational techniques, including virtual screening, *de novo* design, and generative models, play an increasing role in expediting DMTA cycles for modern molecular discovery. However, computationally proposed molecules must be synthetically feasible for laboratory testing. In this perspective, we offer a succinct introduction to the subject, showcase typical workflows to integrate synthesis planning, synthesizability scoring, and molecule generation. Finally, we address limitations and opportunities for future research.

1. Introduction

Virtual screening, and *de novo* design are computational techniques to propose molecules with optimal property profiles for drug, materials, and fine chemicals discovery [1, 2, 3]. Usually, multiple properties, in particular biological, ADMET, and physico-chemical properties have to be optimized and often balanced against each other.

The putative molecules also need to satisfy additional criteria, such as sufficient novelty, and chemical stability. Finally, before any molecule can be tested, it needs to be synthesizable, i.e. successfully made in the lab, or otherwise be accessible from natural sources.

A target molecule can be synthesized if and only if a viable synthesis route of reactions from the starting materials to the target can be found. This depends on the availability of suitable starting materials, (often also called

[1]Manuscript accepted at Current Opinion in Structural Biology
[2]marwinsegler@microsoft.com

building block molecules), and the chemical tractability of the reactions in the route. Also, the viability of a synthesis depends on fulfilling different requirements dependent on the stage of a molecular discovery project. For example, in hit-to-lead discovery or early lead optimization, to achieve short design-make-test cycle times, complex and time-consuming multi-step syntheses that could take weeks or months are usually undesirable. In contrast, in late lead optimization for high-profile first-in-class targets, more involved syntheses might be acceptable. Similarly, low-yielding or expensive reactions are tolerable when it helps to reach a goal quickly in earlier discovery phases, whereas later in scale up for production this is not the case. Taking this a step further, in material science or agro-chemistry, larger amounts of the compounds are often required and material cost needs to be low, hence syntheses typically need to be simpler compared to drug discovery. Thus, synthesizability is not an inherent molecular property, as it can only be indirectly predicted from molecular structure in isolation.

While early computational methods for virtual screening (VS) and *de novo* design (DND) did not take synthesizability into consideration, in the chemoinformatics community the problems of synthesis of algorithmically generated molecules have been recognized since the 1990s [4, 5]. For example, a succinct description is available in Klebe's popular "Drug Design" textbook (2009) [1].

Recently, the interest in *de novo* design has been rekindled by generative machine learning (ML)/AI models [6, 7, 8]. Generative models for molecules complement established chemoinformatics methods. They allow to learn to construct molecules in a probabilistic way directly from data, thus allowing to sample from the distribution of the training data.

Besides summarizing current state of the art, this article also intends to make the concept of synthesizabilty accessible to the machine learning community. We refer the reader also to other reviews on the topic of de novo design and chemical space exploration [3, 9, 10].

The article is organized as follows: We first introduce the concept of chemical spaces, and then provide a unified of how virtual screening and de novo design explore such spaces. We then discuss how synthesizability scoring, synthesis prediction, generative models and synthesis constrained generation can be used to obtain synthesizable chemical spaces and their limitations. Finally, we discuss a few recent examples of prospective validation.

2. Chemical Spaces

Chemical Space usually refers to the set of all possible molecules under a given definition of how these can be constructed, or obtained.

Due to the combinatorial nature of molecules, chemical spaces can be very large [9, 11]. They are almost always too large to enumerate, let alone store. For drug-like molecules, the number has been estimated to be at least 10^{33} [12]. However, current molecules entering the clinic get more complex, and research into new modalities like PROTACs, which leads to molecules with larger molecular mass, indicates the our notion of what constitutes drug-like chemical space needs to be expanded, and previous size estimates are likely too conservative.

Accessibility, of which synthesizability is an important part, differs vastly between chemical spaces (Figure 1). Theoretically, we could generate a very large chemical space of molecules by enumerating all possible combinations of atoms and bonds, subject only to valency constraints. However, most of them would not be synthesizable or even stable, let alone reasonable. In contrast, the molecules in a high-throughput screening (HTS) deck are highly accessible because the compounds are already on stock, and do not have to be made on demand. Virtual molecules, generated on the fly with synthesizability constraints, represent a trade off, as they can often be readily obtained, while at the same time cover large chemical spaces.

3. A unified view on Virtual Screening and De Novo Design

After having introduced Chemical Spaces, we can now discuss how they can be explored.

We consider a general, multi-objective scoring function $f : \mathcal{M} \to \mathbb{R}$ which, given molecule $m \in \mathcal{M}$, yields a numerical score $s \in \mathbb{R}$ as a weighted sum of the objectives. We note that more sophisticated ways for multi-objective scoring and optimization exist, however, this goes beyond the scope of this article [2]. Common objectives could be fingerprint similarity towards a reference molecule, docking, machine learning/Quantitative Structure-Activity Relationships (QSAR), or more sophisticated physics-based simulators such as free energy pertubation (FEP), or computable physicochemical or topological properties, which can be as simple molecular weight, ring counts or the presence or absence of substructures. Second, we assume we have a molecule representation and a construction algorithm, which will be discussed in more

