RSC
Medicinal Chemistry

REVIEW

De novo molecular drug design benchmarking

Cite this: *RSC Med. Chem.*, 2021, **12**, 1273

Received 7th March 2021,
Accepted 24th May 2021

DOI: 10.1039/d1md00074h

rsc.li/medchem

Lauren L. Grant and Clarissa S. Sit *

De novo molecular design for drug discovery is a growing field. Deep neural networks (DNNs) are becoming more widespread in their use for machine learning models. As more DNN models are proposed for molecular design, benchmarking methods are crucial for the comparision and validation of these models. This review looks at recently proposed benchmarking methods Fréchet ChemNet Distance, GuacaMol and Molecular Sets (MOSES), and provides a commentary on their future potential applications in *de novo* molecular drug design and possible next steps for further validation of these benchmarking methods.

Introduction

The prevalence and incidence of multidrug-resistant bacteria has increased drastically in recent decades, while new antibiotic development has lagged.[1] Similarly, drug resistance has become rampant in cancer treatment. Today, 90% of chemotherapy failures are due to drug resistant cancer metastasis and invasion.[2] To find new drug candidates, high-throughput *in vitro* chemical screening has been used to test large physical libraries of up to 10^7 compounds for their biological activity.[3] However, it has been estimated that 10^{30} to 10^{60} potential organic compounds exist in chemical space.[4]

De novo molecular design has the capacity to explore all of chemical space efficiently by generating a small number of molecules using search and optimization procedures.[5] Because of this, *de novo* drug design has the potential to revolutionize medicinal chemistry and drug discovery. *De novo* molecular design can be based on a receptor, often using known protein structures to find molecules that fit well in binding pockets. Alternatively, *de novo* designed molecules can be built "from scratch" from different ligands.[6]

Deep neural networks (DNNs) have existed for a few decades, but their application in *de novo* molecular design is a more recent development. It was only within the past decade that DNNs have been shown to outperform more traditional machine learning methods and, since 2012, DNN based models have won multiple competitions in categories such as image classifications and molecular activity predictions.[7]

DNNs are often defined as having more than three layers.[7] The layers of DNNs fit into three categories: an input layer, hidden layers, and an output layer. Each layer is made up of nodes. The input layer will contain as many nodes as there are features. The output layer, if it is a classifier, contains as many nodes as there are classes. The hidden layers can have different numbers of nodes.[8] Each node will have an activation based on an input signal. This activation will tell a node whether to "fire", in turn sending an input signal to the next node.[8,9] A basic representation of nodes in a deep neural network can be seen in Fig. 1.

Different DNNs feature different network architectures such as recurrent neural networks (RNNs), fully connected neural networks (FCNNs), and convolutional neural networks (CNNs). Different architectures have different basic structures. The way nodes are connected between input, hidden, and output layers will vary with different architectures.[8,10]

All DNN *de novo* molecular generators go through three steps while producing new molecules. First, molecules are created, then they are scored and, finally, new and better molecules are searched for.[5,11]

Molecules are typically created from ligands or fragments.[12] However, methods for molecule construction using *in silico* chemical reactions have also been published with the goal of increasing the synthetic accessibility of proposed *de novo* molecules.[13] Once created, molecules need to be scored by the model.

The way a model scores molecules will depend on whether the model is receptor or ligand based. Models that suggest

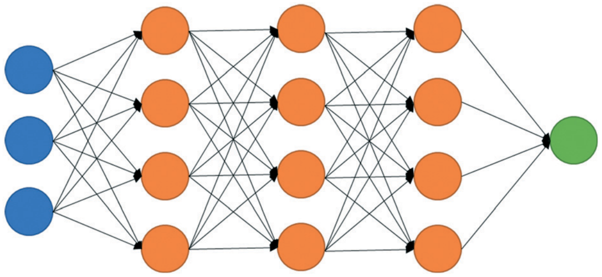

Fig. 1 A basic representation of a deep neural network where the blue circles represent the input layer, the orange circles represent nodes in the hidden layers, and the green circle represents the output layer. Each layer is connected to the next and each node of the previous layer will have an impact on the activation of the nodes in the next layer.

