This is a preprint of the following chapter: Thomas Papastergiou et al., Multiple Instance Learning Based on Mol2vec Molecular Substructure Embeddings for Discovery of NDM-1 Inhibitors, published in Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022), edited by Fdez-Riverola, F., Rocha, M., Mohamad, M.S., Caraiman, S., Gil-González, A.B., 2023, Springer reproduced with permission of Springer (as it appears on the copyright page of the book). The final authenticated version is available online at: https://doi.org/10.1007/978-3-031-17024-9_6.

Multiple Instance Learning Based on Mol2vec Molecular Substructure Embeddings for Discovery of NDM-1 Inhibitors

Thomas Papastergiou[1, 3][0000-0002-0241-080X], Jérôme Azé[1][0000-0002-7372-729X], Sandra Bringay[1,2][0000-0002-2830-3666], Maxime Louet[3][0000-0002-5840-2147], Pascal Poncelet[1][0000-0002-8277-3490], Laurent Gavara[3][0000-0003-0146-1848]

[1] LIRMM, University of Montpellier, CNRS, Montpelier, France
[2] AMIS, Paul Valery University, 34000 Montpellier, France
[3] IBMM, CNRS, University of Montpellier, ENSCM, Montpellier, France
{thomas.papastergiou, jerome.aze, sandra.bringay, pascal.poncelet}@lirmm.fr, {laurent.gavara, maxime.louet}@umontpellier.fr

Abstract. In this paper, we first present a new dataset of NDM-1 biological activities that is compiled by a cleaned version of the NMDI database. A literature review enriched the former database by 741 new compounds, comprising activities against NDM-1 classified in three classes (inactive, weakly and strongly active compounds) by specifying a unifying procedure for the labeling, which covers a range of different activity properties. Second, we restate the classification problem in the Multiple Instance Learning (MIL) setting by representing the compounds as a collection of Mol2vec vectors, each of them corresponding to a specific substructure (either atom or atom including their first neighbors). We observe an amelioration up to 45.7% and 38.47% in respect to balanced accuracy and F1-score, respectively, for the strongly active class in the MIL approach when compared to the classical Machine Learning

paradigm. Finally, we present a classification and ranking framework based on classifiers learned by a k-fold CV procedure, which possess different hyper-parameters per fold, learnt by a Bayes optimization procedure. We observe that the top-3 and top-5 ranked accuracies of the strongly active classified compounds yield 100% for the MIL setting.

Keywords: Machine Leaning, Multiple Instance Learning, Drug Discovery, NDM-1 inhibitors

1 Introduction

New Delhi Metallo-β-lactamase (NDM-1) is a recent bacterial enzyme highly involved in bacterial resistance phenomenon by its capacity to inactivate the main available class of antibiotics: the β-lactam agents [1]. The common way to fight this kind of resistance is the adjuvant strategy, which consists in a combination of a β-lactam agent and a β-lactamase inhibitor [2]. Some combinations are already on the market but sadly are not effective on NDM-1 producers. Due to the specific mode of action of NDM-1, involving zinc atoms into the active site, the design of efficient inhibitors remains an unmet therapeutic need [3]. This major threat on human health has to be addressed to avoid return to the pre-antibiotic era.

The drug discovery process is a very time-consuming (approximately 10-14 years) and costly (1 billion USD magnitude) procedure, characterized by high attrition rates, to reach marketing authorization [4]. Thus, *in silico* strategies, (e.g. Virtual Screening (VS) techniques) are often used as starting point for medical chemistry, for speeding-up the drug discovery process by identifying compounds of high potential against specific targets. VS can be categorized in three main areas: (1) structure-based (requiring knowledge of the 3D structure of the target), (2) ligand-based (requiring knowledge of active ligands) and (3) hybrid approaches [5]. As the number of ligands in openly available databases is constantly increasing (e.g. ZINC 15 [6], ChEMBL [7] etc.) Machine Learning (ML) techniques are used for constructing efficient models used in VS for hit identification (i.e. discovery of small molecules as a starting point for medicinal chemistry programs), drug repurposing, activity scoring [8] or activity prediction [9]. In order to tackle the latter problem in an efficient manner specialized, annotated data are needed, since ligand-activity data that refer to different targets or to general target categories (e.g. antibacterial, anti-cancer, anti-inflammatory etc.) will produce ML models with low efficiency on specified tasks (e.g. discovery of effective NDM-1 inhibitors).