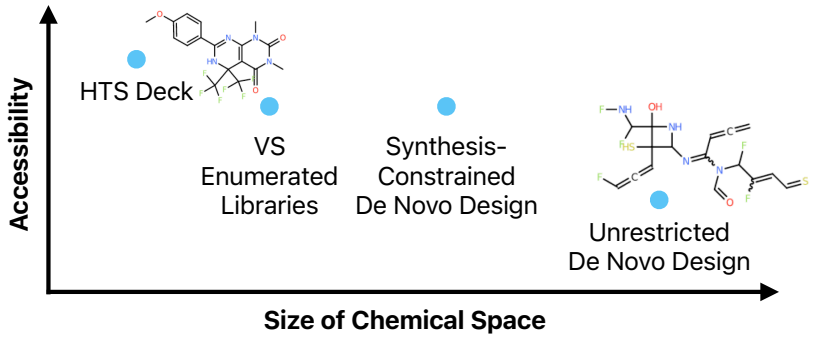

Figure 1: Schematic Representation of Chemical Space. Accessibility here refers to a combination of ease of access and synthesis, but also other important criteria such as stability and chemical reasonableness. On-shelf compounds, for example from a high throughput screening (HTS) collection, are readily accessible. Virtual make-on-demand compound datasets that rely on enumeration of a small number of robust reactions against building blocks represent larger chemical spaces, but suffer from the need to store the full data sets, which becomes intractable quickly. Synthesis-constrained or biased *de novo* design can give molecules as accessible as in enumerated VS libraries, but allows to access much larger spaces as generated molecules do not need to be stored. Unrestricted *de novo* design, which can combine atoms and bonds under valency constraints, provides the largest theoretically possible chemical space. However, most of these molecules would not be synthesizable.

detail in Sec. 7. Third, we need some kind of optimization component, which either returns the best molecules scored so far, or pushes the construction algorithm towards higher scoring molecules. Fig. 2 shows the how the common steps fit together.

Virtual Screening (VS) is named due to its similarity to high-throughput screening. Generally, in VS, first a virtual library, i.e. a dataset, of molecules is generated, and stored.[5, 13] Then, in a second step, all molecules in the library are scored using the scoring function, and eventually the highest scoring molecules are selected. De Novo Design (DND) [14, 15] usually combines three components: molecule generation of novel molecules from scratch, the scoring function, and an optimization, search, or reinforcement learning routine that drives the algorithm to generate higher scoring molecules. The latter drives the molecule generator to create higher-scoring molecules. Hierarchical Virtual Screening,[5, 16] where an informer library is screened and then locally expanded, and Combinatorial Fragment Space Search[17, 18] can be seen as instantiations of DND. Generative Models are machine learning

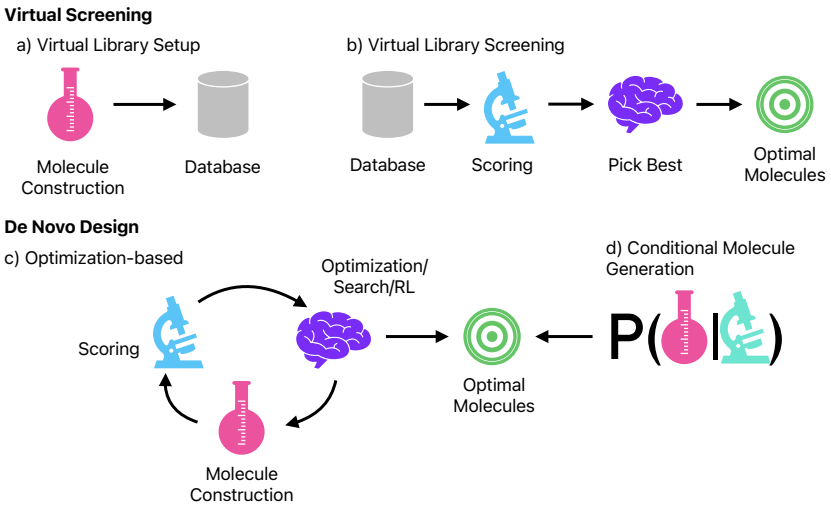

Figure 2: Virtual screening (VS) and *de novo* design (DND) are means to the same end: Obtaining optimal molecules according to an arbitrary scoring function (e.g. a docking engine, MD simulation, ML/QSAR, similarity, and multi-objective combinations). In Virtual Screening, first a database of molecules is constructed (a). Then, each molecule in the database is scored, and the best molecules returned (b). The currently most widely used DND approaches are optimization-based (c). Here, molecules are constructed, scored, and using optimization, search or RL, the algorithm or model is driven to construct molecules with improved scores in the next iteration. Eventually, the best molecules are returned. Conditional Molecule Generation (d) is an emerging approach, where a pre-trained conditional generative model $P(\mathcal{M}|\mathcal{P})$ is used to directly generate molecules $\mathcal{M}$ that likely exhibit properties $\mathcal{P}$.

models that can construct molecules, and have been used as part of DND and VS.

Virtual Screening and *de novo* Design are thus means to the same end: an exploration of chemical space to seek out high-scoring molecules. In practice, VS/DND are often used to generate focused libraries using computationally cheaper scoring functions, which are then narrowed down in a "funnel"-like setup with increasingly expensive scorers.

