Bioinformatics, 37(14), 2021, 2017–2024
doi: 10.1093/bioinformatics/btab054
Advance Access Publication Date: 30 January 2021
Original Paper

Systems biology

Enzyme promiscuity prediction using hierarchy-informed multi-label classification

Gian Marco Visani[1], **Michael C. Hughes**[1] **and Soha Hassoun**[1,2,*]

[1]Department of Computer Science, Tufts University, Medford, MA 02155, USA and [2]Department of Chemical and Biological Engineering, Tufts University, Medford, MA 02155, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on July 31, 2020; revised on December 30, 2020; editorial decision on January 22, 2021; accepted on January 25, 2021

Abstract

Motivation: 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. Our data consists of enzyme-substrate interactions from the BRENDA database. Some interactions are attributed to natural selection and involve the enzyme's natural substrates. The majority of the interactions however involve non-natural substrates, thus reflecting promiscuous enzymatic activities.

Results: We frame this 'enzyme promiscuity prediction' problem as a multi-label classification task. We maximally utilize inhibitor and unlabeled data to train prediction models that can take advantage of known hierarchical relationships between enzyme classes. We report that a hierarchical multi-label neural network, EPP-HMCNF, is the best model for solving this problem, outperforming *k*-nearest neighbors similarity-based and other machine-learning models. We show that inhibitor information during training consistently improves predictive power, particularly for EPP-HMCNF. We also show that all promiscuity prediction models perform worse under a realistic data split when compared to a random data split, and when evaluating performance on non-natural substrates compared to natural substrates.

Availability and implementation: We provide Python code and data for EPP-HMCNF and other models in a repository termed EPP (Enzyme Promiscuity Prediction) at https://github.com/hassounlab/EPP.

Contact: soha@cs.tufts.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Characterizing activities of enzymes plays a critical role in advancing biological and biomedical applications. While enzymes are traditionally assumed *specific*, acting on a particular molecule, many enzymes, if not all, have *promiscuous* activities acting on 'non-natural' substrates, ones other than those that the enzyme evolved to transform (D'Ari and Casadesús, 1998; Khersonsky et al., 2006; Khersonsky and Tawfik, 2010; Nobeli et al., 2009). Despite efforts in understanding types of promiscuity (substrate versus catalytic promiscuity) and cataloguing enzyme activities on various substrates in databases, comprehensive characterization of enzyme promiscuity remains elusive.

Several applications drive the development of computational tools to analyze enzyme promiscuity. The prediction of promiscuous products of Cytochromes P450 is enabled by techniques such as Metaprint2D-react (Adams, 2010), and PROXIMAL (Yousofshahi et al., 2015). Predicting products of metabolism, e.g. the MINEs database (Jeffryes et al., 2015) and BioTransformer (Djoumbou-Feunang et al., 2019), facilitates suggesting chemical identities for compounds collected through untargeted metabolomics. The approach proposed by PROXIMAL (Yousofshahi *et al.*, 2015) was extended to enzymes other than Cytochromes P450 enzymes to facilitate annotation of untargeted metabolomics (Hassanpour et al., 2020) and for suggesting enzymatic activities responsible for catalogued products in metabolomics databases (Amin et al., 2019a, b). Prediction of putative enzymatic links, e.g. Selenzyme (Carbonell et al., 2018), XTMS (Carbonell et al., 2014) and ELP (Jiang et al., 2020), allows the construction of novel biosynthesis or biodegradation pathways. The study of enzyme promiscuity also elucidates the evolution of metabolic networks (Carbonell et al., 2011). Further, enzyme promiscuity can be tuned and exploited for novel protein engineering applications (Chen and Arnold, 2020).

We investigate in this paper data-driven approaches to predict enzyme classes, as defined via the Enzyme Commission (EC) numbers, that are likely to interact with a given query molecule. One approach is based on *k*-nearest neighbor (*k*-NN) similarity, where fingerprint similarity is calculated between the query molecule and the ‘natural’ (native) substrates that are known to be catalyzed by the enzyme under natural selection and under cellular physiological conditions. The remaining four approaches are novel adaptations of several machine-learning models that frame the problem as multi-label classification, where each predicted label corresponds to an EC number. Each model has an architecture that embeds a different information-sharing paradigm across enzymes. Some models exploit hierarchical relationships encoded in the EC Nomenclature, which provides a tree-structured hierarchical classification of enzymes.

While there are works that predict enzymatic reaction similarities (Carbonell and Faulon, 2010; Rahman et al., 2014), or use sequences to predict protein function in terms of Gene Ontology (GO) terms (Feng et al., 2018; Kulmanov et al., 2018; Roy et al., 2012) or EC numbers (Concu and Cordeiro, 2019; Dalkiran et al., 2018; Kumar and Choudhary, 2012; Li et al., 2018; Ryu et al., 2019), or to predict the likelihood of a sequence catalyzing a reaction or to quantify the affinity of sequences on substrates (Mellor et al., 2016), the problem solved herein predicts enzyme classes that act promiscuously on a query molecule. To train our models, we utilize data from the BRENDA database (Schomburg et al., 2017), which lists natural and non-natural substrates interreacting with the catalogued enzyme classes. The listed compounds for each enzyme are predominantly non-natural substrates. As we train our models using this dataset of predominantly promiscuous substrate-enzyme interacting pairs, we refer to this problem as the ‘enzyme promiscuity prediction’ problem. Solving the enzyme promiscuity prediction problem efficiently as proposed herein allows for the quick exploration of enzyme classes that act on molecules. These findings can be further refined by identifying specific sequences that interact with the molecules and that are compatible with the host. Our best-in-class technique, EPP-HMCNF, can therefore be integrated with a variety of tools to explore biotransformation routes (Moura et al., 2013), assess solubility within the host (Amin et al., 2019a,b), or determine the likelihood of an enzyme sequence acting on a molecule borrowing from drug-target interaction prediction techniques (Chen et al., 2018).

A major challenge when addressing this problem is the lack of available data with sufficiently representative examples of negative cases (non-interacting pairs). The BRENDA database provides cases of positive enzyme-molecule interactions. The BRENDA database also lists inhibitor molecules, those that bind with the enzyme (similar to positive cases) but where the catalytic activity is inhibited. Inhibitors thus do not uniformly represent the negative set and cannot be utilized as such. To address the lack of negative data (non-interacting enzyme-substrate pairs), we select molecules from the BRENDA database that are not in the enzyme’s positive nor inhibitor lists, referred to herein as unlabeled molecules, and treat them as negatives during training and testing. This assumption is reasonable as negative interactions far outnumber positive interactions in nature. Importantly, to make judicious use of this unlabeled data during training, we apply a probabilistic weight to each such example to reflect our confidence in the negative label, deriving the weight from the molecule’s structural similarity score to the enzyme’s most similar positive example.

Our results show that the best model for predicting enzyme promiscuity is EPP-HMCNF, a model based on Hierarchical Multi-label Classification Network (known as HMCN-F, where F indicates a feed-forward architecture) (Wehrmann et al., 2018). EPP-HMCNF provides information sharing along the EC hierarchy and across enzymes at each level of the hierarchy. We investigate the performance of EPP-HMCNF against similarity-based methods, finding that, while similarity is a competitive baseline, EPP-HMCNF delivers better performance and stands to offer further improvement as more labeled data becomes available. The main contributions of this work are:

- We develop and evaluate machine-learning classifiers with a range of different patterns of information sharing across enzymes
- We demonstrate the effective utilization of a large dataset of promiscuous interactions. We provide predictions for 983 enzyme classes, a huge increase over previous efforts that made predictions for four specific enzymes (Pertusi et al., 2017) using SVMs and active learning.
- While substrate similarity is widely used for determining promiscuity in metabolomic engineering practice (Pertusi et al., 2015), there is currently no large-scale systematic evaluation of the effectiveness of similarity in predicting enzyme promiscuity across enzymes or enzyme classes. Further, prior works have shown no consensus on a similarity level that deems a query molecule sufficiently similar to a native substrate. Our work offers such a study. Importantly, our work shows that machine learning outperforms similarity-based methods on several metrics including R-Precision (R-PREC), the metric that is most significant from a user’s perspective as it best correlates with downstream usefulness for an experimentalist.
- Our experiments show that inhibitors are hard negative examples and that including inhibitor information during training consistently improves predictive power, particularly for EPP-HMCNF.
- Our results show that all promiscuity prediction models perform worse under a realistic data split (Martin et al., 2017), when compared with a random data split, and when evaluating performance on non-natural substrates compared to natural substrates.

