VAE Architecture

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

![Variational Autoencoder (VAE) Architecture](image)

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.

![A Discrete Representation of Molecules: SMILES Strings](image)

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

Molecular Validity

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

\[
\Pr(C(x)) \geq 1 - \delta
\]

such that the binary classification model is implemented as a probabilistic constraint within a constrained Bayesian optimisation setting, where \(C(x)\) 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

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

where \(f(x)\) 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].

![The Validity Constraint Sought for Decoded SMILES strings - Binary Classification.](image)

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

![Branin-Hoo Function.](image)

Figure 4: Branin-Hoo Function.

![Circular Constraint.](image)

Figure 5: Circular Constraint.

![Minimum Location.](image)

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.

![Cyclopentane.](image)

Figure 7: Cyclopentane.

![Cyclohexane.](image)

Figure 8: Cyclohexane.

References


Contact Information

- rrg27@cam.ac.uk