Enzyme Promiscuity Prediction using Hierarchical Multi-Label Classification

Gian Marco Visani
Department of Computer Science
Tufts University
gian_marco.visani@tufts.edu

Michael C. Hughes
Department of Computer Science
Tufts University
mhughes@cs.tufts.edu

Soha Hassoun*
Department of Computer Science
Department of Chemical and Biological Engineering
Tufts University
soha@cs.tufts.edu

Abstract

As experimental efforts are costly and time consuming, computational characterization of enzyme capabilities is an attractive alternative. We present and evaluate several machine learning models to predict which of 983 distinct enzymes, as defined via the Enzyme Commission (EC) numbers, are likely to interact with a given query molecule. We frame this "enzyme promiscuity prediction" problem as a multi-label classification task. We utilize inhibitor and unlabeled data to train prediction models that can take advantage of known hierarchical relationships between enzymes. We compare these models with alternatives that ignore hierarchical structure and with molecular similarity approaches. We report that hierarchical multi-label neural networks are the best model for solving this problem, outperforming other machine learning and similarity-based methods.

1 Introduction

Characterizing activities of enzymes on substrates promises to play a critical role in advancing biological discovery and biomedical engineering applications. Enzymes are traditionally assumed specific, acting on a specific molecule. The Enzyme Commission (EC) number nomenclature, which assigns EC numbers to enzymes based on experimental evidence of enzyme-catalyzed reactions [1], describes such specificity. Four numbers separated by periods (e.g. 1.2.1.75) provide a tree-structured hierarchical classification of enzyme function. At the top level, there are 7 potential classes of enzymes (numbered 1-7). Classes are further divided into sub-classes, sub-subclasses, and finally individual enzymes at the leaf level of the hierarchy. Despite widespread use, EC classification does not reflect the diverse range of molecules that enzymes catalyze. Importantly, many enzymes, if not all, have promiscuous activities such that they act on substrates other than those they evolved to transform [2-5]. Despite efforts in cataloguing enzyme activities on various substrates in databases, comprehensive characterization of enzyme promiscuity remains elusive. Some prior works on predicting the promiscuity of enzymes are rule-based, e.g., [6-10], predicting enzymatic products for a given a query molecule. More recent efforts utilize support vector machines (SVMs) to evaluate the promiscuity of four specific enzymes, applying active learning to recommend substrates for experimental testing to obtain new labels and thereby improve classification accuracy [11].

*To whom correspondence should be addressed.

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

An established practice in biological engineering is to assess the promiscuity of an enzyme on a query molecule using substrate similarity. Molecular similarity is calculated between the query molecule and the native substrate(s) that is known to be catalyzed by the enzyme [12]. Similarity can be computed on molecular graphs using subgraph isomorphism [13] or on molecular fingerprints, binary feature vectors of predetermined size (e.g., PubChem fingerprint [14], Extended-Connectivity (ECFP) [15]). There is no consensus however on a similarity level that deems a query molecule sufficiently similar to a native substrate. Further, while substrate similarity is widely used in metabolomic engineering practice, there is currently no large-scale systematic evaluation of the effectiveness of this metric in predicting enzyme promiscuity across enzymes or enzyme classes.

We investigate in this paper several data-driven approaches aimed at predicting the enzyme classes, as defined via their Enzyme Commission (EC) numbers, that are likely to interact with a given query molecule. We refer to this problem as the "enzyme promiscuity prediction" problem, and we do not distinguish amongst the types of underlying promiscuity. In addition to similarity-based approaches, we develop a combination of novel as well as established machine-learning models that frame the problem as multi-label classification, where each predicted label corresponds to an EC number. As experimentalists have limited resources to confirm predictions via wet-lab experiments, maximizing precision is a key concern. We therefore evaluate the performance of our models using metrics such as R-precision (R-PREC) that assess the model's correctness among top-ranked molecules that are predicted to interact with a particular enzyme. A model with high R-PREC has a high probability of ranking positive molecules ahead of negative molecules. The top-ranked molecules are naturally the first ones that the user will consider for experimental testing. Therefore, a model with high R-PREC on a test set is most likely to be most useful for end users.

