Towards predicting compound activity by traversing biomedical knowledge graphs

Terence Egbelo[1], Vlad Sykora[2], Michael Bodkin[2], Ziqi Zhang[1], Val Gillet[1]
[1]Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom
[2]In Silico R&D, Evotec (UK) Ltd, 114 Park Drive, Abingdon OX14 4RZ, United Kingdom

Introduction

In a biomedical *knowledge graph* (KG), interesting properties of an entity e.g. a protein or a drug-like compound, are encoded as typed links to other entities. These subject-property-value data units (Figure 1) are also known as KG *triples* and may correlate with more complex patterns within the graph. The task of inferring new triples by exploiting such associations is known as KG completion; this work explores the prediction of compound activity in kinase assays as a KG completion problem.

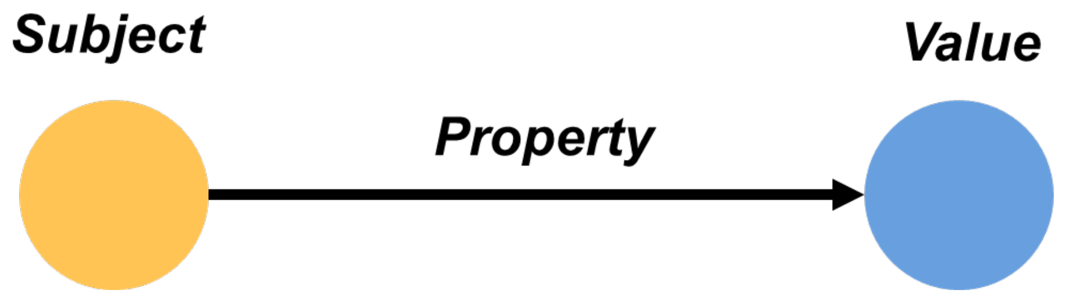


Figure 1: A KG triple

Data

Evotec's Knowledge Graph

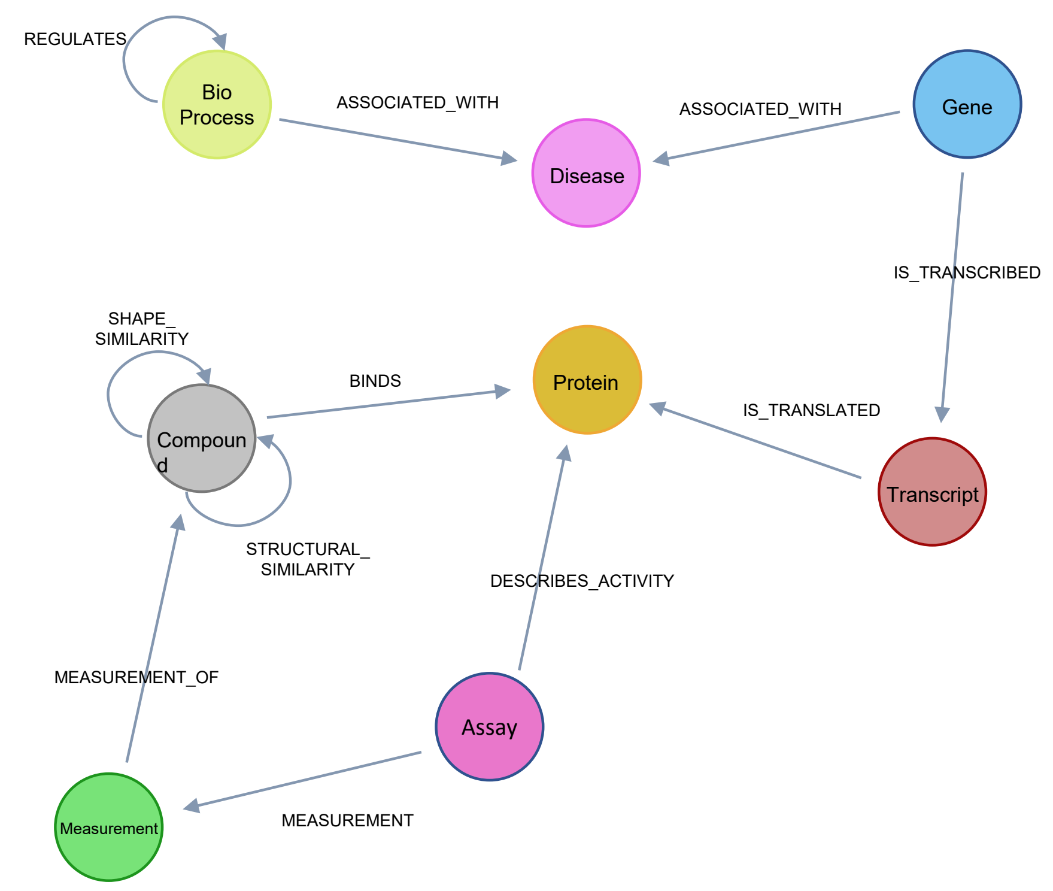


Figure 2: Schema of Evotec's KG

Evotec's KG brings together diverse information from public data sources e.g. Ensembl (Gene and Protein nodes), ChEMBL (Compound, Assay and Measurement nodes), the Experimental Factor Ontology (Disease nodes) and the Gene Ontology (Biological Process nodes), among others.

The KG is constantly being updated; the work presented here uses v1.0 (schema in Figure 2).

Data subset used

The data set used in this workflow is a subset of Evotec's KG and contains ~14,000 kinase assay measurements (Compound, Measurement and Assay nodes) and associated Protein (kinase) nodes as shown in Figure 3.

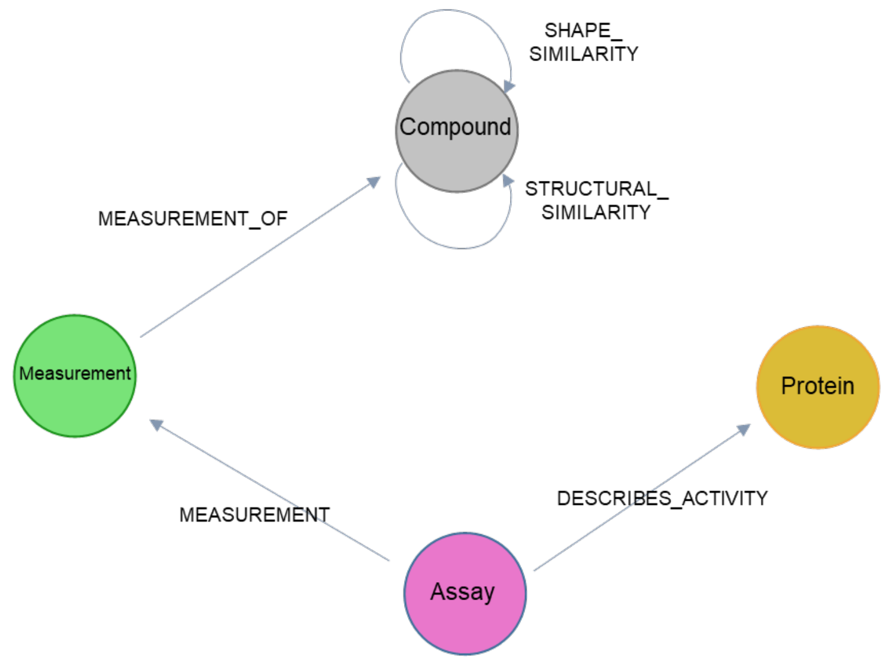
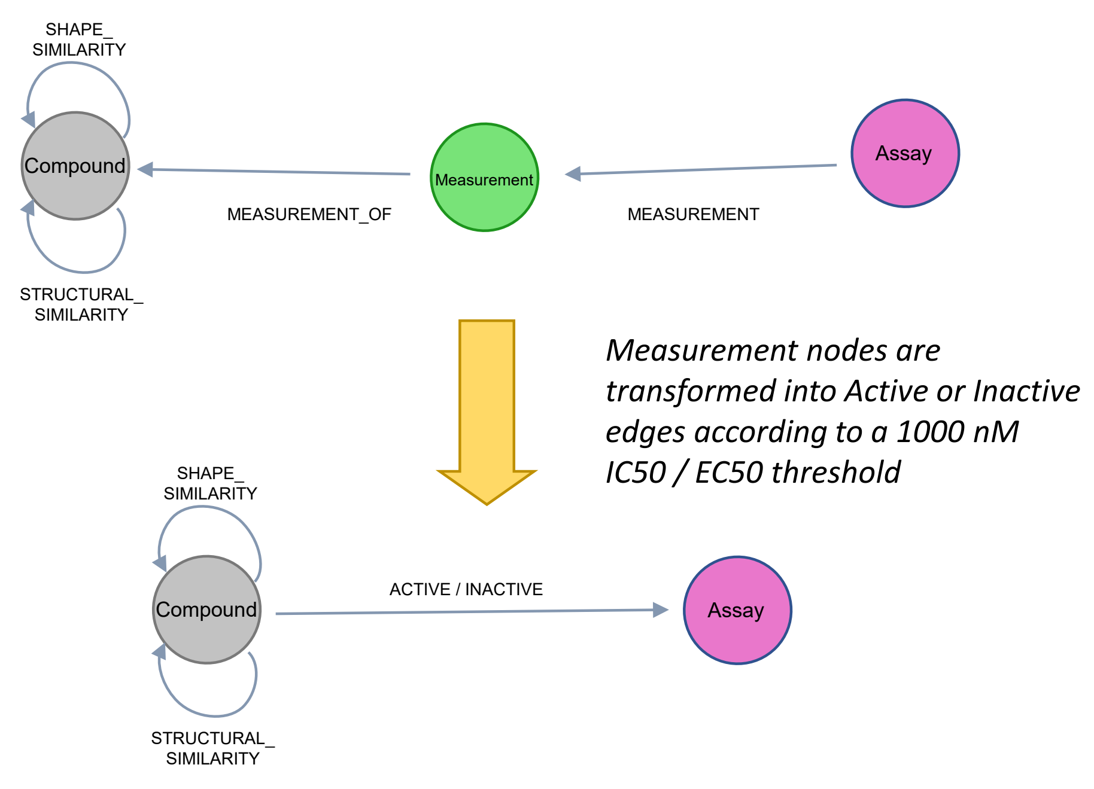