Saint Mary's University, Halifax, NS, Canada. E-mail: clarissa.sit@smu.ca

new molecules to bind to a known 3D structure of a protein will be scored on how well molecules bind to and fit into a protein site.[12,14] These molecular docking scores can be based on molecular force fields, empirical scoring functions, or knowledge-based scoring functions, which are discussed more in depth by Hartenfeller and Schneider.[12]

Ligand based models do not use a protein structure to build new molecules. In order to score molecules, these models can use a reference compound for scoring similarity, with the hypothesis being that similar molecules will have similar pharmacological properties.[12] This is the basis of quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) modelling.[15] Scoring ligand based models can also include methods for predicting physicochemical properties or a combination of these scoring methods.[5]

After scoring molecules, models will look for novel and more effective molecules. This can be done using stochastic optimization methods where a local minimum for the most suitable molecule is found.[12] Models based on the construction of molecules by incrementally adding and scoring ligands have also been developed.[16]

The goal of using DNNs for *de novo* drug design is to create new lead compounds for drug discovery. Ideally, DNNs could be used to propose new compounds in underexplored areas of chemical space. Proposed molecules by DNNs would need to be chemically stable, synthesizable, and bioactive. As well, proposed molecules need to be accepted by medicinal chemists in order to be tested *in vivo* and *in vitro*.

As more *de novo* models for molecular design are proposed, standardized benchmarking procedures are needed to compare the efficacy of these models. Benchmarking models can test *de novo* methods on many different characteristics such as the novelty of the proposed molecules, validity, fragment similarity, internal diversity, and other criteria. Benchmarking will not only allow new *de novo* methods to be tested for their efficacy but provide guidance for future improvement of methods.

Recently, benchmarking methods such as Fréchet ChemNet Distance,[17] MOSES,[18] and GuacaMol[5] have been proposed. They represent a first step towards standardized benchmarking procedures for DNN *de novo* drug design models. However, as will be discussed in this review, there is a lack of research into the biases these benchmarks can suffer from, including biased scoring functions for novelty that Renz *et al.*[19] highlights. Investigation into whether models that scored well on benchmarks produced molecules medicinal chemists can actually synthesize and test is another area that should be explored in the future.

This review will give an overview of the methods used by the benchmarking models Fréchet ChemNet Distance,[17] MOSES,[18] and GuacaMol[5] before discussing where these benchmarks could be improved and where future work is needed to test whether medicinal chemists agree with the scoring outcomes of these benchmarks.

Fréchet ChemNet Distance

Preuer *et al.*[17] developed Fréchet ChemNet Distance (FCD) as a metric to evaluate generative models. FCD was designed to encompass validity, and chemical and biological meaningfulness into one score. FCD was modelled after Fréchet Inception Distance (FID),[20] which is a metric for comparing generative models for images. FCD calculates the distance between the distribution of real-world molecules and the distribution of molecules produced by a generative model. Preuer *et al.*[17] calculated numerical representations of molecules using the penultimate layer of the neural network ChemNet, a long short-term memory (LSTM) RRN that is based on SMILES representations of molecules. SMILES (simplified molecular input line entry system)[21] was designed specifically for computers to represent molecules in two dimensions using only letters to represent atoms and special characters to represent different types of bonds. ChemNet was trained to predict the bioactivities of molecules based on approximately 6000 bioassays found in the drug databases ChEMBL,[22] Zinc,[23] and PubChem.[24]

From the activation of the penultimate layer of ChemNet, mean and covariance are calculated for both the distribution of real-world molecules and the distribution of molecules from a generative model. The two distributions are then compared using Fréchet distance,[25] also called Wasserstein-2 distance.[26] Preuer *et al.*[17] carried out hyperparameter selection and training using two-thirds of available data, with the last third being saved for testing. They state that large enough data sets are needed for both real and generated molecules to estimate mean and covariance.

Preuer *et al.*[17] compared FCD to four common metrics for evaluating generative models. The metrics were mean $\log P$, mean druglikeness, mean synthetic accessibility (SA) score, and the internal diversity score with Tanimoto distance. Preuer *et al.*[17] wanted to show that these metrics have specific flaws and fail to detect biases in generative models. To do this, they manipulated models to produce molecules with either low drug likeness, a high $\log P$, low synthetic accessibility, mode collapse (low internal diversity), or bias toward certain target families for models designed to produce molecules active for a specific target. These manipulated models were compared to real molecules.

FCD was the only evaluation method able to consistently score the manipulated models worse than the real molecules. Preuer *et al.*[17] also tested their hypothesis that chemical information alone is not enough to test a generative model. To do this they tested what they termed the Fréchet Fingerprint Distance (FFD), which was based on 2048 bit ECFP_4 fingerprints and on purely chemical representation of generated molecules. They found that FCD was able to make stronger distinctions between the manipulated and real sets compared to FFD. This was especially true when looking at biologically relevant information such as when the manipulated model was biased towards a certain target family.

Applications of FCD in the literature

FCD has been incorporated into other benchmarking tools such as GuacaMol and MOSES, which are also discussed in this review. On its own FCD has also been used to compare different models for *de novo* drug design. Skalic *et al.*[27] used FCD as part of their evaluation and comparison of their shape-based generative model to other generative models such as an adversarial autoencoder (AAE), a character-level recurrent neural network (CharRNN), a variational autoencoder (VAE), and an objective-reinforced generative adversarial network (ORGAN). The shape-based model was found to be comparable to other methods tested.

