Neural-Based Approaches to Overcome Feature Selection and Applicability Domain in Drug-Related Property Prediction

María Virginia Sabando*, Ignacio Ponzoni, Axel J. Soto

Institute for Computer Science and Engineering, UNS – CONICET, Argentina
Department of Computer Science and Engineering, Universidad Nacional del Sur, Argentina

Abstract

In the fields of pharmaceutical research and biomedical sciences, QSAR modeling is an established approach during drug discovery for prediction of biological activity of drug candidates. Yet, QSAR modeling poses a series of open challenges. First, chemical compounds are represented on a high-dimensional space and thus feature selection is typically applied, although this task entails a challenging combinatorial problem with potential loss of information. Second, the definition of the applicability domain of a QSAR model is a desirable aspect to determine the reliability of predictions on unseen chemicals, which is often difficult to assess due to the extent of the chemical space. Finally, interpretability of these models is also a critical issue for drug designers. The purpose of this work is to thoroughly assess the application of neural-based methods and recent advances deep learning for QSAR modeling. We hypothesize that neural-based methods can overcome the need to perform a descriptor selection phase. We developed three QSAR models based on neural networks for prediction of relevant chemical and biomedical properties that, in the absence of any feature selection step, can outperform the state-of-the-art models for such properties. We also implemented an embedded applicability domain technique based on network output probabilities that proved to be effective; its application improved the

*Corresponding author
Email address: virginia.sabando@cs.uns.edu.ar (María Virginia Sabando)

predictive performance of the model. Finally, we proposed the use of a *post hoc* feature analysis technique based on an aggregation of network weights, which enabled effective detection of relevant features in the model.
Keywords: neural networks, QSAR modeling, model interpretability, applicability domain, feature selection

1. Introduction

The integration of computational sciences into the pharmaceutical and biomedical industry has yielded several applications and technological advances during the last decades [14, 48]. The pharmaceutical industry is primarily headed towards improving the long and costly process of discovery and development of new drugs, which involves several stages including *in vitro* and *in vivo* wet-lab experiments. Computer-aided rational drug design has allowed accelerating drug candidate identification and prioritization while reducing costs and has helped improve the critical attrition rate in drug discovery projects [12, 44]. The purpose of *in silico* drug discovery is to design models for predicting biological activity and physicochemical properties of drug candidate compounds. These models, referred to as Quantitative Structure-Activity Relationship (QSAR), are regression or classification models used in chemical and biological sciences to predict the relationship between features encoding the molecular structure of compounds and the target property or biological activity under study. QSAR models are extensively used for virtual screening and prediction of categorical properties of drug candidates [77].

The development of QSAR models generally involves dealing with high-dimensional data representations. Drug candidates can be described by a large number of features or descriptors, which encode structural properties of the molecules. Feature selection is usually applied prior to the development of a QSAR model in order to harness high dimensionality [16, 59], and mostly because traditional machine learning techniques do not perform properly in this high-dimensional scenario [22]. However, considering the large variety of possi-

ble descriptors that can be calculated from compounds, feature selection represents a difficult combinatorial problem that may neglect valuable information.

Another important aspect of QSAR modeling is determining the reliability of predictions on unseen compounds. The Applicability Domain (AD) of a QSAR model is the molecular subspace where predictions performed by the model are expected to be accurate. AD analysis is a significant step in the process of building a reliable QSAR model [65] and the identification of the AD of a QSAR model remains a current matter of research [36, 32]. A last important aspect of QSAR modeling is interpretability, as such models are managed by medicinal chemists in the process of searching for drug candidates. Being able to gain insight into the features that are the most relevant for prediction is valuable for experts [69], as such interpretation makes it possible to understand the molecular substructures that play a significant role in the biological activity or property of the chemical compound.

Regarding the techniques that have been used for QSAR modeling, two of them stand out. On the one hand, the use of meta-classifiers and consensus approaches has been widespread in QSAR modeling, and these methods have become the state of the art for predicting several physicochemical properties and biological activities [34, 6]. On the other hand, artificial neural networks—a bio-inspired technique [15]—have also been used for QSAR modeling, although their adoption has been criticized due to their lack of generalization and the difficulty in the interpretation of such models in physicochemical terms [7]. Moreover, in recent years QSAR modeling has witnessed the advent of deep learning, which has brought several advantages as well as challenges. Recent advances in Deep Neural Network (DNN) approaches have made neural models less prone to overfitting, and hence more likely to be applied successfully for predicting unseen compounds. In addition, DNN-based models have been found effective for solving large-scale and high-dimensional data analysis problems [8].

The goals of this work are to build QSAR models that incorporate recent advances in deep neural networks for prediction of three relevant properties in biomedical sciences, and to benchmark these models against the state of the

art. In addition, we aim to explore the potential of applying confidence estimation to neural-based models as an effective way for AD assessment, as well as to study the possibility to interpret these models in terms of the molecular features used to represent the chemical data. In order to address these goals, we propose the development of neural-based QSAR models for bioactivity prediction of three different properties. On these models, we evaluate the network output probabilities as a means of performing an AD estimation. In addition, we provide a post-hoc analysis of the most relevant features for each property. The first two properties are Cytochrome P450-drug interaction for isoforms 2C9 and 3A4, which are a family of enzymes involved in the oxidation of compounds. These two isoforms are particularly relevant to drug metabolism. It has been proven that inhibition of CYP enzymes leads to adverse side effects of drug-drug interactions [10], and hence the study of CYP interactions has become of major interest in the fields of drug discovery. The third property is Ready Biodegradability (RB). Biodegradation is highly relevant to biomedical sciences, since the presence of certain substances persisting over an extended period of time has been linked to major health risks, such as cancers, neurological dysfunction and hormonal changes [62, 56]. Biodegradation properties are also relevant in the design of polymeric materials used for biomedical purposes [57]. Therefore, predicting biodegradability properties on chemicals represents a critical aspect for several biomedical areas.

The contributions of this work can be summarized as follows:

- We applied recent advances in deep neural networks to the development of neural-based QSAR models and obtained higher performance compared to state-of-the-art models, while at the same time overcoming the need for a potentially detrimental feature selection phase.

- We proved the effectiveness of using network output probabilities to perform AD estimation, which represents an advancement over consensus-based AD models that merely provide a binary signal with regard to inclusion or exclusion in the applicability domain.