Figure 3: Schema of KG extract used in this work

The Prediction Task



Measurement nodes are transformed into Active or Inactive edges according to a 1000 nM IC50 / EC50 threshold

Figure 4: From Measurement nodes to "Active" and "Inactive" KG edges

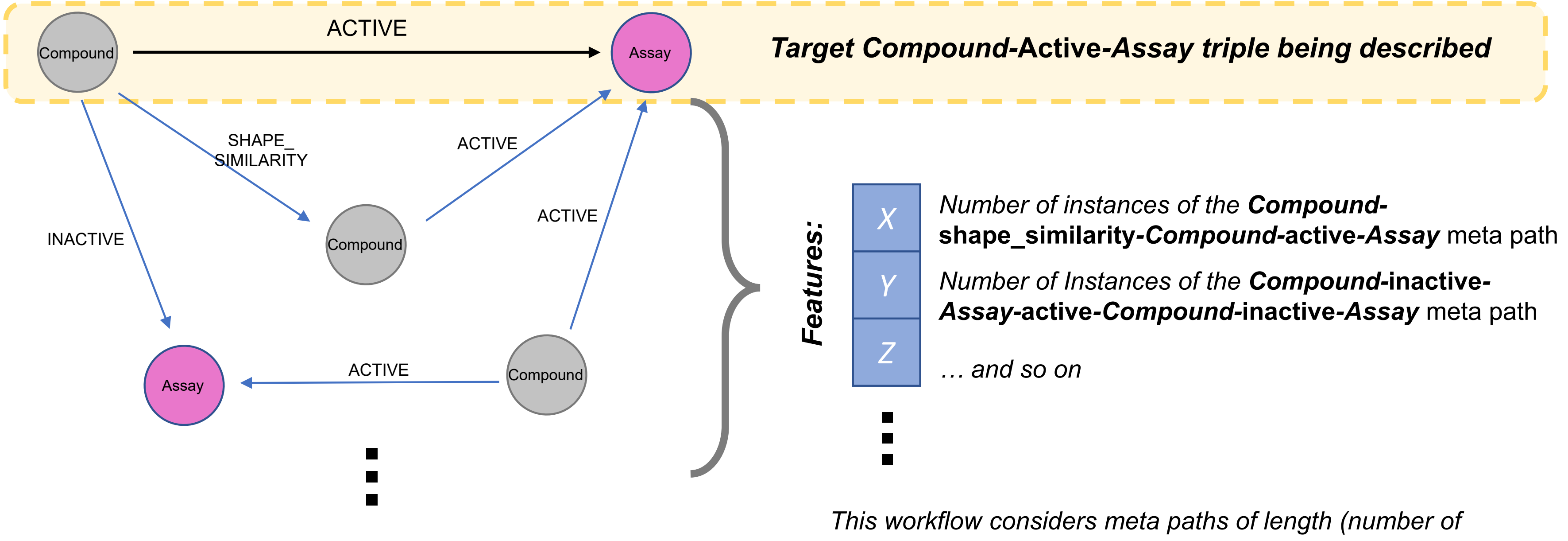
Problem definition: learn classifier models that can separate *Compound*-Active-*Assay* triples from *Compound*-Inactive-*Assay* triples and generalise to correctly infer activity and inactivity in unseen compound-assay pairings.