Grisoni *et al.*[28] also used FCD as part of their evaluation for their new method for SMILES string generation. In their new model, SMILES strings were generated both left-to-right (forwards) and right-to-left (backwards). In their BIMODAL (bidirectional molecule design by alternate learning) model, two RNN's are used. One RNN reads the SMILES strings forward and another backwards. These are then combined for a joint prediction.

For evaluation of their method, Grisoni *et al.*[28] looked at structural novelty defined as "not contained in the training set", uniqueness defined as the percentage of unique SMILES strings generated, and validity defined as the percentage of chemically valid SMILES strings generated. Grisoni *et al.*[28] also looked at scaffold diversity and novelty using Bemis–Murcko scaffolds.[29] Finally, they used FCD to evaluate the chemical and biological relevance of the generated SMILES strings.

For the FCD evaluation, Grisoni *et al.*[28] found the BIMODAL model had a slightly worse FCD score than the traditional forward running RNN when there were 512 hidden units. However, when there were 1024 hidden units, the forward RNN and BIMODAL had comparable scores, with both improving from their 512 scores. This combined with BIMODAL's good scaffold diversity scores made Grisoni *et al.*[28] conclude that the BIMODAL model was worth further investigation.

GuacaMol

GuacaMol[5] uses a set of benchmarks to assess a model's ability to learn from a data set of molecules and to create new molecules with similar properties. GuacaMol[5] splits models for *de novo* molecular design into two categories: distribution-learning and goal-directed. Distribution-learning models aim to generate new molecules based on the chemical distribution of a training set of molecules. Goal-directed models are designed to generate molecules for a specific goal.

GuacaMol[5] evaluates distribution-learning models based on validity, uniqueness, novelty, FCD, and the Kullback–Leibler (KL) divergence. Validity is an assessment of whether the generated molecules are realistic, at least theoretically. The validity benchmark scores molecules lower for incorrect SMILES syntax or if they have an invalid valence. Uniqueness measures the ability of a model to produce unique molecules. If a model produces a molecule more than once, it is penalized.

For novelty, GuacaMol uses the ChEMBL training set to represent a tiny portion of chemical space. Brown *et al.*[5] state that a good model for *de novo* molecular design should be able to explore a large part of chemical space and would, therefore, be unlikely to reproduce molecules from the training set. Models that overfit receive a low score on this task. However, a bad model could potentially score well on this benchmarking task if that model produces many different simple molecules, such as carbon chains.

The FCD is measured as described previously. Lastly, the KL divergence is a measure of how well a probability distribution approximates another distribution. For GuacaMol, KL divergence is used to measure how diverse the generated molecules are from the training set. Low diversity will lead to a low KL divergence score. All scores are on a scale of 0 to 1.

Since goal-directed models are trying to design one optimized molecule to complete a specific task, generated molecules need to be scored individually. For evaluation, GuacaMol[5] has goal-directed models produce a set number of high scoring molecules. The models are allowed to iteratively improve their molecules using the scoring function, but the models do not have access to what the scoring function is explicitly calculating. The optimizing function of a model is evaluated by looking at a combination of the structural features (molecular weight, number of aromatic rings, *etc.*), physicochemical properties, similarity or dissimilarity to other molecules, and presence or absence of substructures, functional groups, or atom types.

For similarity, the model is assessed on its ability to produce molecules that are similar to a target compound. Brown *et al.*[5] describe this as a sort of inverse virtual screening because, instead of looking up molecules in a large data base, molecules are generated based on a target molecule. A rediscovery benchmark is also included. This is like the similarity benchmark, except the goal is to rediscover the target molecule and not produce many molecules like it. An isomer benchmark is also included for which the model is tasked with producing isomers from a given molecular formula. This is used to test the flexibility of the model. Lastly, the median molecules benchmark is included to assess a models ability to produce a molecule that is similar to several different molecules. This task is designed to be conflicting, as it is useful for assessing how a model explores chemical space.

GuacaMol[5] also includes rule sets for assessing the quality of the compounds produced by distribution-learning and goal-directed models. A model that suggests molecules that are potentially unstable, reactive, or too difficult to synthesize will not be used by medicinal chemists. The set of rules used by GuacaMol are designed to determine if suggested molecules could be included in a high-throughput screen. Brown *et al.*[5] note that the list of filters they include for assessing compound quality is probably incomplete. However, it is likely that any molecules filtered out by the rule set would not be chosen for synthesis.