2 Materials and methods

2.1 Dataset

All positive and inhibitor molecules were collected from BRENDA, excluding co-factors because these metabolites are common among many enzymatic reactions. The Morgan fingerprint, with a radius of 2 and 2048 binary features (Rogers and Hahn, 2010), is used to represent each molecule. Compounds in BRENDA that could not be mapped to a specific molecular structure were discarded. 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 8295 unique molecules. Within this same set of molecules, we also identified 13 087 inhibiting interactions, based on 2165 unique inhibitors. Some enzymes had limited positive data. We focused on enzymes that had a minimum of 10 positive examples, so there would be enough data for both training and assessment. Using these inclusion criteria, our dataset consisted of 983 distinct EC numbers, of which 730 had at least one known inhibitor. All 983 EC numbers came from top-level classes 1-6, as we excluded the recently established top-level class 7 due to insufficient data. Of all positive pairings, only approximately 13% are associated with an EC class paired with a natural substrate. These pairings are associated with in vivo metabolism as specified under each EC Number’s ‘Natural Substrates Products’ section in the BRENDA database. The ratio of natural substrate to total interacting positive molecules, listed under the ‘Substrates/Products’ section, varied per EC Number (Fig. 1).

For every EC number, each of the 8295 molecules that were not known to be positive nor inhibitors are considered unlabeled in the BRENDA database. Per EC number, these unlabeled molecules were treated as negative examples during training, as were the inhibitors. Across enzymes, the ratio of positive-to-all molecules was on average 0.0032, with a standard deviation of 0.0044. The 10th percentile of this distribution was 0.0013, while the 90th was 0.0057.

Since inhibitors bind with enzymes, they might be considered ‘closer’ to the positive molecules and thus be more difficult (harder) to classify. In machine-learning applications, including such ‘hard negative’ examples can fine-tune the decision boundary between positives and negatives (e.g. Radenović et al., 2016). We validated that inhibitor molecules across enzymes are, on average, more similar (based on Tanimoto score) to their respective positive molecules than a similar-size randomly selected set of unlabeled molecules (Supplementary File S1, Supplementary Section S1, Supplementary Fig. S1), thus confirming their status as hard negative examples.

Data was further organized in a tree hierarchy to match the structure of the EC nomenclature. There were 6 nodes at the class

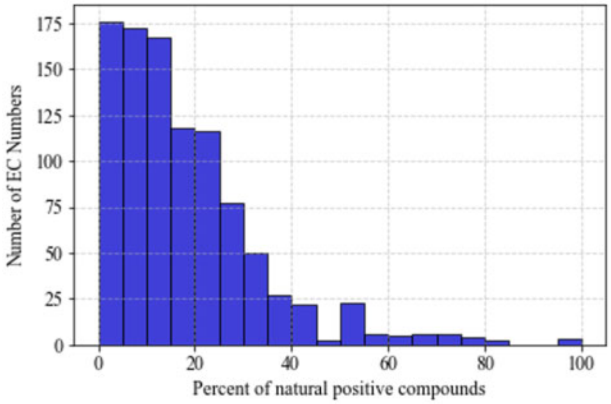

Fig. 1. Distribution of per EC number percentage of natural compounds to all compounds in our dataset

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 labeled 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. In the cases in which a molecule's label is conflicted between positive and inhibitor, which may occur at internal nodes of the hierarchy, only the positive label is assigned.

2.2 Models

We implemented one model based on fingerprint similarity and four machine-learning models. The machine-learning models (Supplementary Fig. S2) are presented in the order corresponding to the amount of information sharing they allow across enzymes: no sharing (predictions for each leaf node are developed independently); top-down hierarchical sharing (each leaf uses learned representations from parents); horizontal sharing (each leaf uses learned representations common across all enzymes); and horizontal-plus-hierarchical sharing (each leaf predictor uses shared representations, as well as representations common to leaves that share a parent).

2.2.1 *k*-NN similarity (no sharing)

Each EC Number-molecule pair in the test set was scored by computing the mean similarity between the test molecule and the *k* most similar positive molecules for the relevant EC Number in the training set. Similarity is evaluated via the Tanimoto score (Bajusz et al., 2015) on the Morgan fingerprints. Hyperparameter *k* was individually optimized for each EC Number via 3-fold cross-validation grid search on the training set, selecting the value with the best average precision score. The search considered the following values of *k*: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 'all', where 'all' indicates that the mean similarity was computed between the test molecule and all positive molecules in the training set. A value of $k = 1$ corresponds to computing the maximum similarity, while the 'all' value corresponds to computing the mean similarity.

2.2.2 No-share RF (no sharing)

An independent Random Forest (RF) binary classifier (Breiman, 2001) was built for every leaf in the tree (distinct EC number). Each RF classifier was composed of 50 decision trees. At each internal decision node, a decision tree was trained to optimize the Gini metric while allowing for the selection of a random subset of features of size proportional to the square root of the number of features. During training, we selected model complexity hyperparameters (the minimum size of any tree's terminal node) by performing a 5-fold cross-validation grid search. The grid search considered minimum terminal node sizes of 1, 3, 5, 10, 20, 50 and 100, selecting the value with best average precision score.

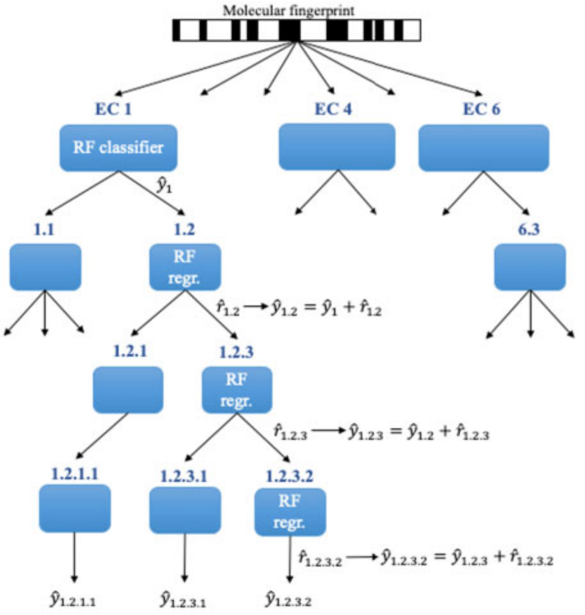
Molecular fingerprint
EC 1
EC 4
EC 6
RF classifier
$\hat{y}_1$
1.1
1.2
6.3
RF regr.
$\hat{r}_{1.2} \rightarrow \hat{y}_{1.2} = \hat{y}_1 + \hat{r}_{1.2}$
1.2.1
1.2.3
RF regr.
$\hat{r}_{1.2.3} \rightarrow \hat{y}_{1.2.3} = \hat{y}_{1.2} + \hat{r}_{1.2.3}$
1.2.1.1
1.2.3.1
1.2.3.2
RF regr.
$\hat{r}_{1.2.3.2} \rightarrow \hat{y}_{1.2.3.2} = \hat{y}_{1.2.3} + \hat{r}_{1.2.3.2}$
$\hat{y}_{1.2.1.1}$
$\hat{y}_{1.2.3.1}$
$\hat{y}_{1.2.3.2}$

Fig. 2. Partial structure of the Hierarchical RF. RF classifiers are trained at the top-level nodes, while RF regressors are trained at the lower-level nodes. Example calculations of probabilistic predictions are shown at some nodes

2.2.3 Hierarchical RF (greedy top-down hierarchical sharing)

This model (Fig. 2) uses a tree-like architecture that mimics the EC nomenclature hierarchy. 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 six 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. 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). The hyperparameters of the RF predictors and RF regressors were selected in the same way as in No-Share RF.

2.2.4 Multi-label NN (horizontal sharing)

This model consists of a fully connected, Multi-Label Neural Network (NN) (Tsoumakas and Katakis, 2007). To make promiscuity predictions, each molecule's input features are fed through a NN with four hidden layers to obtain a learned representation common to all enzymes, and then a final output layer with sigmoid activation produces enzyme-specific probabilistic predictions given this common representation. In this way, all enzymes share a common feature transformation trained to improve performance across all enzymes. However, unlike Hierarchical RF, there is no specific usage of hierarchical information that would prioritize sharing between enzymes of the same class, sub-class or sub-sub-class.

Each layer is fully connected with ReLu activation followed by batch normalization and dropout. Residual connections are present between the hidden layers. Binary cross entropy was used as loss. We used the Adam optimizer with a batch size of 12. The learning rate was manually set to 0.0002, and the model converged after 70 epochs. We cross-validated our model using random search (Bergstra and Bengio, 2012) across 12 pairs of dropout and layer size values that were sampled from uniform distributions between 0.01 and 0.6, and between 128 and 384, respectively. We selected values that maximized the average precision score.

2.2.5 EPP-HMCNF (horizontal plus hierarchical sharing)

We adapt HMCN-F, a state-of-the-art NN architecture for hierarchical multi-label classification (Wehrmann *et al.*, 2018) for our task. HMCN-F consists of one global predictor, which predicts all labels in the hierarchy (producing a probability for each internal node and leaf node), and local predictors for each level in the hierarchy, which predicts labels only for that specific level. The global and local predictions are then merged into a final prediction. In our adaptation (Fig. 3), our model EPP-HMCNF was implemented with four global hidden layers and four local hidden layers, one dedicated to each enzyme class. Each layer is fully connected with ReLU activation followed by batch normalization and dropout. Probabilistic predictions are made via fully connected output layers with sigmoid activation. Residual connections are present between the hidden layers in the global predictor. EPP-HMCNF was trained similarly to the Multi-Label NN, for 70 epochs. The final prediction is obtained by equally weighting the predictions from the global and local predictors. Information is therefore shared across enzymes both in the horizontal fashion of the Multi-Label NN as well as in the vertical fashion across the EC hierarchy, as in Hierarchical RF. Unlike Hierarchical RF, however, training is not greedy across the hierarchy, but happens simultaneously at all levels.