Both the training and validation triples of the above types must first be created by transforming Measurement nodes in the sample graph according to an activity threshold (Figure 4).

.

Meta Path Features

Each *Compound*-Active(Inactive)-*Assay* triple in the data set is characterised by a feature vector whose components are the counts of the distinct path types ("meta paths") in the sample KG that connect the compound and the assay (example in Figure 5).



Number of instances of the ***Compound*-shape_similarity-*Compound*-active-*Assay*** meta path

Number of Instances of the ***Compound*-inactive-*Assay*-active-*Compound*-inactive-*Assay*** meta path

… and so on

Figure 5: Characterising a *Compound*-Active-*Assay* triple via meta path instances

This workflow considers meta paths of length (number of edges) 2, 3 and 4. This yields 74 meta path features given the edge types contained in the data subset used

Supervised Learning

Random Forest (200 trees) and Logistic Regression (L-2 regularised) classifiers were evaluated via 5-fold stratified cross-validation on the original feature matrix as well as on transformations including Z-score normalisation and log (base 10) normalisation (Figure 6).

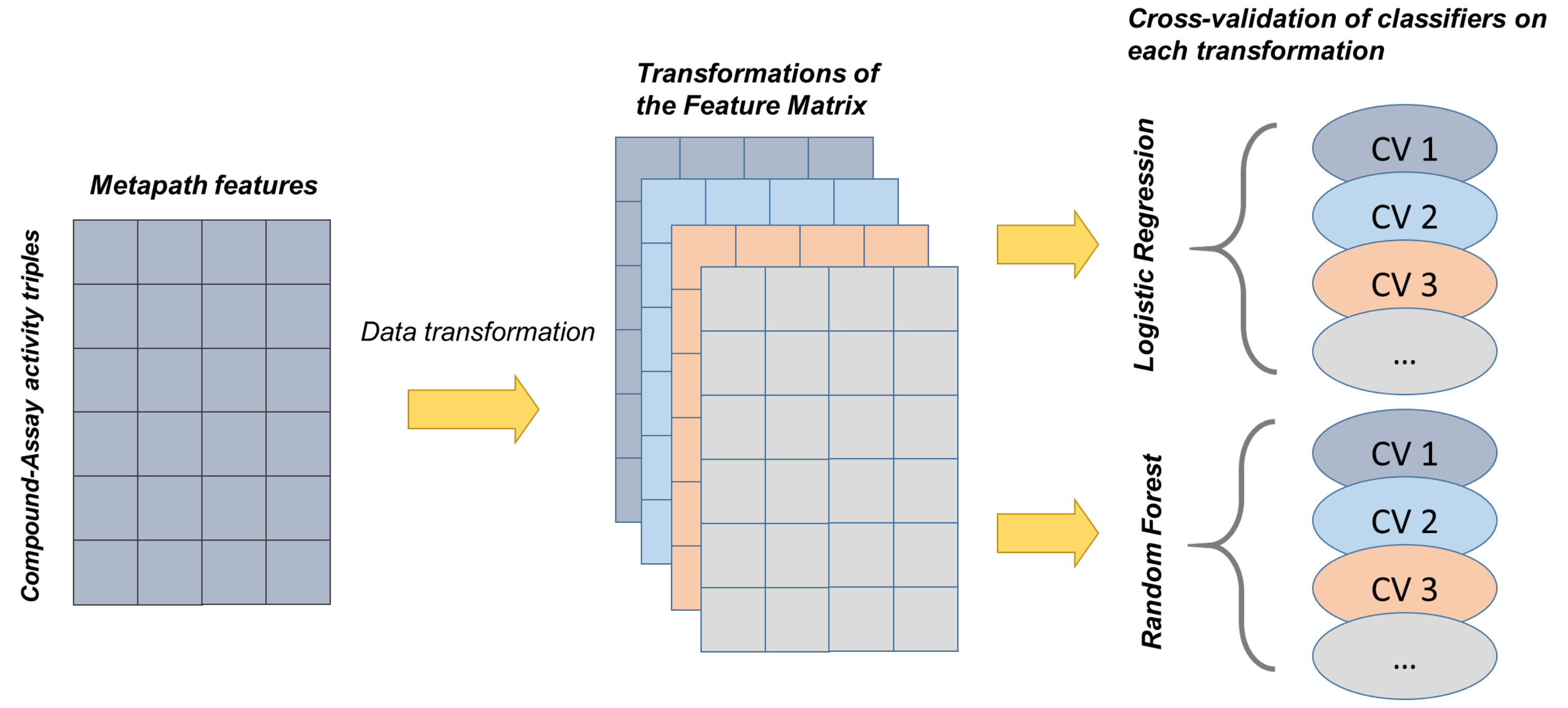

Compound-Assay activity triples
Metapath features
Data transformation
Transformations of the Feature Matrix
Cross-validation of classifiers on each transformation
Logistic Regression
CV 1
CV 2
CV 3
...
Random Forest
CV 1
CV 2
CV 3
...

Figure 6: Overview of the supervised learning pipeline

Model Comparison

The Random Forest classifiers consistently outperform the Logistic Regression classifiers in terms of AUROC under the same cross-validation procedure (Figure 7).

Considering the comparatively limited training data, the general high performance demands further validation on larger and richer samples.

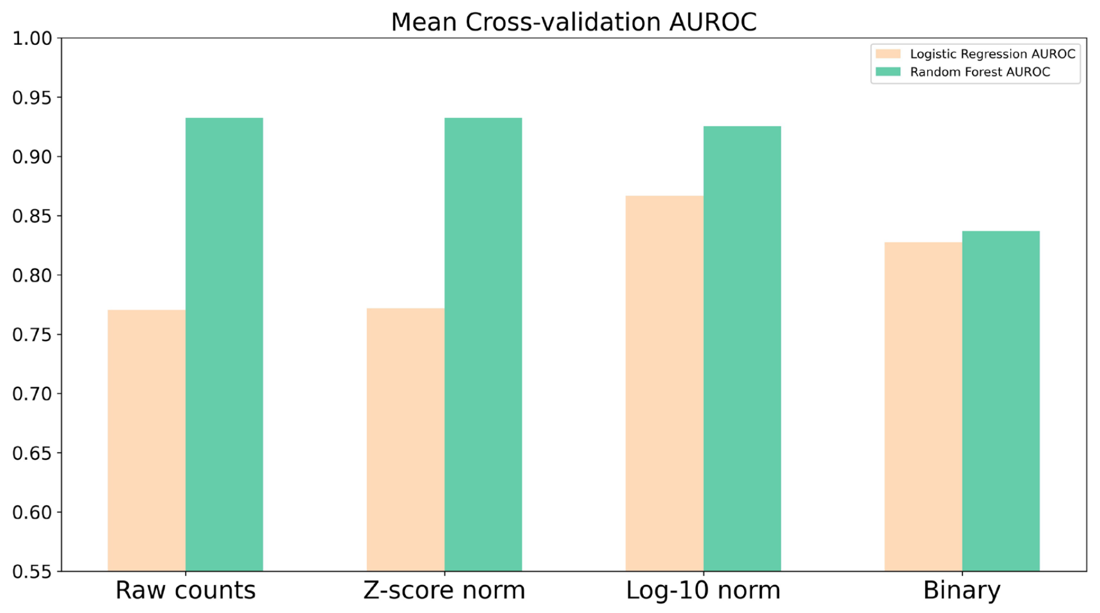


Figure 7: Mean cross-validation AUROC comparison across classifiers and transformations

Selected Feature Analysis

Important features (ranked by impurity reduction, permutation importance and logistic regression coefficients) shared across top-performing classifiers were identified.

