Automatic single- and multi-label enzymatic function prediction by machine learning

Shervine Amidi[1], **Afshine Amidi**[1], **Dimitrios Vlachakis**[2], **Nikos Paragios**[1,3], **and Evangelia I. Zacharaki**[1,3]

[1]**Center for Visual Computing, Department of Applied Mathematics, Ecole Centrale de Paris (CentraleSupélec), 92295 Châtenay-Malabry, France**
[2]**MDAKM group, Department of Computer Engineering and Informatics, University of Patras, 26500, Greece**
[3]**Equipe GALEN, INRIA Saclay, Orsay, Île-de-France, France**

ABSTRACT

The number of protein structures in the PDB database have been increasing more than fifteenfold since 1999. The creation of computational models predicting enzymatic function is of major importance since such models provide the means to better understand the behavior of newly-discovered enzymes when catalyzing chemical reactions. Until now, single-label classification has been widely performed for predicting enzymatic function limiting the application to enzymes performing unique reactions and introducing errors when multi-functional enzymes were examined. Indeed, some enzymes may be performing different reactions and can hence be directly associated with multiple enzymatic functions. In the present work, we propose a multi-label enzymatic function classification scheme that combines structural and amino acid sequence information. We investigate two fusion approaches (in the feature level and decision level) and assess the methodology for general enzymatic function prediction indicated by the first digit of the Enzyme Commission (EC) code (6 main classes) on 40,034 enzymes from the PDB database. The proposed single-label and multi-label models predict correctly any of the actual functional activities in 97.8% and 95.5% respectively, and also predict all possible enzymatic reactions in 85.4% of the multi-labeled enzymes when the number of reactions is unknown. Code and datasets are available at https://figshare.com/s/a63e0bafa9b71fc7cbd7

Keywords:

INTRODUCTION

The ever-growing PDB database contains more than 110,000 proteins that are characterized by different properties including their structure, biological function, chemical composition, or solubility in solvents. Protein classification is important since it allows estimating the properties of novel proteins according to the group to which they are predicted to belong. Enzymes are a type of proteins that are classified according to the chemical reactions they catalyze into 6 primary classes, oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The classes are denoted by the enzyme commission (EC) number (NC-IUBMB (1992)) and have been determined based on experimental evidence. Systematic annotation, reliability and reproducibility of protein functions is discussed in Valencia (2005). Classification of enzymes is a central issue because it helps understanding enzymatic behavior during chemical reactions. While the vast majority of enzymes has been found to perform particular reactions, a non-negligible number of enzymes can perform different reactions and can hence be directly associated with multiple enzymatic functions (Guyon et al. (2006)).

During the last decade, various machine learning techniques have been proposed for both single-label and multi-label enzyme classification on different datasets. Among single-label classification studies, some (Dobson and Doig (2005)) used only structural information and achieved an accuracy of 35% for top-ranked prediction using support vector machine (SVM) with a one-against-one voting scheme on 498 enzymes from the PDB database. Applying SVM on sequence features has also been done by Mohammed and Guda (2015) and achieved an accuracy of 98.39% after training on 150,000+ enzymes with 10-fold cross-validation. Osman and Choong-Yeun Liong (2010) extracted only gene or amino acid sequence

31 information and applied neural networks obtaining an accuracy of 72.94% after training the networks on
32 1,200 enzymes from the PDB database and testing on 2,000 others. Volpato et al. (2013) achieved 96%
33 accuracy with a 10-fold cross-validation scheme on 6,081 entries of the ENZYME database. Sequence
34 structure and amino acid information were also used by desJardins et al. (1997), Kumar and Choudhary
35 (2012) and Lee et al. (2007) who obtained testing accuracies ranging from 74% to 88.2% using the
36 Swiss-Prot database. Combination of sequence, structure and chemical properties of enzymes was also
37 explored by Borgwardt et al. (2005) using kernel methods and SVM on the BRENDA database and
38 achieved an accuracy of 93% with 6-fold cross-validation on information extracted through protein graph
39 models. Multi-label classification using different methods such as RAkEL-RF and MLKNN (Wang et al.
40 (2014)) or MULAN (Zou et al. (2013)) was performed on single- and multi-labeled enzymes. In particular,
41 the latter was assessed on enzymes from the Swiss-Prot database based on their amino acid composition
42 and their physical-chemical property and involved the use of Position-Specific Scoring Matrices. In the
43 best scenario, a macro-averaged precision of 99.31% was obtained on a set of 2,840 multi-functional
44 enzymes after 10-fold cross-validation. Also, a summary of other alignment-free methods used to predict
45 enzyme classes is presented in table 1.

Table 1. Comparative table of several alignment-free approaches

# proteins	Information	Parameters	Classification method			Level	Work
1,371	3D structure	3D-HINT potential	LDA	QSAR	ANN	0-1	Concu et al. (2009c)
4,755		Moments, entropy, electrostatic, HINT potential		MLP			Concu et al. (2009b)
2,276				3D-QSAR			Concu et al. (2009a)
26,632		Global binding descriptors	SVM			1-3	Volkamer et al. (2013)
211,658	Structural	GRAVY				1	Dave and Panchal (2013)
3,095	Sequence	PseAAC, SAAC, GM	ML-kNN				Zou and Xiao (2016)
9,832		FunD, PSSM	OET-kNN			1-2	Shen and Chou (2007)
300,747	Interpro signatures		BR-kNN			1-4	Ferrari et al. (2012)

46 Other work on enzyme classification includes the use of information stemming from topological
47 indices (Munteanu et al. (2008)), peptide graphs (Concu et al. (2009b)), and also includes the machine-
48 learning based ECemble method (Mohammed and Guda (2015)).
49 In this work, a new feature extraction and classification scheme is presented that combines both
50 structural and amino acid sequence information aiming to improve standard classifiers that use only
51 one type of information. Building upon previous work (Amidi et al. (2016)), we investigate a more
52 sophisticated combination approach and assess the performance of the scheme in single-label and multi-
53 label classification tasks. State-of-the-art accuracy is observed as compared to the methods reviewed in
54 the survey by Yadav and Tiwari (2015).