2.3 Confidence weighting of unlabeled data

Providing a per-example weight (a scalar positive value) to make some examples more important during training is a common technique to overcome label balance issues or to account for unlabeled data that may unknowingly contain positive examples (Liu et al., 2003). We assign an overall weight $\boldsymbol{w_{m,e}}$ to each molecule-enzyme pair $(\boldsymbol{m}, \boldsymbol{e})$ that is the product of a scalar similarity weight and a scalar label weight:

$$\boldsymbol{w_{m,e}} = \boldsymbol{SimW_{m,e}} * \boldsymbol{LabelW_{m,e}}$$

The similarity weight $SimW_{m,e}$ is used to denote our confidence in the provided positive or negative label. This value will be set to 1 for positive examples and inhibitors (since we are confident in their negative label). However, when a molecule is unlabeled for an enzyme in BRENDA, we assign it a negative label with a similarity weight between 0 and 1. This value is set to one minus the maximum similarity between the unlabeled molecule and all molecules associated with the corresponding enzyme in the positive set. Similarity is scored using the Tanimoto score for two molecular fingerprints.

To overcome the challenge of negative labels far outnumbering positive examples, all examples are given a label weight $\boldsymbol{LabelW_{m,e}}$, which is the same for all examples of the same label. The label weight is set to 1.0 for positive molecule-enzyme interactions. For negative examples, it is set as:

$$\boldsymbol{LabelW_{m,e}} = \frac{\boldsymbol{NumPositive_{m,e}}}{\sum_m \boldsymbol{SimW_{m,e}}}$$

This weight enforces that the aggregate weight of samples in the negative class is equal to that of the positive class for each enzyme, thus balancing the influence of each binary class during training.

We provide these weights when training RF predictors using the 'sample_weight' keyword argument in Scikit-Learn (Pedregosa et al., 2011). For the NNs, we scaled the contribution of each molecule-enzyme to the loss function with its relative weight. If $\hat{y}_{m,e}$ is the predicted probability and y the true binary label, then:

$$\boldsymbol{loss} = \sum_{\boldsymbol{m,e}} \boldsymbol{w_{m,e}} * \boldsymbol{binaryCrossEntropy}(\boldsymbol{y_{m,e}},\ \hat{\boldsymbol{y}}_{\boldsymbol{m,e}})$$

While we use weighted datasets for training to obtain more robust models, during model evaluation all examples are unweighted.

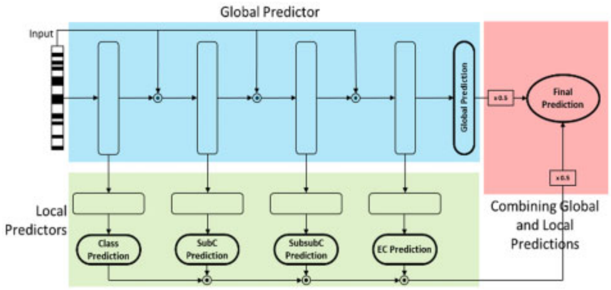

Fig. 3. EPP-HMCNF Structure. Global predictors are for each node in the hierarchy. Local predictors are for each level of the hierarchy. Junctions (circles with a dot inside) denote concatenation

3 Results

We first trained and evaluated our models under the classic random train-test split. Using confidence weighting of unlabeled data, we trained the No-Share RF, Hierarchical RF, Multi-Label NN and HMCN-F. We then evaluated the models, together with k-NN Similarity. We compared our methods against a baseline 'Random' model, where every enzyme-molecule pair in the test set is assigned a likelihood of interaction value that is selected from a uniform distribution between 0 and 1.

We perform our experiments on several test sets. The 'Full Test Set' comprises all enzyme-molecule paired interactions that were saved for testing, corresponding to 20% of the total interactions collected from BRENDA. The 'Inhibitor Test Set' includes only the positive and inhibitor interactions, and none of the unlabeled ones, for each enzyme. The positive and inhibitor interactions were selected from the Full Test Set. Only 885 out of the 983 EC Numbers presented inhibitors, and only 671 EC Numbers presented inhibitor interactions among the data saved for testing. Therefore, only these 671 were considered for testing. The 'Unlabeled Test Set' contains positive and unlabeled examples, but excludes known inhibitors, in the same positive-to-total ratio as the Inhibitor Test Set. For all test sets, unlabeled molecules are uniformly sampled within each EC Number's set. To generate a more competitive and realistic testing scenario, we split the molecular data into training and testing under a realistic split (Martin et al., 2017), and replicated some experiments under this split.

We measured the performance of our models via three metrics: Mean Area Under the Receiver Operating Characteristic (AUROC), Mean Average Precision (AP) and Mean R-Precision (R-PREC) (Manning, 2009). Each overall summary score is respectively computed by averaging the per-EC-Number scores across all EC Numbers in the corresponding dataset. These metrics were selected as they do not require thresholding and instead consider the ranking of positive examples relative to the unknowns in the test data. This is particularly well suited for our purposes, since the similarity models do not return results that can be interpreted as probabilities, and thus thresholding would have required ad hoc and likely unjustifiable assumptions. To compute R-PREC, we are given a list of candidate molecules as input with knowledge that some number R of them is truly relevant, and then we compute the precision among the R top-scoring candidates as ranked by the predictor. A model with high R-PREC has a high probability of ranking positive molecules ahead of unknown 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 useful for end users.

3.1 Comparing the performance of models trained without known inhibitors

To train without including inhibitor information, inhibitors are treated as unlabeled molecules and are assigned similarity-based confidence weights. Figure 4 A–C shows a comparison of all models for each metric. Each score is reported with ±1 standard deviation across all EC Numbers. The results show that k-NN Similarity is a strong baseline for our classification task, thus confirming that substrate similarity is a valuable tool to predict promiscuity. However, similarity is outperformed by most machine-learning techniques. Increasing information-sharing across EC Numbers improves performance, with EPP-HMCNF showing the best mean scores across AP and R-PREC and tying with Multi-Label NN in terms of

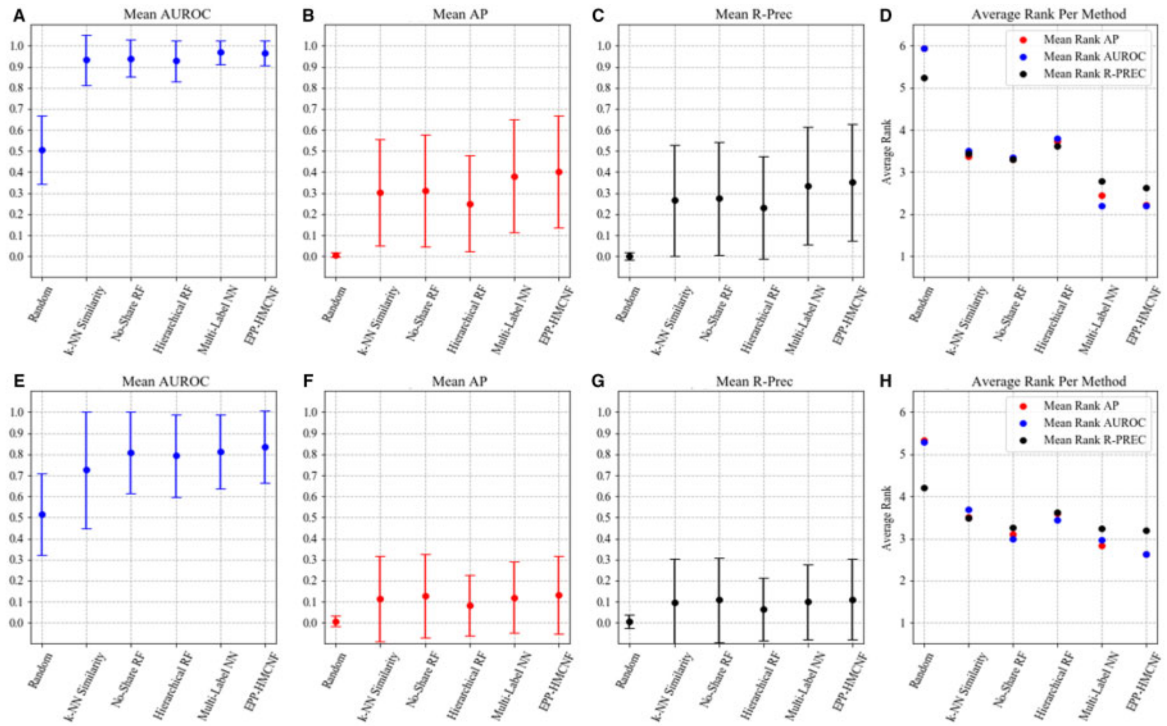

Fig. 4. Evaluation of promiscuity prediction models trained without including inhibitor information under random split (top row), and under realistic split (bottom row), and evaluated using the Full Test Set. (**A, E**) Mean AUROC, (**B, F**) Mean AP, (**C, G**) Mean R-PREC and (**D, H**) Average rank per method. Intervals shown in (A, B, C, E, F, G) represent ±1 standard deviation

AUROC. Hierarchical RF performs worse than No-Share RF with regards to all metrics. The greedy training of Hierarchical RF may limit its ability to fix any mistakes made at higher levels of the hierarchy. Notably, the ±1 standard deviation intervals are overlapping across all models, indicating that the difference in performance across models is smaller than the difference across ECs.