- We applied a *post hoc* interpretability method based on an analysis of the network weights that has never been applied before in the context of QSAR models. We presented the results by means of a novel visualization based on heat maps, and proved that it effectively allows to gain insight into the interpretability of the proposed neural-based models.

- Our models outperformed the current state of the art for three different properties of high relevance in biomedical sciences.

This paper is organized as follows: in Section 2 we conduct a survey on relevant articles in the area and how they relate to our work. In Section 3, the datasets used for our experiments as well as the proposed methods are detailed. We present the results obtained for our proposed models and discuss their implications in Sections 4 and 5. Finally, conclusions and future lines of work are presented in Section 6.

2. Related Work

For the past decades the process of rational drug design has relied on computer modeling techniques, and various *in silico* methods have been widely applied with the aim of both speeding up the discovery process and reduce costs [28, 73, 26]. Traditional techniques, such as Support Vector Machines (SVM), Decision Trees, Naïve Bayes and k-Nearest Neighbors, have been extensively used for building QSAR models because of their relatively good performance and simplicity [74, 3, 47]. Recently, there has been a strong tendency to consensus-based approaches, which consist in assembling different base classifiers to combine their predictions and, as a consequence, increase the prediction capabilities of the model [38, 17, 39, 2, 67]. This type of models are typically among the top performing techniques for the prediction of several chemical properties in QSAR modeling, but at the expense of limited interpretability. Besides, they are constrained by their base models, which usually rely on a feature selection step in order to perform at their best [23, 41], and they are normally not

able to capture complex relationships between descriptors or rule out redundant information [63, 38].

2.1. Prediction techniques for Cytochrome P450-drug interaction and Ready Biodegradability

Extensive research work has been carried out for predicting Cytochrome P450-drug interaction, and the majority of these works usually involve a feature selection process [31, 13, 68]. Jensen et al. [31] presented two Gaussian kernel weighted k-Nearest Neighbors models. It was the first work to incorporate the use of Extended Connectivity Fingerprints (ECFP) and Functional Class Fingerprints (FCFP) [61] as features for CYP2D6 and CYP3A4 inhibition prediction. Cheng et al. [13] developed consensus-based models for prediction of five different CYP isoforms, using SVM, C4.5 Decision Tree, k-Nearest Neighbors and Naïve Bayes as base classifiers, combined by a backpropagation artificial neural network. they also showed an AD estimation that improves prediction accuracy. More recently, Shah et al. [68] developed a joint QSAR model based on feed-forward multi-layer neural networks for prediction of drug metabolism of isoforms 3A4, 2C9 and 2D6 of Cytochrome P450. Fingerprints were used as input features, and the three biological activities were embedded in a multitask deep neural network. Nembri et al. [54] developed two consensus-based models for prediction of isoforms 3A4 and 2C9 inhibition, and also performed an AD analysis. Both of the reported models were constructed upon two different voting approaches and used variations of k-Nearest Neighbors and a classification tree as base classifiers. Each base classifier was constructed employing either ECFP or a small number of molecular descriptors obtained from a two-phase feature selection process. The best performing model reported for isoforms 2C9 and 3A4 was one of the voting approaches (namely *Consensus 1*). Since the work by Nembri et al. [54] reports one of the best prediction performances of biological activity for Cytochrome P450 and also due to the provision of all the data necessary for reproducibility, it constitutes the reference research work on CYP inhibition that we use for comparison.

There are also several research studies for prediction of both Biodegradability and Ready Biodegradability of compounds [46, 18, 11, 5, 36, 21], where two main approaches stand out: consensus and neural-based models. Consensus models are predominant for the prediction of this property, where feature selection techniques are applied in most cases [46, 18, 5, 50]. Involving neural-based techniques, Goh et al. [21] developed a multimodal architecture for biodegradability prediction combining a Convolutional Neural Network with a fully-connected multi-layer perceptron, and using both domain-specific hand-engineered features and learned representations from raw data. In Mansouri et al. [46], two consensus models were proposed for prediction of Ready Biodegradability of compounds over three different base classifiers: Partial least Squares Discriminant Analysis (PLS-DA), SVM and k-Nearest Neighbors. The proposed consensus models were based on two different voting approaches and an AD analysis was carried out on the developed QSAR models. The best voting approach *Consensus 2* is reported as the best predictive model, which represents the best classification performance compared to other published QSAR models on biodegradation, and thus we chose this work Mansouri et al. [46] as our reference method for comparison.

Deep learning has emerged in the last years as a widely used soft-computing technique for the development of QSAR models and other areas in drug discovery research, and it has established itself as the state-of-the-art prediction technique [12, 19]. Although artificial neural networks have already been used for QSAR models in the past [79, 26], there is a recent tendency to adopt new strategies for training neural-based models, such as the application of novel techniques for avoiding overfitting and vanishing/exploding gradients during training. Although the application of deep learning in QSAR modeling is still in its beginnings, several research studies have developed deep learning-based models for various drug discovery problems successfully [43, 40, 37]. In Ma et al. [43], models based on DNNs achieved higher prediction performance than Random Forest on a group of large and diverse QSAR datasets. Lenselink et al. [40] compared five different techniques over a ChEMBL bioactivity benchmark set

and found that DNNs outperformed traditional methods. They also showed that an ensemble of DNNs with additional tuning further improves the performance obtained by more simple DNN-based models. Koutsoukas et al. [37] showed that DNN-based models statistically outperform models based on traditional methods, such as Naïve Bayes, k-Nearest Neighbors, SVM and Random Forest over diverse datasets.

2.2. Applicability domain and interpretability of prediction models

The determination of the applicability domain of a QSAR model is a crucial aspect of the modeling process, since it allows to determine the molecular subspace of compounds where the QSAR model is expected to make reliable predictions [36, 55]. Most research articles in the area address AD determination using a standalone method, where different strategies and statistical measures are adopted to determine AD boundaries [36]. A good number of them focus on defining different molecular similarity criteria for identifying outliers, which are then excluded from the AD of the model [64, 42, 9]. Klingspohn et al. [36] performed a comprehensive study in order to define a taxonomy of AD methods and find the best approaches for estimating the AD of different classification methods. In this article, two main categories of techniques for determining the AD were identified and compared: those based on novelty detection (identification of outliers) and those based on confidence estimation (inferred from the trained classifier). Experiments using six different binary classification techniques on ten datasets were performed. It was concluded that AD measures based on confidence estimation consistently perform better than novelty detection techniques, and thus they are suitable approaches for defining the AD.