2 Methods

2.1 Data collection

All available positive and inhibitor molecules were collected from the BRENDA database, excluding co-factors because these metabolites are common across enzymatic reactions. The Morgan fingerprint (with a radius of 2) was used to represent each molecule, with 2048 binary features [15]. Not all compound names in BRENDA could be mapped to a specific molecular structure. By the end of the conversion process from names in BRENDA to Morgan fingerprints, we identified 25,872 positive pairings between molecules and EC numbers, based on 8,295 unique molecules. Within this set of molecules, we also identified 13,087 inhibiting interactions, based on 2,165 unique inhibitors. Some individual enzymes had limited positive data. We therefore predicted promiscuity for enzymes that had a minimum of 10 positive examples. There was sufficient data to train classifiers for 983 distinct EC numbers. Only 885 of these enzymes had known inhibitors. As there were not many enzymes classified for the recently established top level class (number 7), we only used data for classes 1-6.

The data was further organized in a tree hierarchy to match the structure of the EC nomenclature. There were 6 nodes at the class level (top of hierarchy), 50 nodes at sub-class level, 146 nodes at sub-subclass level, and 983 leaf nodes (distinct EC numbers). At all non-leaf nodes in the hierarchy, positive examples consisted of the union of all positive examples at any child. Similarly, inhibitor (hard negative) data consisted of inhibitors at any child unless already labelled positive due to a positive label from any other child node. At each node, any molecule that is not positive nor an inhibitor is considered unlabeled. Of the 8,295 molecules that were collected from BRENDA, each with 983 labels indicating the molecule's interaction with each enzyme, 20% were picked at random and saved for testing, ensuring that both the positive and the negative classes were represented for each enzyme.

2.2 Models

We developed and evaluated five models with different levels of information sharing for the enzyme promiscuity prediction task.
k-NN Similarity (No sharing) - Each enzyme-molecule pair in the test set was scored by computing the mean structural similarity between the test molecule and the enzyme's k most similar positive molecules in the training set. Hyperparameter k was chosen via a grid search 3-fold cross validation over the training set, maximizing Average Precision.
No-Share RF (No sharing) - an independent Random Forest (RF) classifier with 50 decision trees

was built for every EC Number. Hyperparameters defining the minimum size of any decision tree's terminal node were set via a grid search 5-fold cross validation, maximizing Average Precision.
Hierarchical RF (Greedy Top-Down Hierarchical Sharing) - A hierarchical cascade of random forests (RFs) was trained, with one RF predictor at each internal node and leaf node of the tree. Each of the 6 top-level enzyme categories had a root predictor trained to produce probabilistic predictions given data and binary labels. Then, an RF regressor at each lower-level node was trained to predict the residual error of the estimator at the parent node [16]. The overall probabilistic prediction at a node is thus formed by adding its prediction to those from all preceding levels (thresholding to keep a valid probability in the unit interval). Hyperparameter tuning was performed in the same way as for No-Share RF.
Multi-Label NN (Horizontal Sharing) - A fully connected, 4-layer, Multi-Label Neural Network (NN) was trained to predict a label corresponding to each of the 983 EC Numbers. In this way, all EC Numbers share a common feature transformation trained to improve performance across all EC Numbers. Network hyperparameters (hidden layer size and dropout probability) were chosen via a random search 3-fold cross validation, maximizing Mean Average Precision across EC Numbers.
HMCN-F (Horizontal Plus Hierarchical Sharing) - We implemented a state-of-the-art NN architecture for multi-label classification called the Hierarchical Multi-label Classification Network (known as HMCN-F, where F indicates it is feed-forward) [17]. Hyperparameter tuning was performed in the same way as for Multi-Label NN.

2.3 Training methods – confidence weighting of unlabeled data

Providing a per-example weight (a scalar positive value) used to make some examples more important during training is a common technique to overcome label balance issues or account for unlabeled data that may unknowingly contain positive examples [18]. Ultimately, we assign an overall weight to each molecule-enzyme pair that is the product of a scalar similarity weight and a scalar label weight. The similarity weight is used to denote our confidence in the provided positive or negative label. It is set to 1 for positive examples and inhibitors. Each unlabeled molecule is assigned a negative label together with a similarity weight set to a scalar between 0 and 1 computed as one minus the maximum structural similarity found between the unlabeled molecule and all molecules in the positive set. The label weight is equal for all examples of the same label. It is set to 1 for all positive examples, and for negative examples it is assigned to enforce that the aggregate weight of samples with negative label is equal to aggregate weight of samples with positive label for each EC Number.