We also measure the relative performance of our 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. Whenever two or more models tie in rank on a particular EC Number, we assign to all disputing models the average of the disputed ranks; this occurs more frequently on R-PREC scores, since its computation considers only a handful of predictions per EC Number. We repeat this for each of the three performance metrics (Fig. 4D). EPP-HMCNF ranks best for AP and R-PREC, tying with Multi-Label NN with regards to AUROC, consolidating itself as best model for this task. Overall, the results in Figure 4D are consistent with those in Figure 4A–C: models with higher scores in Figure 4A–C generally have lower (i.e. better) rank in Figure 4D.

We analyzed the average performance of the *k*-NN and EPP-HMCNF models for each EC Class (Supplementary File S1, Supplementary Section S4, Supplementary Fig. S3). We found that some enzymes are easier to characterize than others. Class 3 hydrolase enzymes and Class 5 isomerase enzymes are consistently easiest to classify. Furthermore, we found that EPP-HMCNF outperforms *k*-NN similarity for all classes.

We analyze the average performance of the models across EC Numbers as a function of the number of positive examples. We evaluate the performance of EC Numbers with the most and least number of positive examples (Supplementary File S1, Supplementary Section S5). We show that all models have a better than average R-PREC performance and less R-PREC variability for enzymes with more labeled positive data, indicating that prediction quality is likely to increase as more data becomes available. Hierarchical RF is the model that most benefits from higher data availability, even slightly outperforming *k*-NN. While low data availability for some enzymes is partly responsible for the high variability in scores across enzymes seen in Figure 3A–C, most of the variability is best explained by suggesting that some enzymes are much easier to characterize than others, as indicated by the still high standard deviation of scores when evaluating EC Numbers with the most and least positive examples (Supplementary Fig. S3).

3.2 Training with inhibitor information

To evaluate the impact of including known negative examples in the form of inhibitors, we retrained our models including inhibitor information, setting the similarity weight to 1 for each inhibitor. We then perform two sets of experiments. First, we test the models on the 'Full Test Set' and compare our findings against the result from the no-inhibitor models. Table 1(A) summarizes the differences among the models and reports the percentage of EC Numbers for which the use of each training technique yields a better score. Hierarchical RF and HMCN-F are the models that consistently benefit the most from training with inhibitors, with EPP-HMCNF experiencing a significant boost for all metrics.

Next, we considered the same metrics on the Inhibitor Test Set. The goal of using this set is to show the capacity of each model in distinguishing between positive molecule and inhibitor molecules, the only true negative examples in our dataset. We also tested our models on the Unlabeled Test Set. Since this set contains positive and unlabeled interactions with the same ratio as the ratio between positive and inhibitor interactions of the Inhibitor Test Set, any difference in scores between the Inhibitor and the Unlabeled Test Sets can be attributed entirely to group differences between the inhibitor molecules and unlabeled molecules. Results on the two test sets for all models, trained with and without inhibitor information, are shown in Figure 5. In the Inhibitor Test Set and in the Unlabeled Test Set, the positive-to-total ratios across EC Numbers are much higher than in the Full Test Set. This is important to note since AP and R-PREC scores correlate with the positive-to-total ratio. For each of these sets, the average ratio is 0.579, with a standard deviation of 0.199, a 10th percentile of 0.286, and a 90th percentile of 0.833.

Table 1. Evaluation of training without and with inhibitors

		AUROC				AP				R-PREC			
		No-Share RF	Hierarchical RF	Multi-Label NN	EPP-HMCNF	No-Share RF	Hierarchical RF	Multi-Label NN	EPP-HMCNF	No-Share RF	Hierarchical RF	Multi-Label NN	EPP-HMCNF
						(A) Random Data Split							
Mean score across ECs	**No inhibitors**	0.940	0.928	0.969	0.965	0.312	0.251	0.381	0.402	0.274	0.230	0.334	0.351
	With inhibitors	0.937	0.931	0.967	0.973	0.310	0.264	0.384	0.414	0.267	0.241	0.335	0.359
	% Difference	−0.32	**0.32**	−0.21	**0.83**	−0.64	**5.18**	**0.79**	**2.99**	−2.55	**4.78**	**0.30**	**2.28**
% ECs for which technique is better	**No inhibitors**	49	45	52	41	50	45	49	43	51	48	50	47
	With inhibitors	**51**	**55**	48	**59**	50	**55**	**51**	**57**	49	**52**	50	**53**
						(B) Random Data Split on Reduced Dataset							
Mean score across ECs	**No inhibitors**	0.932	0.921	0.964	0.960	0.291	0.233	0.350	0.367	0.264	0.220	0.312	0.331
	With inhibitors	0.928	0.925	0.962	0.968	0.290	0.247	0.353	0.375	0.255	0.233	0.317	0.334
	% Difference	−0.43	**0.43**	−0.21	**0.83**	−0.34	**6.01**	**0.86**	**2.18**	−3.41	**5.91**	**1.60**	**0.91**
% ECs for which technique is better	**No inhibitors**	49	44	52	42	51	42	49	45	51	47	49	48
	With inhibitors	**51**	**56**	48	**58**	49	**58**	**51**	**55**	49	**53**	**51**	**52**
						(C) Realistic Data Split on Reduced Dataset							
Mean score across ECs	**No inhibitors**	0.809	0.793	0.836	0.849	0.128	0.083	0.120	0.131	0.108	0.063	0.100	0.111
	With inhibitors	0.812	0.811	0.833	0.857	0.133	0.091	0.142	0.149	0.112	0.072	0.119	0.126
	% Difference	**0.37**	**2.27**	−0.36	**0.94**	**3.91**	**9.64**	**18.33**	**13.74**	**3.70**	**14.29**	**19.00**	**13.51**
% ECs for which technique is better	**No inhibitors**	53	44	52	45	50	42	47	46	50	48	47	46
	With inhibitors	48	**56**	48	**55**	50	**58**	**53**	**54**	50	**52**	**53**	**54**

Note: (A) Under random data split on the full test set, (B) under random split on the reduced dataset (i.e. only including ECs for which models could be trained under the realistic split) and (C) under realistic split. Scores in bold indicate that the model trained with inhibitors has better performance for that metric.

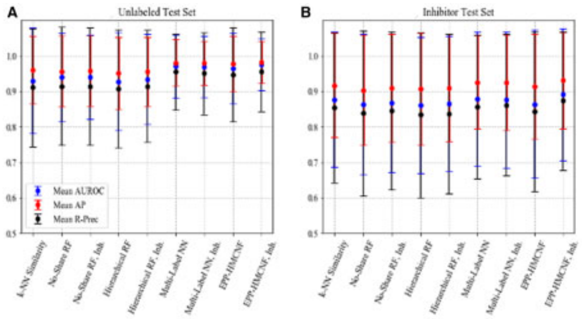

Fig. 5. Results of training with inhibitor information. (**A**) Evaluation on the unlabeled test set and (**B**) evaluation on the inhibitor test set

All scores are lower on the Inhibitor Test Set than on the Unlabeled Test Set, indicating that it is harder for the models to distinguish between positives and unlabeled. The higher standard deviation also shows that the task of discerning inhibitors from positives is harder for some EC Numbers than for others. This is further confirmation that the inhibitors can be designated as 'hard' negative examples. However, EPP-HMCNF trained with inhibitors is the best model at discerning inhibitors from positives (Fig. 5B), confirming EPP-HMCNF as our model of choice.

3.3 Realistic split

We analyzed our models based on a 'realistic data split' that contain potentially novel query compounds in comparison to the training dataset (Martin *et al.*, 2017). We clustered the molecules in our dataset on the molecular fingerprints with an average cluster size of 10. We used UPGMA (the unweighted pair group method with arithmetic mean), a simple bottom-up hierarchical clustering method (Sokal *et al.*, 1958). We saved molecules from singletons and smaller clusters, up to 20% of the total molecules, for testing. This split reduced the similarity between test and training molecules when compared to the similarity obtained under a random split (Fig. 6A and B).

