Defining and Exploring Chemical Spaces

Citation: Coley, Connor W. "Defining and Exploring Chemical Spaces." Trends in Chemistry 3, 2 (February 2021): 133-145. © 2020 Elsevier Inc.

As Published: http://dx.doi.org/10.1016/j.trechm.2020.11.004

Publisher: Elsevier BV

Persistent URL: https://hdl.handle.net/1721.1/131238

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Defining and Exploring Chemical Spaces

Connor W. Coley

MIT Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

ccoley@mit.edu

Abstract

Designing functional molecules with desirable properties is often a challenging, multi-objective optimization. For decades, there have been computational approaches to facilitate this process through the simulation of physical processes, the prediction of molecular properties using structure-property relationships, and the selection or generation of molecular structures. This piece provides an overview of some algorithmic approaches to defining and exploring chemical spaces that have the potential to operationalize the process of molecular discovery. We emphasize the potential roles of machine learning and consideration of synthetic feasibility, which is a prerequisite to "closing the loop". We conclude by summarizing important directions for the future development and evaluation of these methods.

Keywords: molecular discovery, machine learning, optimization, de novo design

Highlights

- Virtual libraries used in molecular discovery are often too large to exhaustively evaluate, warranting the use of algorithms to help with exploration.
- Algorithmic approaches like Bayesian optimization can help efficiently navigate pre-defined chemical spaces in combination with surrogate models.
- On-the-fly molecular generation during exploration enables even larger chemical spaces to be searched, including deep learning-based models, although their chemical spaces are defined only implicitly.
- Emerging approaches to incorporate reactions into machine learning-based generation can ensure molecules are able to be synthesized, similar to preceding algorithms for reaction-based *de novo* design.

Glossary

- **Active learning**: An approach to data acquisition whereby a model iteratively proposes new experiments to perform, i.e., new data points to label.
- **Deep generative model**: A highly-parameterized machine learning model capable of generating structured objects, e.g., strings or graphs.
- **Graph theory**: The study of graph-structured objects comprising nodes and edges.
- **On-the-fly**: Processing or generating data in an ongoing manner as needed, in contrast to preprocessing or enumeration.
- **QSPR model**: A quantitative structure-property relationship model that is able to predict, with some degree of accuracy and reliability, the functions of molecules based on their chemical structure.
- **Reinforcement learning**: A machine learning method whereby an agent learns to optimize its behavior and interaction with an environment.
- **Synthesizability-aware model**: A model that is able to distinguish between chemical structures on the basis of how easily they can be synthesized experimentally.
- **Virtual library**: A digital collection of chemical structures or their representations, as is used in computational screening.

1. Conceptualizing chemical space

Chemical space can be thought of as the set of all possible molecules or materials. We generally consider more narrowly defined chemical spaces that are defined or constrained by the structures or functions of the molecules they contain. For example, "drug-like chemical space" can be used in the context of drug discovery in an attempt to quantify the vast number of molecules that have physical properties similar to those of existing small molecule therapeutics. While quantifying the size of a chemical is rarely useful, it should be noted that there are far more organic molecules thought to be stable than atoms in the solar system, which is unsurprising given the combinatorics of designing molecular graphs. Here, we focus our discussion on small molecules, rather than periodic materials, biomolecules, and polymers, all of which correspond to distinct "chemical spaces". There have been many studies attempting to estimate the size of different chemical spaces [1, 2, 3] and suggest rules for organizing these spaces along important functional axes to improve its visualization and navigability [4, 5, 6, 7].

As we have previously described, the discovery of novel molecules can be framed as a search within chemical space [8, 9]. The goal, often, is to identify which molecule(s) exhibits a set of desirable properties. Besides defining these desirable properties and a strategy to evaluate candidate molecules, the two primary considerations one must make are (1) how to define the space and (2) how to explore the space. Both contribute to the search efficiency and likelihood of finding a good candidate. These two aspects are not decided independently: if you are repurposing FDA-approved drugs, your chemical space is narrow enough that an exhaustive screen may be feasible, but if you have no such restriction, you must employ some strategy to select which molecules to test. These strategies are typically iterative optimization routines (driven by human intuition or driven by quantitative experimental design) with varying degrees of sophistication, as will be discussed later. Navigating chemical space has been extensively written on in the context of (non-algorithmic) drug design [10, 11].

The number of candidate molecules is too large to explore exhaustively, so one often imposes constraints on chemical space depending on the search strategy, application, and practical limitations of cost and time. These constraints can look quite different when candidates are evaluated by physical experiments rather than by computational experiments. In the former case, acquiring new information about the performance of a molecule requires

physically realizing (i.e., synthesizing, purifying, and characterizing) it; considerations of synthesis cost and material availability are paramount. In the latter case, one may be able to postpone these practical considerations until after computational evaluations have identified a putative "optimal" molecule. To bound the computational cost, the search space is still restricted using human expertise or some "prior" on what would make a viable candidate.

This piece examines strategies to define and explore chemical spaces with an emphasis on the role of machine learning and synthetic chemistry constraints (Table 1, Key Table). While this can be performed subconsciously by subject matter experts (e.g., medicinal chemists) in the absence of computer assistance, formalizing these concepts may eventually enable autonomous workflows to produce truly novel, useful outcomes with reduced reliance on human intuition and subjectivity. Elements of concepts we will cover can be found in previous articles, including a recent overview by Lemonick [12]. We do not address visualization and instead refer readers to the work of Reymond and coworkers [5, 7].

2. Defining and exploring enumerated chemical spaces

One approach to molecular discovery is exploring a pre-defined chemical space: an enumerated list of candidate molecules. In this setting, the two stages of (1) defining the space and (2) exploring the space are entirely decoupled. Formally, we might think about this problem as an optimization of an objective function $f(x)$, where x is a molecule belonging to a discrete set $\mathcal{X}$.

2.1. Defining finite chemical spaces

Defining or selecting a finite chemical space often relies on domain expertise. Careful selection of $\mathcal{X}$ can increase the likelihood that it contains a high-performing molecule while minimizing the number of low-performing compounds. Common databases of molecules for computational screening are ZINC [13], a library of commercially-available compounds; PubChem [15], molecules with biological relevance; ChEMBL [14], molecules with bioactivity data; and DrugBank [17], approved or experimental therapeutic molecules. These virtual libraries all represent "general-purpose" chemical spaces with broad biological relevance, and are therefore applied to many problems related to drug discovery [41].

	Unconstrained	Constrained
Pre-defined	ZINC [13], ChEMBL [14], PubChem [15], GDB [16]	DrugBank [17], Enamine REAL [18], WuXi Virtual Library [19], SAVI [20], PGVL [21], PLC [22]
On-the-fly via heuristic methods	Fragment-based GAs [23], GroupBuild [24], BREED [25], GraphGA [26], GEGL [27]	SYNOPSIS [28], Flux [29], MOARF [30], DOGS [31]
On-the-fly via machine learning	SMILES VAE [32], JT-VAE [33], SMILES RNN [34, 35], MolDQN [36]	MoleculeChef [37], ChemBO [38], PGFS [39], REACTOR [40]

Table 1: Categorization of approaches to defining chemical spaces for molecular discovery and an incomplete set of examples for each. Spaces can be defined prior to exploration or defined on-the-fly by evolutionary and/or machine learning-based methods. They can be relatively unconstrained (i.e., only in terms of validity), or constrained by the availability (i.e., in terms of purchaseability or synthseizability).

More focused chemical spaces can be created through a domain-informed enumeration of compounds relevant to a specific application. For example, 1.6M donor-bridge-acceptor trimers for organic electronics [42] or 2.8M transition metal complexes for redox flow batteries [43]. These are exhaustively enumerated chemical spaces with strict constraints on which fragments are included and how they are attached, similar to R-group enumeration methods. Privileged fragments for drug-like molecules have been identified through retrosynthetic analysis and automatic fragmentation [44, 45]; the molecules produced by recombining these fragments are intended to look more realistic than an enumeration based on graph structure alone.