METHODS

Feature extraction

57 Proteins are chains of amino acids joined together by peptide bonds. The three-dimensional (3D)
58 configuration of the amino acids chain is a very good predictor of protein function, thus there has been
59 many efforts in extracting an appropriate representation of the 3D structure (Lie and Koehl (2014)). Since
60 many conformations of this chain are possible due to the possible rotation of the peptide bond planes

relative to each other, the use of rotation invariant features is preferred over features based on cartesian coordinates of the atoms. In this study the two torsion angles of the polypeptide chain were used as structural features. The two torsion angles describe the rotation of the polypeptide backbone around the bonds between N-C$_\alpha$ (angle ϕ) and C$_\alpha$-C (angle ψ). The probability density of the torsion angles ϕ and $\psi \in [-180°, 180°]$ was estimated by calculating the 2D sample histogram of the angles of all residues in the protein. When the protein consisted of more than one chain, the torsion angles of all chains were included together into the feature vector. Smoothness in the density function was achieved by moving average filtering, i.e. by convoluting the 2D histogram with a uniform kernel. The range of angles was discretized using 19×19 bins centered at $0°$ and the obtained matrix of structural features was linearized to a 361-dimensional feature vector for each enzyme representing structural information (X_{SI}).

Although structure relates to amino acid sequence, additional information can be extracted directly from the protein sequences. Assessment of similarities between amino acid sequences of enzymes is usually performed by sequence alignment. The Smith-Waterman sequence alignment algorithm (Smith and Waterman (1981)) has been preferred over the Needleman-Wunsch algorithm (Needleman and Wunsch (1970)) due to the assessment of sequence similarity based on local alignment (in contrast to the global alignment previously performed), which enables possible deletions, insertions, substitutions, matches and mismatches of arbitrary lengths. Optimizing local alignment allows to take into consideration mutations that might have happened in amino acid sequences. The similarity of each pair of sequences i and j can be quantified using the scoring matrix $M_{i,j}$ that is produced by the sequence alignment algorithm. For two sequences i and j, the highest score in the previous matrix, which reflects the success of alignment, is used as similarity criterion $S(i, j)$. Amino acid sequence information is represented in two distinct ways. First, for each one of the six classes, the similarity matrix S of a sequence to all training samples of that class is calculated and summarized as a histogram vector with 10 bins. The six histogram vectors are then concatenated into a 60-dimensional feature vector which is denoted as X_{AA}. Second, the class probabilities $(f_{AA})_x^j$ of a given enzyme j are expressed as the maximum similarities S within each class normalized over all classes:

$$\text{For each class } x, \quad (f_{AA})_x^j = \frac{\displaystyle\max_{\substack{k\in\text{training}\,\cap\,\text{EC}\,x \\ k\neq j}} S(k,j)}{\displaystyle\sum_{l=1}^{6} \max_{\substack{k\in\text{training}\,\cap\,\text{EC}\,l \\ k\neq j}} S(k,j)}$$

Classification and fusion

Two classification techniques have been investigated, the nearest neighbor (NN) and SVM. NN is preferred for its simplicity and its small computation time whereas SVM is useful to find non-linear separation boundaries. The classifiers are trained using a number of annotated examples and then test on novel enzymes. Two types of classification models have been produced: single-label models for the enzymes performing unique reactions and multi-label models for the multi-functional enzymes. Both structural (SI) and amino acid sequence (AA) information are related to the enzymatic activity. In order to take into consideration these two properties, fusion of information is performed in two different ways, in the feature level and in the decision level.

The concept of the feature-level fusion is to concatenate the two sets of 361 structural and 60 amino acid sequence features before performing classification. The feature-level fusion approach is illustrated in figure 1.