Since QSAR models are meant to assist experts during drug discovery, their results should be as interpretable as possible [58]. Consensus-based approaches, in spite of having good predictive performance, tend to lack interpretability since their output result is a combination of different base classifiers. Interpretability of neural network-based models has been studied for several years within the machine learning community [70, 76, 33] and it also remains a matter of research

in drug discovery. Approaches based on *post hoc* interpretability have been explored recently [52, 4, 60]. This type of techniques takes a trained model and makes an attempt to understand predictions in terms of the features used by the model. As opposed to a low-level algorithmic comprehension of the model, which is the most usual approach taken for interpretability analysis, *post hoc* techniques aim to characterize the behavior of the predictive model without attempting to explain its internal representation and operations, but providing a functional understanding of it in terms of its features. In this line of work, Tsang et al. [75] developed a framework to discover statistical interactions between the input features in a feed-forward multi-layer neural network, by direct interpretation of its learned weights. Their method proved to be effective on both synthetic and real-world application datasets, and thus we chose it as our reference paper for post hoc feature analysis.

2.3. Our proposal

Based on the observed limitations and the current state of the art, we propose the development of neural-based methods to model three physicochemical properties, namely: Cytochrome P450-drug interaction for isoforms CYP2C9 and CYP3A4, and Ready Biodegradability. We compare our approach with the consensus models that, to the best of our knowledge, represent the top-performing models that have been published for the prediction of these properties. We present an embedded applicability domain technique, which is derived from our trained models. This approach would be categorized as prediction confidence estimation according to Klingspohn's taxonomy [36]. We propose the application of a *post hoc* interpretability technique based on an aggregative analysis of the weight contributions of the network, which is based on Tsang et al. [75]. To the best of our knowledge, this method has not yet been employed for interpretability of QSAR models. Additionally, the results of this *post hoc* analysis are summarized using a novel visualization based on heat maps. Finally, our models and results are contrasted to those reported in Nembri et al. [54] and Mansouri et al. [46]. The reason for choosing these latter articles as

our baselines for comparison is due to the possibility of reproducing their data and experiments and the high performance attained in the reported results.

The entire workflow of our approach is depicted in Figure 1. First, we pre-processed the three datasets under study (a). Second, we enriched the Original datasets by adding new molecular descriptors (b) and we split the Enriched and Original datasets into the partitions for training and validating our models (c) . Then, we developed our neural-based models by performing an iterative process for hyperparameter tuning and we train the chosen models (d). Next, we evaluated their performance using several metrics and we contrasted these results with different baselines (e). Finally, we developed an AD model based on confidence estimation and applied (f) a *post hoc* feature analysis method, which allows to determine the most influential features to our models (g).

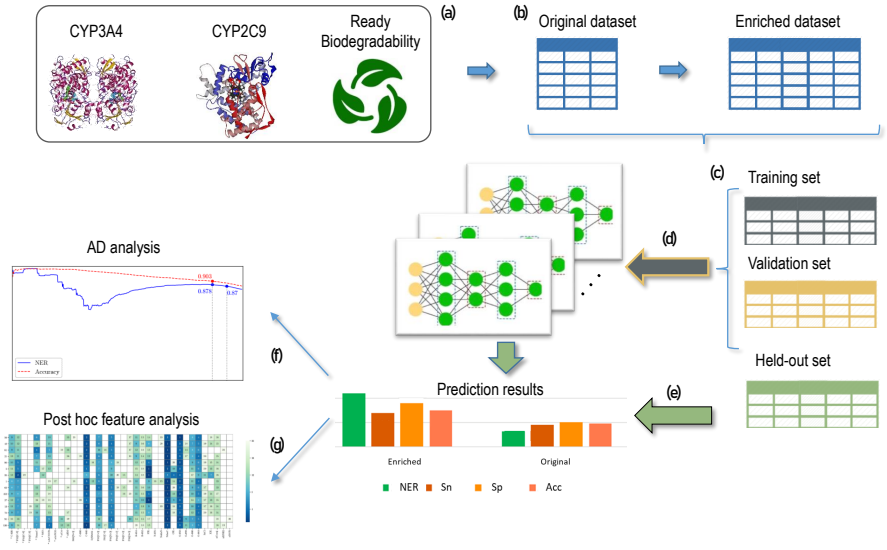


Figure 1: Depiction of the entire workflow of our method: (a) data preprocessing, (b) dataset enrichment, (c) dataset partition, (d) development and training of models, (e) evaluation of models, (f) AD analysis, (g) *post hoc* feature analysis.

3. Materials and Methods

In this section, we provide an overall description of all the techniques used in our work, as well as the preparation of the datasets and the selection of the hyperparameters of the model.

We say that we *enrich* a dataset when we extend its set of features by including new molecular descriptors not previously considered in the Original dataset (Figure 1–b). It is worth noting that the included descriptors are members of the family of descriptors already included in the Original datasets. The reported partitions of the datasets in Train, Validation and Held-out sets[1] were kept the same during the construction of our models (Figure 1–c). For the sake of completeness, we also trained our *Best_E* models using 5-fold Cross Validation. The details of this process and its results are summarized in the Supplementary Material.

All of our models are based on feed-forward multi-layer neural networks. The architecture for each model varies depending on the number of input features or molecular descriptors. As a general approach, larger inputs demand more complex architectures, so the Enriched versions of the datasets yielded models with more nodes than those built for the Original versions of the datasets. Our neural-based models were obtained following a two-phase process (Figure 1–d). The first one consisted in an exploratory phase, where different architectures and optimization strategies were considered. In this phase we developed prototypes and tuned their parameters. The need for an exploratory phase when constructing neural networks has been reported previously in the context of QSAR modeling [81]. The second phase consisted in selecting the best prototype from the first phase. This selection was performed by assessing classification performance on the Validation set. Due to the inherent stochasticity of neural networks, we repeated the training process using a set of fifteen random

[1] Note that the Original datasets refer to these partitions in their papers as *Training*, *Test* and *External Validation*, respectively.

seeds for each dataset. As a result of this two-phase process we obtained fifteen models with the same hyperparameters but initialized differently. The average performance of these fifteen models is reported as the *Average* model, whereas we report as the *Best* model the one that performs the best on the Validation set.

