The Synthesizability of Molecules Proposed by Generative Models

Multiple fields are interested in the discovery of new functional molecules. One of the most popular class of techniques nowadays is the *de novo* molecular generation and optimization, which use deep learning methods (GANs and VAEs included) for the generation of novel chemical structures with certain desired properties. These, combined with certain models such as Bayesian optimization or heuristic optimization algorithms, can even bias generation into producing chemical structures with a desired functionality.

These *de novo* molecular generation methods bypass some of the problems seen before in molecular design strategies such as high-throughput virtual screening: they are less computationally intensive and inefficient due to not requiring the use of extremely large virtual libraries, and they are not limited by the chemical space of existing libraries (i.e. they can produce molecules not present in modern virtual libraries). However, due to the nature of these deep learning generative methods, they can produce molecules optimized against a certain task that are not synthesizable.

In the context of this problem, Coley and Gao analyzed and quantified the rate of synthesizability of these *de novo* generative models and observed what approaches might work to solve this issue; thus, showcasing the possible strategies and limitations of multiple approaches.

According to Coley and Gao, we can quantify synthesizability through structural methods, or through synthetic pathways. In this last one, the approach involves planning synthetic pathways and assess the likelihood that they are valid. One tool for this approach is the use of a *computer-aided synthesis planning program (CASP).* As this tool includes more information, it bypasses the shortcomings of the structural approach. It usually gives us a way to verify why the molecule is believed to be synthesizable, with which building blocks, and in how many steps. As this has not been studied enough, it is what this paper analyzes, using ASKCOS as the CASP.

Additionally, the paper divided the analysis of synthesizability into two evaluations:

1. **Distribution learning tasks (unoptimized molecules)**: meant to generate new molecules with a certain set of properties that resemble a finite set of training molecules.
2. **Goal-directed generation tasks (optimized molecules)**: meant to generate new molecules that maximize a scoring function that represents the suitability of the structure in a drug discovery setting (i.e. an objective). *In the context of generating novel chemical structures with a certain set of properties, this is the type of task we are concerned with.*

Finally, the paper focused in three major approaches to solving the synthesizability problem:

1. **Post hoc filtering.** We narrow down generated molecules as a separate step from generation (i.e. molecule is generated, then we check is it is synthesizable).
2. **A priori biasing.** We bias the training set by only using synthetically available compounds for the generative model.
3. **Heuristic biasing.** We include a proxy heuristic for synthesizability for the objective function (i.e. we include synthesizability in the error function with a score).

The results obtained can be divided in two parts.

1. Examination of quantifiers for synthesizability

First, Coley and Gan checked if ASKCOS usefully correlated with synthesizability by comparing it to a series of compound libraries (MOSES, ChEMBL, ZINC, Sheridan et al., and GDB17). They used a

random set of 3000 molecules from each library. They found that the predicted number of reaction steps by ASKCOS correlated with the expert provided scores. Also, from the trends found, ASKCOS seems to be consistent with the expectations of synthesizability for each library; thus, being appropriate to use for predictions regarding synthesizability.

Afterwards, they compared the ratios of synthesizable molecules of ASKCOS with a series of heuristic scores (SA_Score, SCScore and SMILES string length) that approximate a measure of synthesizability, using the same previous random sets of 3000 molecules; thus, observing how useful these heuristic metrics can be to measure synthesizability. The result was that while none of them can distinguish synthesizable and non-synthesizable compounds perfectly, all of them show a decreasing trend as the scores increase.

2. Evaluation of approaches to solve the synthesizability problem

First, Coley and Gan analyze by distribution learning tasks; at the same time, separating training distribution learning models between ChEMBL and MOSES (i.e. training set biasing, with MOSES consisting of more synthesizable molecules). They apply multiple distribution learning models (SMILES LSTM, VAE, and AAE) on the database used for learning, leading to 300 molecules produced by each distribution learning method. The results were that: (1) the distribution learning methods had similar fraction of synthesizability but did not improve it, (2) a priori biasing is a viable approach for distribution learning methods, as it can be seen from the differences between the results of both databases, (3) post-hoc filtering is not necessarily a bad approach.

For goal-directed tasks, Coley and Gan analyzed the synthesizability of molecules in goal-directed tasks (that is, generating new molecules with certain desired properties). For this, they chose a variety of generative algorithms (SMILES LSTM, SMILES GA, and Graph GA) and 14 multi-property objective functions or MPOs (i.e. functions that measured how "fitting" a molecule is with respect to a certain task in a single scalar value). They also apply post-hoc filtering and a priori biasing using two libraries: MOSES and ChEMBL. Additionally, they compare the results with and without heuristic biasing.

Thus, the optimization is as follows: they generate a series of molecules using a generative algorithm and an objective function and observe: (1) The score of the molecule with the highest objective function and (2) The ratio of synthesizable molecules in the top 100 compounds by objective function score. They repeat this process using different libraries and either including or not including heuristic biasing.

The results were very dependent on the generative methods and MPOs used. In some cases, no synthesizable molecules are produced in the top 100 without heuristic biasing (which may be a result of using post-hoc filters). It is noted, however, that in these cases even after heuristic biasing, the resulting compounds tend to have low objective values. The results also seem to suggest that heuristic biasing does an excellent job at improving synthesizability; although this may come in some cases at the expense of lower objective values. The use of a priori biasing did not have a significant effect on the synthesizability results. They also suggest a strategy: for a small number of candidates desired, optimize without heuristic biasing and then use post-hoc filtering to remove unsynthesizable molecules; and in the case that the top unsynthesizable molecules have a lower objective score that the synthesizable ones, we optimize again with heuristic biasing.

Finally, Coley and Gan present some additional limitations of these methods. As they explain, CASP tools such as ASKCOS are imperfect as they might depend on a series of settings and the database

of chemicals considered buyable. Nevertheless, ASKCOS represents an excellent tool for analyzing synthesizability. Additionally, while ASKCOS can be used instead of a heuristic bias, the tool itself is computationally expensive. As a conclusion, they highlight the need for improved CASP tools for post-hoc filtering, the development of new heuristics, the sampling of CASP oracles to bias generation through reinforcement learning and the design of new algorithms in order to increase the utility of these approaches to increase synthesizability.