The State of Machine Learning-Based Molecular Design

Literature Review

Aaron Tian - aztian@wpi.edu

Massachusetts Academy of Math and Science

Worcester, MA

Abstract

The discovery of small molecule drugs is a costly and time inefficient process, revolving primarily around trial and error. Recent advances in machine learning and deep learning, coupled with rapidly improving computational speeds, suggest artificial intelligence (AI) as a promising solution to the problem. Recent work finds that AI is effective at predictive and generative tasks, outperforming earlier computational methods. However, more research needs to be done on improving accuracy and ensuring synthesizability in the molecules created by machine learning methods in order for the tool to be reliably used in an industry setting. The goal of this review is to introduce the current trajectory of machine learning-based molecular design, as well as areas in need of further research in the future.

The State of Machine Learning-Based Molecular Design

Literature Review

The discovery of new small molecule drugs is a time-consuming and expensive process, costing over $2.5 billion and taking 10 - 15 years on average (DiMasi et al., 2016). To compensate for these costs, pharmaceutical companies rely heavily on the blockbuster model, in which the majority of sales are driven by a small quantity of successful drugs: each generating over $1 billion annually (Malik, 2008). New blockbusters must be discovered when patent protection expires on existing ones, as their prices drop by up to 80% within 6-12 months of expiration. However, new drugs are now being developed at a much slower rate, dropping from 44 on average between 1995-2000 to an average of 33 between 2001-2006. The decline is likely due to all the easy-to-find drugs having already been discovered. This fact, alongside a rise in the use of biologics, suggests a rough future for small molecule therapeutics. Nonetheless, the benefits of small molecule drugs—simple dosing protocols, ease of manufacture, and oral bioavailability—should not be ignored (Ngo & Garneau-Tsodikova, 2018). It may be possible to revive small molecule drug development by automating certain processes within the industry, such as the discovery stages of hit-to-lead and lead optimization.

As a result of advancements in the computational capabilities of modern hardware, deep learning has emerged as a powerful and computationally viable tool to automate a variety of previously-unapproachable problems ranging from image classification to natural language processing. In an effort to extend these technologies to chemistry-related tasks, several methods of encapsulating chemical data have been proposed. The corresponding models that process these data representations are equally diverse and can address a variety of tasks from prediction of quantum chemical properties to goal-directed generation of novel molecular structures, demonstrating that deep learning is a promising approach to handling chemical data. In the past, these tasks were either unapproachable or required expensive processes—such as density functional theory (DFT) in the case of quantum property prediction—to compute. However, more work must be done to make these models relevant in the context of drug discovery; prominent challenges within the field include the inability for models to consistently propose synthetically accessible compounds and the lack of robust evaluation metrics for pharmacologically-relevant molecular properties.

Philosophy

The study of chemistry is largely concerned with the relationship between molecular structures and their physical properties; thus, it is imperative to develop methods that draw concrete mappings between the two. The set of all synthetically accessible molecule configurations is commonly referred to as *chemical space*, and its size is estimated to be around 10^{60} and 10^{100} (Li et al., 2018). Chemical space is innately discontinuous, as small modifications to any given compound can drastically change its physical properties or make it synthetically inaccessible. This property renders traditional machine learning models ineffectual at performing on molecular structures directly, as gradient-based algorithms cannot be used.

A common problem within chemistry-oriented machine learning is to define a mapping between the chemical space and a latent representation in Euclidean space; the majority of machine learning models are designed to operate on the latter. This operation is traditionally performed by training a hidden neural network within the model, though the specific method depends on the data representation used to encode molecular structures—the most common of which include SMILES strings, molecular fingerprints, and discrete graphs (see Appendix A).

With training features in latent space, these machine learning models can then perform regression and generation-oriented tasks.

Predictive Models

The goal of chemistry-oriented predictive models is—given a molecular structure—to predict associated properties which cannot be ascertained deterministically (or which are too computationally expensive to calculate efficiently). Pharmacokinetic properties, which can only be measured via experimentation, and quantum chemical properties, which can be calculated only using expensive simulations such as density-functional theory (DFT), are two of the most sought-after targets in machine learning.