Probably the biggest differences between VS and DND is in the type of guarantees they provide and the size of chemical space they can access. Since screening a given library will score all N molecules, it is guaranteed to find the best-scoring ones. However, in particular with more expensive physics-based scoring, this is limited to smaller Ns. DND on the other hand has no guarantee to return the best molecules in the chemical space accessible by its molecule

generator. However, because it can explore and sample chemical space in a much more focused way due to the additional optimization component, it can potentially find much higher scoring molecules, in particular those that satisfy multiple criteria.

Larger chemical spaces can also lead to unintended consequences: Since most scoring functions, in particular those for predicting biological endpoints, are imperfect proxies, with more fine-grained sampling of the chemical space it becomes more likely that limitations of the scoring function are exploited to give false-positive molecules whose predicted properties to not match wet-lab experiments [19]. This phenomenon is also known as Goodhart's law, often stated as "When a measure becomes a target, it ceases to be a good measure". Alternatively, this can be described as the selection of molecules outside of the applicability domain of the scoring function, or, from the ML perspective, as an unintended adversarial attack on the scoring function, or in the context of reinforcement learning as reward-hacking.

Renz et al. observed that generative models, when used for very fine-grained and long exploration outside of recommended parameter ranges, can be driven to exploit scoring functions and generate unsynthesizable and unreasonable molecules[20]. This was later corrected by Langevin et al. [21], who demonstrated that this observations should be attributed to the imperfections in the scoring function, and can be addressed when using generative models with appropriate parameters.

Therefore, it is often beneficial to restrict the exploration of the chemical space, and, critically, employ some notion of synthesizability either in the scoring, the construction, or the post-processing of the generated molecules.

4. Accounting for synthesizability in Virtual Screening and *de novo* Design

Synthesizability can be accounted for with different techniques, which fall into the following categories:

- Synthesis Planning
- Synthesizability Scoring
- Biased Molecular Generation, using fragments or generative models
- Reaction-Driven Molecule Generation

In Synthesis Planning, complete routes from the building blocks to the target molecule are returned. Synthesizability scoring provides numeric or categorical scores for how easy a molecule can be made. In Biased Molecular Generation, molecules are constructed from fragments derived from sets of successfully synthesized molecules. In Reaction-Driven Molecule Generation, molecules are constructed as they would be made in the lab, using recursive application of virtual reaction rules and building blocks in the forward direction. Fig. 3 shows how these methods can be incorporated into DND workflows. We will now discuss the methods in more detail.

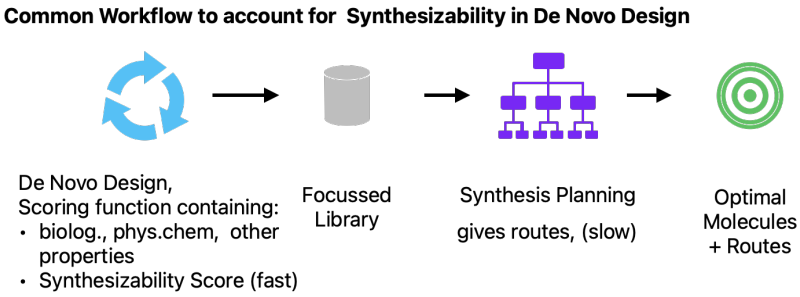

Figure 3: Synthesizability can be incorporated in de novo design workflows by including a fast synthesizability score into the de novo design optimization loop, which is then used to gather a focussed library of the best scoring molecules, biased towards synthesizabilty. Slow full-scale synthesis planning is then used to compute full routes for each molecule in a post-processing step.

5. Synthesizability Scoring

Given a target molecule, synthesizability scoring approaches provide numerical or categorical proxy scores for the ease of synthesis, without performing time-consuming complete synthesis planning. The advantage of such scoring functions is that they are very fast to compute, taking at most milliseconds per molecule, as they are often only based on molecular features of the target molecules. This means that they can be used as part of the scoring function in large scale VS and in the DND loop. Several different approaches to scoring have been proposed. First, scores can be based only on topological features of the molecular graph, and fragment analyses, resting on the assumption that molecules containing common fragments are easier to make, or involve a simplified or truncated, fast retrosynthetic analysis. Boda et al. and

Huang et al. suggested scores combining of structural and reaction features, and simplified retrosynthesis planning. [22, 23] A widely used approach is SAScore, as proposed by Ertl and Schuffenhauer. It penalizes molecules with rare fragments mined from PubChem, and molecules with too many stereo-centers, certain ring-features, and spiro-centers [24]. A disadvantage of such topological approaches is that they capture molecular complexity rather than direct difficulty of synthesis.

More recently, scores based on ML have emerged. SCScore is based on the assumption that the reactants of a reaction should be easier to make than the products, and trains a model for such ranking, which can then be applied to score new molecules [25]. Such scores can also be used to guide synthesis planning algorithms [25, 26]. SYBA is a binary classifier trained to distinguish between real and algorithmically generated, supposedly hard-to-synthesize molecules [27]. More recently, Liu et al and Thakkar et al proposed models to approximate the output of synthesis planning algorithms [28, 29]. They first perform synthesis planning to compute routes for a large test set of molecules, and train the models to predict whether a route for the molecule could be found, or predict scores derived from the routes, such as the number of steps, or cost.