3 Results

3.1 Comparing models without accounting for known inhibitors

We first consider training without including inhibitor information. That is, all inhibitors are treated as unlabeled molecules, and are thus assigned similarity-based confidence weights. We also show results of a baseline 'Random' model, where every enzyme-molecule pair in the test set is assigned a value selected from a uniform distribution between 0 at 1. We measured the performance of our models via three metrics: Mean Area Under the Receiver Operating Characteristic (Mean AUROC), Mean Average Precision (Mean AP) and Mean R-Precision (Mean R-PREC) [19]. Each overall summary score is respectively computed by taking the mean of per-enzyme scores across all enzymes in the corresponding dataset (Figure 1A-C). We also measure the relative performance of the models by ranking them on each of the 983 binary classification tasks (one per EC Number) and then summarize the results by computing the average rank. We repeat this for each of the performance metrics (AUROC, AP and R-PREC) (Figure 1D).

The results show several trends. k-NN Similarity is a strong baseline for our classification task, however, it is outperformed by almost all machine learning models. In general, we see that increasing the degree of information-sharing across EC Numbers increases performance, with HMCN-F (horizontal + hierarchical sharing) showing the best mean scores and the best ranks across all three metrics. The only exception to this rule is that Hierarchical RF performs worse than No-Share RF everywhere except for average R-PREC rank. We believe this to be due to the greedy training of Hierarchical RF, which limits its ability to fix any mistakes made at higher levels of the hierarchy. Overall, these results illustrate that machine learning models are to be strongly preferred to similarity with regards to R-PREC (Figure 1D).

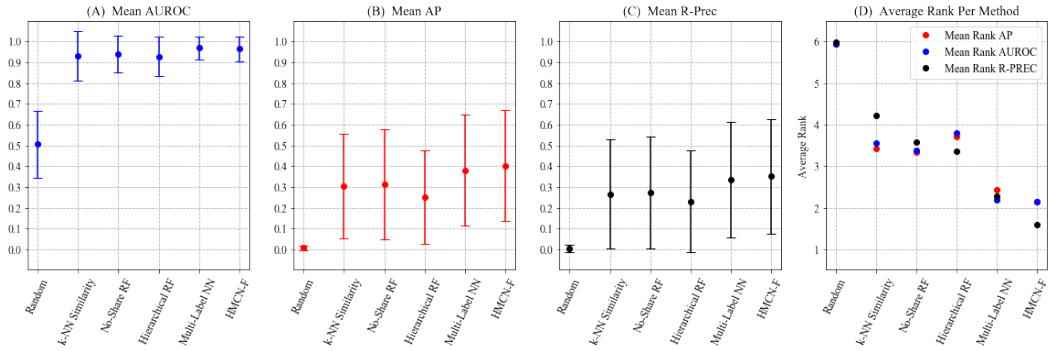

Figure 1: Evaluation of machine learning models trained without including inhibitor information and evaluated using the Full Test Set. (A) Mean AUROC, (B) Mean AP, (C) Mean R-PREC, and (D) Average rank per method (lower is better).

3.2 Comparing models while accounting for inhibitors

To evaluate the impact of including known negative examples in the form of inhibitors, we retrained our models including inhibitor information, thus setting their similarity weight to 1. We test these models using the same framework as in section 3.1, while also providing a comparison between each model trained with and without accounting for inhibitors (Figures S1 and S2). The results show that all four models, when trained with inhibitor information, improve their performance in terms of average R-PREC rank (Figure S2), and most models also improve with regards to the other metrics. Thus, the best model for the enzyme promiscuity prediction task is the HMCN-F trained with inhibitor information.

We also tested all models, trained with and without inhibitors, on an "Inhibitor Test Set" and on an "Unlabeled Test Set". The Inhibitor Test Set contains only positive and inhibitor test interactions, whereas the Unlabeled Test Set contains only positive and unlabeled interactions in the same ratio as in the Inhibitor Test Set. The results show that inhibitors are harder to distinguish from positives than the unlabeled are, thus confirming that they are hard negative examples (Figure S3).