The realistic split affects the distribution of positive data in the training and testing sets. Figure 6C and D plots the distribution of the ratio $r = (positives_{\text{in}} test)/(positivies_{\text{in}} train)$ per EC Number under each splitting strategy. Under the random split, r appears to have a unimodal, symmetric distribution centered at 20%. Under the realistic split, r follows an asymmetric distribution concentrated near 0% and decaying at larger values, similar to an exponential distribution. This distribution's shape results in high number of EC Numbers with either no positive test data, or no or little training data.

Under the realistic split, models could be trained only for 680 EC Numbers, with an EC hierarchy composed of 6 Classes, 45 Subclasses and 122 Sub-subclasses. The relative performance of the models under the random split on this reduced dataset (Table 1B) is consistent with the full test set (Table 1A) across all metrics and in terms of average ranking per method. We compare the performance on this reduced dataset under the two different data splits (Table 1B and C). The relative performance of all models is lower under the realistic split than under the random split, with models trained with inhibitor data demonstrating slightly less degradation in mean AP and mean R-PREC. These results show that it is more difficult to obtain good model performance under the more challenging realistic split. Importantly, mean scores for almost all metrics are improved when training with inhibitors, with Multi-Label NN also experiencing a consistent boost. Crucially, EPP-HMCNF trained with inhibitor information is still the best model for the task (Mean R-PREC 0.126 versus 0.096 of *k*-NN Similarity).

3.4 Predicting on natural versus non-natural query molecules

We evaluated the difference in performance of our best performing classifier, EPP-HMCNF trained with inhibitor information, on enzyme classes acting on natural and non-natural substrates in the Full Test Set (under the random split). For both sets of predictions, we plotted (Fig. 7) the ratio of probabilistic predictions above a series of thresholds between 0 and 1. As all test instances in this experiment are positive, higher probabilistic predictions are desirable, which translate to a higher area under the curve (AUC). The predictions on the interactions with the natural substrates have higher AUC than the predictions on the interactions with the non-natural ones (0.846 versus 0.761). We restricted the comparison to EC numbers that had both natural and non-natural test compounds. For each EC Number, we computed the mean of the predictions on the natural versus non-natural test compounds. The AUC for the former is higher than the AUC of the latter (0.852 versus 0.761). These results suggest that it is easier for EPP-HMCNF to discern enzyme class interactions with natural substrates than with non-natural ones. Importantly, per Figure 7, the model still classifies over 61% of non-natural compounds as positive with a probability above 90%, and over 78% of non-natural as positive with probability above 50%. This illustrates the potential of machine learning models to identify new non-natural compounds, and thus promiscuous interactions, for our EC Numbers.

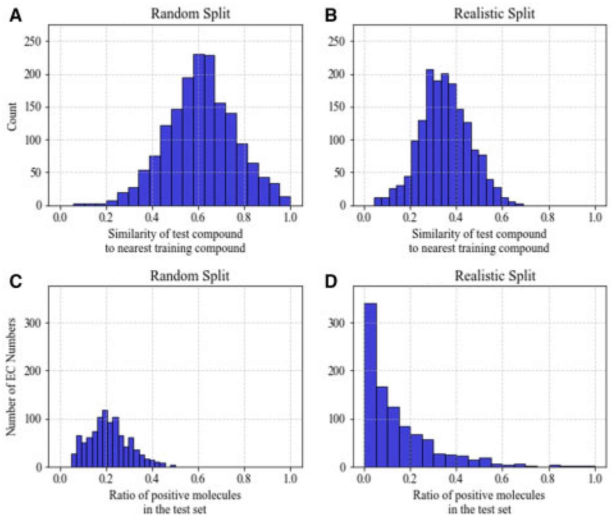

Fig. 6. Data distributions under random versus realistic split. (**A**) Histogram of maximum similarity between one test compound and all training compounds under random split. (**B**) Maximum similarity between one test compound and all training compounds under realistic split. (**C**) Distribution of r, the ratio of number of positive examples in test set over the number of positive examples in the training set per EC Number, under random split. (**D**) Distribution of r under realistic split

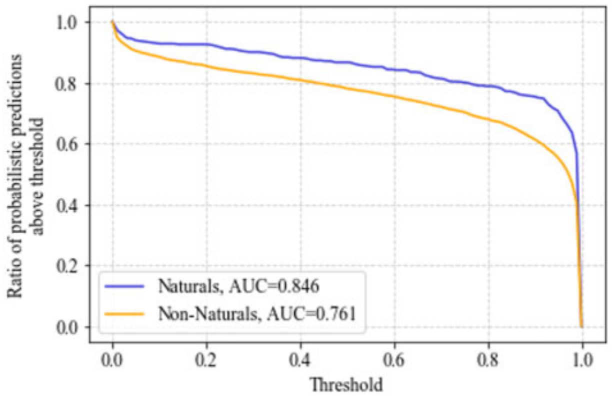

Fig. 7. Sensitivity across thresholds on natural and non-natural test interactions

3.5 Speed and scalability of training and inference

We report training and inference speed of all models, each with their respective optimized set of hyperparameters, and trained for all 70 epochs where applicable (Table 2). The time complexity (discussed in Supplementary File S1, Supplementary Section S6) is also reported. All measurements are made when training the models on the Training set and Full Test set generated via random split. Models were trained and evaluated on an Intel(R) Xeon(R) CPU X5675 with a speed of 3.07 GHz and 64 GB RAM. Further speedups can be obtained via training on GPUs. *K*-NN similarity is expensive during inference, whereas, for all our dataset of 1659 molecules, all machine learning-based models ran under 1.5 min, generating predictions for 983 EC Numbers. The fastest models at inference time are the NN-based models. Their runtime is independent of the number of EC Numbers of interest as the output layer calculates all predictions simultaneously, thus making them appealing for large and comprehensive sets of predictions. As more data becomes available, Multi-Label NN and EPP-HMCNF hold promise to be the fastest and most scalable inference models.

4 Conclusion and discussion

This work proposed and evaluated several machine learning models to predict enzyme promiscuity on a query molecule. Our results show that sharing information both horizontally across enzymes and vertically across the EC hierarchy results in large and consistent gains in prediction quality. EPP-HMNCF implementation trained with inhibitor information is the best model for the enzyme promiscuity prediction task, achieving a Mean R-PREC of 0.359 across 983 EC numbers, against the 0.265 Mean R-PREC of *k*-NN Similarity. For a typical enzyme, slightly more than 1 in 3 wet-lab trials would succeed if we selected the top-ranked molecules using EPP-HMNCF, while a similarity-based method would yield success for 1 in 4 molecules. Thus, although similarity is a competitive baseline, it is outperformed by machine learning methods. Further, similarity is less consistent than EPP-HMCNF across EC Classes. Our data analysis and experiments show that inhibitors are hard negative examples. Indeed, EPP-HMCNF trained with inhibitor information 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. Furthermore, EPP-HMCNF trained with inhibitor information is the best model at discerning known positives from known inhibitors.

Table 2. Training and inference times reported in hh:mm:ss, and time complexity on training and inference

	Train. Time	Infer. Time	Time complexity on training	Time complexity on inference
k-NN Similarity	00:00:00	00:12:54	No training	$\mathrm{O}(ec_i\ \mathrm{p}_{\mathrm{ec}}^{-}\ m)$
No-Share RF	00:12:36	00:00:21	$\mathrm{O}(ec_t\ n\log(n))$	$\mathrm{O}(ec_i\ m\ D)$
Hierar-chical RF	00:20:48	00:01:22	$\mathrm{O}(ec_n\ n\log(n))$	$\mathrm{O}(ec_i\ m\ D)$
Multi-Label NN	04:06:10	00:00:03	$\mathrm{O}(n\ ec_t)$	$\mathrm{O}(m\ ec_t)$
EPP-HMCNF	09:38:40	00:00:06	$\mathrm{O}(n\ ec_n)$	$\mathrm{O}(m\ ec_n)$

Note: n, number of training molecules; m, number of inference molecules, D, maximum tree depth across all random forests; p_{ec}, number of training positive molecules for a specific EC Number; p_{ec}^{-}, average number of training positive molecules across EC numbers; ec_t, number of EC Numbers on which the model is trained; ec_n, number of nodes in the EC hierarchy used in training (note: $ec_t < ec_n \leq 4ec_t$); ec_i, number of EC Numbers for which inference is performed (note: $ec_i \leq ec_t$).

The realistic data split generates a more challenging test set than when using the traditional random split, allowing more 'realistic' assessment of how well these models are likely to generalize to compounds which are different from those in the training set. EPP-HMCNF trained with inhibitor information remains the best model even on this harder task, with a mean R-PREC across 680 EC Numbers of 0.126. Thus, we expect a 1 in 8 wet lab hit rate, in contrast with the 1 in 10 hit rate provided by similarity. While it is important to evaluate models under circumstances that most closely mimic practical scenarios, splitting the data in a non-uniform way inevitably leads to per-label biasing under a multi-label setting, which may bias the models in making more accurate predictions for some labels over others. Future work should further address this issue.