Brown *et al.*[5] used GuacaMol to evaluate and compare different types of distribution-learning and goal-directed models. For the generative-models, a random sampler was included for a baseline comparison. The random sampler took a set number of molecules from the training set at random for evaluation. The random sampler showed high values for KL divergence and FCD, which were considered to be baselines for good models. The GuacaMol[5] benchmark results showed that the SMILES LSTM model was the most consistently high-scoring model across benchmarks.

For comparison of goal-directed models, a "Best of Data Set" was used. The "Best of Data Set" used the highest scoring molecules in the data set. These molecules are used to set the bar for the goal-directed models because the purpose of goal-directed models is to return molecules that are better than the original data set. If models do not perform better than the best data from the data set, then they provide no advantage over virtual screening. Of the tested models, the graph genetic algorithm performed the best on all the benchmarks.

Applications of GuacaMol in the literature

Winter *et al.*[30] tested their proposed method for optimizing molecules, molecular swarm optimization (MSO), using GuacaMol. MSO uses a 'light weight' heuristic optimization method, particle swarm optimization (PSO), proposed by Hartenfeller *et al.*[31] applied to their previously reported continuous chemical representation.[32] Winter *et al.*[30] retrained their model using the same subset of ChEMBL originally used for the GuacaMol benchmark. Winter *et al.*[30] found that their method had a higher average score than the baseline models included in GuacaMol.

Kwon *et al.*[33] proposed an improved VAE method for efficient molecular graph generation. To improve the molecular graph generation, they included three additional learning objectives. These objectives were: approximate graph matching, reinforcement learning, and auxiliary property prediction. Kwon *et al.*[33] tested their generative model for validity, uniqueness, novelty, KL divergence, and FCD. On the tests for validity, uniqueness, and novelty, the proposed method performed as well as or out-performed the baseline SMILES models, LSTM, VAE, AAE, and ORGAN and the molecular graph generation model GraphMCTS. Their model did not perform well on KL divergence or FCD, which suggests that their model was not able to reproduce the property distribution of the training set. Overall, the graph method was superior to the SMILES string generators at producing chemically valid and diverse molecules but struggled to represent the data-distribution of the training set.

EvoMol[34] is another molecular generation model. It sequentially builds molecular graphs using an evolutionary algorithm. This algorithm was developed with the goal to explore both known and unknown chemical space. Leguy *et al.*[34] used the goal-directed benchmarks from GuacaMol to assess EvoMol.

EvoMol performed exceptionally well on the isomer benchmark task in comparison to other molecular generators. EvoMol also outperformed other models on the multi-property objective (MPO) benchmark. Overall, EvoMol performed very well on the GuacaMol benchmark.

Molecular Sets (MOSES)

Molecular Sets (MOSES)[18] is another benchmarking suite for molecular generators. It includes a standardized dataset and evaluation metrics. MOSES compares distribution-learning models using fragment similarity, scaffold similarity, nearest neighbour similarity, internal diversity, and Fréchet ChemNet Distance (FCD). These metrics are used to evaluate how well a generative model approximates an unknown distribution when given a set of training samples from the unknown distribution. MOSES first computes the validity of generated molecules and then only evaluates valid molecules.

Valid molecules are determined using RDKit, which can be used to evaluate molecules for proper atom valency. The authors suggest that, for evaluation, 30 000 molecules should be generated, and molecules labelled valid should make up the generated set for comparison with the reference set.

The benchmarks that compare the generated set to the reference set include novelty. This metric calculates the number of molecules in the generated set that are not present in the reference set. Novelty can be used to evaluate if a model is overfitting. Next, fragment similarity looks at the distribution of BRICS fragments in the generated set compared to the reference set. BRICS is a set of rules for breaking up chemical substructures of biologically active compounds that was developed for medicinal and computational chemists.[35] The fragment similarity metric is large if the generated set has similar fragments to the reference set. If the generated set over (or under) represents fragments from the reference set, then the fragment similarity value will be low. Scaffold similarity is another metric; it operates in the same manner as fragment similarity except using Bemis–Murcko scaffolds instead of BRICS fragments. Bemis–Murcko scaffolds is a list of common drug shapes in a graphical representation based on rings, linker atoms, and sidechains.[29]

MOSES also calculates similarity to a nearest neighbour (SNN) based on Tanimoto similarity between the fingerprint of molecules in the generated set and its nearest neighbour molecule in the reference set. This is used as a precision metric: the lower the score the poorer the precision of the model. MOSES also calculates the FCD for comparison between the generated set and the reference set.