Multiple Instance Learning (MIL) is a paradigm of weakly supervised learning where the samples to be classified (i.e. bags) are represented by multiple vectors (i.e. instances) and labels are only available for the bags. MIL was first introduced by Dietterich et al. in 1997 [10] tackling a musk odor prediction task. In this structure-activity prediction problem, each molecule was represented by their different conformations captured by various feature vectors representing the shape of the molecule in each conformation. The standard MIL assumption was then applied stating that a bag is positive if it contains at least one positive instance (i.e. an active

molecule conformation) and negative otherwise. The MIL paradigm has been used in different application areas including medical imaging classification, frailty prediction using physiological signals [11], natural images classification [12], [13], drug discovery [14] etc.

As the numerical representation of molecules is crucial in order to construct ML models, different approaches have been proposed including Extended-Connectivity Fingerprints (a.k.a. Morgan Fingerprints (MF)), Molecular Graphs or computer learned representations [15] like Mol2vec representation [16], a NLP-inspired technique that considers compound substructures, extracted by MF, as words and compounds as sentences. In this frame, a compound is represented by a collection of vectors, each of which corresponds to a substructure of the molecule, and a vector representation is obtained by adding-up these substructure vectors.

ML have been extensively used in the drug design process for various purposes: prediction on drug-protein interactions, discovering of drug efficacy or ensuring the safety biomarkers, with applications ranging from prediction of protein folding or target identification to hit discovery [8]. More specifically, Shi et al. [17] compiled a NDM-1 activities database, comprising strongly and weakly active compounds of known NDM-1 activities and provided a list of "hypothetical" inactive compounds, based on their physicochemical properties. They have applied classical ML and deep learning models for activity prediction based on physicochemical features extracted by the commercial software MOE2018[1].

In this paper we present a framework to tackle the problem of discovering potential strongly active NDM-1 inhibitors by the use of ML models. For this purpose, (1) we compile a database of 868 compounds of known activity against NDM-1, by collecting compounds from the recent literature and by considering only compounds referring to the NDM-1 enzyme, coming from the NDMI database, proposed by Shi et al. [17]; (2) we establish a unifying set of rules for labelling compounds as inactive, weakly active or strongly active, by considering different experimental properties; (3) we restate the activity classification problem as a MIL problem by representing molecules by a collection of Mol2vec vectors representing molecular substructures; (4) we propose an ensemble classification framework, which is able to rank the classification outputs per predicted class.

The contributions of this paper can be resumed as follows:

1. The compilation of a dataset of known activities against NDM-1 annotated by a set of unifying rules for incorporating different experimental properties;
2. The restatement of the activity classification problem in the MIL paradigm, by representing compounds by Mol2vec representations of their substructures, that shows experimentally better performance than state-of-the-art Mol2vec classical ML models;
3. The introduction of an homogeneous ensemble classifier framework that classifies and ranks the classification results per class, and shows very promising classification and ranking results for the strongly active class in terms of top-5 to

[1] https://www.chemcomp.com/Products.htm

top-15 accuracy, when evaluated on an independent test set showing good generalization capabilities for the MIL ensemble models.