The use of machine learning in chemistry-related tasks was popularized by Gilmer et al. (2017), who formulated molecules as discrete mathematical graphs and developed the Message-Passing Neural Network framework (MPNN) to learn abstract relationships between graph-structured data. MPNN achieved state of the art results on the QM9 dataset (Ramakrishnan et al., 2014), a dataset for the prediction of quantum chemical properties. The MPNN framework was particularly groundbreaking due to its independence of feature engineering, a costly process of calculating structural properties of a given molecule. Since then, various modifications have been made to improve the performance of MPNN.

In a proof-of-concept study by Deng et al. (2021), the output layer of MPNN was replaced by XGBoost, a gradient-boosting algorithm. The study found that the new framework improved performance on 7 of 10 tasks in QM9 before hyperparameter tuning and 10 of 10 tasks after hyperparameter tuning. The study suggests that MPNN has a modular nature, in which components of the framework can be replaced with more optimal algorithms to directly enhance performance.

Li et al. (2021) introduced TrimNet, which utilized a multi-head attention mechanism to reduce unnecessary parameters in the model. The model was formulated to address the heavy computation cost of MPNN, which was found to have many unnecessary parameters that slow down computational speeds. In addition to the novel architecture, TrimNet employed the use of layer normalization. When tested on QM9, TrimNet achieved the new state of the art on all tasks.

Generative Models

At the forefront of computational molecular design, generative models are tasked with developing novel molecular structures. In the context of drug discovery, such models can be utilized to expand virtual screening libraries with novel compounds and optimize the pharmacokinetic properties of lead molecules. For the purposes of this review, the term ‘generative model’ will be used to refer to both novel molecule generators and molecule optimizers. Machine learning-based generative models can be classified under two general paradigms, namely autoregressive and non-autoregressive. Autoregressive models formulate molecule generation as an iterative process, where small modifications are made to a base until a terminal condition is reached, while non-autoregressive models attempt to generate the entire molecule with one computational step.

Evaluation of Generative Models

The quality of molecules proposed by generative models is typically assessed on three criteria: *validity*, referring to the resulting compounds’ adherence to simple chemical rules, such as the octet rule of valency; *novelty*, referring to the amount of generated molecules not present in the training dataset; and *uniqueness*, referring to the amount of variance between the model's output molecules (Xiong et al., 2021). In regard to generative paradigms, a clear tradeoff between the criteria is prevalent. Autoregressive models can achieve high validity scores by incorporating chemical rules to prevent the model from performing invalid modifications, but they suffer from lower novelty and uniqueness scores due to a tendency for the model to converge to a single output. Conversely, non-autoregressive models have higher uniqueness and novelty rates but generate valid compounds at a much lower rate.

Furthermore, optimization-oriented generative models are typically evaluated on the calculated partition coefficient (clogP) and quantitative estimate of drug-likeness (QED) (see Appendix B). These metrics are used in part because of a loose connection to drug evaluation, with logP as a criterion in Lipinski's rule of 5 (Lipinski et al., 1997) and QED as a heuristic estimate of drug quality, but they are mainly used because of a lack of better evaluation metrics. QED and clogP can be calculated inexpensively, which is suitable for the training and evaluation of a generative model, but more relevant properties for drug discovery such as toxicity and ADME (see Appendix B) are difficult to predict for novel compounds. Designing better evaluation metrics is a necessary area of further research in order for molecule generation to be directly applicable to drug discovery (Zhou et al., 2019); however, QED and clogP will suffice as of now for proof-of-concept testing.

Generative Models in Literature

Since autoregressive models are considerably more relevant than non-autoregressive models for synthesizable molecular design, we focus exclusively on this archetype in our review.

