intellegens

WHITE PAPER

Novel deep neural network imputation technique predicts bioactivity from incomplete data

Executive summary

Available experimental bioactivity data is sparse. To overcome this problem, a new deep learning neural network, **Alchemite™**, has been developed to accurately impute assay activities. Unlike traditional machine learning methods, **Alchemite™** can be trained on sparse bioactivity data. This type of data is typical of that found in commercial and public databases, enabling **Alchemite™** to learn directly from correlations between activities measured in different assays. Case studies on public domain data showed that **Alchemite™** significantly outperforms traditional quantitative structure-activity relationship (QSAR) models and other well-known approaches.

Intellegens Ltd., Eagle Labs, Chesterton Road, Cambridge, CB4 3AZ, UK

Introduction

For drug discovery projects, it is essential to have accurate compound bioactivity and property data to base decisions on for the selection of hits. This is also crucial for the efficient progression of compounds through hit to lead and lead optimisation to candidate selection. One of the main approaches for prediction of compound bioactivities are quantitative structure-activity relationship (QSAR) models, which are generated using existing data to identify correlations between the characteristics of compound structures (descriptors) and their biological properties or activities. The resulting models can be applied to new compounds that have not been experimentally tested in order to predict the outcome of the corresponding assays. Several statistical methods have been applied to build QSAR models, from straightforward linear regression to elaborate machine learning approaches such as random forests and support vector machines. An additional approach is the profile-QSAR method, a hierarchical approach that inputs the predictions of individual correlated bioactivity models.

The problem: experimental data is sparse

Traditional machine learning methods cannot directly utilize assay information as input

Available experimental data on potential compounds of interest is sparse. In a typical pharmaceutical company's corporate collection, for all the assay endpoints that are available, only a very small proportion of the possible compound-assay combinations have actually been measured in practice. Public domain databases are also sparse, with the ChEMBL data set being just 0.05% complete. This implies that if only a small fraction of this missing data were to be revealed, the company would benefit from a huge wave of new information. These incomplete experimental datasets could reveal more information regarding the correlations between the endpoints of interest if they could be used as inputs to a predictive model.

Recently, machine learning approaches have been explored for the development of QSAR models. While small improvements in prediction accuracy have been found, these approaches have not generally resulted in significant advances for activity predictions. These methods can train models against multiple endpoints simultaneously, which could allow the model to learn where a descriptor correlates with multiple endpoints, thereby improving the accuracy for all of the corresponding endpoints. However, traditional machine learning methods cannot directly utilize assay information as input because bioactivity data sets are often incomplete so a given activity cannot be relied upon to be present as an input.

The solution: a novel approach to sparse datasets

Alchemite™ applies an expectation-maximization algorithm

A novel deep learning framework has been developed that can learn from and exploit information that is sometimes missing, unlike other machine learning methods. Previously applied to materials discovery, this method can estimate the uncertainty in each individual prediction, enabling it to improve the quality of the predictions by focussing on the most confident results. This new algorithm, initially developed at the University of Cambridge, is now commercialised by Intellegens as **Alchemite™**, and can identify the link between assay bioactivity values, use bioactivity data of other compounds to guide the extrapolation of the model, and also use molecular descriptors as design variables. Typical neural networks require that each property is either an input or output of the network and that all inputs must be provided to obtain a valid output. To handle incomplete data, this method treats the assay bioactivities as both inputs and outputs of the neural network. **Alchemite™** applies an expectation-maximization algorithm, where an initial estimate for the missing data is provided which the neural network iteratively improves.

Alchemite™ outperforms conventional methods

The predictive accuracy of Alchemite™ outperforms the leading approaches while using 80% less computing resource

This novel deep learning method was tested against other popular machine learning methods from the literature. These methods are compared in the figure below, measuring the coefficient of determination (R^2) against the benchmark data set defined by Novartis in Martin et al. The predictive accuracy of **Alchemite™** outperforms the leading approaches while using 80% less computing resource.

Random Forest (QSAR)

Random Forest methods are a popular approach of QSAR analyses and they work by building an ensemble of decision trees to predict individual assay results. However, decision trees require all the input data to be present when being trained, it is not possible to build this kind of model using incomplete bioactivity data as input. In this case, Random Forest must rely on chemical descriptors only.

