Automatic chemical design using a data-driven continuous representation of molecules (Review, 2016.10.31)

DeepRegret: Salil Bhate, Bosh Liu, Scott Longwell, Tyler Shimko, Daniel Thirman, Sashwat Udit

Background

When designing novel drugs the goal is to create new molecules that optimally meet several desired qualities. However, optimizing this problem is hard because the search space is large, discrete, and unstructured: while the search space comprises 10^{23} - 10^{60} possible drug molecules, we have experimentally sampled the space with the meager 10^{8} compounds that have ever been synthesized. This search can be sped by up by first testing molecules computationally and then only synthesizing and testing the best performing ones. Currently, the best methods to explore the search space in simulation, such as genetic algorithms, still have problems. Notably, tuning of design heuristics is required.

The core problem to this approach is how to represent molecular graphs. Since molecular representation is discrete, the graphs can be either undirected graphs, with atoms as nodes and bonds as edges, or as a 3-d arrangement of atoms. These graphs are then converted into numeric representations for cheminformatic purposes. Differentiable, reversible, data-driven representations have several advantages. For example, they obviate the need to hand-specify mutation rules, and they are able to leverage a large set of unlabeled chemical compounds to build a large implicit library which can in turn be used for regression with a small set of labeled examples. New powerful probabilistic generative models have been built that after training on real examples are able to to produce realistic synthetic samples. In this paper the authors propose building a data-driven vector-valued representation for molecules to optimize current compounds and generate new ones.

Data and Model Summary

The authors used two unlabeled datasets. The first contained ~250K drug-like, commercially available compounds. The second contained ~100K molecules, generated in silico, with potential utility in organic light emitting diodes (OLEDs). To be used as input, these molecular structures were serialized as 120 character SMILES strings with blank space padding.

Using Bayesian optimization, the authors explored a variety of autoencoder architectures, which consist of an encoder and decoder. The encoder takes an input SMILES and generates a latent representation, i.e. a fixed dimensional vector; the decoder does the reverse. For their variational autoencoder (VAE), the authors eventually settled on an encoder that used a three-layer convolutional neural net (CNN) with two fully connected layers to produce a latent representation of length 292. Their decoder used a recurrent neural net (RNN) with three layers of gated recurrent units (GRUs).

Additionally, the authors trained a sparse Gaussian process (GP) model to take a latent representation and modify it through Bayesian optimization with an expected improvement (EI) acquisition function to generate 50 new novel latent representations. These novel latent representations were then fed into the decoder to produce corresponding SMILES strings, and their molecular properties were in turn evaluated with the following equation, which takes into account the water-octanol partition coefficient, a synthetic accessibility score, and a penalty for large carbon rings:

$$J(m) = \mathrm{logP}(m) - \mathrm{SA}(m) - \text{ring-penalty}(m)$$

Result Summary

The authors generated projections of their latent representations onto a two-dimensional space. For the set of 250K drug molecules, the projection showed banding and structure, with molecular properties showing a dependence on latent coordinates. A representation that co-localized molecules with similar properties is desirable and this appeared to occur. However, in the dataset of 100K OLED molecules their model produced tighter clustering and a considerable amount of empty latent space. Notably, their VAE failed to regularize the distribution of molecules into a Gaussian shape, likely because of the unique way in which the OLED library was generated. This weakness is possibly highlighted in the ability of the VAE to recover the SMILES sequences of the test set: while the VAE performed well on the drug dataset (95.3% acc.), it performed much more poorly on the OLED set (79.4%).

The authors then demonstrated the use of Bayesian optimization with their autoencoder to generate new molecules with desired molecular properties. Seeding with individual molecules from a list of 1386 FDA-approved drugs, the authors attempted to generate a molecule that would maximize the water-octanol partition coefficient. Their initial attempt produced molecules with unrealistically large numbers of carbon rings. However, after revising their model to take a penalty to the amount of rings into account, it did produce two novel molecules, including one with a partition coefficient higher than any molecule in the training dataset. Similar efforts to discover valid new molecules with the OLED data set were unsuccessful.

Weaknesses

- It is clear that learning a probabilistic, continuous latent representation for parameterizing chemical compounds is of great value for compound design. However, by using a vanilla variational autoencoder, without imposing any structural constraints on the latent representation (which corresponding to chemical insights or constraint), the network has no reason to learn a useful chemical encoding that generalizes.

- That their network indeed fails to learn a useful latent factor is clear from the fact that their Bayesian optimization fails to generalize to the OLED dataset. To solve this problem, they should augment their variational objective with additional latent structure arising from biochemical knowledge.

- In addition, the authors do not present any critical insight into the nature of their dataset or its potential to generalize (except that it fails to do so for the OLED case). There are no insights into selection of a training/test dataset that is truly representative of generalization ability.

- The authors don't present any attempts to regularize their objective toward a useful latent representation. For example, even without additional domain specific knowledge, it would be easy to augment their reconstruction with an adversarial term to produce valid SMILE strings.

- Moreover, while other continuous vector representations of chemical compounds have been published, the authors do not compare their method with existing techniques.

- While the authors report high training and test set reconstruction accuracies (i.e. percentage of correct characters in decoded SMILES strings), they do not state the means by which they split their data into training and test sets. Using "percentage of correct characters in decoded SMILES strings" is also a potentially misleading metric: a single incorrect SMILES character could cause a whole sequence to be invalid, so the percentage of recovered molecular structures is likely lower.

- The VAE employed by the authors is stochastic, which means that a single molecule will not always map to a single latent representation.

To conclude, although the authors proposed pipeline of learning a continuous representation followed by Bayesian optimization is natural and promising, it is clear that domain-specific knowledge imparted into their training objective will be essential for learning a representation that has true utility.

Extensions

To extend the work performed in this paper, the authors could potentially consider adding in information about known protein interactors for existing small molecule drugs. In this way, the neural net could learn features not only of active molecules, but also the bidirectional structural dependencies of both binding partners. To accomplish this, the authors could consider another autoencoder-based approach (Jian-wei, *et al.* 2013) to create a compressed representation of the protein structure, then feed these results into their current architecture to predict novel structures that may also interact with a given protein target. This approach would potentially yield greater precision in binding only "druggable" target proteins.

A second extension that the authors may want to consider is the use of molecular graphs as input for their neural net (Duvenaud, *et al*. 2015). Molecular graph-based neural nets have an advantage over SMILES string-based nets in that they do not have specific input vector size requirements. Instead, molecules of any size can be fed into the network, allowing much larger and more complex molecules to be used in training the authors' network. However, gains in predictive power and applicability may be marginal in this case as the majority of small molecule drugs can likely be handled by the current architecture.

References

Jian-wei, L., Guang-hui, C., Ze-yu, L., Yuan, L., Hai-en, L., & Xiong-Lin, L. (2013). Predicting protein structural classes with autoencoder neural networks (pp. 1894–1899). Presented at the 2013 25th Chinese Control and Decision Conference (CCDC), IEEE. http://doi.org/10.1109/CCDC.2013.6561242

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv.org*, 2224–2232.