There are many classes of deep learning models which fall under the autoregressive category, but reinforcement learning methods (RL) are currently the most prominent. A study by Zhou et al. (2019) implemented MolDQN, a reinforcement learning framework for molecular optimization driven by deep Q-learning and a Markov decision process. MolDQN was designed for multi-objective learning, meaning that the model was capable of optimizing for several molecular properties simultaneously—a highly desirable feature in drug discovery. The algorithm worked by iteratively selecting a modification to perform on a given base structure from a set of actions (atom addition, bond addition, and bond removal). This design led to a 100% validity score because the model could remove invalid modifications to the molecule directly from the action set: an inductive bias for valid molecules. However, the model has a slow convergence speed—an issue experienced by nearly all reinforcement learning-based methods.

Shi et al. (2020) similarly uses reinforcement learning to tune their generative framework but employ a fundamentally different policy network to speed up the training process. GraphAF uses a normalizing flow model as its policy network, which defines an invertible transformation between a base distribution and the chemical space. Using an algorithm rooted in the change of variables formula, the network can effectively make conversions between the two data representations. The design of the algorithm allows for parallel execution, which significantly reduces computational overhead. Similar to MolDQN, the framework for GraphAF uses an autoregressive process to iteratively modify a base structure and is capable of multi-target optimization.

A shared drawback of the approaches mentioned above—and all generative models as a whole—is the inability to account for chemical synthesis. Gao and Coley (2020) claim that more generative models have been developed than synthesizable molecules produced from these models. However, synthetic accessibility is arguably the most important priority when designing compounds, as a structure that cannot be synthesized has no practical use on its own. Recent literature in molecule generation has pivoted toward models that can more effectively account for chemical synthesis.

Chemical Synthesis in Molecule Generation

Chemical synthesis prediction as a standalone task has been addressed using computational methods in the past. AiZynthFinder, developed by Genheden et al., (2020), uses Monte Carlo tree search to recursively search for synthetic pathways to an input molecule. The algorithm identifies full synthesis routes with reasonable accuracy but takes 7.1 seconds on average to find a solution. Analyzing chemical compound libraries with such a tool is clearly inefficient: computation time is strictly dependent on the size of the library, and the produced results cannot be conditionally selected with respect to desired chemical properties.

Gao and Coley (2019) summarize the various approaches that can be used to increase the rate of synthesizability in generative models. *Post hoc* filtering simply evaluates the outputs after being generated and isolates the synthesizable compounds. *A priori* biasing calls for the selective design of a train dataset such that the model only trains on synthesizable molecules in the hope of biasing the algorithm toward synthesizable products. Similarly, heuristic biasing and CASP oracle biasing attempt to bias the model by incorporating synthesizability as an additional optimization parameter using heuristic synthetic accessibility scorers and retrosynthesis tools, respectively.

Since these methods often lack consideration of chemical knowledge, a new archetype of approaches has been proposed to incorporate chemical synthesis as an inductive bias of the model. Bradshaw et al. (2019) proposed MoleculeChef, a generative model that takes an initial bag of reactants from a chemical catalog and applies reaction templates to predict the products of chemical reactions between these molecules. Since molecules proposed by the model are derived from purchasable compounds, their synthesizability is implicitly guaranteed.

Following the philosophy of MoleculeChef, Horwood and Noutahi (2020) created REACTOR, the first reinforcement learning-based approach to generate synthetic pathways. In contrast to previous autoregressive generation models, REACTOR implicitly builds molecules reaction-by-reaction instead of atom-by-atom to guarantee synthesizability. Furthermore, the use of a reinforcement learning framework enabled goal-directed optimization. However, according to Gao et al. (2021), REACTOR's reinforcement learning model does not converge effectively.

A recent study by Gao et al. (2021) formulated molecules as synthetic tree structures, where the root of the tree was the molecule itself and the leaves were purchasable starting compounds. As shown in Figure 1, tracing the leaf compounds through a series of chemical reactions leads to the root molecule. In their framework, the task of generation is broken down into 4 steps, each handled by a dedicated neural network:

1. A modification is selected from the set of actions: Add, Expand, Merge, or End.
2. The first reactant is selected from a compound catalog.
3. From a database of chemical reactions, an appropriate reaction template is selected.
4. Select the second reactant.

