Critical Review 1: Automatic chemical design using a data-driven continuous representation of molecules

Frank Cipollone, David Deriso, Gabriel Maher, Benjamin Nosarzewski, Nikhil Parthasarathy, and Kushal Ranjan

October 31, 2016

1 Problem Context and Background

At a high level, this paper attempts to come up with a method to help perform automatic search and optimization within the super high dimensional space of chemical compounds for the purpose of drug design. In recent years, the field of computer aided drug design (CADD) has gotten much attention because of the great reduction in cost and time it provides for both selecting candidate compounds for testing, and synthesizing novel compounds with therapeutic applications [GSJ14]. These methods are broken down into structure-based or ligand-based (without structure information), but they generally all involve some sort of high throughput similarity search through a virtual library of candidate drug-like molecules, of which there are up to 10^{60} distinct molecules. This search space is clearly too large for brute-force methods. Therefore, heuristics and optimization techniques are required, but these techniques are fundamentally limited by three major issues:

1. Molecule representations are *discrete* which makes things like gradient based optimization much more difficult.
2. Most search techniques today rely on labeling of the chemicals in the databases but there are few techniques that can utilize the vast amount of unlabeled molecule data.
3. Decoding molecules from current numerical feature representations is difficult.

This work specifically focuses on addressing these problems by developing a new continuous learned data representation of molecules that can be integrated into optimization frameworks for discovering new drug-like molecules.
While the high-level objective of this study is presented clearly, the introduction oversimplifies many issues and does not present enough detail in explaining the relevant state of work in this field. Very little reference is made to the other major computational methods that define the current state-of-the-art. For example, this work is extremely closely related to [ea15] as both methods use neural networks to learn new representations from molecule data; however, the authors fail to mention any details about this existing literature and why they do not intend to build directly on the work done to create CNN fingerprints from the molecule graph data. In addition, while it is said that gradient-based optimization will help leverage the geometric information in chemical space, this comment is extremely vague and the justification for why continuous optimization techniques are necessarily more advantageous than discrete optimization algorithms is largely missing.

2 Methods

2.1 Autoencoder

The authors utilize data formatted according the SMILES representation, which is a common text encoding for organic molecules. Given this character-based representation for the data, a convolutional-recurrent encoder-decoder neural network is applied to this data in order to learn a compressed *information bottleneck* representation of the molecule. Encoded strings of up to to 120 characters are used and the encoder network consists of three 1D convolutional layers followed by two fully connected layers. The decoder consists of a fully-connected layer feeding into three layers

of GRU units. The last layer of the decoder defines a distribution over the possible characters at each position of the SMILES string.
The description of the autoencoder structure is extremely brief and almost no intuition is provided as to why the selected architecture was chosen. In the convolutional encoder, for example, relatively large filter sizes of 9 and 11 are used and this is in stark contrast to the small filter sizes used by [NKB14] for sentence modeling. In addition, it seems that no pooling layers were used and this is surprising given the fact that most existing architectures utilize some sort of pooling. At a higher conceptual level, the encoder seems to produce a feature representation of size 292 which is larger than the input string length. As a result, we would expect to need some constraint to enforce sparsity of the weights; however, it seems that no sort of condition is used and no attempts at regularization using dropout were considered. Finally, with regards to the RNN decoder, it is again unclear why the authors choose to include layers of gated recurrent unit networks (GRUs) as opposed to regular recurrent units or LSTM units. The authors mention they performed a hyperparameter search using bayesian optimization; however it does not seem that they really considered the aforementioned issues or how to adapt their models based on understanding of the biology/SMILES data format. For example, the author remarks that the same model architecture was used on both drug-like molecules and OLED molecules without any justification as to why these two datasets would have similar structure or respond similarly to the same model. These decisions matter quite a lot in training neural networks and so it is strange that this kind of discussion/analysis is not considered in the paper.

2.2 Bayesian Optimization of Molecules

A sparse Gaussian process is trained to predict molecule cost from the latent representations. This approach is used to try and predict new molecules that might have better desired properties than in the training set. The optimization is done to maximize an objective consisting of a linear combination of the logP, SA scores and a penalty for large carbon rings. While the logP and SA scores are each individually useful measures (as referenced in the paper), it is unclear why optimization a simple difference of $logP - SA$ is the correct thing to do. No mention of how these values are normalized etc. is provided, and so more justification is needed as to why this is an appropriate objective to use.