2 Materials and Methods

2.1 Dataset collection

In [17], Shi et al. introduced a database of active and "hypothetical" inactive compounds, found in the literature, comprising 511 and 6,358 compounds respectively. The "hypothetical" inactive compounds where specified by considering physiochemical properties of 51,280 compounds of the ZINC database, lacking of activity data against NDM-1. To compile a database comprising only compounds with known activities against NDM-1, we considered only the 511 compounds of NDMI. For each of these compounds, we tried to verify the existence of the publications by performing database searches on the PubMed[2] database using, the provided Digital Identification Number (DOI) of each publication. In a subsequent step, the relevance to NDM-1 inhibitors activities of the publications where checked, and irrelevant entries where discarded. Subsequently, the corresponding Canonical SMILES representation was produced, using the RDKit[3] library, and duplicate entries were discarded. This procedure yielded 127 compounds with known activity scores. Furthermore, a thorough search in the existing literature for compounds with known activities on NDM-1 returned 741 new unique compounds. In total the new NDM-1 activity database comprises 868 unique compounds.

2.2 Labeling the database

The activity against NDM-1 is measured by experimental properties based on enzymatic inhibition: (K_i, IC_{50}, pIC_{50}, enzyme inhibition at a set concentration, or K_d) [18] or in vitro bacterial growth inhibition (MIC) [19]. Our goal is to identify potential strong active compounds against NDM-1. We classify the compounds in the new database in three classes: inactive, weakly active and strongly active compounds, inspired by the classification in [17] but with different, more strict, cut-off values for the strongly active compounds, since the aim is to deliver a classifier that can predict strongly active molecules with high inhibition capacity. We adopt a unifying strategy that comprises all activity properties we include, in contrast to [17], only compounds with known activities against NDM-1 and classify them according to the cut-off values shown in Table 1.

Table 1. Labeling cut-off scores for activity properties

rank		inactive	weakly active	strongly active
1	K_i (μM)	>10	[0.5, 10)	≤0.5

[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://github.com/rdkit/rdkit

2	IC_{50} (μM)	>20	(1, 20]	≤1
3	pIC_{50}	<4.7	[4.7, 6]	≥6
4	%100 μM	<60%	>60%	-
5	K_d (μM)	>10	[0.5, 10)	≤0.5
6	MIC (μg/ml)	>8	(0.5, 8]	≤0.5

As the compounds found in the literature often possess activity measurements for multiple properties and as different papers report different values for the same compound that sometimes leads to different labeling of the same compound, we need a unifying approach to cure these inconsistencies. We adopt a ranking order for the properties and we classify each compound according the property with the highest rank. The ranking of the properties is shown in Table 1. Furthermore, if the classification of a compound according two different publications is ambiguous, respecting the ranking of the properties, the more active label is assigned to the compound, since there is evidence in at least one experiment of the highest activity. We need to note here that when we applied the above procedure to the 127 compounds retained from the NMDI database [17], 51 compounds (40.16%) have changed labels.

The rationale behind the ranking of the activity properties is following the main objective of this work, which is to deliver a classification model for the discovery of active NDM-1 inhibitors. In this sense, properties witch refer to enzymatic inhibition, such as K_i, IC_{50} etc. are placed in higher ranks than activity properties that refer to the NDM-1 agent inhibition (e.g. MIC). In this sense, for compounds that both enzymatic and bacterial inhibition activity are provided, we rely on the enzymatic activity property for their classification. On the other hand, when only the agent's inhibition property is provided, we rely on properties like MIC, although that in vivo experiments tend to possess a higher degree of complexity, than in vitro inhibition experiments on enzymatic assays, and because MIC values are indirect observations. Indeed, it's the concentration of β-lactam agents to have inhibitory effect protected by NDM-1 inhibitor. In this sense, the adopted ranking procedure resolves these ambiguities, in the aforementioned direction, and has a mild effect on the labeling of the dataset, since if the ranking of the activity properties is e.g. reversed only about 3% of the compounds would change labels.