Figure 8 shows the probability densities for each target class for the log10-normalised meta path feature C_sim1_C_sim1_C_a_A which states *"Target compound's shape is similar* (according to Evotec's original thresholding) *to a second compound's shape, whose shape is similar to that of a third compound, which is active in the target assay".*

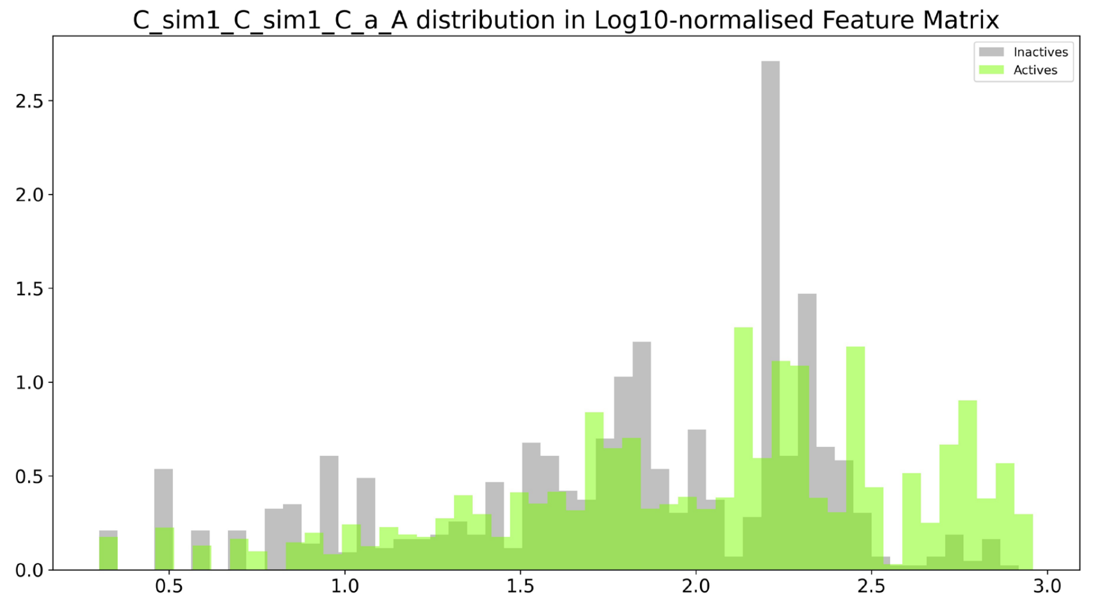


Figure 8: Per-feature target class probability densities

The slight rightward shift in the actives' probability density indicates higher values of the feature for the active class than for the inactive class.

Conclusion & Follow-on Work

The knowledge graph meta path-based approach to compound bioactivity prediction shows promise given the discovery of discriminative paths that occur frequently with the active and inactive compound-assay relationships in the kinase-focused subset of Evotec's KG used here. However, some paths are difficult to interpret and class separation is not always clear. Subsequent iterations of this data pipeline will incorporate a larger and more expressive subset of entities and relations from the KG as well as a greater diversity of protein targets, and will also provide benchmarking against models trained on QSAR chemical descriptor features.

References

1. Lao, N., Mitchell, T., & Cohen, W. (2011, July). Random walk inference and learning in a large scale knowledge base. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 529-539).
2. Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (2011, July). Co-author relationship prediction in heterogeneous bibliographic networks. In 2011 International Conference on Advances in Social Networks Analysis and Mining (pp. 121-128). IEEE.
3. Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., & Bolton, E. (2016). Predicting drug target interactions using meta-path-based semantic network analysis. BMC bioinformatics, 17(1), 1-10.
4. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., ... & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife, 6, e26726.
5. Guo, M. G., Sosa, D. N., & Altman, R. B. (2022). Challenges and opportunities in network-based solutions for biological questions. Briefings in Bioinformatics, 23(1), bbab437.

Email: tegbelo1@sheffield.ac.uk