After the best model for each Enriched dataset was obtained, we built a new model for the Original version of each dataset—namely *Best_O*—using the same hyperparameters as for the Enriched models but using less nodes per layer. The construction of two models, one based on the Enriched dataset and another based on the Original dataset, enable us to compare the performance of our proposed strategy over one same set of compounds with and without the application of a feature selection process, and it allows to analyze the potential of our approach on high-dimensional datasets.

3.1. Datasets

The three datasets used in this work, namely CYP2C9, CYP3A4 [54] and RB [46], are publicly available and were selected taking into account their relevance in QSAR modeling in the context of biomedical data analysis. We made datasets CYP2C9, CYP3A4 and RB in their Enriched versions publicly available[2]. Further information on the calculation of molecular descriptors for the construction of the Enriched datasets can be found in the Supplementary Material.

3.1.1. CYP2C9 and CYP3A4

Datasets CYP2C9 and CYP3A4 have a total of 11940 and 12118 compounds, and the proportion between active/inactive compounds on their Training and Validation sets is 49/100 for CYP2C9 and 66/100 for CYP3A4. As for the Held-out sets, the ratios are 56/100 in CYP2C9 and 98/100 in CYP3A4. CYP2C9 and CYP3A4 share the same compounds in their Training sets as well as in

[2] `https://github.com/VirginiaSabando/DNN-QSAR-2019.git`

their Validation sets. In the Original datasets, CYP2C9 includes ten molecular descriptors, whereas CYP3A4 includes eight molecular descriptors. They also include a 1024-bit ECFP for each compound [61]. For the Enriched versions of both datasets we added a total of 2701 molecular descriptors to the CYP2C9 dataset and 2699 molecular descriptors to the dataset CYP3A4, leading to a total of 2711 and 2707 descriptors, respectively, in addition to the 1024-bit ECFP. We performed the calculation of molecular descriptors and ECFP using Dragon 7 [72].

There were a few compounds on the datasets provided by Nembri et al. [54] whose SMILES codes were not properly formed, and hence we were unable to calculate their molecular descriptors. As a result, one molecule was removed from both Training sets, six molecules were removed from the Held-out set of CYP2C9, and one molecule was removed from the Held-out set of CYP3A4.

3.1.2. Ready Biodegradability (RB)

Dataset RB comprises 1725 compounds, where the ratio between active/inactive compounds is 51/100 for the Training set, 49/100 for the Validation set and 40/100 for the Held-out set. A total of 41 molecular descriptors were provided for dataset RB in its Original form. We calculated additional molecular descriptors to obtain its Enriched version, which gave a total of 1480 molecular descriptors. We computed these descriptors using Dragon 7 [72].

3.2. Model Parameterization

In this section, we describe all the model parameters used for the three Enriched datasets. Parameters for the remaining models and their training can be found in the Supplementary Material. The models were built using Tensorflow 1.7 [1].

The input features for our predictive models were molecular descriptors and ECFP. In the cases of CYP2C9 and CYP3A4, we split the ECFP into 1024 separate bits and then considered each one of these bits as a single input feature. All nodes in the input and hidden layers of the models use rectified linear units

(ReLU) as activation function [53], and the output layer implements a softmax function, which has two nodes for each class output probability. The networks were trained using standard backpropagation [66] and we chose a minibatch size of 200 instances to train the networks. We use cross-entropy with logits as cost function and Adam optimizer [35] for minimizing it.

For weight initialization we experimented with Xavier initialization [20] and He Normal initialization [24], which are both considered state-of-the-art initialization techniques [51, 25]. We also applied Batch Normalization [29] in all layers of our networks for faster training and to avoid exploding/vanishing gradients. In order to avoid overfitting, several regularization techniques were used in each model. We used Dropout [71] with varying dropout rates according to the number of nodes in each layer, and also implemented L2-regularization [27] varying the penalization coefficient λ in each model. We applied early stopping to avoid overtraining the models, hence helping to prevent possible overfitting.

The QSAR model obtained for Enriched dataset CYP2C9 is a feed-forward multi-layer neural network architecture consisting of one input layer of 3735 nodes (for 2711 molecular descriptors plus 1024 bits from ECFP) and three hidden layers of 50, 20 and 5 nodes, respectively. Batch Normalization was applied with a decay value of 0.9 to prevent the weights from growing too large, and we used a learning rate value of 0.00001. We initialized the network weights by using Xavier initialization. A penalization coefficient $\lambda = 0.0001$ was used for L2-regularization. In order to deal with class imbalance we optimized a weighted cost function, which penalized mispredicted instances from the least popular class by increasing the loss by a factor proportional to the class imbalance observed in the Training set.

The architecture of the QSAR model that we developed for Enriched dataset CYP3A4 is similar to that used for CYP2C9, with the difference that the input layer consists of 3731 nodes (for 2707 molecular descriptors plus 1024 bits from ECFP). As in the case of dataset CYP2C9, all layers implement Batch Normalization with a decay value of 0.9. For CYP3A4, we initialized network weights by applying He Normal initialization, and the same regularization criteria than

for CYP2C9 was taken into account for this dataset. For class imbalance mitigation we applied a stratified sampling technique, where an equal number of compounds belonging to each class was drawn to build each of the minibatches during training. The compounds were sampled with replacement from the training set and randomly shuffled before they were fed to the network during the training phase.

Lastly, we developed a QSAR model for Enriched dataset RB based on a less dense feed-forward multi-layer neural network architecture, considering that the input features were fewer than those in the previously described models. The input layer comprises 1480 nodes for molecular descriptors, and the network is also made of three hidden layers of 20, 10 and 5 nodes, respectively. All layers implement Batch Normalization with a decay value of 0.9, as in the case of the previous models. We initialized the network weights by using Xavier initialization, and as regularization techniques we used Dropout and L2-regularization with a penalization coefficient $\lambda = 0.001$. The learning rate was set to 0.0001. We used a stratified sampling technique in order to counteract class imbalance, with the same sampling technique as described in the case of CYP3A4.

3.3. Applicability Domain

The applicability domain (AD) of a QSAR model is the molecular subspace in which the predictions made by the model are expected to be accurate [30, 78]. In other words, the definition of an AD allows the expert to determine whether a prediction on a new compound is likely to be reliable or not.