2.3 Calculating Mol2vec embeddings

ML embeddings

For calculating the embeddings, we used the Mol2vec pre-trained model of [16]. The model was trained on 19.9M compounds of ZINC and ChEMBL databases, as a skip-gram word2vec model, with window size of 10 using radius 1 for the MF (for a more elaborate description of the extraction of the MF refer to [20]). For training the Mol2vec model all MF identifiers of radii 0 and 1 where generated, and considered as words, while each molecule was considered as sentence. The rare identifiers (i.e. identifiers that occurred less than 3 times in the training database) were marked as

"unknown" and were attributed to a special identifier called 'UNK'. After training the word2vec model, with such a specification, the individual vectors of each molecule, that corresponded to the MF substructures, were added up to produce a single vector for each molecule, and thus a molecule is represented by a vector of 300 real values.

MIL embeddings

In order to restate the classification problem in the MIL setting, each molecule (i.e. bag) has to be represented by a collection of the individual's MF substructures vectors (i.e. instances). The labels for each bag are known as the activities of each corresponding molecule are known, but the individual labels for each instance are unknown, since there is no activity information concerning each substructure. Thus, for the molecule to be bound in the target, one or multiple substructures of the molecule must be involved (i.e. active) in the binding affinity.

As the Mol2vec model calculates the embedding vectors of all the substructures of each molecule, up to a specified radius r, after removing all duplicate vectors corresponding to the same substructure, we introduce two different types of MIL representations: (1) each molecule can be represented as a collection of all the substructure vectors of all radii (used in this work), or alternatively (2) each molecule can be represented as a collection of substructure vectors corresponding to a specific radius k, with $0 \leq k \leq r$. In contrast to the Mol2vec model where all the substructure vectors (i.e. vectors corresponding to MF of different radii) are added up to construct a vector representation for each compound, in the MIL representation, each unique substructure vector is explicitly included in the compound's representation. As we will show experimentally, this contributes positive to the performance of the models, since according to the MIL assumption, the inactivity of a molecule suggests that all his substructures must be inactive (i.e. not contributing to the binding affinity) and the weak or strong activity suggests that a portion of his substructures is involved to the binding affinity.

2.4 Classification and ranking frame work

In this section, we introduce a homogeneous classification and ranking framework, which is based on different models acquired by a k-fold Cross Validation (CV) procedure. Let $f_i^{h_i}: \mathbb{R}^m \rightarrow \{cl\_1, ..., cl\_n\}$, and $d_i^{h_i}: \mathbb{R}^m \rightarrow \mathbb{R}$, $i = 1, ..., k$, k classification functions and their corresponding decision functions obtained by a *k*-fold CV procedure, where *n* and *m* are the number of classes and features respectively and $h_i \in \mathbb{R}^l$ are the corresponding hyper-parameters specified by a hyper-parameter optimization procedure for each individual fold. In this sense, we are equipped with *k* homogeneous classifiers trained and evaluated in different training-validation sets having different hyper-parameters. The decision of the ensemble classifier is then given by a voting procedure $g\left(f_0^{h_0}, ..., f_k^{h_k}\right) = c$ and the per class rank of the ensemble's classification output for each sample can be given by $r_c(x) = \underset{i}{\text{mean}}\left\{d_i^{h_i}(x),\ if\ f_i^{h_i}(x) == c\right\}$, $c = cl\_1, ..., cl\_n$. Thus, by calculating for

each sample the mean decision value of these classifiers, which have predicted the decision of the ensemble classifier, we obtain the rank per class of each sample.

3 Results and Discussion

For the evaluation of the proposed methods, we used the NDM-1 activities database described in Section 2.1. The 868 known activities compounds' database, included 345 (39.75%) inactive, 254 (29.26%) weakly active and 269 (30.99%) strongly active molecules, making it a relative balanced dataset.

For representing numerically the compounds of the database for the classical ML paradigm, we generated Mol2vec vectors, employing the 300 dimensional pre-trained model of [16] resulting to 863 unique identifiers and 21 "unknown" structures. After generating the numerical representation for the MIL algorithms, we were equipped by 19,082 instances of radii 0 and 1, from which 1 radius 0 and 55 radius 1 structures were "unknown". Furthermore, we obtained 7,264 and 11,818 instances of radii 0 and 1 respectively. The "unknown" structures were removed from the training and testing sets, since they do not contribute to the representation of a bag, because they represent potential different substructures. The removal of the "unknown" structures does not resulted to bags (i.e. molecules) without representation, since each compound was represented by at least one known substructure.