An important investigation about how different synthesizability scoring methods work when integrated with *de novo* design was provided by Gao et al [30]. They demonstrated that without additional synthesizability scoring, *de novo* design algorithms generated large proportions of unsynthesizable molecules.

In particular when using *de novo* design algorithms that are not synthesis-constrained, we usually recommend to include a synthesizability score into the multi-objective scoring function in particular for *de novo* design.

While cheap to compute, a fundamental limitation of synthesizability scorers is that they do not return full synthesis routes, and do not necessarily capture that a seemingly complex molecule can be easily synthesized if the right building blocks are commercially available.

6. Synthesis Planning and Synthesis Prediction

Computer-Aided Synthesis Planning (CASP) can provide full synthesis routes via multi-step retrosynthesis: Starting with the target molecule, formally reverse reactions and potential precursors for the current molecule are

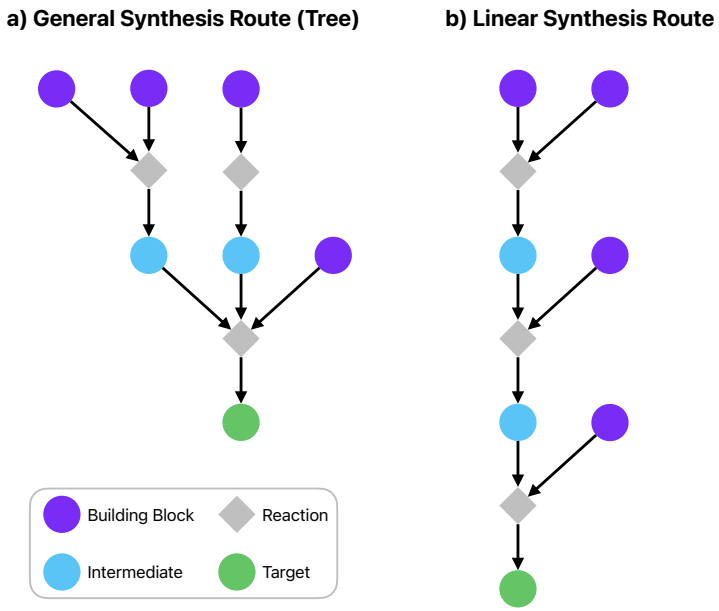

Figure 4: a) Synthesis Planning tools and some reaction driven *de novo* design algorithms return general synthesis routes for target molecules, which have tree structure. Linear routes are a special case of this. b) Some reaction driven *de novo* algorithms can only return linear synthesis routes, which poses a limitation in the chemical space they can explore.

predicted recursively until the molecule is deconstructed into commercially available or known building blocks (see Fig. 4).

CASP has a rich research tradition of 70 years [10]. Until 2017, most systems have relied on hand-coded expert systems or data-mining approaches [10]. The more recent introduction of deep neural networks, and RL-inspired search such as MCTS has led to a qualitative step forward, in particular in how reactions are predicted, as well as how potential steps are prioritized during the search [31, 32]. When combined with conditions and sophisticated forward reaction prediction, these algorithms can also be coupled with robotic synthesis [33]. Early ML-based approaches focused on combining graph-based reaction modeling with ML [31, 33]. Recently, purely seq2seq-approaches have also been presented [34]. Several open source implementations are available [32, 33, 35, 36].

Because they are computationally expensive, requiring at least several seconds for typical lead-like molecules, CASP is usually not used as part

of the VS or DND in-the-loop scoring functions. Rather, they are used to post-process the most interesting molecules from these pipelines. Despite considerable progress in the recent years, and their usefulness as a filtering step to remove clearly problematic molecules, modern CASP tools still have limitations. The reaction models can fail, in particular for less well precedented reactions [31]. It is not fully established whether condition and yield prediction works well enough already[37, 38]. Search algorithms can lead to strategically questionable routes in particular for more complex molecules [31]. Organic chemistry expertise is still required to inspect the results.

7. Fragment- and Synthesis-driven molecular construction and generative models

Molecule construction and generation algorithms in the 1980s and 1990s relied on atom-by-atom construction or manipulation, however, it was recognized early that these methods lead to a large proportion of unsynthesizable as well as unstable molecules [4, 5, 1]. To address this issue, molecules can be constructed from fragments derived from previously synthesized molecules [39, 17]. Here, molecules are fragmented along predefined bonds, for example corresponding to those often build up in chemical reactions [40]. Then, the fragments can be flexibly recombined. Another effective fragment-based approach used is based on matched molecular pairs [41]. While this approach is a major step forward in terms of molecule quality, and often used in practice, it does not come up with proper synthesis routes (see Fig. 5).

A sensible alternative is to conduct virtual reactions (VR) [5]. Here, reaction rules are applied recursively to matching building block molecules in the forward direction, until desired number of steps is reached. Most virtual screening libraries are generated this way. This approach has also been used for *de novo* design. In pioneering work, Vinkers et al. describe an algorithm combining VRs with simulated annealing and search to optimize given scoring functions [42]. Approaches investigating different optimization strategies have been reported as well [43, 44, 45]. Forward VR enumeration can also be combined with retrosynthesis planning to generate focused libraries close to target molecules [46].