Overall, our results suggest that predicting enzyme promiscuity through machine learning techniques that leverage existing knowledge in databases hold promise to advance biological engineering practices. The work presented here can be improved by integrating alternative methods from the PU learning literature (Bekker and Davis, 2020; Zhang and Zuo, 2008) and by considering learned representations that better capture molecular structure than binary fingerprint vectors, e.g. (Jin et al., 2018).

Funding

This research was supported by NSF (National Science Foundation), Award CCF-1909536, and also by NIGMS (National Institute of General Medical Sciences) of the National Institutes of Health, Award R01GM132391. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Adams,S.E. (2010) *Molecular Similarity and Xenobiotic Metabolism*. University of Cambridge, Cambridge.

Amin,S.A. *et al.* (2019a) Towards creating an extended metabolic model (EMM) for *E. coli* using enzyme promiscuity prediction and metabolomics data. *Microb. Cell Factories*, **18**, 109.

Amin,S.A. *et al.* (2019b) Establishing synthesis pathway-host compatibility via enzyme solubility. *Biotechnol. Bioeng.*, **116**, 1405–1416.

Bajusz,D. *et al.* (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.*, **7**, 20.

Bekker,J., and Davis,J. (2020) Learning from positive and unlabeled data: a survey. Mach. Learn., **109**, 719–760.

Bergstra,J. and Bengio,Y. (2012) Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, **13**, 281–305.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Carbonell,P. and Faulon,J.L. (2010) Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics*, **26**, 2012–2019.

Carbonell,P. *et al.* (2011) Origins of specificity and promiscuity in metabolic networks. *J. Biol. Chem.*, **286**, 43994–44004.

Carbonell,P. *et al.* (2014) XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.*, **42**, W389–394.

Carbonell,P. *et al.* (2018) Selenzyme: enzyme selection tool for pathway design. *Bioinformatics*, **34**, 2153–2154.

Chen,K. and Arnold,F.H. (2020) Engineering new catalytic activities in enzymes. *Nat. Catal.*, **3**, 203–213.

Chen,R. *et al.* (2018) Machine learning for drug–target interaction prediction. *Molecules*, **23**, 2208.

Concu,R. and Cordeiro,M. (2019) Alignment-free method to predict enzyme classes and subclasses. *Int. J. Mol. Sci.*, **20**, 5389.

D'Ari,R. and Casadesús,J. (1998) Underground metabolism. *Bioessays*, **20**, 181–186.

Dalkiran,A. *et al.* (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, **19**, 334.

Djoumbou-Feunang,Y. *et al.* (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminf.*, **11**, 1–25.

Feng,S. *et al.* (2018) A hierarchical multi-label classification method based on neural networks for gene function prediction. *Biotechnol. Biotechnol. Equipment*, **32**, 1613–1621.

Hassanpour,N. *et al.* (2020) Biological filtering and substrate promiscuity prediction for annotating untargeted metabolomics. *Metabolites*, **10**, 160.

Jeffryes,J.G. *et al.* (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminf.*, **7**, 44.

Jiang,J. *et al.* (2020) Learning graph representations of biochemical networks and its application to enzymatic link prediction. *Bioinformatics*, **2020**, btaa881.

Jin,W. *et al.* (2018) Junction tree variational autoencoder for molecular graph generation. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:2323–2332.

Khersonsky,O. *et al.*. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology*, **10**, 498–508.

Khersonsky,O. and Tawfik,D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.*, **79**, 471–505.

Kulmanov,M. *et al.* (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.

Kumar,C. and Choudhary,A. (2012) A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP J. Bioinf. Syst. Biol.*, **2012**, 1.

Li,Y. *et al.* (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.

Liu,B. *et al.* (2003) Building text classifiers using positive and unlabeled examples. Proceedings of the Third IEEE International Conference On Data Mining, Melbourne, Florida, 2003, IEEE, pp. 179–188.

Manning,C.D. *et al.* (2009) Evaluation in information retrieval. In: *Introduction to Information Retrieval*. Cambridge University Press, 2008. pp. 151–175.

Martin,E.J. *et al.* (2017) Profile-QSAR 2.0: kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds. *J. Chem. Inf. Model.*, **57**, 2077–2088.

Mellor,J. *et al.* (2016) Semisupervised Gaussian process for automated enzyme search. *ACS Synth. Biol.*, **5**, 518–528.

Moura,M. *et al.* (2013) Computational tools for guided discovery and engineering of metabolic pathways. Systems Metabolic Engineering. Methods in Molecular Biology (Methods and Protocols), vol 985. Humana Press, Totowa, NJ, pp. 123-147.

Nobeli,I. *et al.* (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.*, **27**, 157–167.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pertusi,D.A. *et al.* (2017) Predicting novel substrates for enzymes with minimal experimental effort with active learning. *Metab. Eng.*, **44**, 171–181.

Pertusi,D.A. *et al.* (2015) Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics*, **31**, 1016–1024.

Radenović,F. *et al.* (2016) CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham., pp. 3–20.

Rahman,S.A. *et al.* (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–174.

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.

Roy,A. *et al.* (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–477.

Ryu,J.Y. *et al.* (2019) Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. USA*, **116**, 13996–14001.

Schomburg,I. *et al.* (2017) The BRENDA enzyme information system—from a database to an expert system. *J. Biotechnol.*, **261**, 194–206. BRENDA is available at www.brenda-enzymes.org.

Sokal,R.R. and Michener,C.D. (1958) *A statistical method for evaluating systematic relationships*. University of Kansas Science Bulletin, **38**, 1409–1438.

Tsoumakas,G. and Katakis,I. (2007) Multi-label classification: an overview. *Int. J. Data Warehousing Mining*, **3**, 1–13.

Wehrmann,J. *et al.* (2018) Hierarchical multi-label classification networks. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:5075-5084,pp. 5075–5084.

Yousofshahi,M. *et al.* (2015) PROXIMAL: a method for prediction of xenobiotic metabolism. *BMC Syst. Biol.*, **9**, 94.

Zhang,B. and Zuo,W. (2008) Learning from positive and unlabeled examples: a survey. *2008 International Symposiums on Information Processing*, Moscow, 2008, IEEE, pp. 650–654.

Enzyme Promiscuity Prediction using Hierarchy-Informed Multi-Label Classification

Supplementary File 1

Gian Marco Visani[1], Michael C. Hughes[1] and Soha Hassoun[1,2]

[1]Department of Computer Science, Tufts University, 161 College Ave, Medford, MA, 02155, USA

[2]Department of Chemical and Biological Engineering, Tufts University, 4 Colby St, Medford, MA, 02155, USA

1 Inhibitors are hard negative examples

We investigated the role of inhibitors as hard negative examples during training by comparing structural similarities between positive, inhibitor and unlabeled molecules across enzymes. We considered three pairs of sets: positives-inhibitors, positives-unlabeled, unlabeled-inhibitors. In an effort to obtain results free of statistical bias, the data for each pair was sampled to ensure that the ratio of molecules in the first set to the molecules in the second set is kept constant and equal to the ratio of positives to inhibitors for each individual enzyme. For pairs 2 and 3, we sampled the unlabeled molecules to match the sizes of the inhibitors and of the positives respectively. Furthermore, for this experiment we considered all enzymes for which we had some positive and inhibitor data (2048 of them), not limiting ourselves to considering only enzymes with at least 10 positive examples.

For each enzyme, we computed the average Tanimoto similarity across each possible pair of molecules, where each pair was composed of a molecule from the first set (e.g., positives) and a molecule from the second set (e.g., inhibitors). This was done for each of the three pairs of sets. We show the distribution of the computed per-enzyme average similarity scores on Figure S1.

Comparing Figure S1A with Figure S1C, the mean of the distribution is higher in the former figure (Positive - Inhibitor set) than in the latter (Unlabeled – Inhibitors set). This illustrates that the inhibitors are on average more similar to the positives than they are similar to the unlabeled. The inhibitors therefore can be considered as hard negative examples. Further, the distribution in Figure S1A presents a higher variance (~3x) than the distributions in Figures S1B and S1C. This suggests that the degree of similarity between the positives and the inhibitors varies highly across enzymes.

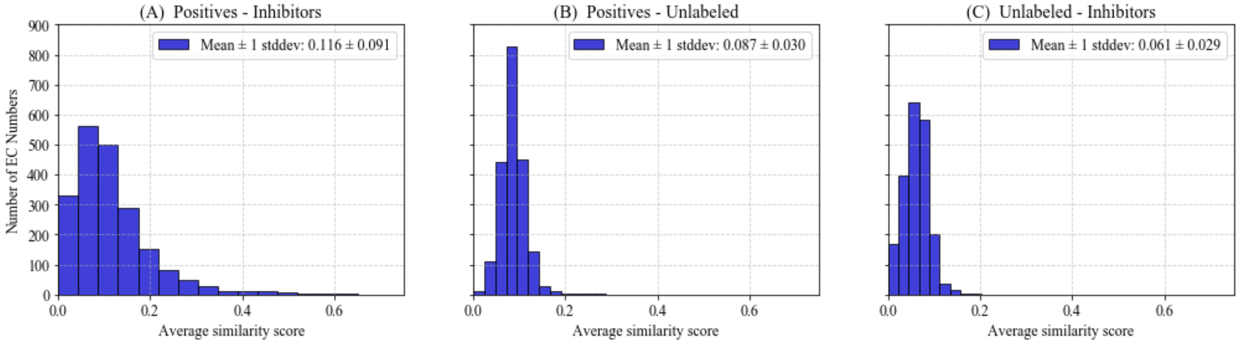

Fig. S1: Distribution of per-enzyme average Tanimoto similarity between molecules from different sets. (A) positive and inhibitor molecules, (B) positive and unlabeled molecules, (C) unlabeled and inhibitor molecules.