MOSES also looks at various property distributions between the generated and reference sets. The molecular weight distribution shows whether the generated set is biased towards either heavier or lighter molecules. $\log P$ and synthetic accessibility (SA) scores are also included. Lastly, the quantitative estimation of drug-likeness (QED) score is included, which evaluates molecules on how likely they are to be viable drug candidates on a scale of 0 to 1. QED is meant to approximate the abstract knowledge held by medicinal chemists on what molecules may be good drug candidates and was proposed by Bickerton *et al.*[36] However, a limitation of QED is that it is only dependent on molecular properties such as molecular weight, octanol–water partition coefficient, number of hydrogen bond donors, and number of hydrogen bond acceptors. Factors such as absorption, distribution, metabolism, excretion, and toxicity (important *in vivo* and *in vitro* properties) are not considered.[37]

Polykovskiy *et al.*[18] applied MOSES to multiple different models that covered many different approaches to molecular generation. The tested models included character-level neural networks (CharRNN), variational autoencoders (VAE), adversarial autoencoders (AAE), junction tree variational autoencoders (JTN-VAE), LatentGAN, and non-neural baselines. The non-neural baselines included a hidden Markov model (HMM), an n-gram model which collects the frequency of n-grams in a training set and uses the distribution for new strings, and a combinatorial generator that uses BRICS fragments of the training set to randomly assemble new molecules.

Polykovskiy *et al.*[18] found that the VAE and AAE models had low novelty scores, which is an indication of overfitting. The VAE had the best SNN score, but that was likely due to overfitting. The best FCD, fragment, and scaffold scores were obtained by the CharRNN model. Because of this, the CharRNN was concluded to be the best model tested. This model did not have a problem with overfitting but was still able to represent the distribution of the training set. Polykovskiy *et al.*[18] conclude that MOSES will allow for fair and comprehensive evaluation of generative models for *de novo* drug design. They also plan to update MOSES with new baseline models and evaluation metrics in the future.

Applications of MOSES in the literature

In addition to calculating FCD as discussed previously, Skalic *et al.*[27] also used the MOSES benchmark to evaluate their shape-based generative model. Skalic *et al.*[27] compared their shape based model to models tested by Polykovskiy *et al.*,[18] including the CharRNN, VAE, AAE, ORGAN, and JTN-VAE. It was found that Skalic *et al.*'s[27] model performed similarly to the other models with good validity and uniqueness scores. The shape-based model produced molecules with the highest level of internal diversity, but this seemed to be at the cost of lower compound desirability and increased reactivity.

Boitreaud *et al.*[38] proposed a new graph to selfies VAE. This graph2selfies model was tested using the MOSES benchmarking platform. Graph2selfies performed well on all benchmarks and outperformed the state-of-the-art graph to graph model JTVAE while also being 18 times faster.

Issues with current benchmarking models

One issue with benchmarking platforms, which has been brought up by Renz *et al.*,[19] is the copy problem. Renz *et al.*[19] benchmarked a model they called AddCarbon. This generative model creates "new" molecules by taking random molecules from the training set and adding one carbon randomly in the SMILES string. As long as the new SMILES string is valid and is not a molecule already in the training set it is used as a new random sample.

The AddCarbon model was tested using GuacaMol and received a perfect score for novelty, validity, and uniqueness and scored high on KL divergence. The AddCarbon model outperformed all the baseline models except for the LSTM model.

Renz *et al.*[19] concluded that since their "useless" model scored so well on the GuacaMol benchmark, it calls into question the usefulness of the benchmark. They suggest that better metrics for quantifying novelty would be beneficial.

Renz *et al.*[19] also looked at the shortcomings of benchmarking goal-directed models. It is difficult to compile all the desirable qualities of a molecule into one score, and molecules generated to optimize a specific score might not be useful. Renz *et al.*[19] provide a few examples of molecules produced by various goal-directed models that score well on GuacaMol's benchmarks, but have unstable, synthetically unrealistic, or highly uncommon substructures.

There were two main biases explored for goal-directed models: model specific biases and data specific biases. Model specific biases exploit unique features of a model while failing to capture the actual desired characteristics that model is supposed to capture. Data specific biases refers to how models will perform much better on data they have been trained on compared to hold-out data.

Renz *et al.*[19] showed both model specific biases and data specific biases by comparing an optimization score to a model control score and a data control score. They took a set of data and split it in half. The optimization score was calculated from a classifier trained on split 1 and optimized. The model control score was calculated from a second classifier trained on split 1 but using a different random seed than the optimized score model. Lastly, the data control score was calculated by training a third classifier using split 2 and then using this model to score optimized molecules from split 1. This was done to see if a model trained on different data scores molecules similarly.

In all cases, the optimized score was higher than model control scores, and data control scores were even lower. This showed that model and data specific biases are likely present in optimization. Renz *et al.*[19] point out that since the goal of *de novo* design is to explore all of chemical space, data specific biases suggests models are failing.

