Constrained Bayesian Optimization for Automatic Chemical Design

Ryan-Rhys Griffiths
Department of Engineering
University of Cambridge
rrg27@cam.ac.uk

José Miguel Hernández-Lobato
Department of Engineering
University of Cambridge
jmh233@cam.ac.uk

Abstract

Automatic Chemical Design leverages recent advances in deep generative modelling to provide a framework for performing continuous optimization of molecular properties. Although the provision of a continuous representation for prospective lead drug candidates has opened the door to gradient-based optimization, some challenges remain for the design process. One known pathology is the model's tendency to decode invalid molecular structures. The goal of this paper is to test the hypothesis that the origin of the pathology is rooted in the current formulation of Bayesian optimization. Recasting the optimization procedure as a constrained Bayesian optimization problem allows the model to produce novel drug compounds consistently ranking in the 100th percentile of the distribution over training set scores.

1 Introduction

The goal of chemical design is to find novel molecular structures that possess desirable properties. This search is complicated by the fact that chemical space is vast. From basic structural rules it is estimated that the space of pharmacologically active molecules satisfying Lipinski's rule of five (Lipinski et al., 1997) is upwards of 10^{60}, a novemdicillion (Reymond and Awale, 2012). Further to this, chemical space is discrete and so existing search methods such as genetic algorithms cannot make use of continuous optimization techniques which use geometrical cues such as gradients to optimize the objective. Recently, however, Gómez-Bombarelli et al. (Gómez-Bombarelli et al., 2016) have developed a model capable of converting a discrete representation of molecules to and from a continuous representation. On conversion to the continuous representation, gradient-based optimization may be undertaken to search for molecules with desirable properties. Optimized continuous representations may then be generated from the model, giving rise to the notion of Automatic Chemical Design.

Although a strong proof of concept for the idea of performing continuous optimization in molecular space, the goal of de novo drug design using Automatic Chemical Design faces some obstacles to realization:

1. Dead regions in the continuous latent space: How to ensure that the Bayesian Optimization procedure and the decoder operate in harmony?
2. What to optimize? How is it possible to encode the property of drug-likeness in a numerical metric?

The principle contribution of the paper, in section 3, will be to present a solution to the first question based on constrained Bayesian optimization. This will rein in the Bayesian optimization process such that it only selects points lying within a region of the latent space where the probability of a successful decoding is high. The second question is discussed in further detail in Griffiths and Hernández-Lobato (2017) in terms of how the model could be used in practice.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Background and Approach

The class of Bayesian Optimization problem is defined and the chosen acquisition function for all experiments is described.

2.1 Constrained Bayesian Optimization

Expressed formally the optimization problem is

$$\max_{\mathbf{x}} \mathbb{E}[f(\mathbf{x})] \text{ s.t. } \Pr(\mathcal{C}(\mathbf{x})) \geq 1 - \delta \tag{1}$$

where $f(\mathbf{x})$ is a black-box objective function, $\Pr(\mathcal{C}(\mathbf{x}))$ denotes the probability that a boolean constraint $\mathcal{C}(\mathbf{x})$ is satisfied and $1 - \delta$ is some user-specified minimum confidence that the constraint is satisfied (Gelbart et al., 2014). The constraint is that a latent point must decode successfully a large fraction of the times decoding is attempted. The black-box objective function is a drug-likeness metric and is noisy because a single latent point may decode to multiple molecules when the model makes a mistake, obtaining different values under the objective.

2.2 Expected Improvement with Constraints

EIC may be thought of as expected improvement (EI) that offers improvement only when a set of constraints are satisfied (Schonlau et al., 1998):

$$\text{EIC}(\mathbf{x}) = \text{EI}(\mathbf{x}) \Pr(\mathcal{C}(\mathbf{x})). \tag{2}$$

The implicit incumbent solution η suppressed in $\text{EI}(\mathbf{x})$, may be set in an analogous way to vanilla expected improvement (Gelbart, 2015) as either:

1. The best observation in which all constraints are observed to be satisfied.
2. The minimum of the posterior mean such that all constraints are satisfied.

The latter approach is adopted for the experiments performed in this paper. If at the stage in the Bayesian optimization procedure where a feasible point has yet to be located, the form of acquisition function used is that defined by (Gelbart, 2015):

$$\text{EIC}(\mathbf{x}) = \begin{cases} \Pr(\mathcal{C}(\mathbf{x}))\text{EI}(\mathbf{x}), & \text{if } \exists \mathbf{x},\ \Pr(\mathcal{C}(\mathbf{x})) \geq 1 - \delta \\ \Pr(\mathcal{C}(\mathbf{x})), & \text{otherwise} \end{cases}. \tag{3}$$

The intuition behind (3) is that if the probabilistic constraint is violated everywhere, the acquisition function selects the point having the highest probability of lying within the feasible region. The algorithm ignores the objective until it has located the feasible region.

