Deep Imputation on Large-Scale Drug Discovery Data

Benedict Irwin[1], Thomas Whitehead[2], Scott Rowland[3], Samar Mahmoud[1], Gareth Conduit[2], and Matthew Segall[1]

[1]Optibrium Ltd
[2]Intellegens Ltd
[3]Takeda Oncology

January 27, 2021

Abstract

More accurate predictions of the biological properties of chemical compounds would guide the selection and design of new compounds in drug discovery and help to address the enormous cost and low success-rate of pharmaceutical R&D. However this domain presents a significant challenge for AI methods due to the sparsity of compound data and the noise inherent in results from biological experiments. In this paper, we demonstrate how data imputation using deep learning provides substantial improvements over quantitative structure-activity relationship (QSAR) machine learning models that are widely applied in drug discovery. We present the largest-to-date successful application of deep-learning imputation to datasets which are comparable in size to the corporate data repository of a pharmaceutical company (678,994 compounds by 1166 endpoints). We demonstrate this improvement for three areas of practical application linked to distinct use cases; i) target activity data compiled from a range of drug discovery projects, ii) a high value and heterogeneous dataset covering complex absorption, distribution, metabolism and elimination properties and, iii) high throughput screening data, testing the algorithm's limits on early-stage noisy and very sparse data. Achieving median coefficients of determination, R^2, of 0.69, 0.36 and 0.43 respectively across these applications, the deep learning imputation method offers an unambiguous improvement over random forest QSAR methods, which achieve median R^2 values of 0.28, 0.19 and 0.23 respectively. We also demonstrate that robust estimates of the uncertainties in the predicted values correlate strongly with the accuracies in prediction, enabling greater confidence in decision-making based on the imputed values.

Hosted file

`Deep Imputation on Large-Scale Drug Discovery Data Preprint.pdf` available at https://authorea.com/users/390563/articles/504854-deep-imputation-on-large-scale-drug-discovery-data