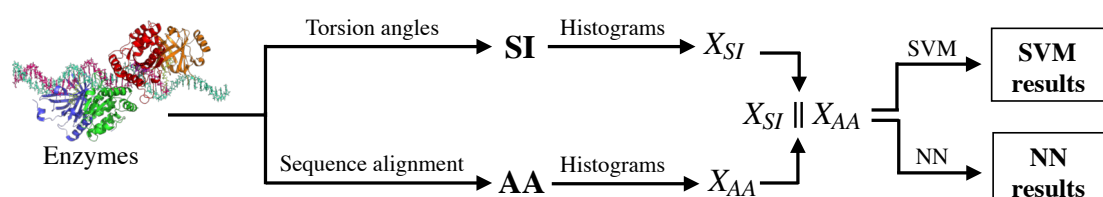
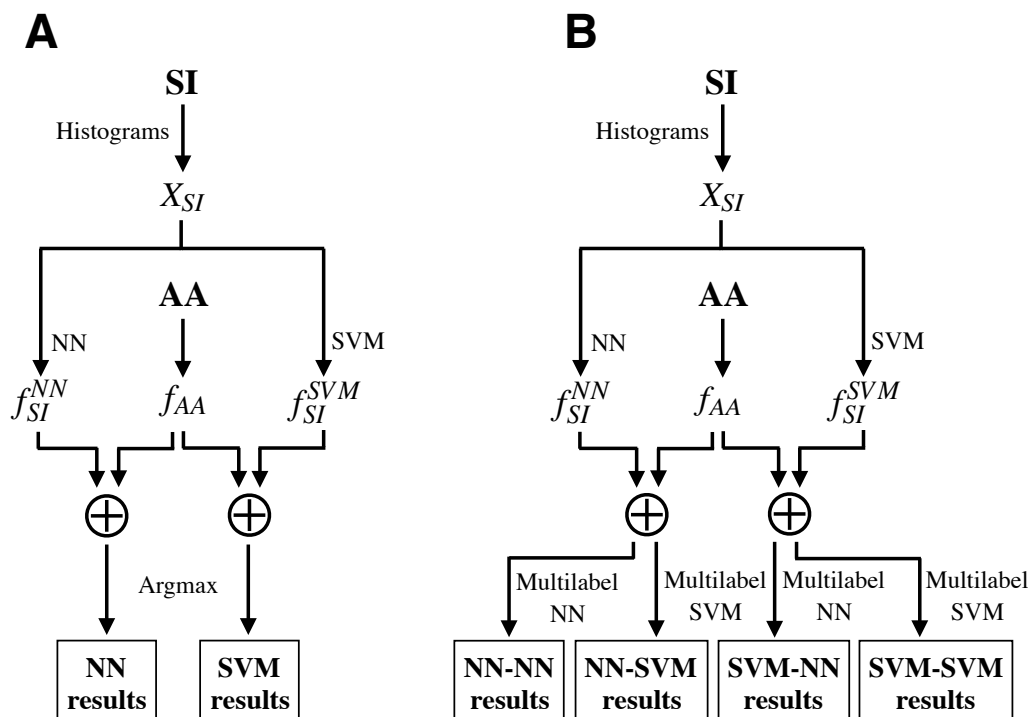
Figure 1. Overview of feature-level fusion

⁹⁹ The decision-level fusion approach associates class probabilities for SI obtained by SVM (Platt (1999))
¹⁰⁰ (f_{SI}^{SVM}) or NN (Atiya (2005)) (f_{SI}^{NN}) with class probabilities for AA (f_{AA}) through a heuristic fusion
¹⁰¹ rule. The applied fusion rule performs weighted averaging of class probabilities using unequal weights.
¹⁰² Thus, the corresponding fused class probability is given by $(1-\alpha)(f_{SI})+\alpha(f_{AA})$. An optimized α is
¹⁰³ empirically obtained for each classification method by maximizing the accuracy over the training data
¹⁰⁴ (Amidi et al. (2016)). A single class is assigned in the single-label classification (figure 2) based on the
¹⁰⁵ maximum probability.

Figure 2. Decision-level fusion for single- and multi-label classification

¹⁰⁶ This hard decision rule cannot be applied to the multi-label scenario. In order to obtain a soft decision
¹⁰⁷ a multi-label classifier is applied on the fused class probabilities to produce the final decision outputs. In
¹⁰⁸ particular, the 6-dimensional class probabilities (fused from AA and SI) are introduced into a multi-label
¹⁰⁹ SVM or multi-label NN which computes a 6-dimensional binary vector where the cth feature is equal to 1
¹¹⁰ if the predicted enzyme belongs to class c and 0 otherwise.

Performance assessment

¹¹² The data have been randomly split into 80% for training and validation and 20% for independent testing.
¹¹³ The training/validation set has been divided into 5 random folds that have been used to determine the
¹¹⁴ optimal parameters. More particularly, the parameters were optimized by a standard 5-fold cross-validation
¹¹⁵ based on classification accuracy in the single-label classification problem and on the subset accuracy in
¹¹⁶ the multi-label classification problem. Upon optimization, the parameters were fixed and remained the
¹¹⁷ same throughout all experiments. Then for both classification problems, performance has been assessed
¹¹⁸ by applying the methods on the independent testing set using the fixed parameters.

Single-label classification

The performance of single-label classification has been assessed based on the confusion matrix whose
elements $C(x,y)$ with $x,y \in [\![1,6]\!]$, indicate the number of enzymes that belong to class x and are predicted
as belonging to class y. Two metrics are based on this definition: the overall accuracy that evaluates the
proportion of correctly classified enzymes among the total number of enzymes and the balanced accuracy

that avoids inflated performance estimates on imbalanced datasets. They are defined by:

$$\text{Overall Accuracy} = \frac{\sum_{x=1}^{6} C(x,x)}{\sum_{x,y=1}^{6} C(x,y)} \quad \text{and} \quad \text{Balanced Accuracy} = \frac{1}{6} \cdot \sum_{x=1}^{6} \frac{C(x,x)}{\sum_{y=1}^{6} C(x,y)}$$

