JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Editorial

Generative Models for Molecular Design

Cite This: J. Chem. Inf. Model. 2020, 60, 5635−5636

Read Online

ACCESS | Metrics & More | Article Recommendations

Downloaded via UNIV OF SOUTHERN CALIFORNIA on March 23, 2022 at 03:36:12 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

The design and discovery of new molecules, materials, and drugs can lead to tremendous scientific, technological, biomedical, and societal advances. Yet it takes on average more than 10 years and over two billion dollars to put a drug on the market. A critical step in molecular design is the generation of a pool of candidates for computational characterization before experimental synthesis. This is a daunting task because the space of possible molecules is not numerable, and thus, cannot be searched completely. For example, the number of potential drug-like compounds has been estimated to be between 10^{23} and 10^{60}, while the number of all compounds that have been synthesized is only on the order of 10^{8}. Machine learning, including deep learning, has had tremendous success in science, engineering, industry, and finance. Deep learning techniques can reveal structure−property relationships and in turn, be used to guide the design of new lead molecules. In molecular design and drug discovery, much effort has been put into developing generative models, such as autoencoders (AEs), generative adversarial networks (GANs),[5] and reinforcement learning (RL). Many of these models are inspired by generative models for image compression, denoising, inpainting, texture synthesis, image translation, natural language processing, and other tasks. From the point of view of machine learning, most generative models in molecular design and drug discovery are based on self-supervised learning (SSL) or unsupervised learning, while semisupervised learning and supervised feature learning have also been developed. Generative models have emerged as a potential game-changer of molecular design and drug discovery.[7] The past few years have witnessed rapid development in generative models, driven by technical challenges, practical needs, and the great promise of these technologies. To help spur further development of generative models in molecular sciences, we introduce a collection of papers that represent the current state-of-art in generative models for molecular design.

Many generative models create simplified molecular-input line-entry system (SMILES) strings by using AEs, which consist of encoders and decoders, equipped with recurrent neural networks (RNNs), related gated recurrent units (GRUs), long short-term memory (LSTM), or bidirectional encoder representations from transformers (BERT). Autoencoders are usually pretrained with a large unlabeled data set to learn the general syntax of SMILES, followed by a fine-tuning training procedure with a small data set to prepare a model for target-specific applications. The statistical distribution learned from the large data set can be transferred to the small data set, which is termed transfer learning. Difficulties arise as to how to create a well-performing transfer learning application and how to select the size of the small data set. These issues have been addressed by Amabilino et al.[1]

Gao et al. propose a generative network complex (GNC) for automated molecular design and generation.[3] Their GNC allows the creation of novel drug-like molecules with desirable chemical, biological, and druggable properties, such as binding affinity, solubility, partition coefficient, and synthesizability. A GRU-based autoencoder is used for unsupervised training of SMILES data. The trained model is used to generate latent space representations of specific drug targets. Druggable properties of latent space representations are optimized with multiple pretrained machine learning predictors that determine the quality of potential drug candidates. Latent-space optimized drug candidates are forwarded to the decoder to generate SMILES strings, which are further filtered or reevaluated by another set of SMILES-based machine learning predictors. Finally, the consensus of the latent space predictors and SMILES-based predictors are used to recommend new drug candidates. Gao et al. have produced thousands of novel alternative drug candidates for eight existing marketed drugs, including Ceritinib, Ribociclib, Acalabrutinib, Idelalisib, Dabrafenib, Macimorelin, Enzalutamide, and Panobinostat. This approach can potentially be used to create future drug molecules with desirable pharmacological and commercial (e.g., cost of manufacture) properties.

Most SMILES-based AE models generate drug candidates without the explicit use of drug target structural information. Boitreaud et al. developed a variational AE (VAE) strategy that leverages target structural information to optimize the binding affinities of potential drug candidates.[2] Their approach couples the VAE and molecular docking software to guide compound optimization in an iterative manner such that the generated compounds fit the target structure better and better, without prior knowledge of bioactivities. Using this docking-enhanced VAE method, the authors can guide the design of molecules toward high docking scores, leading to a higher enrichment of drug leads.

Hierarchical relations among drugs, such as drug targets for organs, disease, or proteins can be a valuable source for computational drug repurposing and drug side-effect prediction. However, such hierarchical relations have not been

Special Issue: Generative Models for Molecular Design

Published: December 28, 2020

ACS Publications © 2020 American Chemical Society

https://dx.doi.org/10.1021/acs.jcim.0c01388
J. Chem. Inf. Model. 2020, 60, 5635−5636

widely utilized in machine learning-based drug design and discovery. Yu et al. develop a semisupervised drug embedding to incorporate hierarchical relations in supervised fine-tune training of a VAE.[8] The hierarchical relations among approved drugs, such as drug−drug similarity, are created by experts, providing valuable information for drug repositioning and the discovery of new side-effects of existing drugs. The proposed method compares favorably with other competitive approaches.

GANs have received tremendous attention in the past few years. This approach can generate novel molecules that are similar to the existing training data set in terms of statistics or distributions determined by the Kullback−Leibler divergence. Obviously, the similarity between distribution can be measured in a variety of ways, leading to various methods, such as the Wasserstein GAN (WGAN) which utilizes the Wasserstein distance. If the training data set is conditioned with certain labels, one arrives at a conditional GAN (CGAN). These approaches can be tailored for specific needs in molecular design, including protein folds with desirable functions. Karimi et al. have developed guided, conditional WGANs (gcWGANs) to design protein sequences for structural folds.[6] They construct a low-dimensional and generalizable representation of the fold space. Ultrafast sequence-to fold prediction is incorporated into WGAN as a loss to guide model training. Sequence data are used in a semisupervised training strategy to better control the protein sequence design.

As discussed above, there are many goal-directed generative models that can be utilized to optimize many druggable properties, such as binding affinity, solubility, distribution coefficient, similarity, etc. in molecular design and drug discovery. However, most methods pay little attention to the synthesizability generated compounds. The synthesizability can be an important obstacle in drug discovery. Gao and Coley have proposed a data-driven computer-aided synthesis planning program to quantify whether a generated molecule can be easily synthesized.[4] They show that synthetic complexity heuristics can successfully generate a set of molecules that are synthetically tractable judged by an open-source computer-aided retrosynthesis analysis tool, ASKCOS. However, some of this set of compounds may not be druggable due to compromises in other important properties.

Kenneth M. Merz, Jr. orcid.org/0000-0001-9139-5893
Gianni De Fabritiis orcid.org/0000-0003-3913-4877
Guo-Wei Wei orcid.org/0000-0002-5781-2937

AUTHOR INFORMATION

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.0c01388

Notes

Views expressed in this editorial are those of the authors and not necessarily the views of the ACS.

REFERENCES

(1) Amabilino, S.; P., Pog'any; Pickett, S. D.; Green, D. V.. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00343

(2) Boitreaud, J.; Mallet, V.; Oliver, C.; Waldispuhl, J.. Optimol: Optimization of binding affinities in chemical space for drug discovery. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00833

(3) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W.. Generative network complex for the automated generation of druglike molecules. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00599

(4) Gao, W.; Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00174

(5) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, 2672−2680.

(6) Karimi, M.; Zhu, S.; Cao, Y.; Shen, Y. De novo protein design for novel folds using guided conditional wasserstein generative adversarial networks. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00593

(7) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361* (6400), 360−365.

(8) Yu, K.; Visweswaran, S.; Batmanghelich, K. Semi-supervised hierarchical drug embedding in hyperbolic space. *J. Chem. Inf. Model.*, **2020**. DOI: 10.1021/acs.jcim.0c00681