Semi-Supervised Bayesian Optimisation for Automatic Chemical Design

Ryan-Rhys Griffiths, José Miguel Hernández-Lobato

June 19, 2017

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

VAE Architecture

The generative model architecture comprises a variational autoencoder with RNN encoder and decoder networks [1].

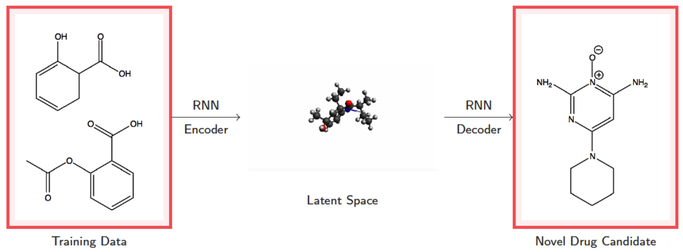

Figure 1: Variational Autoencoder (VAE) Architecture.

Molecules are converted from a discrete to a continuous representation on transitioning to the latent space. Bayesian optimisation in this continuous latent space maximises an objective function that serves as a measure of "drug-likeness" for the latent points.

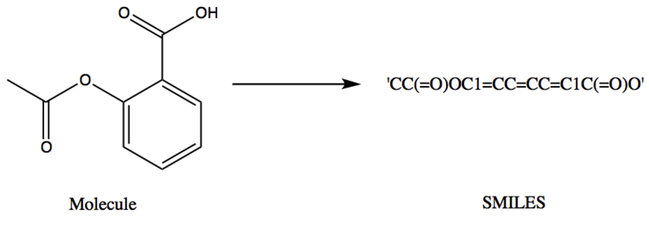

Figure 2: A Discrete Representation of Molecules: SMILES Strings.

Project Objectives

The scope of the project involves the investigation of two sources of improvement to the model of [1]:

1. Molecular Validity - enforcing the constraint that points in the latent space decode to valid molecules.
2. Molecular Quality - Making use of chemistry-related domain knowledge in the design of objective functions for the latent space.

Molecular Validity

Validity may be imposed through the implementation of a binary classification model as a constraint function.

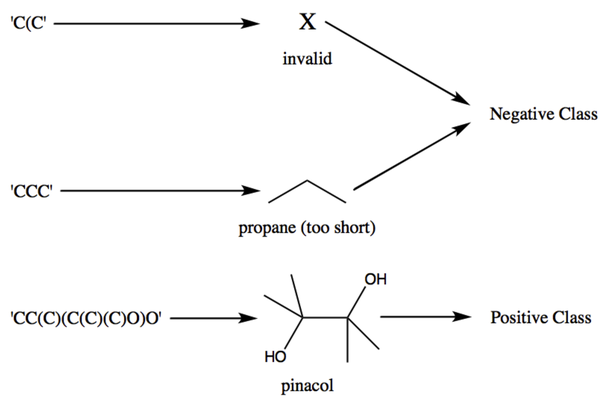

Figure 3: The Validity Constraint Sought for Decoded SMILES strings - Binary Classification.

Such a binary classification model may be implemented as a probabilistic constraint

$$Pr(C(\mathbf{x})) \geqslant 1 - \delta$$

within a constrained Bayesian optimisation setting, where $C(\mathbf{x})$ is a boolean function indicating whether or not the constraint is satisfied and $1-\delta$ is the specified minimum confidence required for the constraint to be considered satisfied. A formulation of this general class of constrained Bayesian optimisation problems is

$$\min_x \mathbb{E}[f(\mathbf{x})] \text{ s.t. } Pr(C(\mathbf{x})) \geqslant 1 - \delta,$$

where $f(\mathbf{x})$ is the objective function. The planned approach of performing constrained Bayesian optimisation with Thompson sampling will be evaluated first for the toy setting of the Branin-Hoo function and compared against alternative improvement criteria such as Expected Improvement. This toy setting is illustrated below, with figures taken from [3].

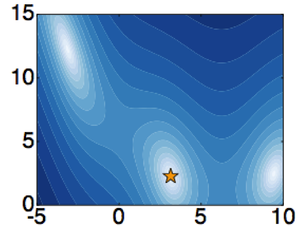

Figure 4: Branin-Hoo Function.

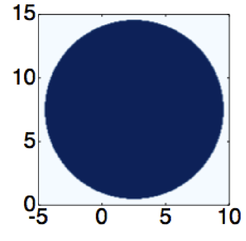

Figure 5: Circular Constraint.

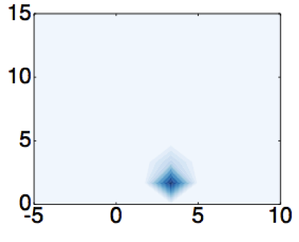

Figure 6: Minimum Location.

Molecular Quality

The latent space objective function

$$J^{\text{QED}}(m) = \text{QED}(m) - \text{SA}(m) - \text{ring-penalty}(m),$$

optimises the linear sum of the quantitative estimate of drug-likeness (QED) and synthetic accessibility (SA) metrics as well as a ring penalty term, where m denotes a molecule. Although the current ring penalty term correctly penalises 7-membered rings and upwards, there is no bias which favours the generation of 5 and 6-membered rings. 5 and 6-membered rings are statistically more abundant in nature relative to other ring sizes owing to the availability of low energy conformations which confer thermodynamic stability. Such conformations include the envelope conformation of cyclopentane as well as the chair conformation of cyclohexane.

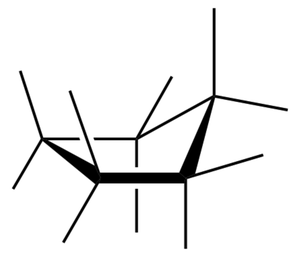

Figure 7: Cyclopentane.

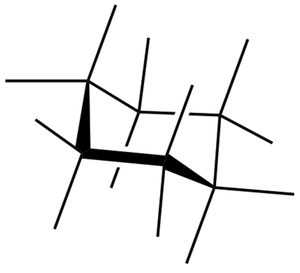

Figure 8: Cyclohexane.

References

[1] Gómez-Bombarelli, R., Duvenaud, D., Hernández-Lobato, J. M., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. and Aspuru-Guzik, A. Automatic Chemical Design using a Data-Driven Continuous Representation of Molecules. *arXiv preprint arXiv:1610.02415* (2016).

[2] Bowman, S. R. *et al.* Generating Sentences from a Continuous Space. *arXiv preprint arXiv:1511.06349* (2015).

[3] Gelbart, M. A., Snoek, J. and Adams, R. P. Bayesian Optimization with Unknown Constraints. *arXiv preprint arXiv:1403.5607* (2014).

[4] Taylor, R. D., MacCoss, M. and Lawson, A. D. G. Rings in Drugs: Miniperspective. *J. Med. Chem* (2014).

Contact Information

- rrg27@cam.ac.uk