Graph-theoretical enumeration of molecular structures has been studied for over a century, starting with simple spaces like that of acyclic alkanes [46, 47]. However, it is only recently that these structures have been recorded, evaluated, and used for discovery. The Chemical Space Project exemplifies modern exhaustive enumeration of all stable molecules containing the most common atom types in organic molecules up to a certain size [16]. Since the original Generated DataBase (GDB) of up to seven heavy atoms [48], Reymond and coworkers have enumerated, analyzed, and released the 166.4 billion structures of up to 17 heavy atoms [49] and published numerous visualizations and analyses thereof.

In addition to the benefits of ensuring $\mathcal{X}$ is relevant to the design objective, pre-defining chemical spaces lets us impose arbitrary constraints on their contents. A practical constraint is the ease of experimental validation. That is, ensuring that any candidate encountered in our chemical space can be physically acquired for experimental testing. In the simplest case, a chemical space can be defined as the set of molecules available in a company's chemical inventory or the set of molecules in-stock at a chemical vendor. Any compound from this list can be acquired rapidly for experimental evaluation.

Accessibility is the primary motivation for make-on-demand libraries, which are chemical spaces defined as the set of molecules that are in-stock or available *and* all molecules that can be produced from those structures through straightforward synthetic protocols. Libraries are often enumerated by applying a small number (< 100) of reaction templates defining common single-step transformations to all possible combinations of starting materials [50, 51, 52, 53] (Figure 1); recursive enumeration generates molecules accessible through multiple synthetic steps. There are numerous implementations of this approach [54] including SAVI [20], efforts within pharmaceutical companies [21, 22], and efforts from commercial vendors [18, 19]. As it becomes

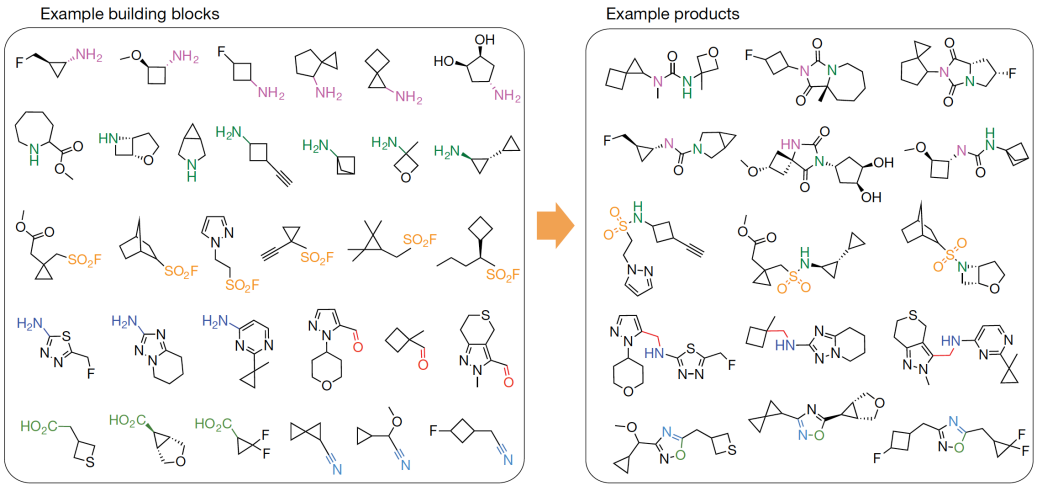

Figure 1: Representative building blocks and enumeated products for the Enamine REAL make-on-demand library [18]. Modular building blocks are combined according to expert-defined reaction rules describing common transformations. Figure reproduced from Lyu et al. [55].

impractical to store such large numbers of compounds due to the combinatorial explosion of reaction products, these spaces may be defined only implicitly.

Whether molecules in these spaces are actually easy to synthesize depends on the robustness of the rules used for enumeration. Lyu et al. cite an 86% synthesis success rate for 51 compounds selected from 170 million in the Enamine REAL library enumerated from 130 reaction types; WuXi estimates a 60-80% success rate for their 1.7B-member collection generated by 30 reaction types [19]. There is an opportunity to improve this success rate through the use of machine learning models for reaction outcome prediction [56, 57], which for common reaction types exhibit accuracies well above 90% on benchmark datasets. These neural models can be directly used to enumerate possible products or used to predict regio/stereoselectivity patterns [58, 59, 60].

2.2. Exploring finite chemical spaces

Once these chemical spaces are defined, there are several approaches for identifying the top-performing molecules within them. The simplest strategy is of course to computationally or experimentally evaluate every candidate molecule exhaustively. The feasibility of this approach depends on the nature

of evaluation and time/cost constraints. It would not be practical to physically test every compound in a database such as ZINC, but it could be for smaller collections like the Drug Repurposing Hub [61] or the NCATS Pharmaceutical Collection [62]. It is worth noting that technologies like DNA-encoded libraries [63] and phage display [64] can be used to physically screen chemical spaces of trillions of molecules, albeit with a sparse and stochastic readout.

If evaluation is computational, practicality is simply a question of computational budget. In one of the largest docking studies reported to date, 138 million and 99 million compounds from the Enamine REAL library were docked against the D_4 receptor and AmpC, respectively [55]; this was enabled by a fast computational pipeline requiring only one second per library compound. More recent studies have since screened over one billion enumerated molecules from the same database [65, 66]. As several make-on-demand libraries exceed this scale by multiple orders of magnitude, we argue that such exhaustive screening techniques are not a viable long-term approach even for inexpensive evaluations like docking.

A popular framework to reduce the overall cost is active learning through iterative, model-guided optimization [67]. This involves selecting subsets of experiments to perform based on predictions from a quantitative structure-property relationship (QSPR) model—a surrogate model, $\hat{f}(x)$—that codifies an approximation to $f(x)$. In Bayesian optimization, predictions both of performance and of model uncertainty are both considered to balance exploration of uncertain candidates and exploitation of candidates likely to be high-performing [68]; simpler optimization schemes may simply perform a greedy search. Examples of this paradigm include the platform Eve for the experimental identification of bioactive molecules [69], a retrospective identification of bioactive compounds using PubChem data [70], computational screening of OLED-relevant molecules through batched greedy optimization [42], and machine learning-guided selection of compounds for docking [71]. There are still many limitations to be addressed related to the surrogate model, $\hat{f}$, in terms of its low-data performance, generalization power, and ability to quantify uncertainty. The development of QSPR models is an area of active research [72], with many emerging methods for learning from graph-structured molecular candidates [73]. Additional settings where algorithmic improvements will be beneficial are those with variable evaluation costs (e.g., the cost of purchasing a compound to physically test it) and those where multiple experiments are run in batches (e.g., parallelized in well plates or over

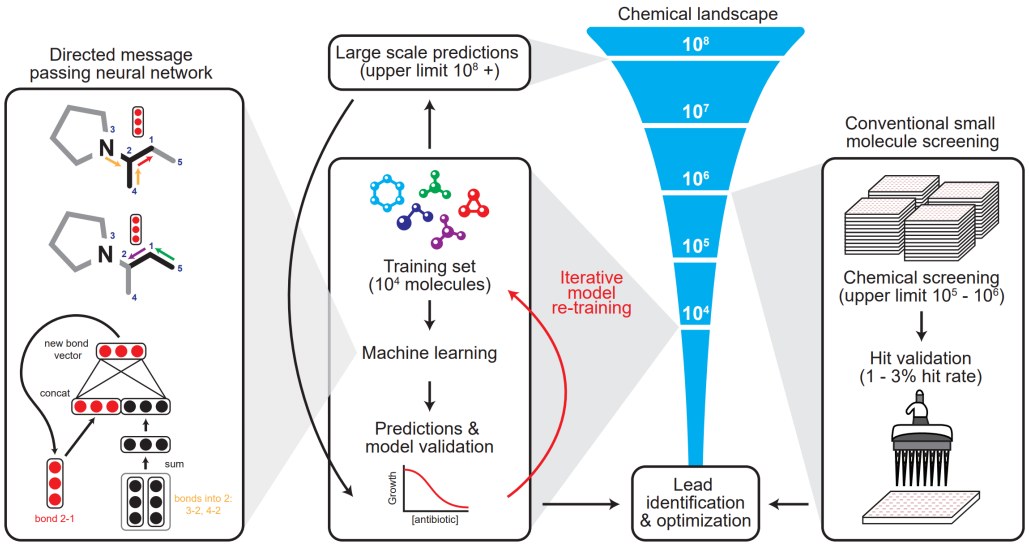

Figure 2: A workflow for machine learning-augmented antibiotic discovery. A pre-defined virtual library (ZINC15) is evaluated using a surrogate machine learning model trained on a set of $\approx 10^4$ experimental measurements to prioritize compounds to test. Figure reproduced from Stokes et al. [74].

multiple CPUs).