3 Experiments

3.1 Implementation

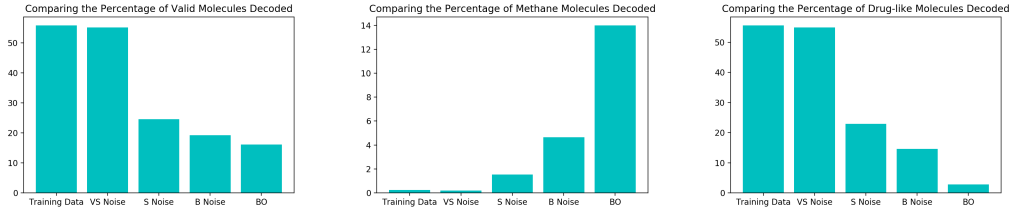
The implementation details of the encoder-decoder network as well as the sparse GP for modelling the objective remain unchanged from (Gómez-Bombarelli et al., 2016). The objective function is

$$J^{\text{logP}}(m) = \text{logP}(m) - \text{SA}(m) - \text{ring-penalty}(m) \tag{4}$$

in all experiments reported, where m denotes a molecule, logP(m) is the water-octanol partition coefficient for that molecule and SA(m) is its synthetic accessibility (Ertl and Schuffenhauer, 2009). The ring penalty term is as featured in (Gómez-Bombarelli et al., 2016). For the constrained Bayesian optimization algorithm, the BNN is constructed with 2 hidden layers each 100 units wide with ReLU activation functions and a logistic output. Minibatch size is set to 1000 and the network is trained for 5 epochs with a learning rate of 0.0005. 20 iterations of parallel Bayesian optimization were performed using the Kriging-Believer algorithm in all cases, collecting data in batch sizes of 50. The same training set as (Gómez-Bombarelli et al., 2016) was used.

3.2 Diagnostic Experiments and Labelling Criteria

A set of experiments were designed in order to investigate the hypothesis that points collected by Bayesian optimization lie far away from the training data in latent space and hence give rise to a large number of invalid decodings. The resulting observations are summarized in Figure 1.



(a) Valid Molecules (b) Methane Molecules (c) Drug-like Molecules.

Figure 1: Experiments on 5 disjoint sets comprising 50 latent points each. Very small (VS) Noise are training data latent points with approximately 1% noise added to their values, Small (S) Noise have 10% noise added to their values and Big (B) Noise have 50% noise added to their values. All latent points underwent 500 decode attempts and the results are averaged over the 50 points in each set. The percentage of decodings to: **a)** valid molecules **b)** methane molecules. **c)** drug-like molecules.

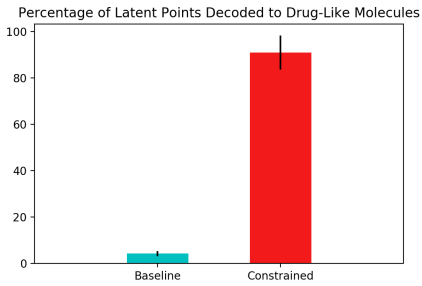
There would appear to be a noticeable decrease in the percentage of valid molecules decoded as one moves further away from the training data in latent space. Points collected by Bayesian optimization do worst in terms of their percentage of valid decodings. This would suggest that these points lie farther from the training data than even the B Noise set. The decoder would seem to over-generate methane molecules when far away from the data. Given that methane, although valid, has far too low a molecular weight to be a suitable drug candidate, a third plot in Figure 1c, shows the percentage of decoded molecules such that the molecules are both valid and have a tangible molecular weight. The definition of a tangible molecular weight was interpreted somewhat arbitrarily as a SMILES length of 5 or greater. Henceforth, molecules that are both valid and have a SMILES length greater than 5 will be referred to as drug-like.

As a result of these diagnostic experiments, it was decided that the criteria for labelling latent points to initialize the binary classification neural network for the constraint would be the following: if the latent point decodes into drug-like molecules in more than 20% of decode attempts, it is classified as a positive (drug-like) point and negative otherwise. The following experiments show the effect the constraint has on the validity and quality of the novel molecules generated.

