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.

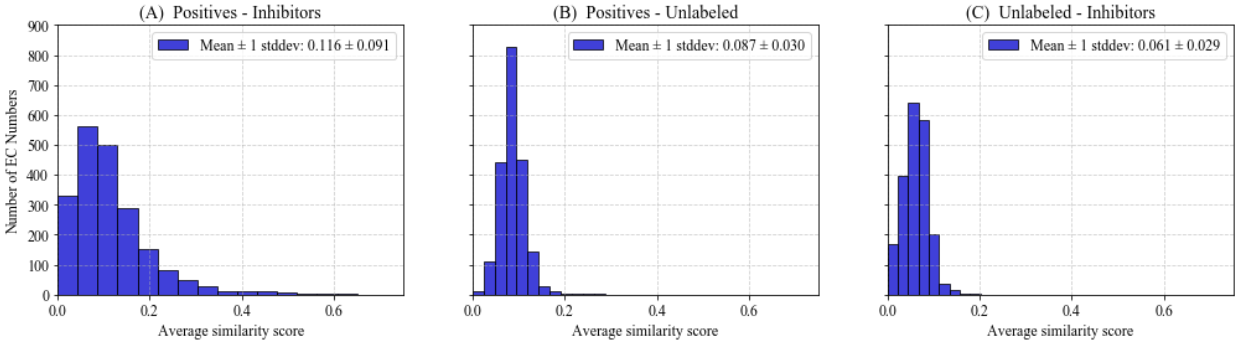

Fig. S1: Distribution of per-enzyme average Tanimoto similarity between molecules from different sets. (A) positive and inhibitor molecules, (B) positive and unlabeled molecules, (C) unlabeled and inhibitor molecules.

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

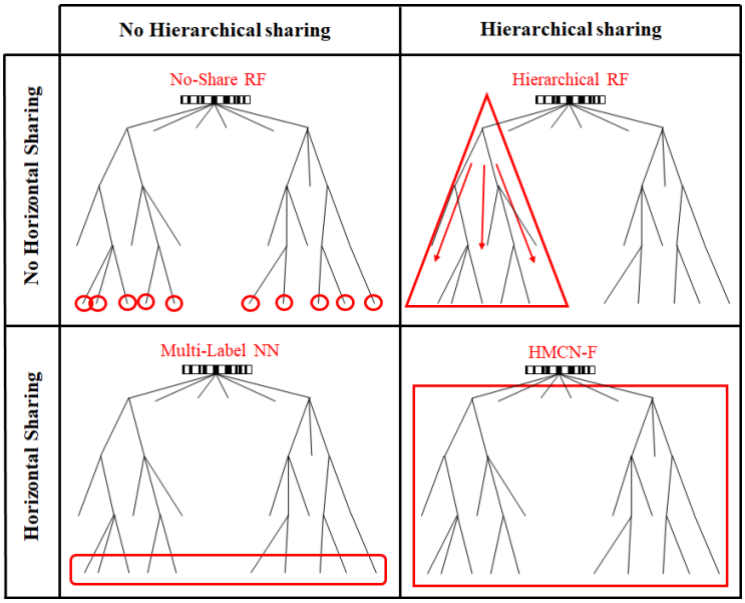

Fig. S2. Illustration of the sharing strategies used by the four machine-learning models. The hierarchical structure shown in each box indicates the EC hierarchy. Predictions for nodes inside a red box are made by using information from all other nodes in the box.

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.