Multi-label classification

In the case of multi-label classification, the labels of an enzyme i are represented by a 6-dimensional binary vector L_i where the value 1 at a position $j \in [\![1,6]\!]$ indicates the positivity of class j and 0 otherwise. Also, we denote N the total number of enzymes, as well as L_i^{true} and L_i^{pred} the sets of true and predicted labels of enzyme i respectively. The performance of the multi-label classifiers cannot be assessed using the exact same definitions as for the single-label classifiers. Various multi-label metrics defined in previous works (Zhang and Zhou (2006), Tsoumakas and Katakis (2007) and Madjarov et al. (2012)) have been considered in our study. Here, we introduce the Kronecker delta δ, the symmetric difference Δ, the binary union $\cup$ and intersection $\cap$ operations, as well as the l_1-norm $|\ |$. The following metrics have been chosen to assess the performance of our new method:

- *Hamming-Loss* assesses the frequency of misclassification of a classifier on a given set of enzymes. This index is averaged over all classes and all enzymes. Also, we will note 1-Hamming-Loss the complementary of this indicator so that the worst-case value is 0 and the best one 1. Conversely to the Hamming-Loss index, the latter assesses the average over all enzymes of the proportion of binary class memberships that are correctly predicted.

$$\text{Hamming-Loss} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{6} |L_i^{pred} \Delta L_i^{true}|$$

- *Accuracy* averages over all enzymes the Jaccard similarity coefficient of the predicted and true sets of labels. This index reflects the averaged proportion of similar class membership between those two sets.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{|L_i^{pred} \cap L_i^{true}|}{|L_i^{pred} \cup L_i^{true}|}$$

- *Precision, recall and F1 score*, which have been adapted for multi-label classification. The two first metrics respectively reflect the proportion of detected positives that are effectively positive, and the proportion of positives samples that are correctly detected. Finally, the F1 score balances the information provided by these two indexes through the computation of an harmonic mean.

$$\text{Precision} = \frac{1}{N} \sum_{i=1}^{N} \frac{|L_i^{pred} \cap L_i^{true}|}{|L_i^{pred}|} \quad \text{Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{|L_i^{pred} \cap L_i^{true}|}{|L_i^{true}|} \quad \text{F1} = \frac{2}{N} \sum_{i=1}^{N} \frac{|L_i^{pred} \cap L_i^{true}|}{|L_i^{pred}| + |L_i^{true}|}$$

- *Subset accuracy* considers that a given enzyme is correctly classified if and only if all class memberships are correctly predicted. This metric is the strictest of this study, since it requires the sets of true and predicted labels to be identical in order for an enzyme to be considered as correctly classified.

$$\text{Subset accuracy} = \frac{1}{N} \sum_{i=1}^{N} \delta(L_i^{pred}, L_i^{true})$$

- *Macro-precision, recall and F1* compute respectively precision, recall and F1-score separately for each class, and then average the values over the 6 classes. These indexes are crucial for us, as they will highlight the incidence of our method on small-populated labels. In the following definitions, TP_j and FP_j respectively represent the number of true positives and false positives, and Precision_j and Recall_j are those associated to class $j \in [\![1,6]\!]$.

$$\text{M-precision} = \frac{1}{6} \sum_{j=1}^{6} \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j} \quad \text{M-recall} = \frac{1}{6} \sum_{j=1}^{6} \frac{\text{TP}_j}{\text{TP}_j + \text{FN}_j} \quad \text{M-F1} = \frac{2}{6} \sum_{j=1}^{6} \frac{\text{Precision}_j \times \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j}$$

5/12

- *micro-precision, recall and F1* are similar to the single-label definition of those three quantities, whereas here they rely on the values of the sum over all classes of true positives, false positives and false negatives. The *micro* indexes will indicate whether the majority of the enzymes are correctly classified, regardless if they belong to low- or high-populated classes.

$$\text{m-precision} = \frac{\sum_{j=1}^{6} \text{TP}_j}{\sum_{j=1}^{6} \text{TP}_j + \sum_{j=1}^{6} \text{FP}_j} \qquad \text{m-recall} = \frac{\sum_{j=1}^{6} \text{TP}_j}{\sum_{j=1}^{6} \text{TP}_j + \sum_{j=1}^{6} \text{FN}_j} \qquad \text{m-F1} = 2 \cdot \frac{\text{m-precision} \times \text{m-recall}}{\text{m-precision} + \text{m-recall}}$$

130 **Data**

131 The method has been applied on data from the PDB database that include one set of single-labeled
132 enzymes (table 2) and one set of multi-labeled enzymes (table 3).

Table 2. Dataset I: 39,251 single-labeled enzymes

Class	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
Name	Oxidoreductase	Transferase	Hydrolase	Lyase	Isomerase	Ligase
Number	7,256	10,665	15,451	2,694	1,642	1,543

133

Table 3. Dataset II: 783 multi-labeled enzymes

Number of classes	2												3			4
EC numbers	1	1	1	1	2	2	2	2	3	3	3	4	1	1	1	1
	2	3	4	5	3	4	5	6	4	5	6	5	2	2	4	2
													3	4	5	4
																5
Number of enzymes	62	44	14	2	217	160	45	15	82	23	73	28	1	7	6	4

The total number of enzymes with 2, 3 and 4 labels each, are 765, 14, 4 respectively.

134 **RESULTS**

135 **Single-label classification**

136 Classification via decision-level fusion has been performed using $\alpha = 0.95$ for the SVM method and
137 $\alpha = 0.99$ for the NN method. Overall and balanced accuracies obtained with each method on the testing
138 set are detailed in table 4.