The usefulness of molecules generated with VR enumeration hinges on the quality of the underlying reaction rules. In particular, for the simpler molecules required in particular in hit expansion and hit to lead scenarios, virtual screening on such libraries can work well and is now routinely used in

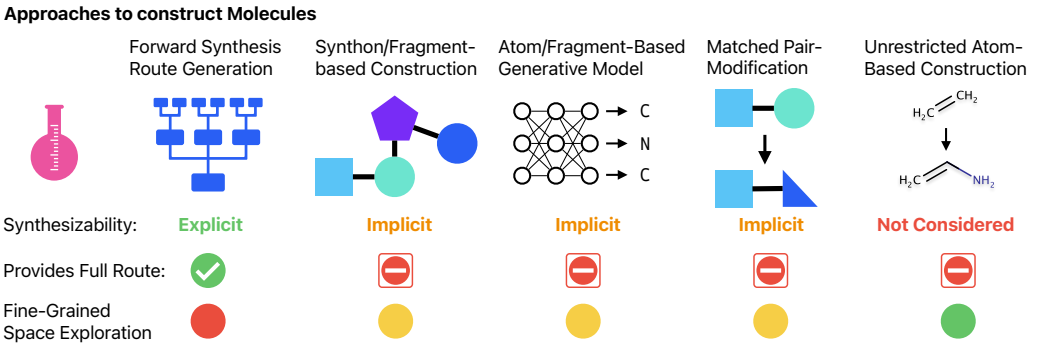

Figure 5: Approaches for Molecule Construction take synthesizability into account differently. Forward synthesis models, which can be based on generative models or also rule-based, apply reactions to building blocks until a target molecule is returned, which allows to generate molecule and their corresponding routes, thus handling synthesizability explicitly. It does not allow for a fine-grained exploration of chemical space. Fragment-based algorithms, which include synthon-based approaches, pretrained generative models, and algorithms based on matched molecular pairs are all based on data-mining of or training on data sets of known, synthesizable models, and thus implicitly capture the synthesizability of the underlying data. This restricts chemical space exploration to some extent. Unrestricted molecule construction, which can combine any atom and any bond, subject to basic valency constraints, allows for a very fine grained exploration of chemical space, but does not consider synthesizability even implicitly. However, depending on the use case, all of these approaches have their place.

academia and industry: VR-enumerated libraries compiled by vendors such as Enamine have been very successfully used in prospective VS campaigns, leading to an ≈80% success rate in synthesis [47, 48, 16, 49, 18]

More recently, generative models for molecules have been suggested as construction algorithms. These refer to a variety of machine learning models, for example autoregressive models, (variational) autoencoders, normalizing flows or diffusion models, which can be trained on target chemical spaces, and then sample new molecules from that space as if one would sample from probability distribution over molecules.[6, 7, 3] Generative models for molecules can be trained on different representations, for example SMILES, molecular graphs, fragment graphs, or 3D molecular structures, and also perform *de novo* design, for example when coupled with fine-tuning or RL algorithms.[6, 50, 51, 52, 53, 54, 55, 15, 3] Despite their simplicity, SMILES-based autoregressive models (also called Chemical Language Models, CLMs), have shown robust performance to generate reasonable and synthesizable

molecules when trained on previously synthesized, drug-like molecules, and have been experimentally validated (see below). However, in particular when used in de novo design, generative models based on atomistic representations can be pushed towards generating unreasonable or hard to synthesize molecules, even though to a lesser degree than previous, unconstrained de novo design algorithms, if the exploration is not carefully restricted.[15, 30]

A promising newer direction are generative models for forward synthesis routes, which similar to the VR approach generate molecules and synthesis routes, leading to much more reasonable structures[56, 57, 58, 45, 59]: Bradshaw et al. proposed generative models that construct multi-step forward reaction routes where the model learns to pick building blocks and intermediates, which are then submitted to a reaction predictor[56, 57]. They demonstrated competitive performance of their algorithm on the Guacamol benchmark[15] compared to less restricted generative models, while maintaining synthesizability. By switching the reaction model to reaction rules, and changes to the tree generation process, Gao et al. recently reported a generative model for reaction routes that does not use a reaction predictor, but can learn to choose which reactions to run. They demonstrated competitive performance to strong baselines on the TDC generative benchmark, while maintaining synthesizability and molecule quality. They also demonstrated that their model could generate synthesizable analogs using conditional molecule generation, as well as forward synthesis planning.[58] However, this approach has not been scaled up yet to more complex rule-bases, thus limiting the size the chemical space the model can explore. Gottipatti et al. proposed a reinforcement rearning-based approach to generate routes in the forward direction, however, their algorithm is limited to generating linear reaction routes, instead of more general tree-shaped routes [60]. Recently, also hybrid algorithms combining atom-based generative models with reaction-driven generation emerged.[61, 59, 62, 63]

Despite this progress, and proven uses for simpler, early stage molecules, current reaction-driven construction algorithms, both enumeration-based and generative models, still have a number of limitations: The predicted routes depend on the quality of single step reaction models or reaction rule-base. Current reaction models, for example the Molecular Transformer, will always predict a product, or even hallucinate completely incorrect structures, even when a reaction is not going to proceed. On the other end, compiling and maintaining high quality rule-bases is time-consuming, and does not scale [31]. The breadth and complexity of chemistry captured by these models is still

limited, leading to a restricted exploration of chemical space. Improvements could be made by incorporating models that can be trained on positive and negative reactions, such as the in-scope filter models introduced by Segler and Coley [31, 33], even though significantly more effort needs to be made to gather the required training data. It is also an open question to what extent reaction-driven generation, which is discrete, limits the exploration of chemical space, making optimization during *de novo* design harder than compared to models which can more flexibly exchange atoms in a molecule.