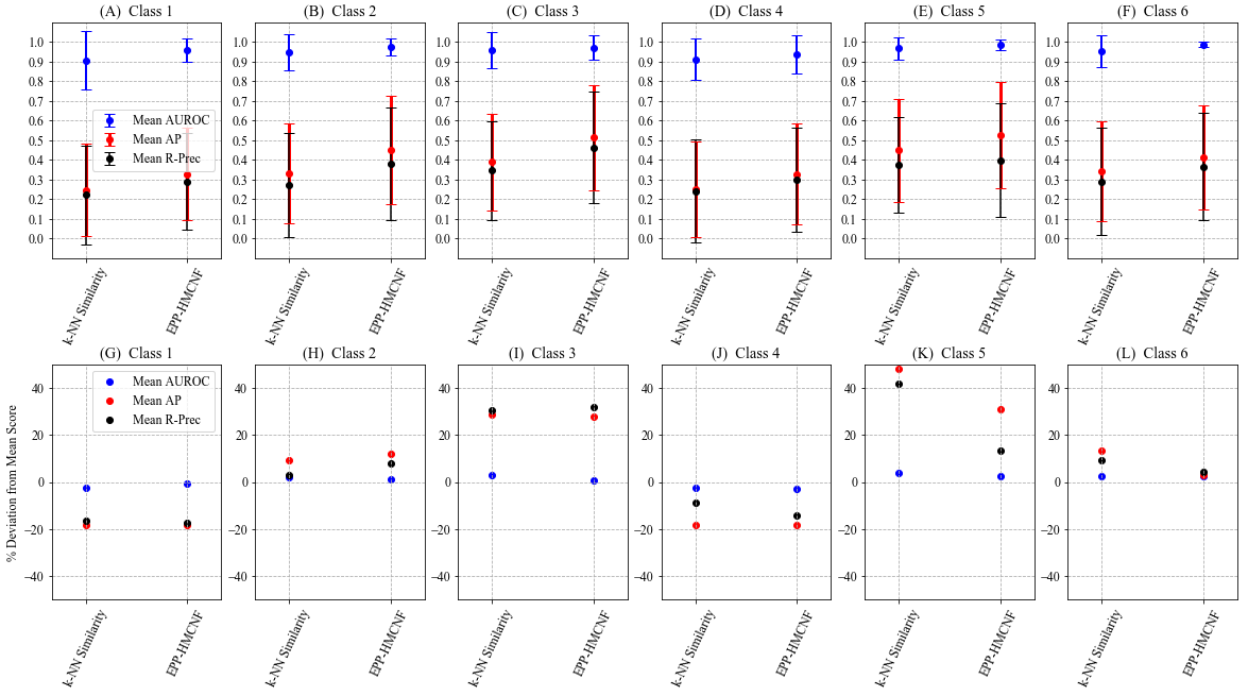

Fig. S3. Mean scores per Class for *k*-NN similarity and EPP-HMCNF models for all EC Classes (A-F), and Percent deviation per Class from the mean score across all EC Numbers (G-L). The vertical bars in A-F indicate +/- 1 standard deviation.

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 5^{th} – 95^{th} 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

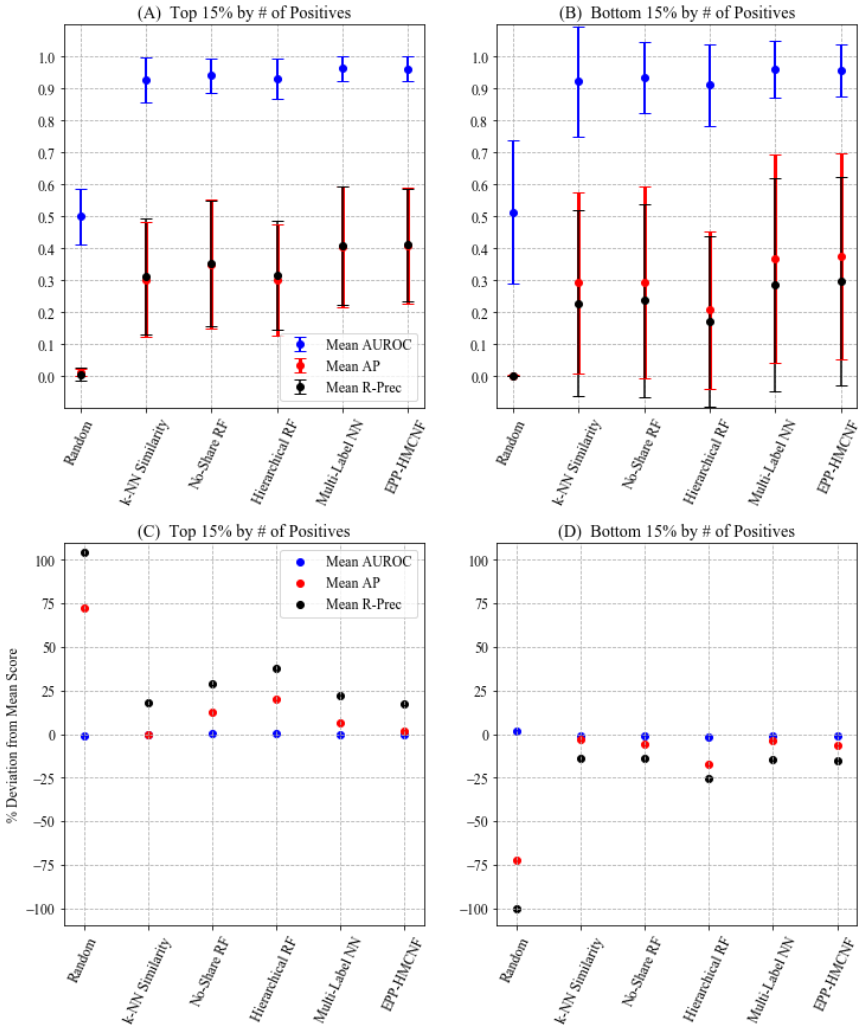

Fig. S4. Mean score across all EC Numbers with (A) most (top 15%), and (B) least (bottom 15) number of positive examples, and percent deviation from the mean across all EC Numbers with (C) most (top 15%), and (D) least (bottom 15) number of positive examples.
Results are shown for all models, with +/-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