OSM R2 Generative Modelling Writeup

Here are my generated first pass molecules for the second stage of the Open Source Malaria QSAR Competition Round 2. I have selected this molecule based on calculated LogS ranking and model consensus, which I will elaborate upon in the methodology. While my methodology could not generate a compound with a triazolopyrazine scaffold, I have also included the other generated molecules in case their scaffolds are of interest. This will hopefully be addressed in time for deadline for the hard deadline next week.

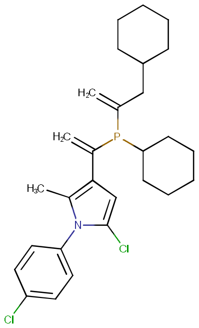
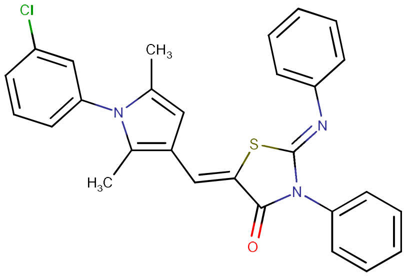
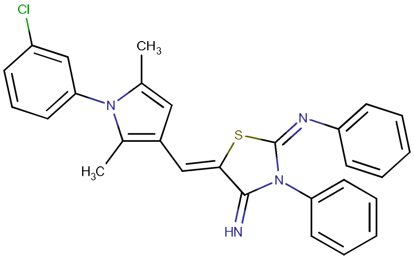
Methodology

A 22,000-molecule dataset annotated with EC50 (uM) values was compiled from the public domain through the ChEMBL database with the aim of generating a state of the art variational autoencoder. However, time constraints resulted in the repurposing of this dataset to construct an additional regression model instead. This model used a genetic algorithm to optimise 100 models relating a Mordred descriptor subset to *Plasmodium falciparium* EC50 values over ten generations. The OSM Round 2 blind prediction test set was annotated with experimental results for external validation and to enable comparison to the submitted model. All models were trained with a maximum predictive threshold of 2.5 uM in this project. This model produced 0.376 normalised MAE which was better than the 0.469 normalised MAE of the submitted model which enabled the use of a consensus approach to rank generated compounds.

An evolutionary natural language-based generative approach used the 349-molecule training dataset from the first stage of the OSM Round 2 modelling competition to generate 14 SMILES structures of similar molecules. This approach generated and optimised 3000 potential ligands using a score accounting for synthetic accessibility and lipophilicity with an aromatic ring penalty over 1000 generations. There was an attempt to include solubility as an optimisation target, however, this resulted in the generation of very small molecules such as pentane. Only the top scoring structures were selected for further processing resulting in the preparation of 14 generated molecules for consensus model assessment. This entailed the calculation of Mordred and quantum mechanical Hartree Fock with 3 corrections descriptors, in addition to JCLogS values calculated at pH 7.4 to quantify solubility. Predictions from both models were averaged to produce consensus predictions. The ranking criteria of generated molecules consisted of solubility followed by normalised activity.

Results and Discussion

Molecule 11.mol2 featured the lowest normalised consensus predicted potency, however, it features peculiar functional groups such as multiple cyclohexane rings and alkene groups. Molecules 6.mol2 and 5.mol2 features structures that are more similar to the those in the literature, specifically arylpyrroles, display low predicted EC50 values in the secondary model while featuring no predicted activity in the submitted model. While it is likely that these models can be mispredicting, it is also possible that these compounds are acting on other mechanistic targets that a large-scale phenotypic model is better suited at predicting with orders of magnitude more data than the submitted model. SMILES structures and additional generated ligands are available in the supplementary files.

Name	Structure	Submitted Model (uM)	Secondary Model (uM)	Averaged Consensus (uM)	JCLogS (pH 7.4)
11.mol2		0.53	1.07	0.80	-7.68
5.mol2		2.48	1.15	1.82	-7.69
6.mol2		2.47	0.91	1.69	-7.86