Surrogate models can help explore pre-defined chemical spaces outside of iterative active learning by providing an inexpensive approximation to $f(x)$. Identifying the optimal molecules to test at each iteration is equivalent to exhaustive screening using $\hat{f}(x)$in place of the true design objective. While multiple iterations lead to improved surrogate models, a one-iteration approach to finding an optimal molecule can still be very effective. This approach was recently used to identify a novel antibiotic from a drug repurposing collection with fewer experiments than an exhaustive screen [74] (Figure 2); a similar one- and few-iteration screen was also used to identify kinase inhibitors, including an essential Mtb kinase [75]. Both studies used machine learning models as their surrogates—a directed message passing network and a Gaussian process using compound representations from unsupervised learning, respectively.

3. Defining and exploring chemical spaces on-the-fly

If we are not interested in exploring a chemical space exhaustively, we may not need to enumerate it upfront. Implicitly-defined virtual libraries exceed trillions of molecules, and even the vaguest estimates of the size of biologically-relevant chemical space (10^{20}-10^{60}) are clearly too large to enumerate. Even when using Bayesian optimization, selecting "optimal" experiments from billions or trillions of molecules requires an equal number of surrogate model predictions; this constrains the size of pre-enumerated chemical spaces one can consider with a fixed computational budget. Instead, we perform on-the-fly generation and exploration simultaneously.

3.1. Genetic algorithms

One popular class of techniques for finding optimal molecules in an implicitly-defined chemical space is genetic algorithms (GAs). GAs are model- and derivative-free optimization routines that "evolve" candidate solutions based on their performance and those of other candidates through stochastic mutation and crossover events. They have a long history of use in cheminformatics [76] including molecular design [23]. In an exemplary study, Venkatasubramanian et al. define chemistry-informed operators that allow for two parent molecules to crossover (i.e., AB and $A'B'$ yields AB' and $A'B$), two to be merged, one to randomly permute its backbone or side chains, and one to have functional groups inserted, removed, or translocated [23].

Subsequent studies refined this strategy of generating novel molecular structures by applying mutation operators [25] on molecular graphs or string representations, in some cases demonstrating interpolation between two known structures [77, 78], exploration of chemical space around a known active compound [79], and adaptation to atom- and bond-level mutations in combination with a Monte Carlo tree search strategy [26], reinforcement learning [27], or a neural network to improve diversity [80]. As discussed by Jensen and coworkers, a relatively small number of generations are required to transmute any one molecule into another [81], making them remarkably efficient for exploration. These GAs are closely related to fragment-based design, which generates molecular structures piece-by-piece through addition operations alone, e.g., [24].

The initial pool of starting materials and the set of mutation operators defines the chemical spaces that can be accessed by the algorithm: for example,

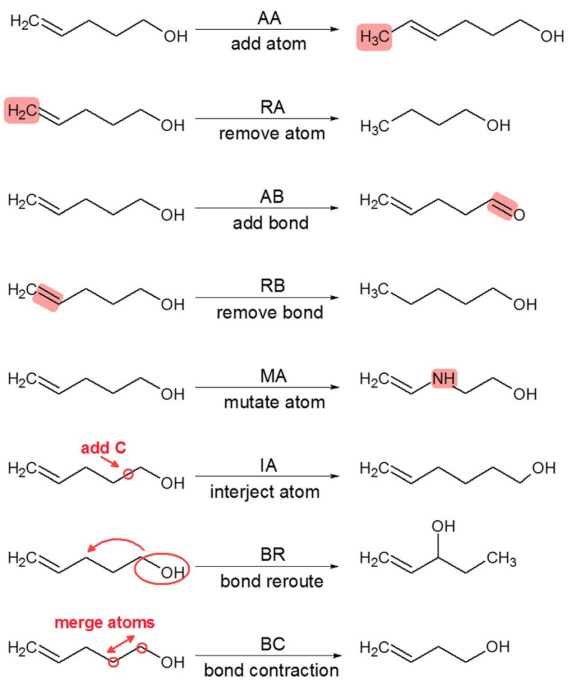

Figure 3: Mutation operators used in the genetic algorithm-based Molpher. Atom- and bond-level modifications to previously-generated structures yield new structures as candidates for evaluation. Over the course of many GA generations, a combinatorially-large chemical space can be accessed. Figure reproduced from Hoksza et al. [78].

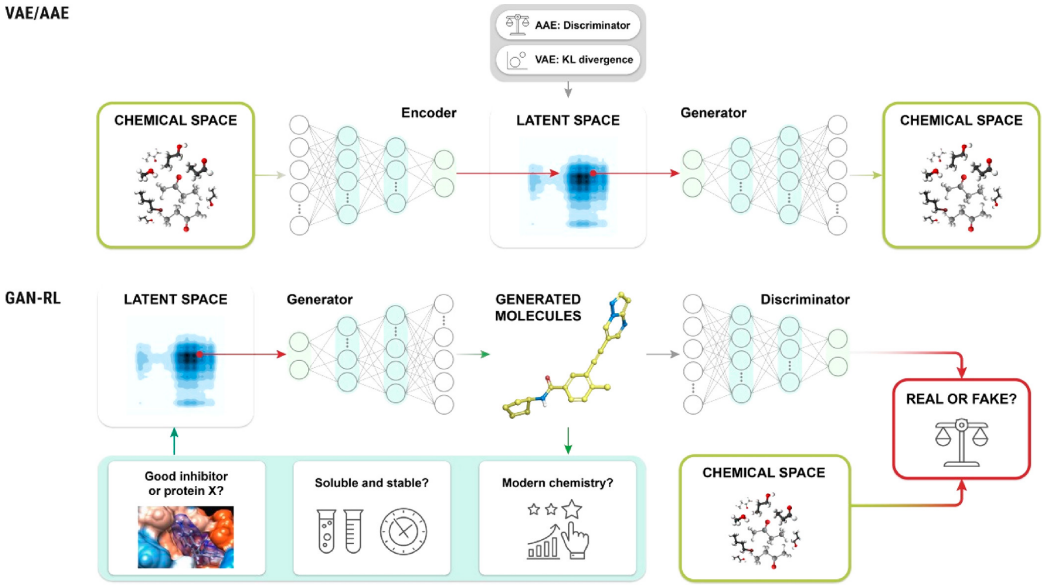

Figure 4: Overview of some deep learning architectures used for molecular generation. Exploration and optimization can occur through navigation in the numerical latent space. VAE: variational autoencoder; AAE: advertsarial autoencoder; GAN: generative-adversarial network; RL: reinforcement learning. Figure reproduced from Vanhaelen et al. [86].

a combinatorial chemical space of 230 billion septa-substituted dihydroazulenes [82]. Because we can perform arbitrarily many addition operations, chemical spaces can become astronomically large. It is not terribly instructive to try to quantify their size in these cases, since the *relevant* chemical space (surrounding the optimal solution) may be a small fraction of its theoretical size.

3.2. Deep generative models

Deep generative models likewise maintain an implicit definition of chemical space and have shown tremendous promise for molecular design as reviewed elsewhere [83, 84, 85, 86] (Figure 4). Prototypical frameworks for generation include variational autoencoders (VAEs) operating on SMILES string representations [32], recurrent neural networks also operating on SMILES string [34, 35, 87], VAEs operating on molecular graphs or junction trees [33, 88], reinforcement learning (RL) agents composing graphs atom-by-atom [36, 89, 90], and many adaptations and improvements thereof.

