This article was published in Innovations in Pharmaceutical Technology magazine, Autumn/Winter 2020

Data Imputation through Deep Learning: Expanding the data available for drug discovery by up to 100x

Matthew Segall*, Benedict Irwin*, Thomas Whitehead†, Samar Mahmoud*, Greg Shields*, Graham Turner*, Alex Elliott*, Stefan-Bogdan Marcu*, Robert Parini†, Edmund Champness*, Gareth Conduit†‡
*Optibrium Ltd., Cambridge, UK, info@optibrium.com, † Intellegens Ltd., Cambridge, UK, info@intellegens.ai, ‡Cavendish Laboratory, University of Cambridge, Cambridge UK

Synopsis

Machine learning (ML) methods are routinely used in drug discovery to build models that can predict the properties of compounds directly from their chemical structure. These quantitative structure-activity relationship (QSAR) models take ‘features’ of chemical structures (often referred to as ‘descriptors’) as input to predict one or more properties, including activities against biological targets or in phenotypic assays and a broad range of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. However, even the most sophisticated ML methods can struggle to produce high-quality predictions, due to the limitations of drug discovery data: The number of compounds with data for any given experimental endpoint is small when compared with machine learning data sets in many other fields; the overlap of compounds measured in different endpoints is even smaller; and the data generated by biological assays are noisy due to experimental variability.

Imputation methods take a different approach, using the limited property data that are available as *inputs*, to ‘fill in the gaps’ where measured values are not yet available. An example of an imputation method is Alchemite™, which applies deep learning to both compound descriptors and sparse assay data, as illustrated in Figure 1. The resulting model ‘learns’ directly from correlations between experimental endpoints, in addition to relationships between structural features of compounds and the experimental data. This approach makes better use of the sparse and noisy data in drug discovery, to produce more accurate predictions than QSAR models, which enables better targeting of the most promising compounds.

You can download this article as a PDF