We propose using class probability provided by the output layer of our models to estimate their AD. This leads to AD models which are embedded into the prediction models. The embedded AD models were evaluated as follows. First, we computed class probability values using the network softmax layer for every compound. Then, we sorted these values in decreasing order to elaborate a ranking. Finally, in order to evaluate the goodness of the ranking of confident predictions, we computed Mean Average Precision (MAP) [45], where several metrics (Accuracy, *NER*, etc) were calculated on the k-highest ranked

compounds, where k is varied from 1 to n, and n being the total number of compounds. All these different metrics computed for different number of compounds are averaged. This AD approach (Figure 1–f) allows to evaluate the performance of the models at any desired threshold of membership to the AD.

3.4. Post hoc *feature analysis*

As QSAR models are tools for the benefit of chemists and drug developers alike, it remains an important asset to provide means of interpretability for any proposed models [49, 58]. For domain specialists it is useful to know the features that make a particular family of compounds to show some degree of activity regarding a property of interest, since this allows to reduce the search space during drug discovery.

We propose a *post hoc* feature analysis technique as a way of providing interpretability to our neural-based models, so that domain experts can determine the most relevant molecular descriptors in the context of a prediction model (Figure 1–g). By analyzing the network weight contributions in an aggregative manner, it is possible to gain insight into the descriptors that are more influential on the predicted target value.

The proposed *post hoc* feature analysis technique is described as follows: once training is completed, we calculate a score for every descriptor by taking into account the sequence of contributions from the input to the output nodes. For a given feature, these contributions are calculated by aggregating the weights of the neural model that are connected to this feature. More formally, for any layer, the score of a node j is computed using

$$S(n_j) = \frac{1}{k} \sum_{i=1}^{k} |w_{j,i}| \, S(n_i), \tag{1}$$

which is the average of the k products between the weights connecting node j with the k nodes in the following layer and their scores. Given that this is a recursive definition, by setting the score corresponding to the node of the output layer to 1, we can compute all node scores by starting from the output nodes and going backwards until the scores corresponding to the input nodes are

computed. We considered the absolute values of the weights, as a way to analyze quantitative impact of the descriptors on the output, regardless of whether they contribute in a positive or negative manner on the result. The rationale is that input features exhibiting high scores are likely to be more relevant than those showing low scores, as slight changes in their values would have greater impact on the outcome of the network.

4. Results

We performed an evaluation of each of our models by comparing them against *Consensus 1* and *Consensus 2*, which are the top-performing methods ever reported for the three datasets under study [54, 46] (Figure 1–e). To account for a fair comparison, we used the same metrics as reported in Nembri et al. [54], Mansouri et al. [46], i.e., Sensitivity (Sn), Specificity (Sp) and NER, as well as Accuracy (Acc) and MAP, as described in Section 3.3. Sensitivity and Specificity quantify the accuracy in predicting the active and inactive class, respectively, while NER is the arithmetic mean of Sn and Sp. Additionally, other performance metrics can be found in the Supplementary Material.

Since our neural-based models are inherently stochastic, fifteen different trials were run to train and test the models, each one using a different random seed. Therefore, we report both the average performance of the developed models, i.e., taking into account all trials, and the best performance in terms of *NER*, i.e., the model obtained from the best seed.

Regarding the AD analysis, we report both *NER* and the percentage of compounds that are not within the AD—which are referred to as *Not Assigned* compounds ($\%na$)—as it was also done by Nembri et al. [54] and Mansouri et al. [46]. We report *NER* when $\%na$ is fixed to the value of the best reported method in the referenced articles. Similarly, we report $\%na$ when *NER* is fixed to the same values reported in the referenced papers.

4.1. CYP2C9

The results for the Original and Enriched versions of CYP2C9 are presented in Table 1. We also include the results of the best model reported by Nembri et al. [54], i.e., *Consensus 1*. It can be seen that for both Validation and Held-out sets our best Enriched model, namely *Best_E*, performs better than the best model on the Original dataset, namely *Best_O*, and both of them outperform *Consensus 1*. In addition, all of our models achieve equivalent or better *NER* values than *Consensus 1* when keeping $\%na$ at the same value.

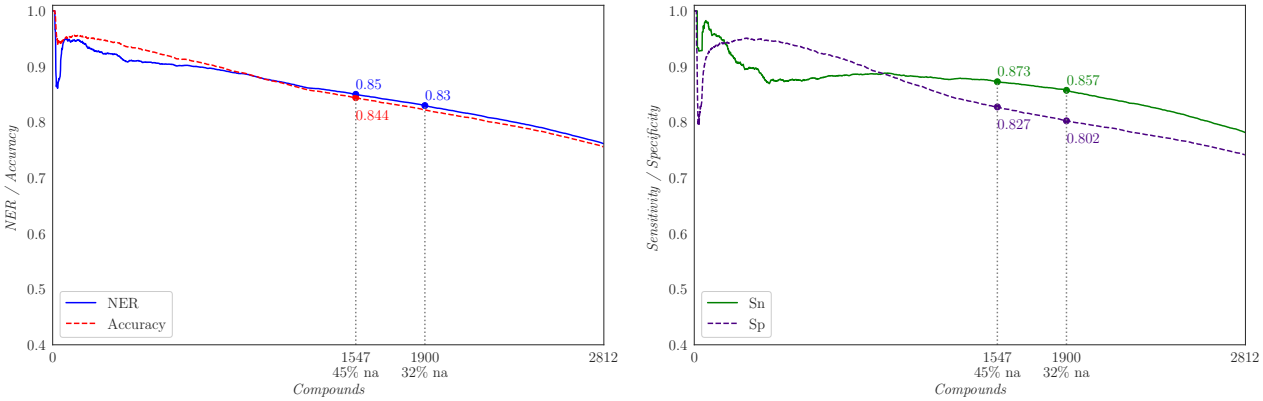
CYP2C9		Validation set				Held-out set			
		NER	Sn	Sp	Acc	NER	Sn	Sp	Acc
Original	Consensus 1	0.89	0.89	0.88	-	0.83	0.85	0.82	-
	Average_O	0.91	0.90	0.92	0.91	0.83	0.79	0.88	0.86
	Best_O	0.92	0.92	0.92	0.92	0.85	0.82	0.87	0.86
Enriched	Average_E	0.92	0.93	0.91	0.92	0.85	0.87	0.83	0.84
	Best_E	**0.93**	**0.94**	**0.93**	**0.93**	**0.87**	**0.89**	**0.86**	**0.87**

Table 1: Results on the Validation and Held-out sets of CYP2C9. Consistently with the results reported by Nembri et al. [54] the percentage of not assigned compounds ($\%na$) was set to 40% for Validation set, and 45% for Held-out set.