Another major issue with *de novo* molecular design benchmarking is the synthesizability of molecules proposed by high scoring models. A model may preform well on benchmarking platforms such as GuacaMol[5] and MOSES,[18] but if proposed molecules are synthetically inaccessible, they are useless. Gao and Coley[39] found that, especially for goal-directed models, a high score on benchmarking does not mean the model will produce synthetically accessible molecules.

Specifically, of the multiple goal-directed models tested by Brown *et al.*,[5] the graph genetic algorithm model scored the best. This model was found by Gao and Coley[39] to produce no synthesizable molecules when tested by their data-driven computer-aided synthesis planning program. The SMILES genetic algorithm also scored well when tested by GuacaMol,[5] but did not produce any synthetically accessible molecules when tested by Gao and Coley.[39]

Since benchmarking methods such as GuacaMol[5] and MOSES[18] are new, there is not a lot of research evaluating their efficacy. Therefore, it is difficult to determine if medicinal chemists would agree with the scores given by these benchmarks to various *de novo* molecular generators.

In another study, Bush *et al.*[40] evaluated three molecular generators on their ability to produce molecules deemed acceptable by medicinal chemists. To do this, they used three tests.

For the first test, 13 medicinal chemists were asked to suggest 20 molecules to explore the structure–activity relationship based on four hit molecules. These suggested molecules were meant to capture human "ideation". The four hit molecules were then fed into the molecular generators being tested. The molecular generators were then evaluated on their ability to propose the "ideal" molecules suggested by medicinal chemists.

The second test had medicinal chemists evaluate molecules suggested by the algorithms. For each hit molecule, medicinal chemists were presented with 75 ideas generated algorithmically and 25 ideas designed by medicinal chemists. Molecules were categorized as "like" or "dislike" based on whether the medicinal chemist would consider each molecule for synthesis.

For the last test, molecular generators were evaluated on their ability to generate molecules from six drug patents. One molecule was chosen as a seed molecule, typically the marketed drug, from each patent. The seed molecules were fed to each algorithm. Generated molecules were compared to the patent. Any generated molecules that matched molecules in the patent were refed to the algorithm. This was done in an iterative process meant to mimic design-make-test cycles.

Bush *et al.*[40] designed these tests to assist medicinal chemists in picking the best algorithms for generating new molecules. If a molecular generator is supported by medicinal chemists, it will hopefully produce more synthetically accessible molecules with better biochemical and physicochemical properties. Indeed Bush *et al.*[40] found that one of the algorithms tested was beneficial. This algorithm could potentially be incorporated into a computational based medicinal chemistry design procedure.

Of the three molecular generators tested by Bush *et al.*,[40] none were also tested by Preuer *et al.*,[17] Brown *et al.*,[5] or Polykovskiy *et al.*[18] In the future it would be interesting to compare results using methods proposed by Bush *et al.*[40] to the results of benchmarking models. If GuacaMol[5] or MOSES[18] are found to score molecular generators similarly to how medicinal chemists score them, this would be evidence that these benchmarks are practical and will be useful for assessing molecular generators developed in the future.

Conclusions

Herein we have discussed multiple benchmarking methods for *de novo* molecular design, which is a growing field in the realm of drug discovery. As such, there is a need for valid methods for comparing new models.

Preuer *et al.*[17] proposed the benchmark FCD. They also showed that including both chemical and biological information is important for testing generative models. FCD was incorporated into the benchmarking models GuacaMol and MOSES.

GuacaMol[5] and MOSES[18] seem to be, overall, more complete benchmarking platforms. Not only do they include datasets for testing, but also multiple different benchmarking metrics for comparing models. GuacaMol evaluates both distribution-learning and goal-directed models, while MOSES focuses primarily on distribution-learning models.

Both GuacaMol and MOSES mention the need for future improvement of their benchmarking platforms. Brown *et al.*[5] note that some of their benchmarking tasks were too easily solved by baseline models. This indicates a need for harder benchmarks. Polykovskiy *et al.*[18] note that in the future they will expand MOSES' repository with new baseline models and evaluation metrics.

The shortcomings of benchmarking techniques pointed out by Renz *et al.*[19] should be considered for future updates of GuacaMol and MOSES, and for any new benchmarking models proposed in the future.

To validate benchmarking models in the future it would be beneficial to compare how medicinal chemists score molecular generators using methods such as those proposed by Bush *et al.*[40] to scores generated by benchmarking models. One of the more difficult areas for benchmarking molecular generators is assessing synthetic accessibility. Generated molecules may score well on benchmarks if they are valid chemical structures that fit well into the desired distribution of molecules. However, if chemists cannot physically make suggested molecules for testing, then they are useless.