DNN

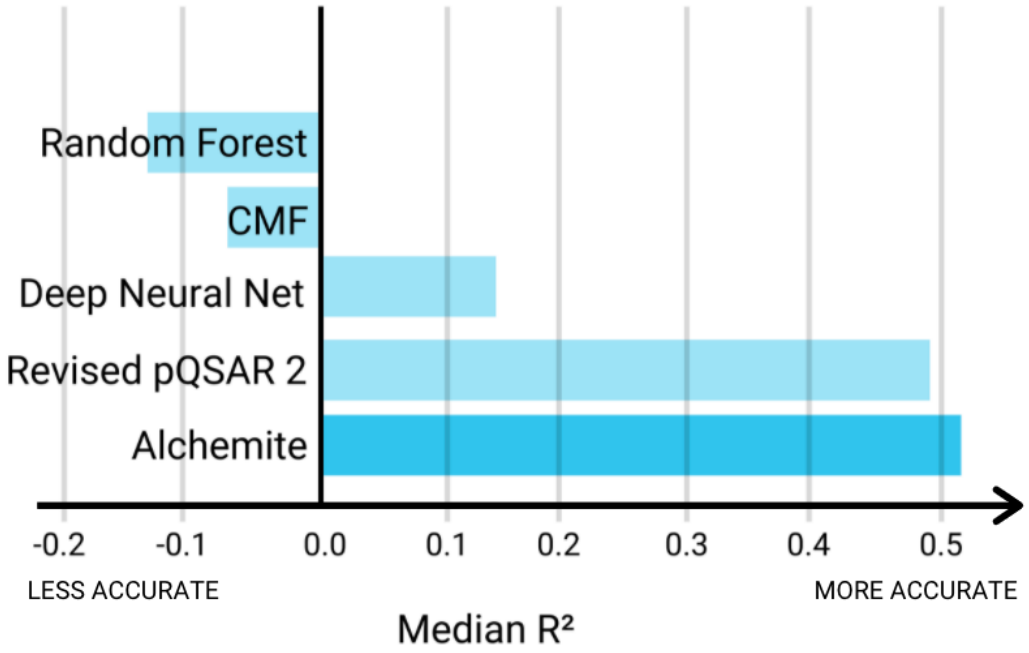
A traditional multi-target deep neural network (DNN) model.

Collective Matrix Factorisation (CMF)

This model makes effective use of the available chemical descriptors as well as bioactivity data.

profile-QSAR 2.0 (Martin et al)

This model first proposed by Martin et al. and revised by Intellegens, builds a linear partial least squares (PLS) model of assay bioactivities from the predictions of random forest models for each assay individually.



Conclusions

Alchemite™ is able to predict bioactivity from incomplete data. It improves the quality of predictions by using correlations between both different bioactivity assays and also between molecular descriptors and bioactivities, which results in an important improvement in the accuracy of prediction over other conventional QSAR models, even those that implement deep learning. Moreover, this method can also accurately predict the confidence in each individual prediction, which allows for attention to be focussed on the most confident results.

Alchemite™ is quicker to train than conventional neural networks, training in minutes rather than days on conventional data sets

Alchemite™ has been designed to be computationally efficient for application to pharma-scale datasets. Alchemite™ is quicker to train than conventional neural networks, training in minutes rather than days on large datasets, and uses less memory than random forest methods. This is a broadly applicable method that can be applied beyond binding assay data, for example for the identification of additional active compounds within a database, prediction of selectivity profiles, and recognition of the most influential chemical properties. It can also make accurate predictions on other values such as chemical absorption, distribution, metabolism, excretion and toxicity properties, which play key roles in drug discovery and development.

References

T. M. Whitehead, B. W. J. Irwin, P. Hunt, M. D. Segall, and G. J. Conduit (2019). Imputation of Assay Bioactivity Data Using Deep Learning. *Journal of Chemical Information and Modeling, 59*(3), 1197-1204. DOI: 10.1021/acs.jcim.8b00768

Martin, E.; Valery R. Polyakov, V.; Tian, L.; Perez, R. (2017). Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *Journal of Chemical Information and Modeling, 57*(8), 2077-2088.

About Intellegens

Intellegens has developed a unique deep learning engine, **Alchemite™** for training neural networks from the sparse and noisy data typical of real-world science and business challenges. The technique was first developed at the University of Cambridge where it has been used to develop aerospace alloys, guide the design of new drugs, and design next-generation battery technology. The tool is now being used to solve a wide range of industrial customer problems, optimising products and processes, saving time and cost in discovery and development, and enabling breakthrough insights.

www.intellegens.ai | info@intellegens.ai | @intellegensai