A more comprehensive evaluation of the performance of our models on the CYP2C9 Held-out set is presented in Figures 2 and 3. These figures show different performance measures when different cutoff values for the AD are considered. In Figure 2 we present the mean of all trials—i.e., *Average* performance, whereas in Figure 3 the results for the best trial are presented. Both figures correspond to the models trained on the Enriched version of CYP2C9. The horizontal axis represents the number of compounds sorted by class probability, so that the left-most compounds are the most confidently predicted ones. The vertical axis represents different performance measures evaluated over the set.

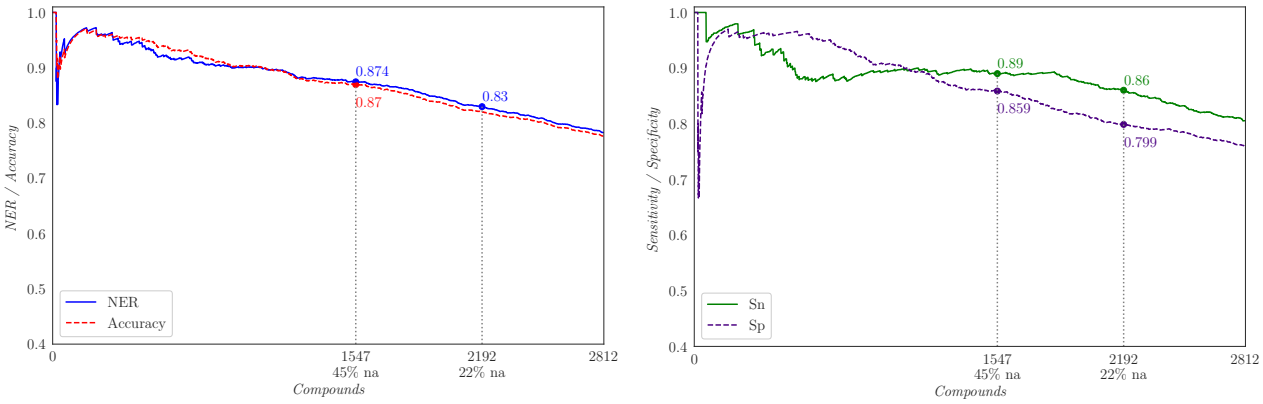
By looking at these plots it is possible to set any cutoff point in the horizontal axis in order to evaluate performance when the compounds with least certain

prediction—those to the right of the cutoff point—are discarded. In particular, two cutoff points are marked as noteworthy; these are the cutoff values where $\%na$ and *NER* match with the ones reported by Nembri et al. [54]. MAP results for the Validation sets of the three datasets can be found in the Supplementary Material.



(a) (b)

Figure 2: Average MAP performance of all trials on the Held-out set of Enriched CYP2C9. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.



(a) (b)

Figure 3: MAP Performance of the best trial on the Held-out set of Enriched CYP2C9. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.

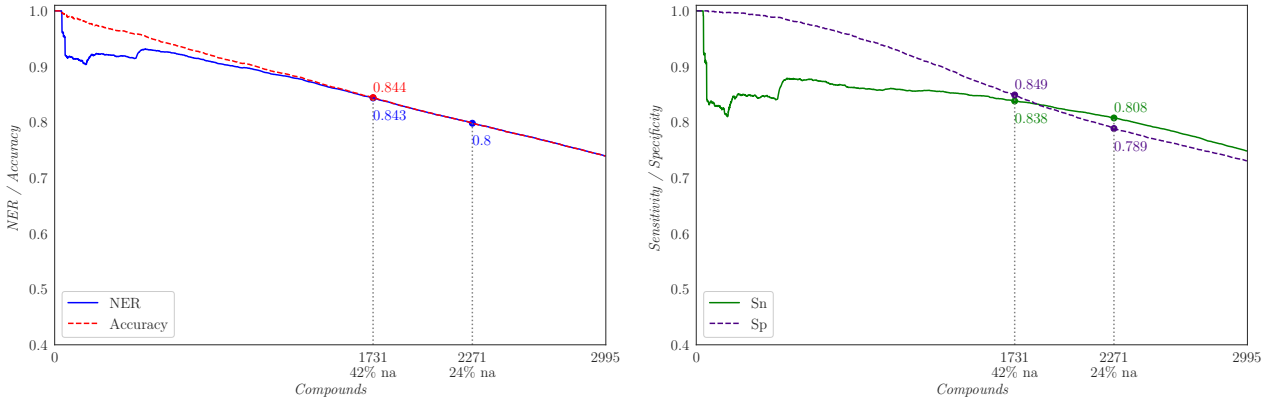
4.2. CYP3A4

We present the results for CYP3A4 in Table 2. For both Validation and Held-out sets our best Enriched model show better performance than the best model trained on the Original version of the dataset, which in turn overcomes the results reported for *Consensus 1*. All of our models obtain higher Non-Error Rate for the same number of discarded compounds than the reference model, yet ours exhibiting balanced values of Sensitivity and Specificity.

CYP3A4		Validation set				Held-out set			
		NER	Sn	Sp	Acc	NER	Sn	Sp	Acc
Original	Consensus 1	0.88	0.92	0.83	-	0.80	0.89	0.70	-
	Average_O	0.89	0.83	0.94	0.91	0.82	0.76	0.88	0.83
	Best_O	0.91	0.91	0.91	0.91	0.83	0.84	0.82	0.83
Enriched	Average_E	0.92	0.89	0.94	0.92	0.84	0.84	0.85	0.84
	Best_E	**0.93**	**0.91**	**0.94**	**0.93**	**0.85**	**0.86**	**0.84**	**0.85**