In the future, benchmarking suites could help computational medicinal chemists evaluate and improve models for *de novo* molecular drug design. This is crucial as computational models for drug discovery have the potential to unlock so much of chemical space not currently being explored by other methods such as high-throughput screening. However, further studies are needed to validate the efficacy of benchmarking models such a GuacaMol[5] and MOSES[18]

Author contributions

L. L. Grant was responsible for conceptualization and writing of the original draft. C. S. Sit was responsible for supervision, funding acquisition and writing (reviewing and editing) of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

L. L. Grant is supported by a Nova Scotia Graduate Scholarship. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference numbers 510963 and RGPIN-2017-05813].

References

1 M. Frieri, K. Kumar and A. Boutin, Antibiotic Resistance, *J. Infect. Public Health*, 2017, **10**(4), 369–378, DOI: 10.1016/j.jiph.2016.08.007.

2 B. Mansoori, A. Mohammadi, S. Davudian, S. Shirjang and B. Baradaran, The Different Mechanisms of Cancer Drug Resistance: A Brief Review, *Adv. Pharm. Bull.*, 2017, **7**(3), 339–348, DOI: 10.15171/apb.2017.041.

3 N. van Hilten, F. Chevillard and P. Kolb, Virtual Compound Libraries in Computer-Assisted Drug Discovery, *J. Chem. Inf. Model.*, 2019, **59**(2), 644–651, DOI: 10.1021/acs.jcim.8b00737.

4 W. P. Walters, Virtual Chemical Libraries: Miniperspective, *J. Med. Chem.*, 2019, **62**(3), 1116–1124, DOI: 10.1021/acs.jmedchem.8b01048.

5 N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, GuacaMol: Benchmarking Models for de Novo Molecular Design, *J. Chem. Inf. Model.*, 2019, **59**(3), 1096–1108, DOI: 10.1021/acs.jcim.8b00839.

6 B. Sattarov, I. I. Baskin, D. Horvath, G. Marcou, E. J. Bjerrum and A. Varnek, De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping, *J. Chem. Inf. Model.*, 2019, **59**(3), 1182–1196, DOI: 10.1021/acs.jcim.8b00751.

7 D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, Deep Learning for Molecular Design—a Review of the State of the Art, *Mol. Syst. Des. Eng.*, 2019, **4**(4), 828–849, DOI: 10.1039/C9ME00039A.

8 M. Sewak, S. K. Sahay and H. Rathore, An Overview of Deep Learning Architecture of Deep Neural Networks and Autoencoders, *J. Comput. Theor. Nanosci.*, 2020, **17**(1), 182–188, DOI: 10.1166/jctn.2020.8648.

9 P. Sibi, S. A. Jones and P. Siddarth, Analysis of Different Activation Functions Using Back Propagation Neural Networks, *Journal of Theoretical and Applied Information Technology*, 2005, **47**(3), 1264–1268.

10 T. Bouwmans, S. Javed, M. Sultana and S. K. Jung, Deep Neural Network Concepts for Background Subtraction:A Systematic Review and Comparative Evaluation, *Neural Netw.*, 2019, **117**, 8–66, DOI: 10.1016/j.neunet.2019.04.024.

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

12 M. Hartenfeller and G. Schneider, Enabling Future Drug Discovery by de Novo Design, *WIREs Comput. Mol. Sci.*, 2011, **1**(5), 742–759, DOI: 10.1002/wcms.49.

13 M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby and S. A. Renner, Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design, *J. Chem. Inf. Model.*, 2011, **51**(12), 3093–3098, DOI: 10.1021/ci200379p.

14 V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco and G. Melagraki, Advances in De Novo Drug Design: From Conventional to Machine Learning Methods, *Int. J. Mol. Sci.*, 2021, **22**(4), 1676, DOI: 10.3390/ijms22041676.

15 C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna and V. Prachayasittikul, A Practical Overview of Quantitative Structure-Activity Relationship, *EXCLI J.*, 2009, **8**, 74–88, DOI: 10.17877/DE290R-690.

16 J. Degen and M. Rarey, FlexNovo: Structure-Based Searching in Large Fragment Spaces, *ChemMedChem*, 2006, **1**(8), 854–868, DOI: 10.1002/cmdc.200500102.

17 K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter and G. Klambauer, Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery, *J. Chem. Inf. Model.*, 2018, **58**(9), 1736–1741, DOI: 10.1021/acs.jcim.8b00234.

18 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, *Front. Pharmacol.*, 2020, **11**, 565644, DOI: 10.3389/fphar.2020.565644.

19 P. Renz, D. Van Rompaey, J. K. Wegner, S. Hochreiter and G. Klambauer, On Failure Modes in Molecule Generation and Optimization, *Drug Discovery Today: Technol.*, 2019, **32–33**, 55–63, DOI: 10.1016/j.ddtec.2020.09.003.

20 M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, ArXiv Prepr., 2017, ArXiv170608500.

21 D. Weininger, SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Model.*, 1988, **28**(1), 31–36, DOI: 10.1021/ci00057a005.