2 Datasets statistics

We provide statistics about the composition of our training set and various test sets, for both random and the realistic splits. All splits are based on dividing the molecules, and not the interactions, between training and test sets. Each molecule is assigned to either the training set or the full test set. The Inhibitor Test Set is a subset of the full test set which only includes inhibitors as negative examples (no unlabeled examples). In contrast, the Unlabeled Test Set is a subset of the full test set which only includes unlabeled data as negative examples (excludes all inhibitors). We did not create inhibitor or unlabeled test sets under the realistic split.

	Random split				Realistic split	
	Training Set	Full Test Set	Inhibitor Test Set	Unlabelled Test Set	Training Set	Full Test Set
Count of EC Numbers	983	983	671	671	680	680
Count of EC Number – molecule pairs	6,523,188	1,630,797	8,378	8,378	4,512,480	1,128,120
Positive molecules per EC Number	**min:** 1 **median:** 14 **max:** 681	**min:** 1 **median:** 4 **max:** 192	**min:** 1 **median:** 4 **max:** 192	**min:** 1 **median:** 4 **max:** 192	**min:** 5 **median:** 16 **max:** 739	**min:** 1 **median:** 3 **max:** 134
Negative molecules per EC Number	**min:** 5,955 **median:** 6,622 **max:** 6,631	**min:** 1,467 **median:** 1,655 **max:** 1,658	**min:** 1 **median:** 2 **max:** 24	**min:** 1 **median:** 2 **max:** 24	**min:** 5,897 **median:** 6,620 **max:** 6,631	**min:** 1,525 **median:** 1,656 **max:** 1,658
Inhibitor molecules per EC Number	**min:** 0 **median:** 6 **max:** 88	**min:** 0 **median:** 2 **max:** 24	**min:** 1 **median:** 2 **max:** 24	**min:** 0 **median:** 0 **max: 0**	**min:** 0 **median:** 4 **max:** 85	**min:** 0 **median:** 3 **max:** 37

Table S1. Summary statistics of the training and test sets used in our experiments under random and realistic splits. The Inhibitor Test Set and the Unlabeled Test Set have fewer enzymes.

3 Data sharing strategies for our machine-learning models

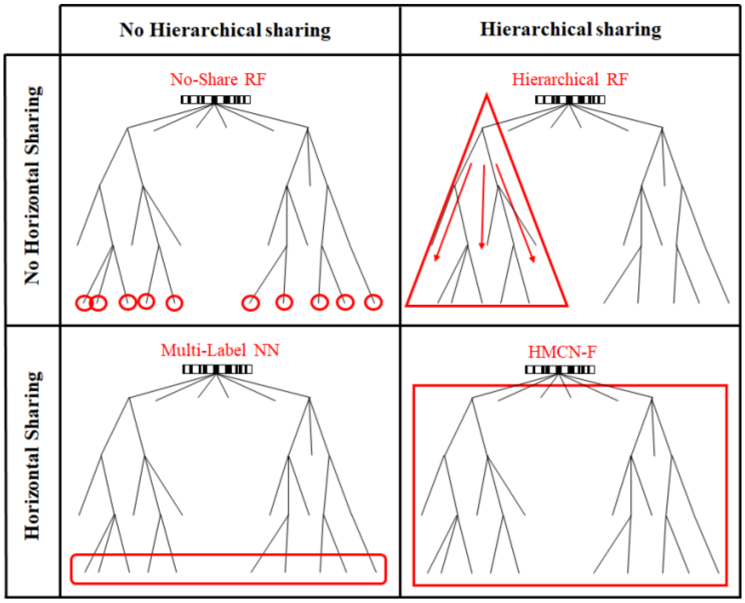

Fig. S2. Illustration of the sharing strategies used by the four machine-learning models. The hierarchical structure shown in each box indicates the EC hierarchy. Predictions for nodes inside a red box are made by using information from all other nodes in the box.

4 Results by EC Class

We investigate the performance of our best performing model, EPP-HMCNF, against similarity models per enzyme Class. We divided the 983 EC Numbers by their respective Class and analyzed the average performance of the models on each Class. The EC Numbers are partitioned in classes as follows: Class 1 (Oxidoreductases) – 385 EC Numbers; Class 2 (Transferases) – 285; Class 3 (Hydrolases) – 165; Class 4 (Lyases) – 91; Class 5 (Isomerases) – 20; Class 6 (Ligases) – 37.

EPP-HMCNF is consistently better than *k*-NN Similarity across metrics and classes (Figure S3, A-F), making EPP-HMCNF preferable to similarity for every EC Class. Furthermore, despite the differences in number of enzymes per EC Class, variances are similar across EC Classes, indicating that there is high variability within each EC Class.

To better analyze the relative performance of the models across different EC Classes, we computed the percent deviation of the mean score per Class (Figure S3, A-F) from the mean score across all EC Numbers (Figure 4). The deviations are shown in Figure S3, G-L. Some classes are easier to characterize than others. In particular, Class 3 EC Numbers, which catalyze the formation of two products from one substrate by hydrolysis, and Class 5 EC Numbers, which catalyze intramolecular rearrangements, are easiest to classify since Class 3 and Class 5 have the highest scores across all metrics and models. Class 1 and 4 EC Numbers are harder to classify than average for all models, as indicated by their relative negative scores. Class 2 EC Numbers are easier to classify (higher scores) than average, but less so than classes 3 and 5. Class 6 EC Numbers are easier to classify for similarity but have an average (scores around zero) or slightly larger than average score for EPP-HMCNF. Figure S3, G-L also shows that Mean AUROC varies less from the mean than MAP and Mean R-PREC, indicated by scores closer to zero than the other two metrics. All three metrics however generally follow the same trend: they either all have negative or positive scores, with few exceptions having scores concentrated around zero. Last, we see that scores deviate from their relative means more for the similarity models than for EPP-HMCNF, with the exception, though small, of class 2. This suggests that EPP-HMCNF has more consistent performance than *k*-NN similarity across the enzyme Classes.

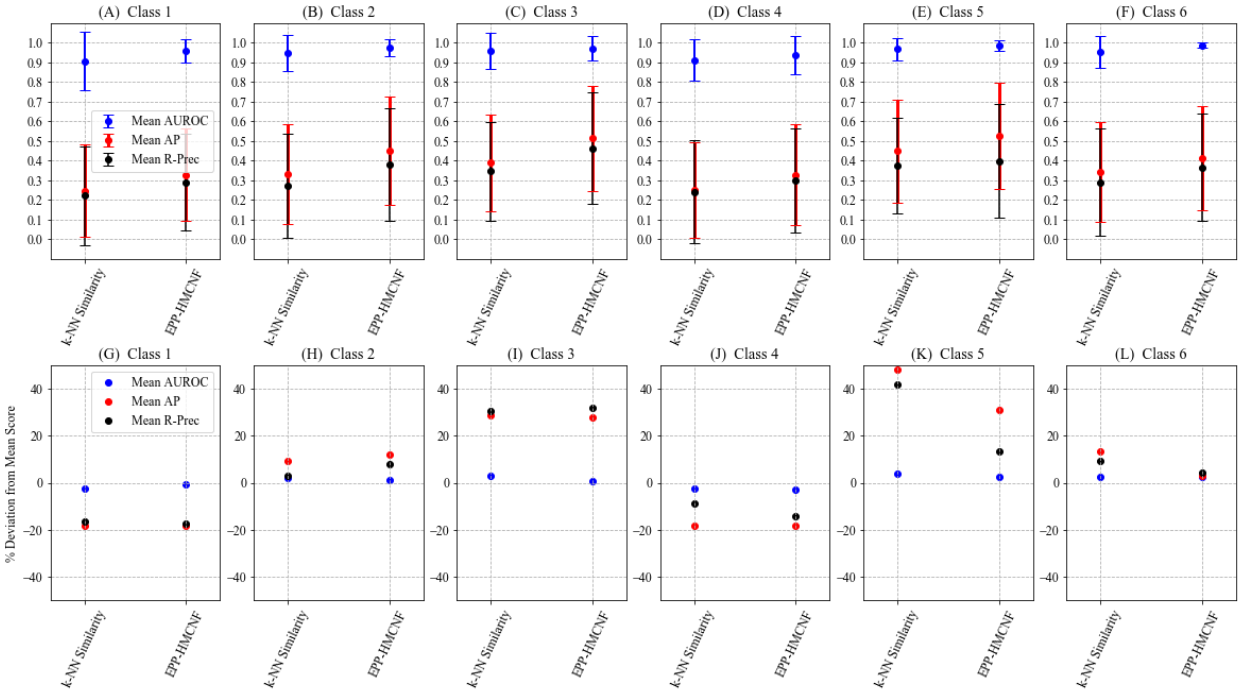

Fig. S3. Mean scores per Class for *k*-NN similarity and EPP-HMCNF models for all EC Classes (A-F), and Percent deviation per Class from the mean score across all EC Numbers (G-L). The vertical bars in A-F indicate +/- 1 standard deviation.