4 Conclusion and Discussion

This work proposed and evaluated several methods to predict enzyme promiscuity on a query molecule. Our results show that sharing information both horizontally across EC Numbers and vertically across the EC hierarchy results in the highest gains in prediction quality. Indeed, the HMCN-F trained with accounting for inhibitor information is the best model for the enzyme promiscuity prediction task, achieving a Mean R-PREC of 0.359 across 983 EC numbers. An R-PREC of 0.359 represents a useful hit rate for wet-lab experimentation: for a typical enzyme, slightly more than 1 in 3 wet-lab trials would succeed if we selected the top-ranked molecules in our test set. Furthermore, our experiments show that inhibitors are hard negative examples. Indeed, HMCN-F performs with a Mean R-PREC of 0.955 on the Unlabeled Test set, while yielding a Mean R-PREC of 0.873 on the Inhibitor Test Set, indicating that inhibitors are harder to distinguish from positives when compared to unlabeled molecules.

The work presented here can be improved by integrating alternative methods from the PU learning literature [20, 21] and by considering learned representations that better capture molecular structure than binary fingerprint vectors, e.g. [22]. While there are works that use sequences to predict protein function in terms of GO terms [23, 24] and to predict EC numbers [25], the problem solved within predicts enzyme classes that act on a query molecule. This problem is pressing in synthetic biology and biological engineering applications when constructing biochemical conversion routes to synthesize valuable specialty chemicals such as biofuels, solvents, and polymers [26]. This problem is also important when constructing biodegradation pathways of environmental pollutants and xenobiotics. Our tool can be combined with route exploration and construction tools [27] to allow for novel transformation steps that are not currently documented in existing databases.

References

1. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Available from: https://web.archive.org/web/20060219074423/http://www.chem.qmul.ac.uk/iubmb/enzyme/.
2. D'Ari, R. and J. Casadesus, Underground metabolism. Bioessays, 1998. 20(2): p. 181-6.
3. Nobeli, I., A.D. Favia, and J.M. Thornton, Protein promiscuity and its implications for biotechnology. Nature biotechnology, 2009. 27(2): p. 157-67.
4. Khersonsky, O., C. Roodveldt, and D.S. Tawfik, Enzyme promiscuity: evolutionary and mechanistic aspects. 2006.
5. Khersonsky, O. and D.S. Tawfik, Enzyme promiscuity: a mechanistic and evolutionary perspective. Annual review of biochemistry, 2010. 79: p. 471-505.
6. Greene, N., et al., Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res, 1999. 10(2-3): p. 299-314.
7. Marchant, C.A., K.A. Briggs, and A. Long, In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. Toxicol Mech Methods, 2008. 18(2-3): p. 177-87.
8. Adams, S.E., Molecular similarity and xenobiotic metabolism. 2010, University of Cambridge.
9. Yousofshahi, M., et al., PROXIMAL: a method for Prediction of Xenobiotic Metabolism. BMC systems biology, 2015. 9(1): p. 94.
10. Jeffryes, J.G., et al., MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of Cheminformatics, 2015.
11. Pertusi, D.A., et al., Predicting novel substrates for enzymes with minimal experimental effort with active learning. Metab Eng, 2017. 44: p. 171-181.
12. Pertusi, D.A., et al., Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics, 2015. 31(7): p. 1016-24.
13. Hattori, M., et al., SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic acids research, 2010. 38(suppl_2): p. W652-W656.
14. Kim, S., et al., PubChem substance and compound databases. Nucleic acids research, 2015. 44(D1): p. D1202-D1213.
15. Rogers, D. and M. Hahn, Extended-connectivity fingerprints. Journal of chemical information and modeling, 2010. 50(5): p. 742-754.
16. Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.
17. Wehrmann, J., R. Cerri, and R. Barros, Hierarchical Multi-Label Classification Networks, in Proceedings of the 35th International Conference on Machine Learning, D. Jennifer and K. Andreas, Editors. 2018, PMLR: Proceedings of Machine Learning Research. p. 5075–5084.
18. Liu, B., et al. Building Text Classifiers Using Positive and Unlabeled Examples. in ICDM. 2003. Citeseer.
19. Manning, C.D., Prabhakar Raghavan, and Hinrich Schutze, Evaluation in information retrieval, in Introduction to information retrieval. 2009.
20. Bekker, J. and J. Davis, Learning from positive and unlabeled data: A survey. arXiv preprint arXiv:1811.04820, 2018.
21. Zhang, B. and W. Zuo. Learning from positive and unlabeled examples: A survey. in 2008 International Symposiums on Information Processing. 2008. IEEE.
22. Jin, W., R. Barzilay, and T. Jaakkola, Junction tree variational autoencoder for molecular graph generation. arXiv preprint arXiv:1802.04364, 2018.
23. Kulmanov, M., M.A. Khan, and R. Hoehndorf, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics, 2017. 34(4): p. 660-668.
24. Feng, S., P. Fu, and W. Zheng, A hierarchical multi-label classification method based on neural networks for gene function prediction. Biotechnology & Biotechnological Equipment, 2018. 32(6): p. 1613-1621.
25. Rousu, J., et al., Kernel-based learning of hierarchical multilabel classification models. Journal of Machine Learning Research, 2006. 7(Jul): p. 1601-1626.
26. Chubukov, V., et al., Synthetic and systems biology for microbial production of commodity chemicals. npj Systems Biology and Applications, 2016. 2: p. 16009.
27. Moura, M., L. Broadbelt, and K. Tyo, Computational tools for guided discovery and engineering of metabolic pathways, in Systems metabolic engineering. 2013, Springer. p. 123-147.