3 Results and Conclusions

3.1 Qualitative Analysis of Latent Features

Figure 3 provides the first main results regarding the latent representation discovered by the VAE. This figure projects the training data onto a two dimensional latent space and shows that for both drug-like molecules and a generic set of organic LED molecules, the deep autoencoder is able to map the discrete molecule representation to a continuous representation that preserves molecular similarity according to some predefined molecule properties. The main evidence for this conclusion is the banding structure seen in the graphs that shows molecules with similar properties clustered together. While this result is interesting, from a clarity perspective, the paper fails to explain what the two-dimensional latent space actually is. The encoder seems to map a SMILE structure to a 292 latent feature vector and it is not evident how the two-dimensional projection is obtained from this representation. In addition, the representation is justified by coloring the points according to "a chemical property that is relevant to their function", but this choice of chemical property as the final basis of comparison is not substantiated. For example, the water-octanol partition coefficient (logP) is one score that is used to measure the drug-likeness of molecules, but no literature is provided that shows how the logP clustering in this latent feature space compares to clustering according to more standard fingerprints or other standard features. In addition, no results are presented that show the equivalent clustering generalizes to other chemical properties such as the other parts of Lipinski's rule of five [AL16]. Finally, the author glosses introduces "the combinatorial donor-bridge-acceptor way in which they were generated" as a potential reason for why the OLED molecules could not be effectively compressed to a 2D representation in the latent space, but does not explain the intuition behind or provide any evidence for that conclusion.

3.2 Reconstruction Accuracy

The next result presented is a table of the reconstruction accuracies for various latent dimensions for the two datasets defined earlier. The authors conclude that "with a large enough latent dimension [we can] achieve perfect reconstruction"; however, the actual results seem to suggest that the performance of the naive autoencoder loss with only 56 latent dimensions is in fact better than using larger latent dimensions. This inconsistency between results and conclusions is puzzling and must be an error. In addition, because there is no report of how this accuracy is calculated and what a "significant deviation" in accuracy would be, it is hard to interpret any claims using these results.

3.3 VAE Perturbation Experiments

In this section the authors perform a number of numerical experiments to investigate different applications of the VAE autoencoder. For the first experiment the authors select 8 points in the latent space and present multiple decondings for each point. It is claimed that the decodings result in "mid-sized realisitc molecules", however mid-sized and realistic are left undefined and as such it is not possible to evaluate this claim. In the next experiment multiple molecules are selected from FDA approved drugs listed in the drugbank 4.0 database [VCY+14]. Each selected drug is encoded using the VAE, whereafter multiple decodings are sampled. The decodings result in different variants of the original drug, however it is unclear to what extent all variations of a particular drug are discovered and which variants are missed. For the third experiment the authors generate a random 2-dimensional subspace of the latent space and decode the latent representation at each integer grid point in the space. The description of this method is left quite vague; while the paper's Figure 2 shows the end result of this process, an illustration of the 2-dimensional latent space subspace would be helpful. For the final experiment grid interpolation is used to interpolate between a group of chosen drugs. A number of drugs are generated using the grid interpolation procedure, however no comment is made as to whether the results are notworthy. The experiments show that the properties of the VAE allow for many different methods of drug generation. The main criticism is that the evalaution of the results is either vague or missing altogether.

3.4 Bayesian Optimization

In this section the authors propose an optimization procedure to generate drugs with desired properties. To demonstrate the method, a cost function is formulated using the log of the water-octanol partition coefficient, synthetic accesibility and a penality term for molecules having carbon rings of size larger than 6. Bayesian optimization is then selected as the optimization algorithm to maximize the cost function. Using this procedure the authors generate 500 latent feature vectors, of which more than half (the precised amount is unspecified) decode to valid SMILES string. Additionally it is unspecified how exactly the authors obtain the 500 feature vectors from the optimization procedure. Of the obtained molecules, two have objective values higher than any example in the training set, and additionally also have higher logP scores. A distribution of the objective function over the training data set is shown, however no distribution is shown for the generated molecules making it unclear how well the optimization procedure performs overall. Another criticism is that the choice of Bayesian optimization is left unargued.

3.5 OLED Experiments

The Bayesian Optimization procedure is then extended towards generating OLED molecules. A dataset of 150,000 OLED molecules, generated using fragment combination, is used to train the autoencoders. However the generated latent features obtained for the OLED dataset, either do not decode to a valid SMILES string, or decode into training examples. It is claimed that the failure to generate new OLED molecules is due to the fact that the training set contains many similar molecules, however it is unclear how this conclusion was reached. Furthermore it is not clear whether the a comparison was made between using the regular autoencoder verus the VAE during optimization.

References

[AL16] Mohammad Sayed Alam and Dong-Ung Lee. Synthesis, biological evaluation, drug-likeness, and in silico screening of novel benzylidene-hydrazone analogues as small molecule anticancer agents. *Archives of Pharmacal Research*, 39(2):191–201, 2016.

[ea15] K. D. Duvenaud et al. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems*, pages 2215–2223, 2015.

[GSJ14] Jens Meiler Gregory Sliwoski, Sandeepkumar Kothiwalke and Edward W. Lowe. Jr. Computational methods in drug discovery. *Pharmacological Reviews*, 66(1):334–395, 2014.

[NKB14] E. Grefenstette N. Kalchbrenner and P Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[VCY+14] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamhee S, Dame ZT, Han B, Zhou Y, and Wishat DS. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42:1091–1097, 2014.