22 A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak and S. McGlinchey, The ChEMBL Bioactivity Database: An Update, *Nucleic Acids Res.*, 2014, **42**(D1), D1083–D1090.

23 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, ZINC: A Free Tool to Discover Chemistry for Biology, *J. Chem. Inf. Model.*, 2012, **52**(7), 1757–1768.

24 Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He and J. Zhang, Pubchem Bioassay: 2017 Update, *Nucleic Acids Res.*, 2017, **45**(D1), D955–D963.

25 M. Fréchet, Sur La Distance de Deux Lois de Probabilité, *C. R. Hebd. Seances Acad. Sci.*, 1957, **244**(6), 689–692.

26 L. N. Wasserstein, Markov Processes over Denumerable Products of Spaces, Describing Large Systems of Automata, *Probl. Peredachi Inf.*, 1969, **5**(3), 64–72.

27 M. Skalic, J. Jiménez, D. Sabbadin and G. De Fabritiis, Shape-Based Generative Modeling for de Novo Drug Design, *J. Chem. Inf. Model.*, 2019, **59**(3), 1205–1214, DOI: 10.1021/acs.jcim.8b00706.

28 F. Grisoni, M. Moret, R. Lingwood and G. Schneider, Bidirectional Molecule Generation with Recurrent Neural Networks, *J. Chem. Inf. Model.*, 2020, **60**(3), 1175–1183, DOI: 10.1021/acs.jcim.9b00943.

29 G. W. Bemis and M. A. Murcko, The Properties of Known Drugs. 1. Molecular Frameworks, *J. Med. Chem.*, 1996, **39**(15), 2887–2893, DOI: 10.1021/jm9602928.

30 R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé and D.-A. Clevert, Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space, *Chem. Sci.*, 2019, **10**(34), 8016–8024, DOI: 10.1039/C9SC01928F.

31 M. Hartenfeller, E. Proschak, A. Schüller and G. Schneider, Concept of Combinatorial De Novo Design of Drug-like Molecules by Particle Swarm Optimization, *Chem. Biol. Drug Des.*, 2008, **72**(1), 16–26, DOI: 10.1111/j.1747-0285.2008.00672.x.

32 R. Winter, F. Montanari, F. Noé and D.-A. Clevert, Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations, *Chem. Sci.*, 2019, **10**(6), 1692–1701, DOI: 10.1039/C8SC04175J.

33 Y. Kwon, J. Yoo, Y.-S. Choi, W.-J. Son, D. Lee and S. Kang, Efficient Learning of Non-Autoregressive Graph Variational Autoencoders for Molecular Graph Generation, *Aust. J. Chem.*, 2019, **11**(1), 70, DOI: 10.1186/s13321-019-0396-x.

34 J. Leguy, T. Cauchy, M. Glavatskikh, B. Duval and B. Da Mota, EvoMol: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de Novo Molecular Generation, *Aust. J. Chem.*, 2020, **12**(1), 55, DOI: 10.1186/s13321-020-00458-z.

35 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces, *ChemMedChem*, 2008, **3**(10), 1503–1507, DOI: 10.1002/cmdc.200800178.

36 G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, Quantifying the Chemical Beauty of Drugs, *Nat. Chem.*, 2012, **4**(2), 90–98, DOI: 10.1038/nchem.1243.

37 L. Guan, H. Yang, Y. Cai, L. Sun, P. Di, W. Li, G. Liu and Y. Tang, ADMET-Score – a Comprehensive Scoring Function for Evaluation of Chemical Drug-Likeness, *MedChemComm*, 2019, **10**(1), 148–157, DOI: 10.1039/C8MD00472B.

38 J. Boitreaud, V. Mallet, C. Oliver and J. Waldispühl, OptiMol: Optimization of Binding Affinities in Chemical Space for Drug Discovery, *J. Chem. Inf. Model.*, 2020, **60**(12), 5658–5666, DOI: 10.1021/acs.jcim.0c00833.

39 W. Gao and C. W. Coley, The Synthesizability of Molecules Proposed by Generative Models, *J. Chem. Inf. Model.*, 2020, **60**(12), 5714–5723, DOI: 10.1021/acs.jcim.0c00174.

40 J. T. Bush, P. Pogany, S. D. Pickett, M. Barker, A. Baxter, S. Campos, A. W. J. Cooper, D. Hirst, G. Inglis, A. Nadin, V. K. Patel, D. Poole, J. Pritchard, Y. Washio, G. White and D. V. S. A. Green, Turing Test for Molecular Generators, *J. Med. Chem.*, 2020, **63**(20), 11964–11971, DOI: 10.1021/acs.jmedchem.0c01148.