Appendix

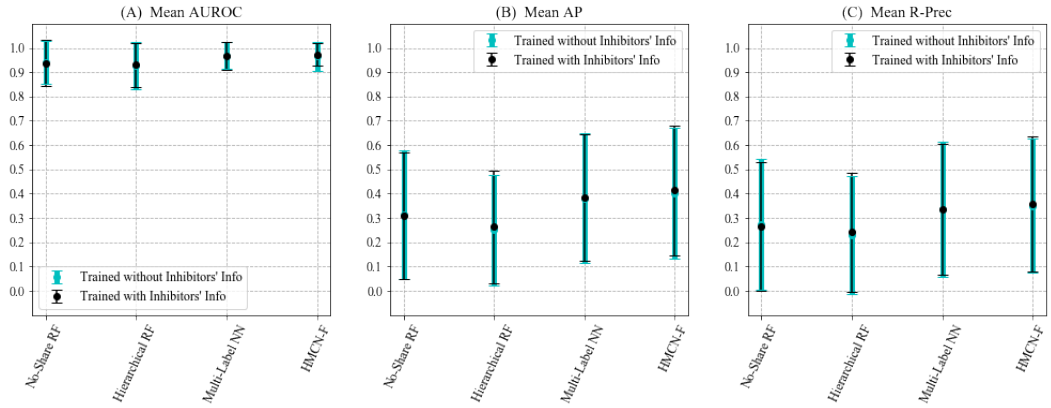

Figure S1: Mean scores showing results both for models trained with and without inhibitor information. (A) Mean AUROC, (B) Mean AP, (C) Mean R-Prec. The vertical bars indicate +/- 1 standard deviation.

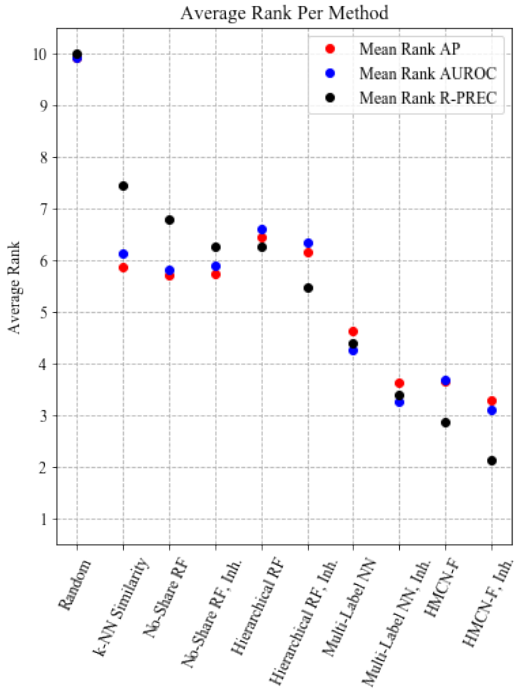

Figure S2: Average rank across metrics for all proposed models.