The performance evaluation of the ranking and ensemble classification framework was performed by an independent Test Set (TS), acquired by a stratified 90% Training (TrS)-10% (TS) split of the database, while the evaluation of the classifiers was performed by 10-fold CV on the TrS split. Support Vector Machines (SVM) with Radial Basis Kernel (RBF), Linear Discriminant Analysis (LDA) and Random Forest (RF) [8] have been used as representatives of classical ML algorithms and TensMIL [11] and TensMIL2 [12] as MIL state-of-the-art algorithms, from which we decoupled the feature extraction by tensor decomposition phase, since our data are of 2D nature, and used only the classification procedure.

TensMIL and TesnMIL2 consist of two inference phases: in the first phase, a score for each instance (i.e. a substructure) is calculated and the bags' scores distribution are estimated. These distributions are then fed to a bag classifier who yields the classification result. The difference of TensMIL2 is that in the first phase it incorporates an instance selection procedure for selecting the most informative instances (i.e. substructures) per bag.

For tuning the hyper-parameters for each algorithm, a Bayes optimization approach was adopted like in [11], using as objective function the mean 2-fold CV balanced accuracy (Bacc) on a validation set. The hyper-parameters were tuned separately for each one of the 10-folds, resulting thus to 10 different classifiers with different sets of hyper-parameters. The hyper-parameters tuned for each classifier were: C and γ for SVM, *nrOfForestTrees* for RF, ϑ_H and ϑ_p for TensMIL and q and p for TensMIL2, where ϑ_H corresponds to the number of the histogram bins for the distribution estimation, ϑ_p and p to the variance retained of the PCA applied to the instances' feature matrix and q to the quantile defining the threshold for the instance selection

procedure of TensMIL2. For the required ϑ_H parameters of TensMIL 2 we used for each experiment the $\left\lceil \underset{i}{\text{mean}}\, \theta_H^i \right\rfloor$, *i*=1, …, 10, where ϑ_H^i is the parameters acquired by TensMIL on the *i-th* fold of the corresponding experiment. For the LDA algorithm none hyper-parameter was tuned. For discussion on the hyper-parameters, the interested reader may refer to the corresponding publications.

For the ensemble classifier, we used a majority voting approach in the sense that the class predicted by the majority of the classifiers is attributed to the corresponding sample.

The metrics used for evaluating the ML, MIL and ensemble classifiers were the mean of 10-fold CV accuracy, balanced accuracy, precision-, recall- and F1-score-per class. For the evaluation of the ranking procedure, we used the per class top-k accuracy: $TopAcc_c^{(k)} = \frac{\#top\text{-}k\ ranked\ True\ Positives}{k}$, with *c* being the corresponding class.

3.1 Results

Classification and generalization evaluation

Since we are interested in discovering strongly active NDM-1 inhibitors, special attention on the presentation of the results will be given to the strong active class. In Table 2 we compare the classification performance of the ML and MIL paradigms. The reported Precision (Prec.) and Recall metrics refer to the corresponding metrics of the strong active class.

Table 2. Comparison between classical ML and MIL for NDM-1 activity classification.

	10-fold CV				Ensemble classifier			
	Acc.	**Bacc.**	**Prec.**[4]	**Recall**[3]	**Acc.**	**Bacc.**	**Prec**[3]**.**	**Recall**[3]
SVM	52.25%	50.83%	64.38%	66.40%	39.08%	32.76%	0.00%	0.00%
LDA	51.09%	50.24%	58.95%	68.00%	35.63%	38.00%	33.77%	**96.30%**
RF	52.76%	51.15%	62.41%	62.28%	40.23%	33.71%	0.00%	0.00%
TensMIL	72.08%	70.81%	80.65%	**80.53%**	**75.86%**	**73.16%**	74.19%	85.19%
TensMIL2	**74.40%**	**73.20%**	**82.57%**	80.52%	73.56%	70.79%	**80.00%**	74.07%

