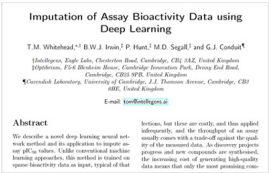
This paper was printed in the Journal of Chemical Information and Modeling.

**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 $pIC_{50}$ 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. Whitehead,*,† B.W.J. Irwin,† P. Hunt,† M.D. Segall,‡ and G.J. Conduit¶

†Intellegens, Eagle Labs, Chesterton Road, Cambridge, CB4 3AZ, United Kingdom
‡Optibrium, F14 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, United Kingdom
¶Cavendish Laboratory, University of Cambridge, J.J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom

E-mail: tom@intellegens.ai

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.

lections, but these are costly, and thus applied infrequently, and the throughput of an assay usually comes with a trade-off against the quality of the measured data. As discovery projects progress and new compounds are synthesized, the increasing cost of generating high-quality data means that only the most promising com-

You can download this paper as a PDF