When applied to discovery, these approaches explore chemical space by biasing generation to candidates that are high-performing. These models are usually pretrained on enumerated chemical spaces to learn basic principles of molecular generation, chemical validity, and what "typical" molecules look like. For autoencoder-based methods that encode/decode molecular structures to/from numerical representations, numerical latent space optimization (LSO) can be used to optimize the molecules that latent vectors correspond to; this is typically done with a fixed decoder network, but the decoder can also be updated on-the-fly [91]. For reinforcement learning methods where an agent learns a policy for generating molecules through a sequence of actions, the value of $f(x)$ can be treated as a reward to update the agent's behavior directly.

While there are hundreds of publications developing and evaluating these methods on computational benchmarks (e.g., [92]), experimental validations of deep generative model predictions are rare. A few exceptions are worth noting. Polykovskiy, Zhebrak, Vetrov, et al. used a generative model to propose 300,000 molecules as potential JAK3 kinase inhibitors, which were filtered to 5000 using docking, clustering, and medchem filters, filtered again to 100 using molecular dynamics simulations, and finally *one* molecule was hand selected and validated [93]. A second study from the same company filtered 30,000 generated structures down to 40, then selected six to test on the basis of synthesizability [94]. In an application to materials discovery, Sumita et al. used RL-based generation and DFT calculations to propose 3200 molecules with targeted maximum absorption wavelengths, of which 86 passed DFT evaluation and six with known synthetic routes were synthesized [95].

The chemical spaces deep generative models explore is defined by the structures they are able to propose as SMILES strings, molecular graphs, or otherwise. Autoregressive generation limits molecules' sizes (e.g., number of SMILES tokens or atom-addition actions) but still produce a massive chemical space. The initial chemical space used for training and any natural inductive biases of the models influences the structures proposed. Benchmarks for the "distribution learning" setting where we are not trying to optimize $f(x)$ show that these models are very effective at proposing novel molecules that exhibit similar properties to their training set [96].

4. Defining and exploring synthetically-constrained chemical spaces on-the-fly

Part of why experimental validation is missing from many computer-aided molecular design studies, particularly showcasing deep generative models, is the expense of physical experiments. Molecules sampled from chemical spaces defined on-the-fly may be challenging, time-consuming, and/or costly to synthesize and evaluate. We previously demonstrated the severity of this problem for generative models using a data-driven retrosynthetic planning tool to assess synthesizability [97]. The status quo is to perform a *post hoc* filtering of molecules according to manual assessment. Here, we discuss a category of techniques for *synthetically-aware* definition and exploration of chemical spaces on-the-fly by incorporating explicit building block and synthetic constraints (Figure 1g of [97]).

4.1. Synthetically-aware genetic algorithms

Synthetic constraints can be incorporated into GAs by restricting allowable mutation operations or by taking advantage of the structured nature of a synthetic route. Methods for the former case include deriving "chemically reasonable mutations" from synthetically accessible compounds [98] and defining retrosynthesis-inspired rules for fragmenting molecules [29, 30, 99]. For the latter, a common approach is reaction-based *de novo* design, where candidates are generated through expert-encoded reaction templates and evaluated (i.e., an on-the-fly equivalent of make-on-demand libraries) [28, 31, 100, 101]. These approaches tend to use model-free algorithms for optimization like GAs to make incremental changes to molecular structures. An early example of synthetically-aware molecular optimization through a GA is Weber et al.'s 1995 experimental optimization of thrombin inhibitors that explored a 160,000-member chemical space implicitly defined by 10 isocyanides, 40 aldehydes, 10 amines, 40 carboxylic acids, and an Ugi-type reaction template [102].

4.2. Synthetically-aware deep generative models

Generative models operating on SMILES tokens or individual atoms are especially prone to generating synthetically-challenging structures. In the past two years, however, methods have emerged that integrate concepts from reaction-based *de novo* design with machine learning-based generation. These approaches impose constraints on the process of generating candidates

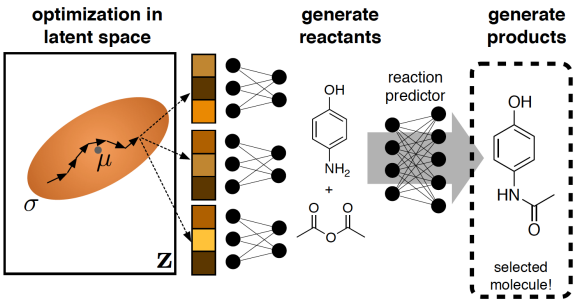

Figure 5: The MoleculeChef model for synthetically-constrained molecular optimization. Chemical space is explored by navigating the latent space of a multi-reactant VAE, generating reactant products from a list of commercially-available starting materials, and running those reactants through a reaction prediction model. Figure reproduced from Bradshaw et al. [37].

directly by reframing molecular optimization as the optimization of building block and/or reaction type selection. For VAEs or generative adversarial networks, this requires modifying the decoder to produce multiple reactant molecules and/or reaction steps; for RL approaches, this requires changing the action space to the selection of reactants and reaction types or conditions.

A naive approach to attempt to constrain molecules generated by a VAE would be to impose constraints on the latent space itself and restrict which points LSO is able to visit. While MoleculeChef uses a VAE, its VAE operates on *sets* of molecules rather than individual molecules [37]. These sets are proposed by a recurrent decoder that selects available starting materials from a discrete list of options. A SMILES-based reaction predictor [57] anticipates the product that would be formed in a reaction between molecules in the set after one synthetic step; the properties of this product are what the model optimizes for (Figure 5). ChemBO does not learn a latent space over reactant sets, but instead uses a graph-based reaction predictor [56] to perform a random walk on a synthesis graph by sampling starting materials and simulating reaction outcomes [38]. The product predicted to be most optimal according to a surrogate model $\hat{f}(x)$ is selected for full evaluation by $f(x)$. This approach is similar to Bayesian optimization within an enumerated chemical space, but the space is continually growing through on-the-fly reaction prediction at each iteration.

Two other studies, released within one week of each other, formulate chemical space exploration as a Markov decision process (MDP)—a sequence

of actions that correspond to reaction steps in a linear synthesis—and train RL agents to learn a policy that yields optimal products [39, 40]. Both begin their optimizations with a random starting material and select from a list of expert-defined reaction templates. The first, PGFS, also selects the other reactant required for bimolecular reactions [39], while the second, REACTOR, enumerates all possible second reactants and chooses the one with the highest reward [40]. As training progresses and more products are evaluated, the agent's policies (parameterized as neural networks) are updated to reward strategies that yield higher-performing molecules.

The chemical spaces these synthetically-constrained approaches can cover are the same as make-on-demand libraries when both are extended to multistep synthetic sequences. But because there is no pre-enumeration of compounds, there is no need to impose a strict limit on the number of allowable reaction steps. The reliability of synthetic pathways (i.e., the actual synthesizability of molecules they propose) comes with the same caveats as make-on-demand libraries, so improving machine learning models for outcome prediction and/or pathway-level evaluation will improve their robustness.

5. Concluding Remarks

Underlying much of our discussion around defining and exploring chemical spaces has been the desire to efficiently navigate chemical spaces to avoid exhaustive screening. We also have slightly biased our discussion towards discovery workflows that ultimately involve *experimental* validation. The approach one should select for chemical space exploration will depend on the nature of the evaluation function, $f(x)$, the extent to which domain expertise can narrow down the problem-relevant chemical space, and the time/cost budget. There are several unresolved questions and factors to be considered when further developing these techniques (see Outstanding Questions).