As presented in Table 2, the MIL approach resulted in an amelioration from 38.43% up to 45.7% in terms of balanced accuracy (Bacc.) with respect to the ML approach. Precision (Prec.) and Recall for the strong activity class was augmented up to 40.07% and 29.30% respectively in the case of the MIL setting in comparison to the ML paradigm. The improvement of the classification performance could be attributed to the compounds' MIL representation. Instead of representing each compound by the sum of the vectors corresponding to each substructure, as is the case of the ML paradigm, each molecule is represented by the set of vectors of their substructures. As, in the frame of MIL, the individual activity labels of each instance (i.e. substructure) are unknown, and as the binding of a ligand to a target is a subject of specific substructures of a compound (i.e. the binding site of the ligand may concern a

[4] Refers to the strong activity class metric.

part of the compound having special structures and binding properties) the MIL representation has be proven beneficial to the activity classification performance.

Furthermore, for evaluating the generalization ability of the models as well as for assessing the ranking performance of the ensemble classification framework introduced in Section 2.4, we evaluated their performance in an independent test set that was not subject of the training, hyper-parameter tuning and CV evaluation of the models. As presented in Table 2, the ensemble classifier in the frame of classical ML models performs worse than the individual classifiers, suggesting that the generalization ability of these classifiers are poor. In contrast, in the MIL setting we see that the ensemble classification framework in the case of TensMIL performs better than the individual classifiers, in terms of Acc., Bacc. and Recall for the strong active class, and in the case of TensMIL2 it performs slightly worse than the initial classifiers, suggesting the generalization ability of the initial classifiers. The ensemble classifiers in the MIL setting perform in terms of balanced accuracy from 86.29% to 123.32% better than in the classical ML setting. In the case of the LDA model, the ensemble classifier displays a 96% recall, but only 34% precision for the strong activity class, suggesting that in this case a significant amount of compounds are predicted as strongly active and thus the False Positive predictions are relative high.

For further assessing the performance of the classifiers and their generalization ability the 10-fold CV and on the independent TS F1-scores of the classifier and the ensemble classification framework are presented in Table 3. Over all, the MIL classifiers perform better in comparison to the classical ML classifiers. More specifically, in the MIL setting we have from 24.95% to 37.07%, from 138.84% to 190.24% and from 31.36% to 38.47% better F1-scores respectively for the inactive, weakly active and strongly active classes, for the 10-fold CV evaluation. For the MIL algorithms the F1-score performance is better or slightly worse for the ensemble classifier on an independent TS, in contrast to the ML algorithms. Furthermore, we observe that the ensemble classifier based on the SVM and RF algorithms was not able to predict samples of the strong activity class. Finally, in general, we observe lower performances in respect to the F1-score for the weakly activity class, than for the inactive and strongly active classes.

Finally, comparing the results in [17], where handcrafted features and “hypothetical” inactive compounds were used, to our experiments, we conclude that in general the classification performance, with respect to the F1 score in [17] is better for the inactive and weakly active class. In contrast, TensMIL2 performs from 15.36% to 41.66% better than the models in [17] for the strongly active class. Although the two experiments are not fully comparable, we can conclude that the use of Mol2vec representations in the MIL setting and the stricter labeling for the strongly active class had a positive effect in the performance of the classification of the strongly active class.

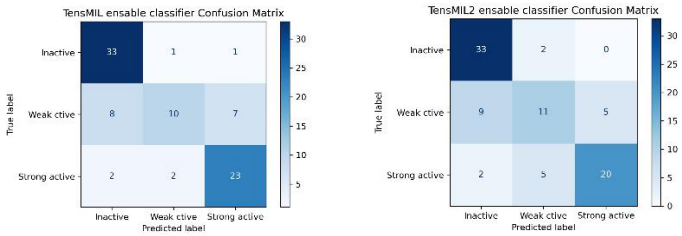

Fig. 1. Confusion Matrices of TensMIL and TensMIL2, for the ensemble classifier on a independent test set

Table 3. Per class CV and ensemble classifier F1-scores.