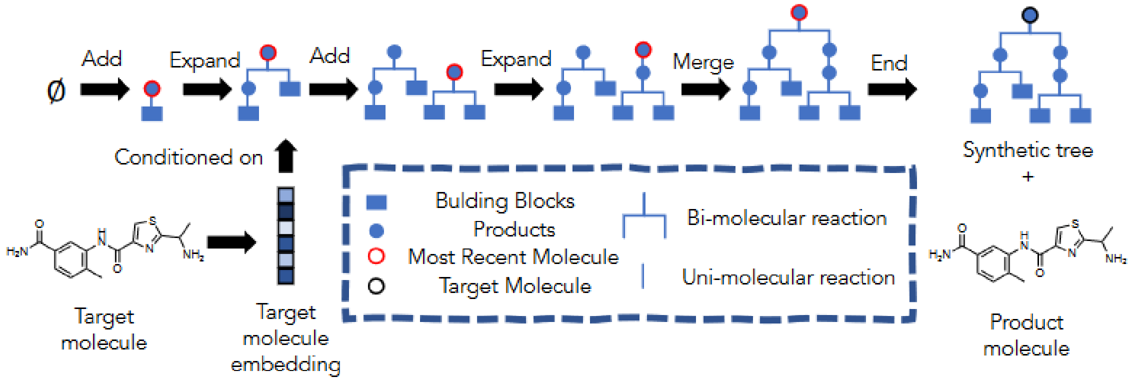

Figure 1: Visualization of synthetic pathway generation via modification of a synthetic tree. A synthetic tree with purchasable leaf nodes is theoretically guaranteed to lead to a synthesizable compound, as the tree itself presents a synthesis pathway to the root node. (Gao et al., 2021).

The primary benefit of this model is its low computational overhead; since each network performs a relatively simple task, the collective framework is significantly faster than a reinforcement learning approach. However, the initial reactant selection network becomes a bottleneck in the model due to the lack of information about the desired molecule and the exceedingly large set of possible compounds to choose from.

Conclusion

Recent advances in machine learning on chemical data have sparked interest in automating the discovery stages of drug development with computational methods. Currently, these models are able to reliably predict quantum molecular properties and generate novel molecular structures with 100% theoretical chemical validity. However, there exist challenges in both predictive and generative paradigms that hinder the application of these models in the context of drug discovery. In the case of prediction, quantum chemical properties are significantly less relevant in comparison to ADME and toxicity, as these metrics directly influence the evaluation of a drug compound. Generative models face a different problem: low synthetic accessibility of generated compounds. This issue has been addressed in literature, but the approaches face a critical flaw. Synthesis-oriented generative models assume that the generated outputs may be further optimized using manual methods (Horwood and Noutahi, 2020), but this defeats the purpose of pursuing synthetically accessible outputs; changing the structure of a compound means that a synthesis route is no longer guaranteed. In practice, the output of a synthesis-oriented model must be plausible enough to be brought directly to the preclinical testing phase. To address this issue, combining generative and predictive models into a unified framework is a promising direction of future research. In the future, predictive models may be able to compensate for the lack of deterministic methods to compute pharmacokinetic properties and, when integrated into the generation process, can direct the model to create more pharmacologically plausible structures for clinical testing. Furthermore, collaboration with pharmaceutical companies may supply the data needed to develop these predictive models for ADME and toxicity properties.

References

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, *4*(2), 90–98. https://doi.org/10.1038/nchem.1243

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 [Physics, Stat]*. http://arxiv.org/abs/1906.05221

Deng, D., Chen, X., Zhang, R., Lei, Z., Wang, X., & Zhou, F. (2021). XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties. *Journal of Chemical Information and Modeling*, *61*(6), 2697–2705. https://doi.org/10.1021/acs.jcim.0c01489

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Gao, W., & Coley, C. W. (2020). The Synthesizability of Molecules Proposed by Generative Models. *ArXiv:2002.07007 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/2002.07007

Gao, W., Mercado, R., & Coley, C. W. (2021). Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. *ArXiv:2110.06389 [Cs, q-Bio]*. http://arxiv.org/abs/2110.06389

Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O., & Bjerrum, E. (2020). AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, *12*(1), 70. https://doi.org/10.1186/s13321-020-00472-1

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. *ArXiv:1704.01212 [Cs]*. http://arxiv.org/abs/1704.01212

Horwood, J., & Noutahi, E. (2020). Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. *ACS Omega*, *5*(51), 32984–32994. https://doi.org/10.1021/acsomega.0c04153

Li, P., Li, Y., Hsieh, C.-Y., Zhang, S., Liu, X., Liu, H., Song, S., & Yao, X. (2021). TrimNet: Learning molecular representation from triplet messages for biomedicine. *Briefings in Bioinformatics*, *22*(4), bbaa266. https://doi.org/10.1093/bib/bbaa266

Li, Y., Zhang, L., & Liu, Z. (2018). Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, *10*(1), 33. https://doi.org/10.1186/s13321-018-0287-6

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *23*(1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1

Malik, N. N. (2008). Drug discovery: Past, present and future. *Drug Discovery Today*, *13*(21–22), 909–912. https://doi.org/10.1016/j.drudis.2008.09.007

Ngo, H. X., & Garneau-Tsodikova, S. (2018). What are the drugs of the future? *MedChemComm*, *9*(5), 757–758. https://doi.org/10.1039/C8MD90019A

Ramakrishnan, R., Dral, P. O., Rupp, M., & von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, *1*(1), 140022. https://doi.org/10.1038/sdata.2014.22

Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., & Tang, J. (2020). GraphAF: A Flow-based Autoregressive Model for Molecular Graph Generation. *ArXiv:2001.09382 [Cs, Stat]*. http://arxiv.org/abs/2001.09382

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, *28*(1), 31–36. https://doi.org/10.1021/ci00057a005

Xiong, J., Xiong, Z., Chen, K., Jiang, H., & Zheng, M. (2021). Graph neural networks for automated de novo drug design. *Drug Discovery Today*, *26*(6), 1382–1393. https://doi.org/10.1016/j.drudis.2021.02.011

Zhou, Z., Kearnes, S., Li, L., Zare, R. N., & Riley, P. (2019). Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports*, *9*(1), 10752. https://doi.org/10.1038/s41598-019-47148-x

Appendix A

Molecular Data Representations

Three major classes of representations have typically been used in literature to describe molecular structures to computers: canonical SMILES strings, molecular fingerprints, and discrete graphs.

Canonical SMILES String Simplified Molecular Input Line Entry System, or SMILES, is a system developed by Weininger et al. (1988) to methodically convert from a molecular structure to a sequence of letters, numbers, and characters. The term *canonical SMILES* refers to a string which has been standardized to a canonical ordering of terms so as to create a one-to-one mapping between SMILES and molecular structures. Canonicalized forms are more preferable over loosely ordered ones for machine learning tasks, as one-to-one relationships are intuitively easier to learn. Since SMILES strings are essentially text, traditional language-processing models such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) can operate directly on them; however, this approach lacks consideration of basic chemical principles.

Discrete Graph In the study of discrete mathematics and graph theory, graphs refer to collections of objects, commonly known as vertices or nodes; and edges, which serve as connections between two endpoint vertices. Formally, a graph G can be expressed as $G = (V, E)$, where $V$is the set of vertices, and E is the set of edges. By formulating the atoms of a molecule as vertices and edges as bonds between those atoms, discrete graphs serve as a highly intuitive molecular representation.

Machine learning models to process discrete graphs were previously discussed, but Gilmer et al. (2017) were the first to demonstrate their effectiveness as molecular representations.

Appendix B

Evaluation Metrics

ADME is a fundamental drug design principle and broadly refers to 4 concepts: absorption, distribution, metabolism, and excretion.

clogP refers to a molecule's calculated partition coefficient and is a measure of lipophilicity.

QED stands for quantitative estimate of drug-likeness and is a heuristic measure of a drug molecule's "chemical beauty," developed by Bickerton et al., (2012).