5.1. Closed-loop physical experimentation and robotic laboratories

The integration of computational experimental design algorithms and robotic laboratories—"closing the loop"—is frequently discussed as a paradigm for accelerated scientific discovery. However, the ability of automated platforms to make truly novel and impactful discoveries has been limited by their inability to generate and test hypotheses without human intervention [8, 9]. In the context of chemical space exploration and molecular discovery, experimental testing requires that proposed candidates are able to be physically

Outstanding Questions

- What are appropriate ways to quantify the size and diversity of a chemical space? Are these useful measures that correlate in some way to the success of molecular discovery efforts?
- When is it better to use a model-free optimization strategy (e.g., genetic algorithms) compared to a model-based optimization strategy (e.g., Bayesian optimization)?
- What computational benchmarks can be used to compare different exploration algorithms that reflect the true complexity of molecular design and optimization? Current objectives are dominated by heuristics and do not explicitly penalize sample inefficiency.
- Can generative algorithms be modified to enable the precise specification of stereoisomers or other configurational isomers when designing ligands for synthesis or protein binding?
- How can considerations of cost and synthetic feasibility be incorporated into the selection of optimal experiments? These are essential factors when designing algorithms to control closed-loop, automated experimental platforms.

obtained: for manual evaluation, that they be purchseable or synthesizable from commercial starting materials; for automated evaluation, that they be available or synthesizable from an on-hand chemical inventory given experimental constraints.

If we have access to a very large chemical library without any on-demand synthesis capabilities, an automated platform must use an algorithm for exploring the pre-defined chemical space this library represents (e.g., a high throughput screening facility [103]). If we only have the ability to perform simple one-step chemistries (e.g., [104]), it likely makes sense to exhaustively enumerate this space and use Bayesian optimization to select from the list of candidates. Platforms for automated multi-step synthesis with intermediate purification are still in the proof-of-concept phase [105, 106, 107], but would theoretically have access to a much larger chemical space than what can be practically enumerated. In this case, it makes sense to use an approach that defines the chemical space on-the-fly constrained by the starting materials and reactions compatible with the platform.

5.2. Considerations beyond synthesizability for molecular design

The *number* of experiments has been the predominant measure of cost for optimal experimental design algorithms. Factors considered by humans when determining what molecules to test are of course far more complex than the number of unique molecules (and a binary assessment of their synthesizability). As computer-aided synthesis planning and predictive chemistry tools become increasingly sophisticated, detailed considerations of synthesis time, ease of parallelization, and utilization of common intermediates might be able to be factored into batched molecular design. For settings where compounds will be purchased or outsourced, cost-sensitive Bayesian optimization frameworks can help quantify tradeoffs between the information a new experiment will provide and its price.

5.3. Diversity of chemical spaces

As or more important than the size of a chemical space is its diversity and whether it contains molecules that satisfy our design objectives. While "diversity" is ill-defined, one could consider a diverse chemical space as one where a high-performing molecule can be found for multiple distinct discovery tasks. Chemical spaces accessible for experimental testing are defined both by the availability of commercial compounds and by the scope of known synthetic methods to transform them. Selecting diverse building

blocks with privileged or interesting structural motifs is one way to enable construction of a diverse chemical space [108]. Medicinal chemistry is dominated by a small number of reaction types [109] and as illustrated through a recent study by Tomberg and Boström, focusing on building block diversity might let us use simple chemistries more conducive to automation without sacrificing product diversity. However, over-reliance on routine chemistries may make us fated to repeat the boom and bust of combinatorial chemistry, which did not yield results at a rate commensurate with the sheer number of compounds tested. Diversity-oriented synthesis (DOS) is an orthogonal approach [111, 112] where the reactions themselves introduce a high degree of structural complexity. Clever modulation of reaction conditions can lead to dozens of unique products even given very simple amine and carboxylic acid building blocks [113]. A combination of both will likely lead to the most useful compound sets. Designing optimal screening collections as small predefined chemical spaces is an ongoing topic of research, particularly for drug discovery applications [114, 115, 116].

5.4. Simplification of molecular structures for on-the-fly generation

An underappreciated limitation of most algorithms for on-the-fly generation, particularly using deep generative models, is their ability to handle stereoisomerism. We know that stereoisomers can exhibit drastically different properties (e.g., the BINAP family of ligands), and yet SMILES and graph representations are fundamentally unable to distinguish configurational isomers defined by more than tetrahedral chirality and cis/trans bond isomerism (e.g., atropisomers, folded polypeptides). One could argue that models operating on these representations do not meaningfully understand point chirality either. Fragment-based methods that explicitly operate in 3D coordinates or generative models designed to propose individual conformers [117] theoretically overcome this limitation, but there have been few (if any) evaluations of on-the-fly molecular generation where the design objective is sensitive to atropisomerism. Even representing these structures as comformational ensembles for the sake of property prediction (i.e., as would be used to build a surrogate model $\hat{f}(x)$) is a broader, and perhaps more immediate, challenge.

5.5. Simplification of design objectives

Computational approximations to physical properties are rarely able to replace physical testing (e.g., docking scores as approximations of binding affinity). This makes it difficult to benchmark algorithms for chemical space

exploration, as benchmarks' design objectives do not reflect the true complexity of the problem and may not reveal certain failure modes [118]. However, experimental data can only be used in a retrospective setting to test algorithms for exploring pre-defined spaces (e.g., [70, 103]).

Design objectives tend to focus on optimizing simple heuristics calculated by fragment-contribution approaches or based on similarity to a target structure; several algorithms already achieve near perfect results on these tasks [92]. A notable exception is a study by Aumentado-Armstrong that incorporates a docking scores in the design objective [119]; although a simple proxy for bioactivity, optimizing a docking score can still prove challenging for some algorithms [120]. We require a new generation of benchmark tasks that contain (a) complex design objectives that are less smooth with respect to molecular structure, exhibit more local optima, and require tradeoffs between competing design objectives, (b) tasks requiring the generation of individual stereoisomers, (c) metrics related to sample efficiency, and (d) budget-limited settings where not all compounds carry the same experimental cost.

5.6. Human involvement in computer-aided chemical space exploration

As a closing thought, we should acknowledge the role of human expertise in computational workflows for chemical space exploration. Computational workflows should not be used for their own sake, but because they enable discoveries that are otherwise inaccessible to human researchers, whether that is due to the complexity of the optimization task or simply speed or throughput. We should not eschew human expertise if it is possible to restrict ourselves to a smaller chemical space (that is easier to search) through domain knowledge. Human filters and manual selection of compounds has almost always been an intermediate step between computational predictions and physical experiments and should remain so until we can capture those considerations algorithmically. As we build that ability, the influence of human subjectivity can be reduced and the process of chemical space exploration can be further operationalized.

Acknowledgements

We thank David Graff for commenting on the manuscript and the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium for inspiring conversations.

References

[1] Bohacek, R. S., McMartin, C., & Guida, W. C. (1996). The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.*, *16*(1), 3–50. https://doi.org/10.1002/(SICI)1098-1128(199601)16:1$\langle$$3::AID-MED1$\rangle$3.0.CO;2-6

[2] Drew, K. L. M., Baiman, H., Khwaounjoo, P., Yu, B., & Reynisson, J. (2012). Size estimation of chemical space: How big is it? *J. Pharm. Pharmacol.*, *64*(4), 490–495. https://doi.org/10.1111/j.2042-7158.2011.01424.x

[3] Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.*, *27*(8), 675–679. https://doi.org/10.1007/s10822-013-9672-4

[4] Oprea, T. I., & Gottfries, J. (2001). Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.*, *3*(2), 157–166. https://doi.org/10.1021/cc0000388

[5] Reymond, J.-L., & Awale, M. (2012). Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.*, *3*(9), 649–657. https://doi.org/10.1021/cn3000422

[6] Awale, M., & Reymond, J.-L. (2016). Web-based 3D-visualization of the DrugBank chemical space. *J. Cheminform.*, *8*. https://doi.org/10.1186/s13321-016-0138-2

[7] Probst, D., & Reymond, J.-L. (2020). Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.*, *12*(1), 12. https://doi.org/10.1186/s13321-020-0416-x

[8] Coley, C. W., Eyke, N. S., & Jensen, K. F. (2020a). Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201909987

[9] Coley, C. W., Eyke, N. S., & Jensen, K. F. (2020b). Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201909989

[10] Dobson, C. M. (2004). Chemical space and biology [Number: 7019 Publisher: Nature Publishing Group]. *Nature*, *432*(7019), 824–828. https://doi.org/10.1038/nature03192