Table 4. Testing performance of dataset I

Type	SI		AA		Decision fusion	Feature fusion	
Classifier	SVM	NN	NN	SVM	NN	SVM	NN
Overall accuracy	0.830	0.828	0.976	0.977	0.978	0.942	0.878
Balanced accuracy	0.755	0.788	0.968	0.966	0.968	0.910	0.856

139 The decision-level fusion classification increased the overall accuracy by 0.2% compared to the best
140 results obtained by either AA only or SI only. The balanced accuracy achieved is 96.8%, which is the

same as the one achieved by NN classification using AA only. Also, SVM classification via feature-level fusion achieves 11.2% better overall accuracy than classification via SI only but 3.4% less overall accuracy than NN classification on AA only. In general, classification using SVM tends to achieve better overall accuracy than with NN (0.2% and 6.4% respectively for SI only and feature-level fusion), whereas excepted for the feature-level fusion, NN tends to achieve better balanced accuracy than SVM (0.2% and 3.3% respectively for decision-level fusion and SI only).

Multi-label classification

As described in the methods' section, the optimal fusion parameter α was empirically determined for each dataset (single- or multi-functional) and fusion scheme. The optimal values are shown in figure 3 from which it can be seen that the values of α for the decision-level fusion in multi-label classification ($\alpha = 0.69, 0.73, 0.76, 0.80$ respectively for the SVM-NN, SVM-SVM, NN-NN and NN-SVM methods) are approximately 20% smaller compared to the values obtained in single-label classification ($\alpha = 0.95, 0.99$ respectively for the SVM and NN methods). This shows that structural information plays a more significant role in differentiating enzymatic activity in the case of multi-labeled enzymes than in single-label classification (which is mostly based on amino-acid sequence information).

Figure 3. Testing subset accuracy for dataset II

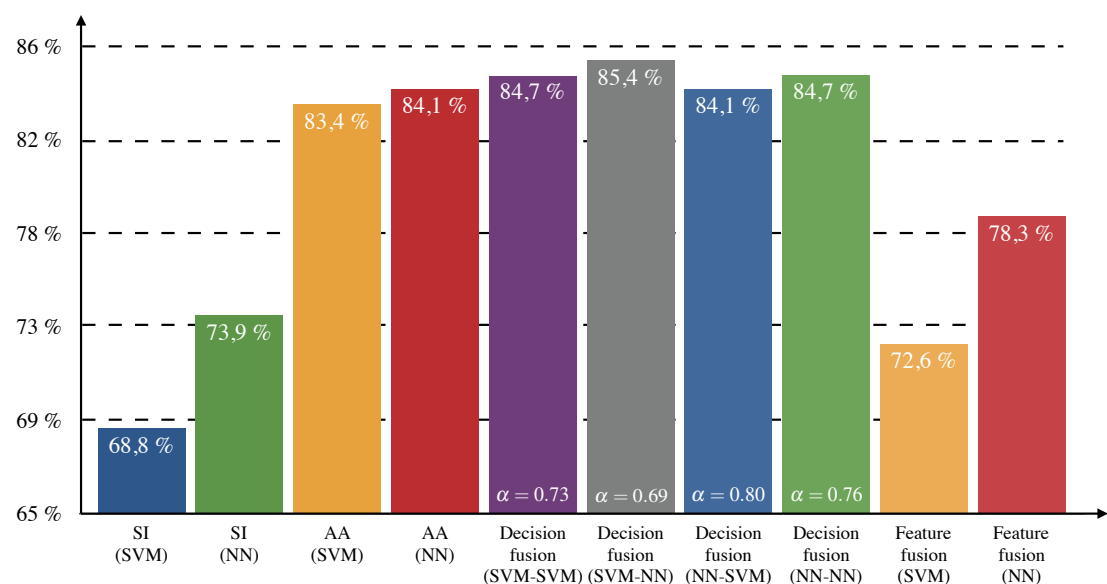



Figure 3 shows the subset accuracy for the testing set obtained for each approach in multi-label classification. Both SVM and NN classifiers achieved approximately 10% less subset accuracy when using only SI than when using only AA. Combining SI and AA according to the feature-level fusion scheme leads to intermediate values (between the ones achieved by only SI and AA) of subset accuracy. However, the combination of information based on the decision-level fusion scheme increased the subset accuracy by up to 1.3% compared to the best approach using AA only. The best results were obtained with the SVM-NN classification scheme. The overlap and discrepancy in correct predictions using SI (SVM), AA (NN) and the decision-level fusion scheme with SVM-NN are illustrated in figure 4.

We observed that 65.6% of the enzymes in the testing test were correctly predicted by all compared approaches (SI only, AA only and decision-level fusion). Also, out of 29 enzymes correctly predicted by AA but not by SI, 28 are also correctly predicted by the SVM-NN decision-level fusion scheme. This shows that the decision-level fusion incorporates the relevant information provided by AA which was missed by SI. Conversely, out of 5 enzymes correctly predicted by SI and not by AA, 2 of them are correctly predicted by the decision-level fusion scheme. This could be related to the chosen values of α that assigns a larger weight to the class probabilities calculated by AA than the ones extracted from SI.