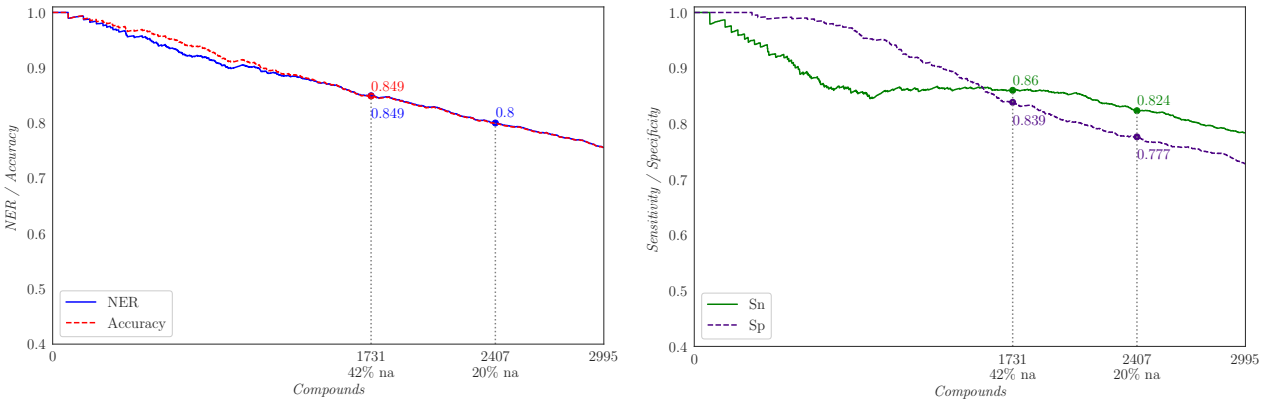
Table 2: Results on the Validation and Held-out sets of CYP3A4. Consistently with the results reported by Nembri et al. [54] the percentage of not assigned compounds (%*na*) was set to 36% for Validation set, and 42% for Held-out set.

The plots showing the comprehensive performance of the QSAR models with regard to its AD model for CYP3A4 Held-out set are presented in Figures 4 and 5. Similarly as it was done for CYP2C9, Figure 4 shows the results for the mean of all of our trials, while Figure 5 reports the results when our best model is considered.



(a) (b)

Figure 4: Average MAP performance of all trials on the Held-out set of Enriched CYP3A4. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.



(a) (b)

Figure 5: MAP Performance of the best trial on the Held-out set of Enriched CYP3A4. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.

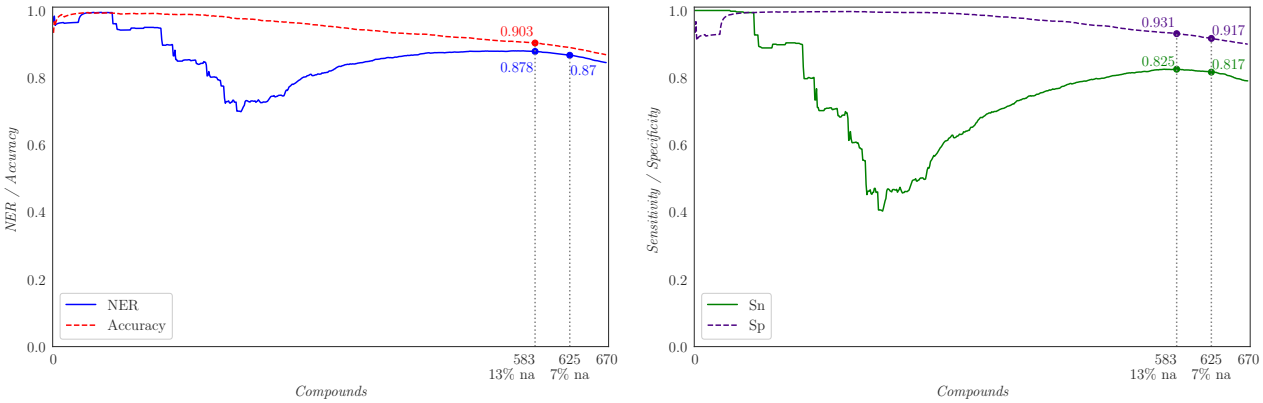
4.3. Ready Biodegradability

Table 3 shows the results for RB. Our best Enriched model, i.e., *Best_E*, exhibits higher *NER* in both Validation and Held-out sets than *Best_O*, our best model trained on the Original dataset. Both of these models show higher *NER* values than for *Consensus 2*.

Ready Biodegradability (RB)		Validation set				Held-out set			
		NER	Sn	Sp	Acc	NER	Sn	Sp	Acc
Original	Consensus 2	0.91	0.88	0.94	-	0.87	0.81	0.94	-
	Average_O	0.92	0.94	0.90	0.91	0.88	0.85	0.91	0.90
	Best_O	0.91	0.91	0.91	0.91	0.88	0.85	0.92	0.90
Enriched	Average_E	0.94	0.93	0.90	0.94	0.88	0.83	0.93	0.90
	Best_E	**0.94**	**0.95**	**0.92**	**0.93**	**0.89**	**0.85**	**0.93**	**0.91**

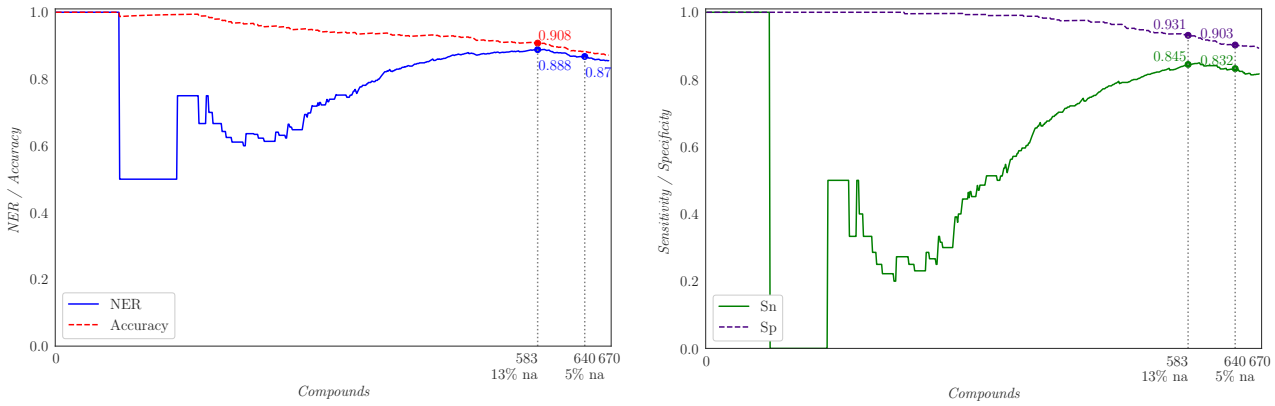
Table 3: Results on the Validation and Held-out sets of Ready Biodegradability (RB). Consistently with the results reported by Mansouri et al. [46] the percentage of not assigned compounds (%*na*) was set to 15% for Validation set, and 13% for Held-out set.

