AI/ML AND VIRTUAL HUMAN PLATFORMS FOR THREAT AGENT HAZARD ASSESSMENT AND MEDICAL COUNTERMEASURE DISCOVERY AND DRUG DEVELOPMENT

Deep Learning Qsar Modeling For Fraction Unbound In Human Plasma

Michael Riedl Battelle **Sayak Mukherjee** Battelle **Mitch Gauthier** Battelle **Robert Moyer** Battelle

Background: The existing body of QSAR work relies on extracting a set of structure based chemical descriptors or fingerprints, followed by subset selection and training of a machine learning model using multivariate data analysis. Problematically, performance of these models varies widely depending on the software used to calculate the descriptors as well as the algorithms used for subset selection of the most relevant descriptors.
Purpose: The objective of this work is to develop a radical paradigm that can circumvent calculation and down selection of candidate descriptors by predicting chemical activity directly from the simplified molecular-input line-entry system (SMILES) strings without compromising accuracy.

Method: First, SMILES strings were embedded in an abstract latent space using the state-of-the art natural language processing model Bidirectional Encoder Representations (BERT). The embedding was accomplished using a “multi head attention” framework that incorporated positional information of all valid tokens (a portion of the SMILES string) with respect to all the other tokens and pre-trained on a set of SMILES tokens that were masked and not seen by the model. The pretraining dataset was comprised of ~9.8 million commercially available SMILES strings, allowing the model to learn the optimal embedding in an unsupervised fashion from a large dataset. Then, the pre-trained embeddings and added prediction layers were further “fine-tuned” in an additional step to optimize predictions for the endpoints of interest.

Preliminary Results: Here, we report our initial performance on the task of predicting fraction unbound in human plasma, a very important parameter for chemical absorption, distribution, metabolism, and excretion (ADME). We trained our model on a publicly available dataset (DW) of 1858 chemicals that resulted in a prediction RMSE of 0.396 on a holdout test set of 461 chemicals and outperformed other models that previously used the same data set. Additionally, as a proof of generalizability and robustness, we re-trained our model on another dataset (DO) used by an open-source software OPERA and compared prediction accuracies of our re-trained model and OPERA on DW. Our model and OPERA performed comparably with a test RMSE of ~ 0.6 for both.

DTRA Relevance: We built a list of 126 medical counter measures (MCMs) (referred to as DDTRA), including 101 antibiotics, 6 statins, 12 surfactants, and 7 other drugs and generated predictions for those chemicals. Our prediction RMSE was ~ 0.491.
Conclusion: Rapid development and deployment of MCMs against novel threats demands accurate in silico predictions of physiochemical properties. So far, QSAR models relied on calculation of molecular descriptors and subset selection. Our approach eliminates those steps and makes predictions from the SMILES string alone without compromising accuracy. Using fraction unbound in human plasma as an example, we showed that our approach outperformed other models, and our prediction performance was robust. Our approach required BERT pre-training only once, and therefore could be easily adopted to other endpoints without additional computational burden.

We acknowledge Battelle Memorial Institute for supporting this work