5 Comparing performance on enzymes with most vs least positive data

We further analyze the performance of our models as a function of the number of positive examples available per enzyme. To this end, we perform two sets of experiments. First, we extract the results for two subsets of enzymes: those with most positive examples, and those with the least amount. We considered the top and bottom 15% of enzymes by number of positive examples, where each group consisted of 147 enzymes. The median number of positives examples are 55 and 10 for the top-15% and bottom-15% groups, respectively, while the numbers in the 5th – 95th percentile for these two groups were 39-164, and 10-11, respectively. Figure S4, A-B displays the mean scores for the top 15% and bottom 15% groups, together with +/- 1 standard deviation, for all models, while Figure S4, C-D shows the percent deviation of the mean scores for each group (Figure S4, A-B) from the mean scores across all EC Numbers (Figure 4).

Mean AUROC scores do not deviate much from their means for either group (Figure S4, C-D). Thus, the analysis from this point forward will focus on AP and R-PREC. The top 15% enzymes present Mean AP and Mean R-PREC scores that are higher than average for all models, whereas the trend is reversed for the bottom 15% enzymes (Fig. S4, C-D). Furthermore, Hierarchical RF presents the highest deviation from average in both the top 15% and bottom 15% scores across models (excluding Random), even slightly outperforming *k*-NN Similarity. Figure S4 shows that the top 15% enzymes have less variability in scores, and thus they produce more consistent results than the bottom 15% enzymes. However, the top 15% enzymes still present a high enough variability in scores (0.236 – 0.588 R-PREC range between +/- 1 standard deviation for EPP-HMCNF, Fig. S4A), leading to the conclusion that some enzymes are significantly easier to characterize than others despite high data availability. Overall, these results suggest that training with more known positive data improves performance, and thus that the machine learning models are likely to perform better as more data becomes available.

The performance of the random-guessing model (Random) appears better on the top 15% EC Numbers than on the bottom 15% EC Numbers for AP and R-PREC (Figure S4, C-D). Clearly, this cannot be due to the EC Numbers having more positive training data. Instead, this difference in performance is due to a bias in the evaluation of AP and R-PREC on different EC Numbers. The EC Numbers with more positive data in training are also the EC Numbers with more positive data in testing, since training and testing are created via a random split. The EC Numbers with more positive data also have a higher positive-to-total (P/T) ratio in testing. Specifically, the mean P/T test ratio is 0.0091 for the top 15% EC Numbers, and 0.0014 for the bottom 15% EC Numbers. AP and R-PREC are biased positively by the P/T test ratio. This biasing can be very intuitively explained for R-PREC: the performance of a random-guesser (i.e. with no predictive skill) follows exactly the P/T test ratio. Indeed, for the typical EC Number in the top 15% group, where the mean P/T test ratio is 0.0091, the random-guesser will each time pick a true positive with 0.0091 probability, resulting in a R-PREC of 0.0091. By the same logic, for the typical EC Number in the bottom 15% group, where the mean P/T test ratio is 0.0014, the random-guesser will each time pick a true positive with 0.0014 probability, resulting in a R-PREC of 0.0014. A similar argument can be made for AP. This analysis suggests that part of the increase in performance by the machine learning models on the top 15% EC Numbers vs. the bottom 15% EC Numbers may be accounted for by the difference in P/T test ratios between the two groups. However, we believe that the difference in P/T test ratio between the two groups (0.0077 difference of the two mean ratios) is too small to account for differences in score as big

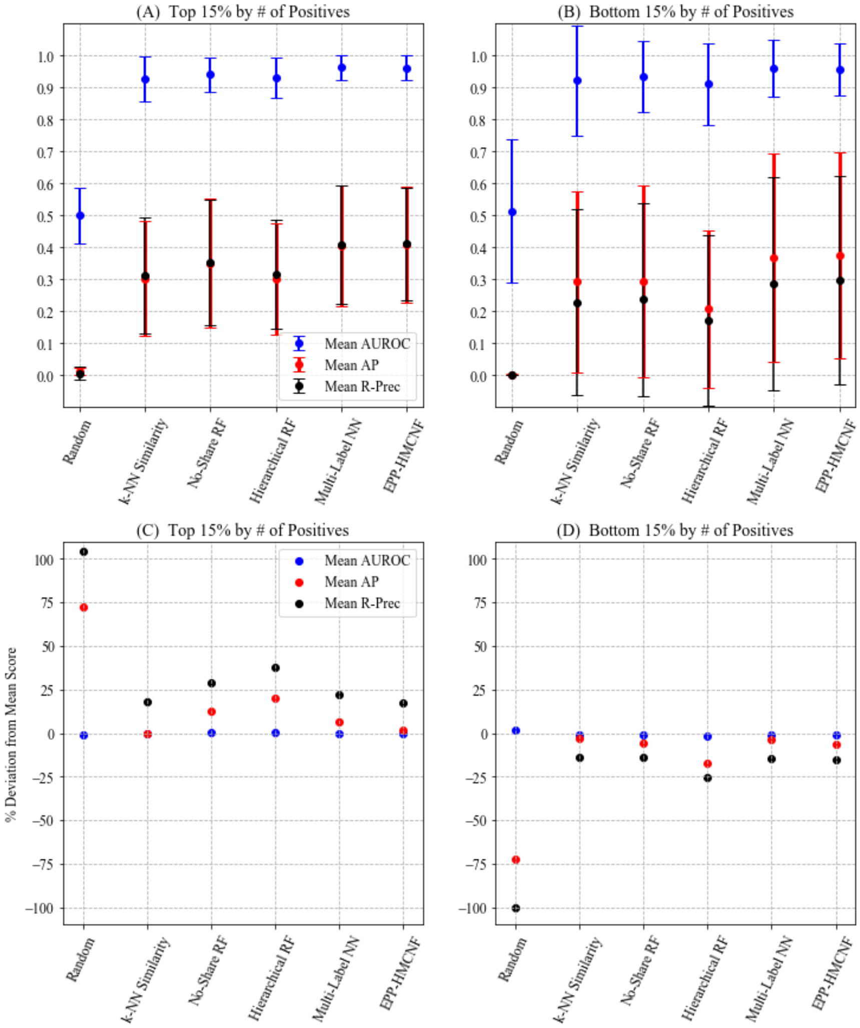

Fig. S4. Mean score across all EC Numbers with (A) most (top 15%), and (B) least (bottom 15) number of positive examples, and percent deviation from the mean across all EC Numbers with (C) most (top 15%), and (D) least (bottom 15) number of positive examples. Results are shown for all models, with +/- 1 standard deviation.

as 0.1 in both AP and R-PREC (Figure S4, A-B). Thus, the explanations for these large differences must be due to differences in the number of positive examples in training.

6 Discussion of time complexity of the various models

k-NN Similarity: As this model has no “training” step, training data is stored and used at inference time. At inference, the model computes, for each EC Number, the similarity between each inference sample and each positive example in training. The runtime is thus:

$$O\left(m \sum_{j=1}^{ec_i} p_j\right) = O(m\; ec_i\; \overline{p_{ec}})$$

No-Share RF: There is one RF for each EC number. The theoretical runtime for training each RF is $O(n^2 \log(n))$ assuming continuous variable values. The scikit-learn implementation avoids sorting the feature values at each splitting node and has a time complexity of $O(n \log(n))$ (Raschka, 2018). The runtime for inference for each RF is $O(mD)$, where the maximum tree depth D is the minimum of the number of features and the number of training examples.

Hierarchical RF: There is one RF for each node in the EC hierarchy. The runtimes for training and inference for each RF is $O(n \log(n))$ and $O(mD)$, respectively.

Multi-Label NN: The runtime is dominated by the size of the training and test datasets. The number of EC Numbers, ec_t, has little impact on the overall training time because ec_t only modifies the size of the output layer. During inference, this model makes predictions for ec_t EC Numbers at once; ec_i has no effect on the runtime.

EPP-HMCNF: The runtimes for training and inference for this model is similar to those of the multi-label NN. However, the size of the output layers is proportional to the number of EC hierarchy nodes, ec_n.

References

Raschka, S. (2018). STAT 479: Machine Learning Lecture Notes, Department of Statistics, University of Wisconsin-Madison.