Computation of 1-Hamming-Loss for each approach is shown in figure 5. All decision-level fusion schemes achieved higher values than the approaches using only AA or only SI. The decision-level scheme that performed best in terms of Hamming-Loss is SVM-SVM with an increase of 1.8% compared to

Figure 4. Repartition of correctly predicted enzymes with respect to subset accuracy

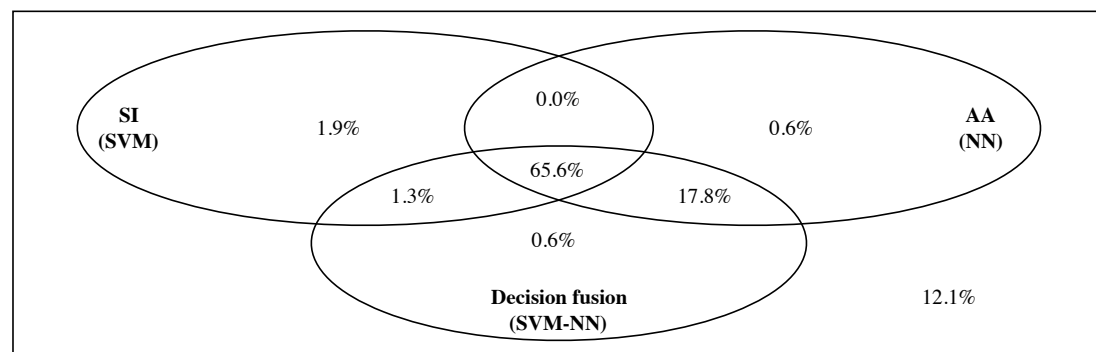



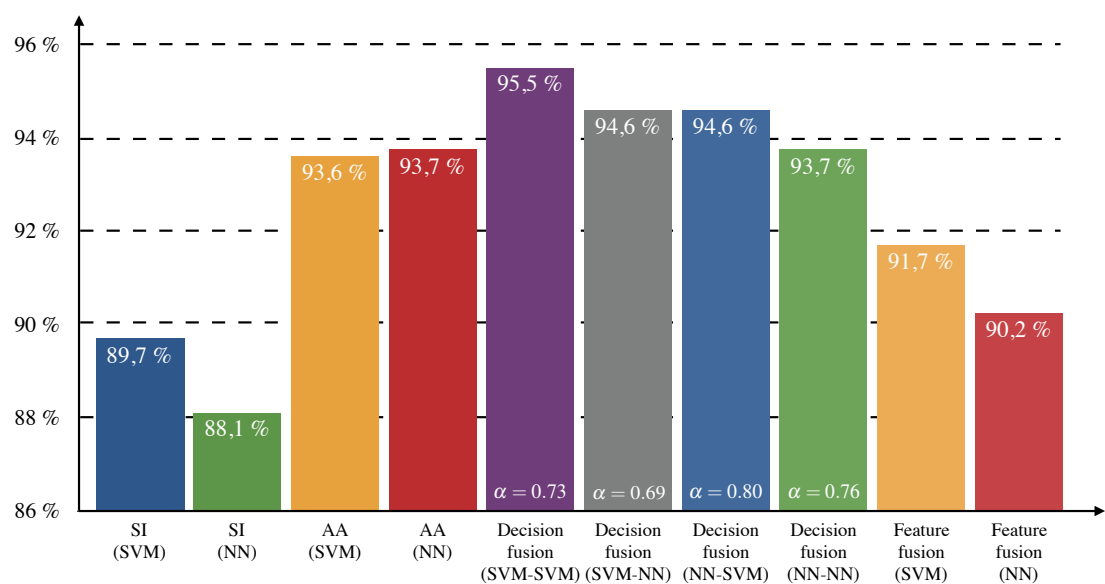
Figure 5. Testing 1-Hamming-Loss for dataset II

174 AA (NN). The comparison of 1-Hamming-Loss per class for each best method (SI only, AA only and
175 decision-level fusion) is shown in table 5.

Table 5. Comparison of 1-Hamming-Loss per class with SVM-SVM

Classifier	1-Hamming-Loss per class					
	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
SI SVM only	0.962	0.834	0.860	0.822	0.943	0.962
AA NN only	0.962	**0.930**	0.898	0.885	0.962	**0.987**
Decision fusion SVM-SVM	**0.968**	0.917	**0.949**	**0.943**	**0.968**	**0.987**

The SVM-SVM method achieves for each class except for the transferases up to 5.8% higher 1-Hamming-Loss than the maximum accuracy achieved by the best classifier of a single type of information (SI or AA). There is an increase in the performance regardless of the size of the class. In particular, classification of a large class such as the hydrolases had a 5.1% increase in 1-Hamming-Loss, whereas small classes like the lyases and isomerases were classified respectively with 5.8 and 0.6% better performance after fusion than with SI or AA only.

Table 6. Testing performance of dataset II

Type		SI		AA		Decision fusion				Feature fusion	
						SVM		NN			
Classifier		SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN
Alpha						0.73	0.69	0.80	0.76		
Hamming-Loss		0.103	0.119	0.064	0.063	**0.045**	0.054	0.054	0.063	0.083	0.098
Accuracy		0.790	0.800	0.883	0.885	**0.906**	0.898	0.879	0.889	0.823	0.831
Precision		0.857	0.829	0.901	0.906	0.942	0.918	**0.947**	0.907	0.889	0.856
Recall		0.825	0.831	0.908	0.908	**0.924**	0.920	0.885	0.911	0.847	0.856
F1 score		0.835	0.829	0.904	0.906	**0.928**	0.919	0.893	0.908	0.859	0.855
Subset accuracy		0.688	0.739	0.834	0.841	0.847	**0.854**	0.841	0.847	0.726	0.783
Macro	precision	0.921	0.744	0.940	0.941	0.962	0.945	**0.967**	0.903	0.927	0.806
	recall	0.741	0.777	0.881	0.871	**0.887**	0.879	0.854	0.881	0.791	0.787
	F1	0.801	0.758	0.902	0.897	**0.921**	0.905	0.905	0.889	0.844	0.794
micro	precision	0.864	0.822	0.904	0.907	0.943	0.919	**0.953**	0.904	0.901	0.857
	recall	0.829	0.832	0.910	0.910	**0.925**	0.922	0.885	0.913	0.850	0.857
	F1	0.846	0.827	0.907	0.908	**0.934**	0.921	0.918	0.909	0.875	0.857