[11] Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine [Number: 7019 Publisher: Nature Publishing

Group]. *Nature*, *432*(7019), 855–861. https://doi.org/10.1038/nature03193

[12] Lemonick, S. (2020). Exploring chemical space: Can AI take us where no human has gone before? [Library Catalog: cen.acs.org]. Retrieved August 6, 2020, from https://cen.acs.org/physical-chemistry/computational-chemistry/Exploring-chemical-space-AI-take/98/i13

[13] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G. (2012). ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, *52*(7), 1757–1768. https://doi.org/10.1021/ci3001277

[14] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, *40*(Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777

[15] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.*, *47*(D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033

[16] Reymond, J.-L. (2015). The Chemical Space Project. *Acc. Chem. Res.*, *48*(3), 722–730. https://doi.org/10.1021/ar500432k

[17] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, *34*(Database issue), D668–D672. https://doi.org/10.1093/nar/gkj067

[18] *REAL Compounds - Enamine* [Accessed 2019-07-25]. (2019). https://enamine.net/library-synthesis/real-compounds

[19] LabNetwork [Accessed 2020-08-06]. (2020). https://www.labnetwork.com/frontend-app/p/%5C#!/library/virtual

[20] Patel, H., Ihlenfeldt, W., Judson, P., Moroz, Y. S., Pevzner, Y., Peach, M., Tarasova, N., & Nicklaus, M. (2020). Synthetically Accessible Virtual Inventory (SAVI). https://doi.org/10.26434/chemrxiv.12185559.v1

[21] Hu, Q., Peng, Z., Kostrowicki, J., & Kuki, A. (2011). LEAP into the Pfizer Global Virtual Library (PGVL) space: Creation of readily synthesizable design ideas automatically. *Methods in Molecular Biology*

(Clifton, N.J.), *685*, 253–276. https://doi.org/10.1007/978-1-60761-931-4_13

[22] Nicolaou, C. A., Watson, I. A., Hu, H., & Wang, J. (2016). The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.*, *56*(7), 1253–1266. https://doi.org/10.1021/acs.jcim.6b00173

[23] Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1994). Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.*, *18*(9), 833–844. https://doi.org/10.1016/0098-1354(93)E0023-3

[24] Rotstein, S. H., & Murcko, M. A. (1993). GroupBuild: A fragment-based method for de novo drug design. *J. Med. Chem.*, *36*(12), 1700–1710.

[25] Pierce, A. C., Rao, G., & Bemis, G. W. (2004). BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease. *J. Med. Chem.*, *47*(11), 2768–2775. https://doi.org/10.1021/jm030543u

[26] Jensen, J. H. (2019). A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.*, *10*(12), 3567–3572. https://doi.org/10.1039/C8SC05372C

[27] Ahn, S., Kim, J., Lee, H., & Shin, J. (2020). Guiding Deep Molecular Optimization with Genetic Exploration. *arXiv:2007.04897 [cs, q-bio, stat]*. Retrieved July 14, 2020, from http://arxiv.org/abs/2007.04897

[28] Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F. D., Heeres, J., Koymans, L. M. H., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., & Janssen, P. A. J. (2003). SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.*, *46*(13), 2765–2773. https://doi.org/10.1021/jm030809x

[29] Fechner, U., & Schneider, G. (2006). Flux (1): A virtual synthesis scheme for fragment-based de novo design. *J. Chem. Inf. Model.*, *46*(2), 699–707. https://doi.org/10.1021/ci0503560

[30] Firth, N. C., Atrash, B., Brown, N., & Blagg, J. (2015). MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. *J. Chem. Inf. Model.*, *55*(6), 1169–1180. https://doi.org/10.1021/acs.jcim.5b00073

[31] Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., & Schneider, G. (2012). DOGS: Reaction-Driven de novo Design of Bioactive Compounds. *PLOS Computa-*

tional Biology, *8*(2), e1002380. https://doi.org/10.1371/journal.pcbi.1002380

[32] 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. (2018). Automatic chemical design using a data-driven continuous representation of molecules [arXiv: 1610.02415]. *ACS Cent. Sci.*, *4*(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572

[33] Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv:1802.04364*. https://arxiv.org/abs/1802.04364

[34] Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.*, *4*(1), 120–131. https://doi.org/10.1021/acscentsci.7b00512

[35] Olivecrona, M., Blaschke, T., Engkvist, O., & Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. Cheminform.*, *9*(1), 48. https://doi.org/10.1186/s13321-017-0235-x

[36] Zhou, Z., Kearnes, S., Li, L., Zare, R. N., & Riley, P. (2018). Optimization of Molecules via Deep Reinforcement Learning [arXiv: 1810.08678]. *arXiv:1810.08678 [cs, stat]*. Retrieved November 2, 2018, from http://arxiv.org/abs/1810.08678

[37] Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S., & Hernández-Lobato, J. M. (2019). A Model to Search for Synthesizable Molecules [arXiv: 1906.05221]. *arXiv:1906.05221 [physics, stat]*. Retrieved December 19, 2019, from http://arxiv.org/abs/1906.05221

[38] Korovina, K., Xu, S., Kandasamy, K., Neiswanger, W., Poczos, B., Schneider, J., & Xing, E. P. (2019). ChemBO: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations [arXiv: 1908.01425]. *arXiv:1908.01425 [physics, stat]*. Retrieved August 10, 2019, from http://arxiv.org/abs/1908.01425

[39] 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. (2020). Learning To Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning. Retrieved May 1, 2020, from https://arxiv.org/abs/2004.12485v1

[40] Horwood, J., & Noutahi, E. (2020). Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. Retrieved May 1, 2020, from https://arxiv.org/abs/2004.14308v1

[41] Walters, W. P. (2018). Virtual Chemical Libraries. *J. Med. Chem.* https://doi.org/10.1021/acs.jmedchem.8b01048

[42] Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M., ... Aspuru-Guzik, A. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, *15*(10), 1120–1127. https://doi.org/10.1038/nmat4717

[43] Janet, J. P., Ramesh, S., Duan, C., & Kulik, H. J. (2020). Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.*, *6*(4), 513–524. https://doi.org/10.1021/acscentsci.0c00026

[44] Lewell, X. Q., Judd, D. B., Watson, S. P., & Hann, M. M. (1998). RECAP–retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inform. Comput. Sci.*, *38*(3), 511–522. https://doi.org/10.1021/ci970429i

[45] Ertl, P. (2003). Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inform. Comput. Sci.*, *43*(2), 374–380. https://doi.org/10.1021/ci0255782

[46] Cayley, E. (1875). Ueber die analytischen Figuren, welche in der Mathematik Bäume genannt werden und ihre Anwendung auf die Theorie chemischer Verbindungen. *Berichte der deutschen chemischen Gesellschaft*, *8*(2), 1056–1059.

[47] Henze, H. R., & Blair, C. M. (1931). The number of isomeric hydrocarbons of the methane series. *J. Am. Chem. Soc.*, *53*(8), 3077–3085.