8. Benchmarking Synthesis-aware generative models

Computational benchmarks are important to drive the progress of computational methods. Measuring enrichment factors in virtual screening, or benchmarks like Guacamol represent sensible benchmarks and have been useful to make progress, however, also here "Goodhart's law" has to be invoked, as they represent only proxies, and still do not capture the full intricacies of molecular discovery. We nevertheless recommend the following steps for benchmarking the approaches described in this work:

1. Can the virtual screening or de-novo design algorithm find molecules that maximize given objectives? Example Benchmarks: Guacamol [15], VS benchmarks
2. Use CASP tools to predict routes for generated molecules or synthesizabilty scores: Gao et al.[30]
3. Are the generated molecules reasonable? Quality filters like in Guacamol[15] can be applied. ML Publications should always contain visualizations of non-cherry picked, random samples of molecules.
4. Synthesis planning algorithms should be evaluated quantitatively and qualitatively.[31]

On the other hand, it is an open question whether small increases on computational benchmarks translate to meaningful improvements in real world drug discovery, where data is often sparse, and projects can operate on quite diverse chemical matter, with distributions shifts, where robustness is key. However, we do not believe experimental validation should ever become a mandatory requirement to benchmark computational methods, as it introduces hard to compare multi-step processes, and not every computational group has the resources for wet-lab validation. We encourage the community to continue to work on refining appropriate benchmarks.

9. Recent Examples of Prospective Validation

Here, we will present recent examples of the successful validation of virtual screening and *de novo* design. Some caution about visibility has to be noted, as these are mostly academic contributions. Applications in industrial drug discovery can usually not be published for several years after a project has been concluded.

In several high-profile works, Virtual Screening using large enumerated on-demand libraries has repeatedly demonstrated its utility, in particular in earlier discovery stages.[48, 16, 49] Seumer et al. used a genetic DND algorithm in conjuction with SAscore in the loop and synthesis planning post-processing to discover a novel organocatalyst [64]. Grisoni and colleagues demonstrated

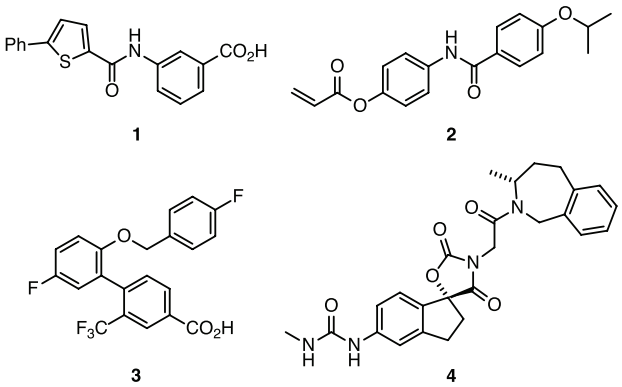

Figure 6: Merk et al synthesized compounds 1-3, demonstrating activity against PPAR and RXR[52, 53]. Yang et al reported activity against p300 for compound 4.[65] All structures were generated with pretrained SMILES-LSTM CLM models.

successful syntheses by combining CLMs with a simplified retrosynthetic analysis based on a small number of robust reaction schemes geared towards lab on a chip technology [51]. Merk and colleagues demonstrated several examples of using CLMs to prospectively design molecules for nuclear receptor targets (see also Fig. 6) [52, 53, 66].

It is noteworthy that despite their simplicity, generative SMILES-based language models[6] using behavioural cloning, fine-tuning, or RL, are currently one of the most successfully validated *de novo* design algorithms in a prospective setting, where actual compounds have been synthesized and tested [52, 53, 65, 51, 54].

These successes indicate that it may be time to update our textbook knowledge on the challenges of synthesis of de novo designed molecules[1].

Despite the importance of these milestone results, currently published prospective examples are still relatively simple molecules, or relatively close analogs of well-explored target classes, such as kinases[67]. This should not be downplayed, given that the generation of close analogues to explore structure-activity-relationships is routine task in medicinal chemistry, and doing this successfully with algorithms is a win for the machine.

We also anticipate that with further use, and the publication lag in industry, we will start to see the results on more ambitious targets and novel chemical spaces in the next couple of years.

10. Conclusion

Digital Medicinal Chemists now have a variety of increasingly robust methods at their disposal, which can be productively used with reasonable effort in the hunt for small molecule ligands and materials. However, to be most useful, current algorithms still require basic knowledge in ML and chemoinformatics, as well as some expertise in organic and medicinal chemistry to run, and should be seen as idea generators or assistants. Also, we currently do not see a single algorithm emerging that works best across different tasks. Often different algorithms work similarly well, which is why we currently recommend practitioners to choose the algorithms they feel most comfortable with, depending on task constraints.