Table 6 shows the results of the ten methods, according to all of the metrics that have been assessed for multi-label classification. With respect to all of the indexes, we observe that the decision-level fusion schemes outperform those carrying only one type of information. More particularly, each of the SVM-SVM, SVM-NN and NN-SVM techniques provide a distinct advantage in the process of multi-label classification. First of all, the SVM-SVM scheme is best in terms of 1-Hamming-Loss with a testing value of 95.5%, which surpasses other methods by at least a 1% margin. Also, this scheme proves to be the best in terms of the three definitions of recall, meaning that if an enzyme belongs to a certain class, SVM-SVM will be the more likely to detect it. In terms of predicting exact matches of the true labels, the

SVM-NN method will be the best one to consider with a testing value of 85.4%, which is at least 1.3% ahead of the performance achieved considering only one type of information. One of the most impressive rises in performance stems from the NN-SVM method, which proves to outperform SI and AA methods by +4.1%, +2.6%, and +4.6% in terms of precision, M-precision and m-precision respectively. Not only does it show that the relevance of class predictions is improved overall, but also and more importantly that small-populated classes benefit from this progression as well.

The code, which was written in Matlab and Python languages, is freely and publicly available at https://figshare.com/s/a63e0bafa9b71fc7cbd7. Running on a single Intel Xeon X5650 processor, the average prediction time of the enzymatic function(s) of a new enzyme was less than 3 seconds. Computations were achieved using High Performance Computing (HPC) resources from the "mesocentre" computing center of Ecole Centrale de Paris (http://www.mesocentre.ecp.fr) supported by CNRS.

DISCUSSION AND CONCLUSION

The results of both single-label and multi-label classifications showed that the combination of information leads to more accurate enzyme class prediction than the individual structural or amino acid descriptors. Among fusion approaches, the decision-level fusion performed better than the feature-level fusion. In the multi-label case, the SVM-NN fusion scheme achieved the best subset accuracy by predicting correctly the labels of 85.4% of the enzymes. The NN-NN fusion scheme also performed well (84.7%) and required the least computational time during the training phase. Structural information seems to be more important in the case of multi-label classification than in single-label, since the optimal relative weight of amino acid sequence features during fusion was found to be smaller in multi-labeled enzymes ($\alpha \in [0.69, 0.80]$) compared to single-labeled enzymes ($\alpha \in [0.95, 0.99]$).

In all examined cases, AA was more informative than SI in respect to the prediction of enzymatic activity. The same trend has been observed in Zou et al. (2013) where their study showed an increase of 0.81% with sequence related features, compared with structural features. However, it should be noted that we examined only general functional characteristics indicated by the first digit of EC code. A study assessing the relationship between function and structure (Todd et al. (2001)) revealed 95% conservation of the fourth EC digit for proteins with up to 30% sequence identity. Similarity, Devos and Valencia (2000) concluded that enzymatic function is mostly conserved for the first digit of EC code whereas more detailed functional characteristics are poorly conserved.

The single- and multi-label classification models have been trained and tested on enzymes assumed to perform single or multiple reactions, correspondingly. However, the single-label enzymes might be associated with other reactions not detected yet and in fact be multi-label. In order to assess the method in a more general scenario, we mixed both single- and multi-label information during training phase and observed a slight improvement in prediction accuracy. Specifically, we chose to examine the NN-NN fusion scheme because of its small computation time, and merged SI and AA probabilities obtained by both datasets I and II. This model achieved 89.2% subset accuracy and 95.8% accuracy for the multi-label dataset (by cross-validation) indicating an increase of 4.5% and 2.1% in respect to the results obtained with the NN-NN scheme trained only on multi-labeled data (shown in figure 3 and figure 5). This also corresponds to an increase of 3.8% and 0.3%, respectively, from the best fusion schemes.

Moreover, since it is unknown for new (testing) enzymes if they perform unique reactions, they have to be treated as multi-label. In order to estimate the performance of the single-label model in the case of unknown enzymes, we tested the best single-label classifier (i.e the NN on the decision level) on the multi-label dataset. For 93.0% of the enzymes the model predicted correctly one of their actual labels, whereas the prediction of all actual labels (by selecting the classes with the highest probability scores) was correct in 44.8% of the enzymes.

Furthermore, we investigated techniques dealing with imbalanced classes but did not observe any conclusive outcome. In particular, ADASYN improved overall accuracy on the single-label dataset by 0.1% but reduced balanced accuracy by 1.1%.

In conclusion, computational models calculated from experimentally acquired annotations of large datasets provide the means for fast, automated and reproducible prediction of functional activity of newly discovered enzymes and thus can guide scientists in deciphering metabolic pathways and in developing potent molecular agents. Future work includes the representation of the whole 3D geometry using additional structural attributes and the incorporation of deep learning architectures that have proven to be

powerful tools in supervised learning. The main advantage of deep learning techniques is the automatic exploitation of features and tuning of performance in a seamless fashion, that optimizes conventional analysis frameworks.