	10-fold CV F1-score			Ensemble classifier F1-score		
	Inactive class	**Weakly active class**	**Strong active class**	**Inactive class**	**Weakly active class**	**Strong active class**
SVM	0.5946	0.2021	0.6107	0.5546	0.0714	Inf.
LDA	0.5794	0.2243	0.6082	0.1053	0.1875	0.5
RF	0.6216	0.2142	0.5861	0.5667	0.0741	Inf.
TensMIL	0.7767	0.5357	0.8023	**0.8462**	**0.5263**	**0.7931**
TensMIL2	**0.7942**	**0.5866**	**0.8116**	0.8354	0.5116	0.7692

Ranking evaluation.

The results of the ranking procedure are displayed in Table 4 where the top-3, 5, 10 and 15 ranked compounds accuracy per class is presented.

Table 4. Ranking performance (top-k accuracy) of the ensemble classifiers per class

	Inactive Class				Weak active class				Strong active class			
	Top-3	**Top-5**	**Top-10**	**Top-15**	**Top-3**	**Top-5**	**Top-10**	**Top-15**	**Top-3**	**Top-5**	**Top-10**	**Top-15**
SVM	0.333	0.6	0.6	0.6	0.333	0.2	0.1	0.0667	0	0	0	0
LDA	0.667	0.4	0.2	0.133	**0.667**	0.6	0.3	0.2	0.333	0.6	0.7	0.6
RF	0.333	0.4	0.5	0.467	0.333	0.2	0.1	0.067	0	0	0	0
TensMIL	**1**	**0.8**	**0.9**	0.867	**0.667**	0.6	**0.8**	**0.667**	**1**	**1**	**0.9**	**0.933**
TensMIL2	**1**	**0.8**	**0.9**	**0.933**	**0.667**	**0.8**	0.6	0.6	**1**	**1**	**0.9**	**0.933**

The improvement of the MIL algorithms in comparison to the classical ML algorithms in terms of the top-k ranking accuracy for the inactive class is from 1.33x to 7x (top-15 accuracy), for the weakly active compounds up to 10x and for the active class up to 3x. TensMIL and TensMIL2 are displaying 100% top-3 and top-5 accuracy, meaning that the top-5 ranked compounds are strongly active. In contrast, the ranking based on the ensembles of RF and SVM algorithms did not yield strongly active compounds in the top-15 ranks. In the evaluation of the classification performance of the ensemble classifier, MIL algorithms display better ranking accuracy than ML algorithms and furthermore, their performances on inactive and strongly active class are better than on the weakly active class.

4 Conclusion

To conclude, the compilation of a new database comprising compounds with known activities against NDM-1 (excluding “hypothetical” inactive compounds), as well as the unifying labeling procedure, that comprises a stricter, in comparison to former approaches, rules for the strongly active compounds, can be beneficial for discovering strongly active compounds against NDM-1. Furthermore, the restatement of the classification problem in the MIL framework, by representing a compound as a bag of vectors corresponding to their substructures, showed promising results, in terms of the efficiency in the three class classification problem, as, often, a part of the molecule corresponding to certain substructures, is responsible for the binding of the ligand to the target. The introduction of the homogeneous ensemble classifier and the ranking procedure, especially if MIL algorithms are used, showed promising results, as in the case of TensMIL and TensMIL2 classifiers ensembles, where the top-3 and top-5 ranked strongly active predicted compounds belong to the strongly active class, as predicted on an independent test set. This fact suggests that a screening for active compounds could reveal strongly active compounds among the top ranked results of the ensemble classifier. Finally, the classification evaluation of the ensemble classifier on an independent test set showed a great generalization ability for the MIL classifiers.

Acknowledgments. This project was publicly funded through ANR (the French National Research Agency) under the "Investissements d’avenir" programme with the reference ANR-16-IDEX-0006.