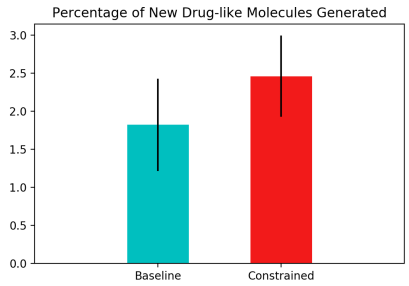
3.3 Molecular Validity

The BNN for the constraint was initialized with $117,440$ positive class points and $117,440$ negative class points. The positive points were obtained by running the training data through the decoder assigning them positive labels if they satisfied the criteria outlined in the previous section. The negative class points were collected by decoding points sampled uniformly at random across the design space. Each latent point undergoes 100 decode attempts and the most probable SMILES string is retained. The relative performance of constrained Bayesian optimization and unconstrained Bayesian optimization (baseline) (Gómez-Bombarelli et al., 2016) is compared in Figure 2a.

The results show that greater than 80% of the latent points decoded by constrained Bayesian optimization produce drug-like molecules compared to less than 5% for unconstrained Bayesian optimization. One must account however, for the fact that the constrained approach may be decoding the training data. There is a tradeoff between exploring the latent space and wandering into a dead region. If the constrained region is too tight, then the decoding process will produce molecules that have been seen in training, yet when there is no constraint, the optimization process will decode points in dead regions of the latent space. One means of examining the optimal balance between these two extremes is to look at the number of new drug-like molecules generated by the processes as in Figure 2b. One may observe that constrained Bayesian optimization outperforms unconstrained Bayesian optimization in terms of the generation of new molecules, but not by a large margin. A manual inspection of the SMILES strings collected by the unconstrained optimization approach showed that there were many strings with lengths marginally larger than the cutoff point, which is suggestive of partially decoded molecules. As such, a fairer metric for comparison should be the quality of the new molecules produced as judged by the scores from the black-box objective function. This is examined next.



(a) Drug-like Molecules



(b) New Drug-like Molecules

Figure 2: **a)** The percentage of latent points decoded to drug-like molecules. **b)** The percentage of new drug-like molecules generated. The results are from 20 iterations of Bayesian optimization with batches of 50 data points collected at each iteration (1000 latent points decoded in total). The standard error is given for 5 separate train/test set splits of 90/10.

3.4 Molecular Quality

The results of Figure 3 indicate that constrained Bayesian optimization is able to generate higher quality molecules relative to unconstrained Bayesian optimization. Table 1a gives the percentile that the averaged score of the new molecules found by each process occupies in the distribution over training set scores. In order to ensure that the averaged results aren't vulnerable to outliers, the scores of the best new drug-like molecules are provided in Table 1b. The constrained optimization procedure in every run produced new drug-like molecules ranked in the 100th percentile of the distribution over training set scores.

Figure 3: The best scores for new molecules generated from the baseline model (blue) and the model with constrained Bayesian Optimization (red). The vertical lines show the best scores averaged over 5 separate train/test set splits of 90/10. The histogram is presented against the backdrop of the top 10% of the training data scores for reference.

Run	Baseline	Constrained
1	49th	86th
2	51st	97th
3	12th	90th
4	37th	93rd
5	29th	94th

(a) Average Percentile Score of New Molecules

Run	Baseline	Constrained
1	88th	100th
2	98th	100th
3	76th	100th
4	96th	100th
5	95th	100th

(b) Percentile Score of Single Best Molecule

Table 1: **a)** Percentile of the Averaged New Molecule Score Relative to the Training Data. **b)** Percentile of the Best New Molecule Score Relative to the Training Data. The Results of 5 Separate Train/Test Set Splits of 90/10 are Provided in Both Cases.

4 Conclusions

The reformulation of the search procedure in the Automatic Chemical Design model as a constrained Bayesian optimization problem has led to concrete improvements on two fronts:

1. Validity - The number of valid molecules produced by the constrained optimization procedure offers a marked improvement over the original model.
2. Quality - For five independent train/test splits, the scores of the best molecules generated by the constrained optimization procedure consistently ranked in the 100th percentile of the distribution over training set scores. This approach is liable to work for a large range of objectives encoding countless desirable molecular properties.

References

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

Jean-Louis Reymond and Mahendra Awale. Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience*, 3(9):649–657, 2012.

Rafael Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv preprint arXiv:1610.02415*, 2016.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501*, 2017.

Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

Matthias Schonlau, William J Welch, and Donald R Jones. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pages 11–25, 1998.

Michael Adam Gelbart. *Constrained Bayesian Optimization and Applications*. PhD thesis, Harvard University, 2015.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.