ACKNOWLEDGMENT

The authors wish to thank Prof. V. Megalooikonomou from the MDAKM group, Department of Computer Engineering and Informatics, University of Patras, for his earlier collaboration on structural similarity. The authors would also like to thank Chloé-Agathe Azencott for useful discussion about the Smith-Waterman algorithm.

REFERENCES

Amidi, A., Amidi, S., Vlachakis, D., Paragios, N., and Zacharaki, E. I. (2016). A machine learning methodology for enzyme functional classification combining structural and protein sequence descriptors. In *IWBBIO*.

Atiya, A. (2005). Estimating the posterior probabilities using the k-nearest neighbor rule. *Neural Computation*.

Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S. V. N., Smola, A. J., and Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *ISMB*, 21:47–56.

Concu, R., Dea-Ayuela, M., Perez-Montoto, L., Bolas-Fernandez, F., Prado-Prado, F., Podda, G., Uriarte, E., Ubeira, F., and Gonzalez-Diaz, H. (2009a). Prediction of enzyme classes from 3d structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of leishmania proteins. *J Proteome Res.*, 8(9):4372–4382.

Concu, R., Dea-Ayuela, M., Perez-Montoto, L., Uriarte, F. P.-P. E., Bolas-Fernandez, F., Podda, G., Pazos, A., Munteanu, C., Ubeira, F., and Gonzalez-Diaz, H. (2009b). 3d entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in leishmania parasites. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1794(12):1784–1794.

Concu, R., Podda, G., Uriarte, E., and Gonzalez-Diaz, H. (2009c). Computational chemistry study of 3d-structure-function relationships for enzymes based on markov models for protein electrostatic, hint, and van der waals potentials. *J Comput Chem.*, 30(9):1510–1520.

Dave, K. and Panchal, H. (2013). Enzpred-enzymatic protein class predicting by machine learning. *Current Topics in Medicinal Chemistry*, 13(14):1674–1680.

desJardins, M., Karp, P. D., Krummenacker, M., Lee, T. J., and Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*.

Devos, D. and Valencia, A. (2000). Practical limits of function prediction. *Proteins*.

Dobson, P. D. and Doig, A. J. (2005). Predicting enzyme class from protein structure without alignments. *J Mol Biol*.

Ferrari, L. D., Aitken, S., van Hemert, J., and Goryanin, I. (2012). Enzml: multi-label prediction of enzyme classes using interpro signatures. *BMC Bioinformatics*.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction, Foundations and Applications*. Springer.

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

Lee, B. J., Lee, H. G., Lee, J. Y., and Ryu, K. H. (2007). Classification of enzyme function from protein sequence based on feature representation. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference*, pages 741–747. IEEE.

Lie, J. and Koehl, P. (2014). 3d representations of amino acids-applications to protein sequence comparison and classification. *Computational and Structural Biotechnology Journal*, 11:47–58.

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Dzeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*.

Mohammed, A. and Guda, C. (2015). Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism. *BMC Genomics*, 16.

Munteanu, C., Gonzalez-Diaz, H., and Magalhaes, A. (2008). Enzymes/non-enzymes classification model complexity based on composition, sequence, 3d and topological indices. *Journal of Theoretical Biology*, 254(2):476–482.

NC-IUBMB (1992). *Enzyme Nomenclature*. Number ISBN 0-12-227164-5 (hardback), 0-12-227165-3 (paperback). Academic Press, San Diego, California.

Needleman, S. B. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(443-453).

Osman, M. H. and Choong-Yeun Liong, I. H. (2010). Hybrid learning algorithm in neural network system for enzyme classification. *ICSRS*, 2(ISSN 2074-8523).

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press.

Shen, H. and Chou, K. (2007). Ezypred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Commun.*, 364(1):53–59.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Todd, A., Orengo, C., and Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, pages 1–13.

Valencia, A. (2005). Automatic annotation of protein function. *Current Opinion in Structural Biology*, 15:267–274.

Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2013). Predicting enzymatic function from global binding site descriptors. *Proteins*, 81(3):479–489.

Volpato, V., Adelfio, A., and Pollastri, G. (2013). Accurate prediction of protein enzymatic class by n-to-1 neural networks. *BMC Bioinformatics*.

Wang, Y., Jing, R., Hua, Y., Fu, Y., Dai, X., Huang, L., and Li, M. (2014). Classification of multi-family enzymes by multi-label machine learning and sequence-based descriptors. In *Analytical Methods*, volume 6, pages 6832–6840.

Yadav, S. K. and Tiwari, A. K. (2015). Classification of enzymes using machine learning base approaches: a review. *Machine Learning and Application: An International Journal*, 2.

Zhang, M.-L. and Zhou, Z.-H. (2006). Multi-label neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transaction*, 18:1338–1351.

Zou, H. and Xiao, X. (2016). Classifying multifunctional enzymes by incorporating three different models into chou's general pseudo amino acid composition. *J Membr Biol.*, 249(4):551–557.

Zou, Q., Chen, W., Huang, Y., Liu, X., and Jiang, Y. (2013). *Identifying Multi-Functional Enzyme by Hierarchical Multi-Label Classifier*, volume 10. American Scientific Publishers.