Despite this progress, more work is needed to faithfully predict the outcome of reactions, synthesis routes, synthesis execution instructions, and learn distributions over molecules. Major challenges also still exist the scoring. Setting up multi-objective scoring functions is still somewhat a dark art, and requires trial and error to get right. The accurate prediction of biological properties is still limited both with machine learning, as well as physics-based approaches, and requires further improvements. Also, the community should put more effort in providing unified benchmarking methods. Finally, on an organizational level, to harness automated molecular design, computational methods need to be incorporated further in the discovery process, which requires strong collaboration from different groups of medicinal and computational chemists, machine learning researchers, software engineers, and data scientists.

Nevertheless, an increasing data-first culture in medicinal chemistry, as well as integration with automated experimentation, which will lead to much improved data, will allow reaction-driven algorithms to become even better, or even self-improve. This will allow automated molecular design to play an increasing role in assisting medicinal chemists to discover new drugs to address unmet clinical need.

11. Acknowledgements

We thank Austin Tripp, Krzysztof Maziarz, Richard Lewis, and Nadine Schneider for helpful discussions, and the anonymous reviewers for valuable and constructive feedback.

12. Conflict of Interest

The authors declare no conflict of interest.

References

[1] Klebe, G. *Wirkstoffdesign: Entwurf und Wirkung von Arzneistoffen*; Springer-Verlag, 2009.

[2] Luukkonen, S.; van den Maagdenberg, H. W.; Emmerich, M. T.; van Westen, G. J. Artificial intelligence in multi-objective drug design. *Current Opinion in Structural Biology* **2023**, *79*, 102537.

[3] Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discovery Today* **2021**, *26*, 2707–2715.

[4] Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspectives in Drug Discovery and Design* **1995**, *3*, 34–50.

[5] Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening—an overview. *Drug discovery today* **1998**, *3*, 160–178.

[6] Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*.

[7] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*.

[8] Bjerrum, E. J.; Threlfall, R. Molecular generation with recurrent neural networks (RNNs). *arXiv preprint arXiv:1705.04612* **2017**,

[9] Coley, C. W. Defining and exploring chemical spaces. *Trends in Chemistry* **2021**, *3*, 133–145.

[10] Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H.; Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews* **2020**, *49*, 6154–6168.

[11] Reymond, J.-L. The chemical space project. *Accounts of Chemical Research* **2015**, *48*, 722–730.

[12] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* **2013**, *27*, 675–679.

[13] Walters, W. P. Virtual chemical libraries: miniperspective. *Journal of medicinal chemistry* **2018**, *62*, 1116–1124.

[14] Hartenfeller, M.; Schneider, G. Enabling future drug discovery by de novo design. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 742–759.

[15] **Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096 – 1108, Provides a standard benchmark for de novo design algorithms.

[16] *Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F., et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452–459.

[17] Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 497–520.

[18] Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M. Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *Journal of Chemical Information and Modeling* **2022**, *62*, 553–566.

[19] Lyu, J.; Irwin, J. J.; Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology* **2023**, 1–7.

[20] Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies* **2019**, *32*, 55–63.

[21] **Langevin, M.; Vuilleumier, R.; Bianciotto, M. Explaining and avoiding failure modes in goal-directed generation of small molecules. *Journal of Cheminformatics* **2022**, *14*, 1–13, Key investigation of the interplay of scoring functions and de novo design algorithms, showing how to mitigate scoring function exploitation.

[22] Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput. Aided Mol. Des.* **2007**, *21*.

[23] Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768–2777.

[24] Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **2009**, *1*.

[25] Coley, C. W.; Rogers, L.; Green, W.; Jensen, K. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58 2*.

[26] Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. *Journal of Cheminformatics* **2023**, *15*, 6.

[27] Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminformatics* **2020**, *12*.

[28] **Liu, C.-H.; Korablyov, M.; Jastrzebski, S.; Włodarczyk-Pruszynski, P.; Bengio, Y.; Segler, M. RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software. *Journal of Chemical Information and Modeling* **2022**, *62*, 2293–2300, Regression model–based synthesizability score trained on the output of a synthesis planning algorithm.

[29] **Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore)–rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical science* **2021**, *12*, 3339–3349.

[30] **Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, Key investagtion of how to integrate synthesizability into the de novo design process. Also first modern analysis of how some generative algorithms have issues with synthesizablity.

[31] Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*.

[32] **Liu, G.; Xue, D.; Xie, S.; Xia, Y.; Tripp, A.; Maziarz, K.; Segler, M.; Qin, T.; Zhang, Z.; Liu, T.-Y. Retrosynthetic Planning with Dual Value Networks. *arXiv preprint arXiv:2301.13755* **2023**, Current state of the art synthesis planning algorithm, at the time of writing.

[33] Coley, C. W.; Thomas, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*.

[34] Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020**, *11*, 3316–3325.

[35] Tripp, A.; Maziarz, K.; Lewis, S.; Liu, G.; Segler, M. Re-Evaluating Chemical Synthesis Planning Algorithms. NeurIPS 2022 AI for Science: Progress and Promises. 2022.

[36] Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics* **2020**, *12*, 70.

[37] Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS central science* **2018**, *4*, 1465–1476.

[38] Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2021**, *2*, 015016.

[39] Schneider, G.; Lee, M.-L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *Journal of computer-aided molecular design* **2000**, *14*, 487–494.

