

Imputation of Assay Bioactivity Data using Deep Learning

Written by Matt Segall

Thursday, 14 February 2019 10:24 - Last Updated Wednesday, 05 June 2019 09:03

This paper was printed in the Journal of Chemical Information and Modeling.

Imputation of Assay Bioactivity Data Using Deep Learning

Whitehead TM*, Irwin BWJ, Hunt P, Segall MD, Conduit GJ** (*Intellegens, **Cavendish Laboratory)

J. Chem. Inf. Model. (2019) 59(3) pp. 1197-1204

Abstract

We describe a novel deep learning neural network method and its application to impute assay pIC₅₀ values. Unlike conventional machine learning approaches, this method is trained on sparse bioactivity data as input, typical of that found in public and commercial databases, enabling it to learn directly from correlations between activities measured in different assays.

In two case studies on public domain data sets we show that the neural network method outperforms traditional quantitative structure-activity relationship (QSAR) models and other leading approaches. Furthermore, by focussing on only the most confident predictions the accuracy is increased to $R^2 > 0.9$ using our method, as compared to $R^2 = 0.44$ when reporting all predictions.

Imputation of Assay Bioactivity Data using Deep Learning

T.M. Whitehead¹, B.W.J. Irwin², P. Hunt², M.D. Segall¹ and G.J. Conduit¹

¹Intellegens, Judge Lane, Chesterton Road, Cambridge, CB1 3RL, United Kingdom
²Department, Fitz Willshire House, Cavendish Laboratory, Park Street East Road, Cambridge, CB2 3RQ, United Kingdom

*Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB1 3RQ, United Kingdom

Email: mds@intelligens.co.uk

Abstract

We describe a novel deep learning neural network method and its application to impute assay pIC₅₀ values. Unlike conventional machine learning approaches, this method is trained on sparse bioactivity data as input, typical of that found in public and commercial databases, enabling it to learn directly from correlations between activities measured in different assays.

In two case studies on public domain data sets we show that the neural network method outperforms traditional quantitative structure-activity relationship (QSAR) models and other leading approaches. Furthermore, by focussing on only the most confident predictions the accuracy is increased to $R^2 > 0.9$ using our method, as compared to $R^2 = 0.44$ when reporting all predictions.

You can download this paper as a [PDF](#)