[48] Fink, T., & Reymond, J.-L. (2007). Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes,

and Drug Discovery. *J. Chem. Inf. Model.*, *47*(2), 342–353. https://doi.org/10.1021/ci600423u

[49] Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.*, *52*(11), 2864–2875. https://doi.org/10.1021/ci300415d

[50] Cramer, R. D., Patterson, D. E., Clark, R. D., Soltanshahi, F., & Lawless, M. S. (1998). Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inform. Comput. Sci.*, *38*(6), 1010–1023. https://doi.org/10.1021/ci9800209

[51] Nikitin, S., Zaitseva, N., Demina, O., Solovieva, V., Mazin, E., Mikhalev, S., Smolov, M., Rubinov, A., Vlasov, P., Lepikhin, D., Khachko, D., Fokin, V., Queen, C., & Zosimov, V. (2005). A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput. Aided Mol. Des.*, *19*(1), 47–63. https://doi.org/10.1007/s10822-005-0097-6

[52] Cramer, R. D., Soltanshahi, F., Jilek, R., & Campbell, B. (2007). AllChem: Generating and searching 1020 synthetically accessible structures. *J. Comput. Aided Mol. Des.*, *21*(6), 341–350. https://doi.org/10.1007/s10822-006-9093-8

[53] Patel, H., Bodkin, M. J., Chen, B., & Gillet, V. J. (2009). Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *J. Chem. Inf. Model.*, *49*(5), 1163–1184. https://doi.org/10.1021/ci800413m

[54] Hoffmann, T., & Gastreich, M. (2019). The next level in chemical space navigation: Going far beyond enumerable compound libraries. *Drug Discov. Today*. https://doi.org/10.1016/j.drudis.2019.02.013

[55] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., & Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 1. https://doi.org/10.1038/s41586-019-0917-9

[56] Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R., & Jensen, K. F. (2019). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, *10*(2), 370–377. https://doi.org/10.1039/C8SC04228D

[57] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular Transformer: A Model for Uncertainty-

Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.*, *5*(9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576

[58] Tomberg, A., Johansson, M. J., & Norrby, P.-O. (2019). A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *The Journal of Organic Chemistry*, *84*(8), 4695–4703. https://doi.org/10.1021/acs.joc.8b02270

[59] Beker, W., Gajewska, E. P., Badowski, T., & Grzybowski, B. A. (2019). Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem. Int. Ed.*, *58*(14), 4515–4519. https://doi.org/10.1002/anie.201806920

[60] Struble, T. J., Coley, C. W., & Jensen, K. F. (2020). Multitask prediction of site selectivity in aromatic C–H functionalization reactions. *React. Chem. Eng.*, *5*(5), 896–902. https://doi.org/10.1039/D0RE00071J

[61] Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C. C., Subramanian, A., & Golub, T. R. (2017). The Drug Repurposing Hub: A next-generation drug library and information resource. *Nat. Med.*, *23*(4), 405–408. https://doi.org/10.1038/nm.4306

[62] Huang, R., Zhu, H., Shinn, P., Ngan, D., Ye, L., Thakur, A., Grewal, G., Zhao, T., Southall, N., Hall, M. D., Simeonov, A., & Austin, C. P. (2019). The NCATS Pharmaceutical Collection: A 10-year update. *Drug Discov. Today*, *24*(12), 2341–2349. https://doi.org/10.1016/j.drudis.2019.09.019

[63] Clark, M. A., Acharya, R. A., Arico-Muendel, C. C., Belyanskaya, S. L., Benjamin, D. R., Carlson, N. R., Centrella, P. A., Chiu, C. H., Creaser, S. P., Cuozzo, J. W., Davie, C. P., Ding, Y., Franklin, G. J., Franzen, K. D., Gefter, M. L., Hale, S. P., Hansen, N. J. V., Israel, D. I., Jiang, J., ... Morgan, B. A. (2009). Design, synthesis and selection of DNA-encoded small-molecule libraries [Number: 9 Publisher: Nature Publishing Group]. *Nature Chemical Biology*, *5*(9), 647–654. https://doi.org/10.1038/nchembio.211

[64] Smith, G. P., & Petrenko, V. A. (1997). Phage Display. *Chemical Reviews*, *97*(2), 391–410. https://doi.org/10.1021/cr960065d

[65] Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I.,

Wagner, G., & Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, *580*(7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z

[66] Acharya, A., Agarwal, R., Baker, M., Baudry, J., Bhowmik, D., Boehm, S., Byler, K., Coates, L., Chen, S. Y.-C., Cooper, C. J., Demerdash, O., Daidone, I., Eblen, J., R. Ellingson, S., Forli, S., Glaser, J., Gumbart, J. C., Gunnels, J., Hernandez, O., ... Zanetti-Polzi, L. (2020). Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. https://doi.org/10.26434/chemrxiv.12725465.v1

[67] Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[68] Frazier, P. I. (2018). *A Tutorial on Bayesian Optimization*. Retrieved June 6, 2020, from https://arxiv.org/abs/1807.02811v1

[69] Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L. N., Grave, K. D., Ramon, J., Clare, M. d., Sirawaraporn, W., Oliver, S. G., & King, R. D. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of The Royal Society Interface*, *12*(104), 20141289. https://doi.org/10.1098/rsif.2014.1289

[70] Kangas, J. D., Naik, A. W., & Murphy, R. F. (2014). Efficient discovery of responses of proteins to compounds using active learning. *BMC Bioinformatics*, *15*(1), 143. https://doi.org/10.1186/1471-2105-15-143

[71] Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., & Cherkasov, A. (2020). Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* https://doi.org/10.1021/acscentsci.0c00229

[72] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chem. Soc. Rev.* https://doi.org/10.1039/D0CS00098A

[73] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks [arXiv: 1901.00596].

IEEE Transactions on Neural Networks and Learning Systems, 1–21. https://doi.org/10.1109/TNNLS.2020.2978386

[74] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, *180*(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

[75] Hie, B., Bryson, B. D., & Berger, B. (2020). Learning with uncertainty for biological discovery and design [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.08.11.247072. https://doi.org/10.1101/2020.08.11.247072

[76] Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: A review. *J. Chemom.*, *15*(7), 559–569. https://doi.org/10.1002/cem.651

[77] van Deursen, R., & Reymond, J.-L. (2007). Chemical Space Travel. *ChemMedChem*, *2*(5), 636–640. https://doi.org/10.1002/cmdc.200700021

[78] Hoksza, D., Škoda, P., Voršilák, M., & Svozil, D. (2014). Molpher: A software framework for systematic chemical space exploration. *J. Cheminform.*, *6*(1), 7. https://doi.org/10.1186/1758-2946-6-7

[79] Kawai, K., Nagata, N., & Takahashi, Y. (2014). De Novo Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach. *J. Chem. Inf. Model.*, *54*(1), 49–56. https://doi.org/10.1021/ci400418c

[80] Nigam, A., Friederich, P., Krenn, M., & Aspuru-Guzik, A. (2020). Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space [arXiv: 1909.11655]. *arXiv:1909.11655 [physics]*. Retrieved August 13, 2020, from http://arxiv.org/abs/1909.11655

[81] Henault, E. S., Rasmussen, M. H., & Jensen, J. H. (2020). Chemical Space Exploration: How Genetic Algorithms Find the Needle in the Haystack. https://doi.org/10.26434/chemrxiv.12152661.v1

[82] Koerstz, M., Christensen, A. S., Mikkelsen, K. V., Nielsen, M. B., & Jensen, J. H. (2020). High Throughput Virtual Screening of 230 Billion Molecular Solar Heat Battery Candidates. https://doi.org/10.26434/chemrxiv.8003813.v2

[83] Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, *361*(6400), 360–365. https://doi.org/10.1126/science.aat2663

[84] Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, *4*(4), 828–849. https://doi.org/10.1039/C9ME00039A

[85] Schwalbe-Koda, D., & Gómez-Bombarelli, R. (2019). Generative Models for Automatic Chemical Design [arXiv: 1907.01632]. *arXiv:1907.01632 [physics, stat]*. Retrieved July 16, 2019, from http://arxiv.org/abs/1907.01632

[86] Vanhaelen, Q., Lin, Y.-C., & Zhavoronkov, A. (2020). The Advent of Generative Chemistry. *ACS Medicinal Chemistry Letters*. https://doi.org/10.1021/acsmedchemlett.0c00088

[87] Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design [Publisher: American Association for the Advancement of Science Section: Research Article]. *Science Advances*, *4*(7), eaap7885. https://doi.org/10.1126/sciadv.aap7885

[88] Liu, Q., Allamanis, M., Brockschmidt, M., & Gaunt, A. (2018). Constrained Graph Variational Autoencoders for Molecule Design (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Eds.). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Neurips 31*. Curran Associates, Inc.