[40] Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A. Synthi: a new open-source tool for synthon-based library design. *Journal of Chemical Information and Modeling* **2021**, *62*, 2151–2163.

[41] Polishchuk, P. CReM: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics* **2020**, *12*, 1–18.

[42] Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. SYNOPSIS: SYNthesize and OPtimize system in silico. *Journal of medicinal chemistry* **2003**, *46*, 2765–2773.

[43] Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comp. Biol.* **2012**, *8*, e1002380.

[44] Spiegel, J. O.; Durrant, J. D. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *Journal of cheminformatics* **2020**, *12*, 1–16.

[45] Makara, G. M.; Kovács, L.; Szabó, I.; Pőcze, G. Derivatization design of synthetically accessible space for optimization: in silico synthesis vs deep generative design. *ACS Medicinal Chemistry Letters* **2021**, *12*, 185–194.

[46] *Dolfus, U.; Briem, H.; Rarey, M. Synthesis-Aware Generation of Structural Analogues. *Journal of Chemical Information and Modeling* **2022**, *62*, 3565–3576, Using Computer–Aided Synthesis Planning to generate easy to synthesize libraries.

[47] Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling* **2020**, *60*, 6065–6073.

[48] Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K., et al. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224–229.

[49] *Kaplan, A. L.; Confair, D. N.; Kim, K.; Barros-Álvarez, X.; Rodriguiz, R. M.; Yang, Y.; Kweon, O. S.; Che, T.; McCorvy, J. D.; Kamber, D. N., et al. Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity. *Nature* **2022**, *610*, 582–591, Demonstration of Large Scale Docking for ligand discovery.

[50] Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *Journal of chemical information and modeling* **2020**, *60*, 5918–5922.

[51] Grisoni, F.; Huisman, B. J.; Button, A. L.; Moret, M.; Atz, K.; Merk, D.; Schneider, G. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Science Advances* **2021**, *7*, eabg3338.

[52] Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **2018**, *37*, 1700153.

[53] Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Nat. Commun. Chem.* **2018**, *1*, 68.

[54] Moret, M.; Pachon Angona, I.; Cotos, L.; Yan, S.; Atz, K.; Brunner, C.; Baumgartner, M.; Grisoni, F.; Schneider, G. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications* **2023**, *14*, 114.

[55] Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology* **2023**, *79*, 102527.

[56] Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M.; Hernández-Lobato, J. M. A Model to Search for Synthesizable Molecules. Advances in Neural Information Processing Systems 32. 2019.

[57] **Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M.; Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis DAGs. Advances in Neural Information Processing Systems 33. 2020.

[58] **Gao, W.; Mercado, R.; Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. *arXiv preprint arXiv:2110.06389* **2021**, Currently state of the art synthesis–tree based generative model for molecules.

[59] Hua, Y.; Fang, X.; Xing, G.; Xu, Y.; Liang, L.; Deng, C.; Dai, X.; Liu, H.; Lu, T.; Zhang, Y., et al. Effective reaction-based de novo strategy for kinase targets: a case study on MERTK inhibitors. *Journal of Chemical Information and Modeling* **2022**, *62*, 1654–1668.

[60] Gottipati, S. K.; Sattarov, B.; Niu, S.; Pathak, Y.; Wei, H.; Liu, S.; Thomas, K. M. J.; Blackburn, S.; Coley, C. W.; Tang, J.; Chandar, S.; Bengio, Y. Learning To Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning. International Conference on Machine Learning. 2020.

[61] Fialková, V.; Zhao, J.; Papadopoulos, K.; Engkvist, O.; Bjerrum, E. J.; Kogej, T.; Patronov, A. LibINVENT: reaction-based generative scaffold

decoration for in silico library design. *Journal of Chemical Information and Modeling* **2021**, *62*, 2046–2063.

[62] Nguyen, D. H.; Tsuda, K. Generating reaction trees with cascaded variational autoencoders. *The Journal of Chemical Physics* **2022**, *156*, 044117.

[63] Seo, S.; Lim, J.; Kim, W. Y. Molecular Generative Model via Retrosynthetically Prepared Chemical Building Block Assembly. *Advanced Science* **2023**, 2206674.

[64] Seumer, J.; Hansen, J. K. S.; Brøndsted Nielsen, M.; Jensen, J. H. Computational evolution of new catalysts for the Morita–Baylis–Hillman reaction. *Angewandte Chemie International Edition* **2022**, e202218565.

[65] Yang, Y.; Zhang, R.; Li, Z.; Mei, L.; Wan, S.; Ding, H.; Chen, Z.; Xing, J.; Feng, H.; Han, J., et al. Discovery of Highly Potent, Selective, and Orally Efficacious p300/CBP Histone Acetyltransferases Inhibitors. *J. Med. Chem.* **2020**,

[66] **Ballarotto, M.; Willems, S.; Stiller, T.; Nawa, F.; Marschner, J. A.; Grisoni, F.; Merk, D. De Novo Design of Nurr1 Agonists via Fragment-Augmented Generative Deep Learning in Low-Data Regime. *Journal of Medicinal Chemistry* **2023**, Prospective Use of Generative Models to generate active ligands.

[67] Walters, W. P.; Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nature biotechnology* **2020**, *38*, 143–145.