References

1. M. F. Mojica, R. A. Bonomo, and W. Fast, « B1-Metallo-β-Lactamases: Where Do We Stand? », *Curr. Drug Targets*, vol. 17, nº 9, p. 1029-1050, May 2016.
2. C. González-Bello, « Antibiotic adjuvants – A strategy to unlock bacterial resistance to antibiotics », *Bioorg. Med. Chem. Lett.*, vol. 27, nº 18, p. 4221-4228, Sept. 2017.
3. P. Linciano et al., « Ten Years with New Delhi Metallo-β-lactamase-1 (NDM-1): From Structural Insights to Inhibitor Design », *ACS Infect. Dis.*, vol. 5, nº 1, p. 9-34, Jan. 2019.
4. J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, « Innovation in the pharmaceutical industry: New estimates of R&D costs », *J. Health Econ.*, vol. 47, p. 20-33, May 2016.
5. A. U. Khan, « Virtual Screening Strategies: A State of Art to Combat with Multiple Drug Resistance Strains », *MOJ Proteomics Bioinforma.*, Mar. 2015.
6. T. Sterling and J. J. Irwin, « ZINC 15 – Ligand Discovery for Everyone », *J. Chem. Inf. Model.*, vol. 55, nº 11, p. 2324-2337, Nov. 2015.
7. A. Gaulton *et al.*, « The ChEMBL database in 2017 », *Nucleic Acids Res.*, vol. 45, nº D1, p. D945-D954, Jan. 2017.
8. S. Dara et al., « Machine Learning in Drug Discovery: A Review », *Artif. Intell. Rev.*, vol. 55, nº 3, p. 1947-1999, Mar. 2022.
9. H. C. S. Chan et al., « Advancing Drug Discovery via Artificial Intelligence », *Trends Pharmacol. Sci.*, vol. 40, nº 8, p. 592-604, Aug. 2019.

10. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, « Solving the multiple instance problem with axis-parallel rectangles », *Artif. Intell.*, vol. 89, nº 1, p. 31-71, Jan. 1997.
11. T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, « Tensor Decomposition for Multiple-Instance Classification of High-Order Medical Data », *Complexity*, vol. 2018, p. 1-13, Dec. 2018.
12. T. Papastergiou, E. I. Zacharaki, and V. Megalooikonomou, « TensMIL2: Improved Multiple Instance Classification Through Tensor Decomposition and Instance Selection », in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sept. 2019, p. 1-5.
13. E. Branikas et al., « Instance Selection Techniques for Multiple Instance Classification », in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece, July. 2019, p. 1-7.
14. M.-A. Carbonneau et al., « Multiple instance learning: A survey of problem characteristics and applications », *Pattern Recognit.*, vol. 77, p. 329-353, May 2018.
15. D. S. Wigh, J. M. Goodman, and A. A. Lapkin, « A review of molecular representation in the age of machine learning », *WIREs Comput. Mol. Sci.*, Feb. 2022.
16. S. Jaeger, S. Fulle, and S. Turk, « Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition », *J. Chem. Inf. Model.*, vol. 58, nº 1, p. 27-35, Jan. 2018.
17. C. Shi et al. , « Applications of machine-learning methods for the discovery of NDM-1 inhibitors », *Chem. Biol. Drug Des.*, vol. 96, nº 5, p. 1232-1243, Nov. 2020.
18. B. T. Burlingham and T. S. Widlanski, «An Intuitive Look at the Relationship of Ki and IC50: A More General Use for the Dixon Plot», *J. Chem. Educ.*, vol. 80, nº 2, p. 214, Feb. 2003.
19. J. M. Andrews, «Determination of minimum inhibitory concentrations», *J. Antimicrob. Chemother.*, vol. 48, nº suppl_1, p. 5-16, July. 2001.
20. D. Rogers and M. Hahn, « Extended-Connectivity Fingerprints », *J. Chem. Inf. Model.*, vol. 50, nº 5, p. 742-754, May 2010.