[89] Li, Y., Vinyals, O., Dyer, C., Pascanu, R., & Battaglia, P. (2018). Learning Deep Generative Models of Graphs [arXiv: 1803.03324]. *arXiv:1803.03324 [cs, stat]*. Retrieved August 11, 2020, from http://arxiv.org/abs/1803.03324

[90] You, J., Liu, B., Ying, R., Pande, V., & Leskovec, J. (2019). Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation [arXiv: 1806.02473]. *arXiv:1806.02473 [cs, stat]*. Retrieved August 11, 2020, from http://arxiv.org/abs/1806.02473

[91] Tripp, A., Daxberger, E., & Hernández-Lobato, J. M. (2020). Sample-Efficient Optimization in the Latent Space of Deep Generative Models via Weighted Retraining [arXiv: 2006.09191]. *arXiv:2006.09191 [cs, stat]*. Retrieved June 22, 2020, from http://arxiv.org/abs/2006.09191

[92] Brown, N., Fiscato, M., Segler, M. H. S., & Vaucher, A. C. (2018). GuacaMol: Benchmarking Models for De Novo Molecular Design [arXiv: 1811.09621]. *arXiv:1811.09621 [physics, q-bio]*. Retrieved June 10, 2019, from http://arxiv.org/abs/1811.09621

[93] Polykovskiy, D., Zhebrak, A., Vetrov, D., Ivanenkov, Y., Aladinskiy, V., Mamoshina, P., Bozdaganyan, M., Aliper, A., Zhavoronkov, A., & Kadurin, A. (2018). Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharm.*, *15*(10), 4398–4405. https://doi.org/10.1021/acs.molpharmaceut.8b00839

[94] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., ... Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, *37*(9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x

[95] Sumita, M., Yang, X., Ishihara, S., Tamura, R., & Tsuda, K. (2018). Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.*, *4*(9), 1126–1133. https://doi.org/10.1021/acscentsci.8b00213

[96] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Nikolenko, S., Aspuru-Guzik, A., & Zhavoronkov, A. (2018). Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models [arXiv: 1811.12823]. *arXiv:1811.12823 [cs, stat]*. Retrieved July 21, 2019, from http://arxiv.org/abs/1811.12823

[97] Gao, W., & Coley, C. W. (2020). The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.0c00174

[98] Polishchuk, P. (2020). CReM: Chemically reasonable mutations framework for structure generation. *J. Cheminform.*, *12*(1), 28. https://doi.org/10.1186/s13321-020-00431-w

[99] Schneider, G., Lee, M.-L., Stahl, M., & Schneider, P. (2000). De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.*, *14*(5), 487–494. https://doi.org/10.1023/A:1008184403558

[100] Beccari, A. R., Cavazzoni, C., Beato, C., & Costantino, G. (2013). LiGen: A High Performance Workflow for Chemistry Driven de Novo Design. *J. Chem. Inf. Model.*, *53*(6), 1518–1527. https://doi.org/10.1021/ci400078g

[101] Pophale, R., Daeyaert, F., & W. Deem, M. (2013). Computational prediction of chemically synthesizable organic structure directing agents for zeolites. *J. Mater. Chem. A*, *1*(23), 6750–6760. https://doi.org/10.1039/C3TA10626H

[102] Weber, L., Illgen, K., & Almstetter, M. (1999). Discovery of New Multi Component Reactions with Combinatorial Methods. *Synlett*, *1999*(3), 366–374. https://doi.org/10.1055/s-1999-2612

[103] Paricharak, S., IJzerman, A. P., Bender, A., & Nigsch, F. (2016). Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. *ACS Chem. Bio.*, *11*(5), 1255–1264. https://doi.org/10.1021/acschembio.6b00029

[104] Desai, B., Dixon, K., Farrant, E., Feng, Q., Gibson, K. R., van Hoorn, W. P., Mills, J., Morgan, T., Parry, D. M., Ramjee, M. K., Selway, C. N., Tarver, G. J., Whitlock, G., & Wright, A. G. (2013). Rapid Discovery of a Novel Series of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform. *J. Med. Chem.*, *56*(7), 3033–3047. https://doi.org/10.1021/jm400099d

[105] Godfrey, A. G., Masquelin, T., & Hemmerle, H. (2013). A remote-controlled adaptive medchem lab: An innovative approach to enable drug discovery in the 21st Century. *Drug Discov. Today*, *18*(17), 795–802. https://doi.org/10.1016/j.drudis.2013.03.001

[106] Baranczak, A., Tu, N. P., Marjanovic, J., Searle, P. A., Vasudevan, A., & Djuric, S. W. (2017). Integrated Platform for Expedited Synthesis–Purification–Testing of Small Molecule Libraries. *ACS Medicinal Chemistry Letters*, *8*(4), 461–465. https://doi.org/10.1021/acsmedchemlett.7b00054

[107] Coley, C. W., Thomas, D. A., Lummiss, J. A. M., Jaworski, J. N., Breen, C. P., Schultz, V., Hart, T., Fishman, J. S., Rogers, L., Gao, H., Hicklin, R. W., Plehiers, P. P., Byington, J., Piotti, J. S., Green, W. H., Hart, A. J., Jamison, T. F., & Jensen, K. F. (2019). A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, *365*(6453), eaax1566. https://doi.org/10.1126/science.aax1566

[108] Goldberg, F. W., Kettle, J. G., Kogej, T., Perry, M. W. D., & Tomkinson, N. P. (2015). Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today*, *20*(1), 11–17. https://doi.org/10.1016/j.drudis.2014.09.023

[109] Roughley, S. D., & Jordan, A. M. (2011). The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.*, *54*(10), 3451–3479. https://doi.org/10.1021/jm200187y

[110] Tomberg, A., & Boström, J. (2020). Can "Easy" Chemistry Produce Complex, Diverse and Novel Molecules? https://doi.org/10.26434/chemrxiv.12563231.v1

[111] Schreiber, S. L. (2000). Target-Oriented and Diversity-Oriented Organic Synthesis in Drug Discovery. *Science*, *287*(5460), 1964–1969. https://doi.org/10.1126/science.287.5460.1964

[112] Gerry, C. J., & Schreiber, S. L. (2020). Recent achievements and current trajectories of diversity-oriented synthesis. *Curr. Opin. Chem. Biol.*, *56*, 1–9. https://doi.org/10.1016/j.cbpa.2019.08.008

[113] Mahjour, B., Shen, Y., Liu, W., & Cernak, T. (2020). A map of the amine–carboxylic acid coupling system [Number: 7801 Publisher: Nature Publishing Group]. *Nature*, *580*(7801), 71–75. https://doi.org/10.1038/s41586-020-2142-y

[114] Huggins, D. J., Venkitaraman, A. R., & Spring, D. R. (2011). Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Bio.*, *6*(3), 208–217. https://doi.org/10.1021/cb100420r

[115] Baell, J. B. (2013). Broad Coverage of Commercially Available Lead-like Screening Space with Fewer than 350,000 Compounds. *J. Chem. Inf. Model.*, *53*(1), 39–55. https://doi.org/10.1021/ci300461a

[116] Yang, Z.-Y., He, J.-H., Lu, A.-P., Hou, T.-J., & Cao, D.-S. (2020). Application of Negative Design To Design a More Desirable Virtual Screening Library. *J. Med. Chem.*, *63*(9), 4411–4429. https://doi.org/10.1021/acs.jmedchem.9b01476

[117] Simm, G. N. C., & Hernández-Lobato, J. M. (2019). A Generative Model for Molecular Distance Geometry [arXiv: 1909.11459]. *arXiv:1909.11459 [cs, stat]*. Retrieved October 1, 2019, from http://arxiv.org/abs/1909.11459

[118] Renz, P., Van Rompaey, D., Wegner, J. K., Hochreiter, S., & Klambauer, G. (2020). On Failure Modes of Molecule Generators and Optimizers. https://doi.org/10.26434/chemrxiv.12213542.v1

[119] Aumentado-Armstrong, T. (2018). Latent Molecular Optimization for Targeted Therapeutic Design [arXiv: 1809.02032]. *arXiv:1809.02032 [cs, q-bio]*. Retrieved September 11, 2018, from http://arxiv.org/abs/1809.02032

[120] Cieplinski, T., Danel, T., Podlewska, S., & Jastrzebski, S. (2020). We Should at Least Be Able to Design Molecules That Dock Well [arXiv: 2006.16955]. *arXiv:2006.16955 [cs, q-bio, stat]*. Retrieved July 3, 2020, from http://arxiv.org/abs/2006.16955