The plots displaying the performance of the QSAR models with regard to its AD model for the Enriched model on the RB Held-out set are presented for the mean of all of our trials (Figure 6), and for the best trial, i.e., *Best_E* (Figure 7).



(a) (b)

Figure 6: Average MAP performance of all trials on the Held-out set of Enriched RB. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.



(a) (b)

Figure 7: MAP Performance of the best trial on the Held-out set of Enriched RB. (a) *NER* and *Accuracy* are shown. (b) *Sensitivity* and *Specificity* are shown.

4.4. Post hoc Feature Analysis

We performed a *post hoc* feature analysis in order to gain insight on which molecular descriptors are the most relevant to our models. We propose a novel visualization for summarizing main patterns of features that were found to be relevant across multiple trials by means of a heat map. The heat maps that encode the results corresponding to the feature analysis for datasets CYP2C9, CYP3A4 and RB are presented in Figures 8, 9 and 10, respectively. Each row corresponds to a different trial of our model using its own seed for initialization of weights and random variables. These trials are sorted by performance, where the top row represents the best trial. Each column on the maps represents molecular descriptors, where only the 20 most relevant descriptors of each model according to our measure were considered. Descriptors marked with an asterisk are also part of the Original version of the dataset. Descriptor names starting with 'ECFP' represent Extended Connectivity Fingerprint fragments, which are followed by a number that represents the location of the bit that was identified by our method as relevant. The rank that a descriptor occupies in the relevance order of a particular trial is encoded with the number inside the

corresponding heat map cell. Likewise, darker colors are applied to cells depicting higher relevance descriptors in a specific trial, whereas lighter colors apply to less relevant descriptors.

5. Discussion

In this section, we review the results presented previously in Section 4. We discuss the performance of the models, as well as the results of our embedded AD model and *post hoc* feature analysis technique.

5.1. Neural-based Classifiers versus Consensus-based Classifiers

As it is shown in Table 1, our models for prediction of CYP2C9 drug interaction outperform the results reported by Nembri et al. [54]. The *NER* values of *Average_E* and *Average_O* are consistently superior to the reference results. On the internal Validation set of CYP2C9, the average *Sn* and *Sp* values in both the Original and the Enriched version of the dataset are higher and more balanced than those achieved by *Consensus 1*, which indicate that our model has successfully overcome the class imbalance of the dataset as it was able to correctly predict both active and inactive compounds with similar accuracy. When evaluated on the Held-out set of CYP2C9, our models also improve the performance of the reference results, although the differences between our models and *Consensus 1* are smaller than those obtained on the Validation set. Cohesively, a mild imbalance between *Sn* and *Sp* is observed, which is consistent with the results reported by Nembri et al. [54] on the Held-out set. The best trial on the Enriched version of the CYP2C9 dataset, i.e.,*Best_E*, attained a *NER* value of 0.93 for the internal validation set and a value of 0.87 when tested on the Held-out set, which is an improvement of 0.04 over the same results for *Consensus 1.*

In Table 2 the predictive performance of our models exceed the results reported using *Consensus 1.* Similarly to what it was observed for CYP2C9, the average *NER* of our models is higher than that reported for *Consensus 1* in

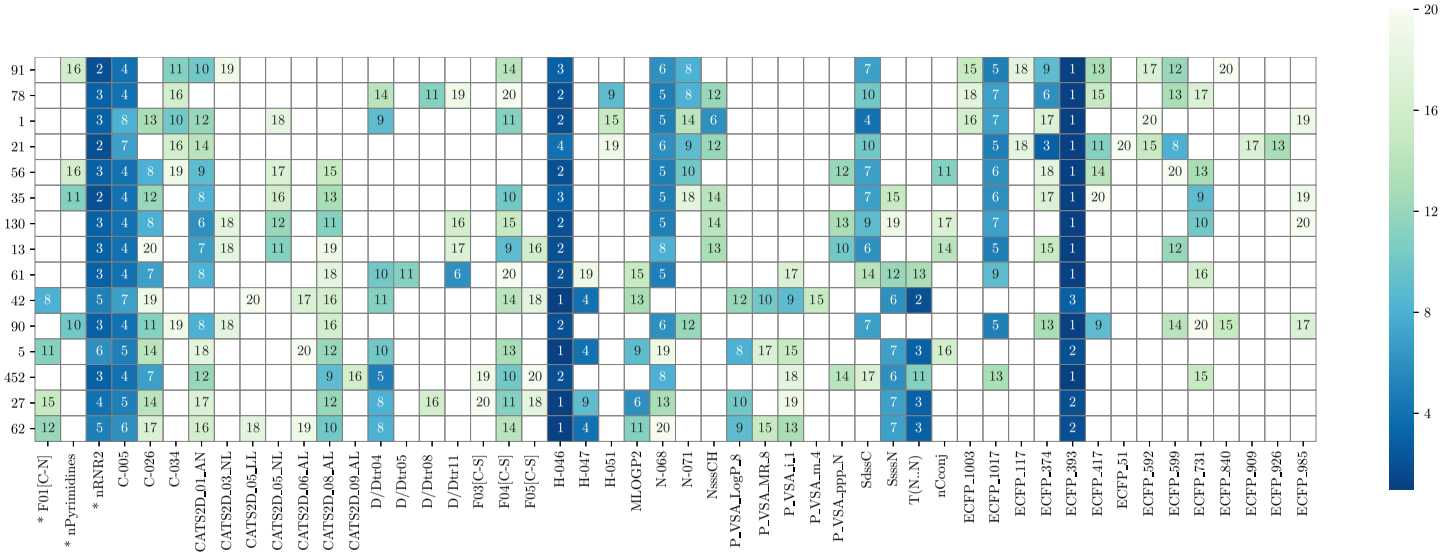


Figure 8: *Post hoc* feature analysis on CYP2C9. The rows represent different trials of our model sorted by performance, and the columns represent the 20 most relevant molecular descriptors. Both the color and the value in the cells represent the rank of a descriptor in terms of its relevance for a particular trial.

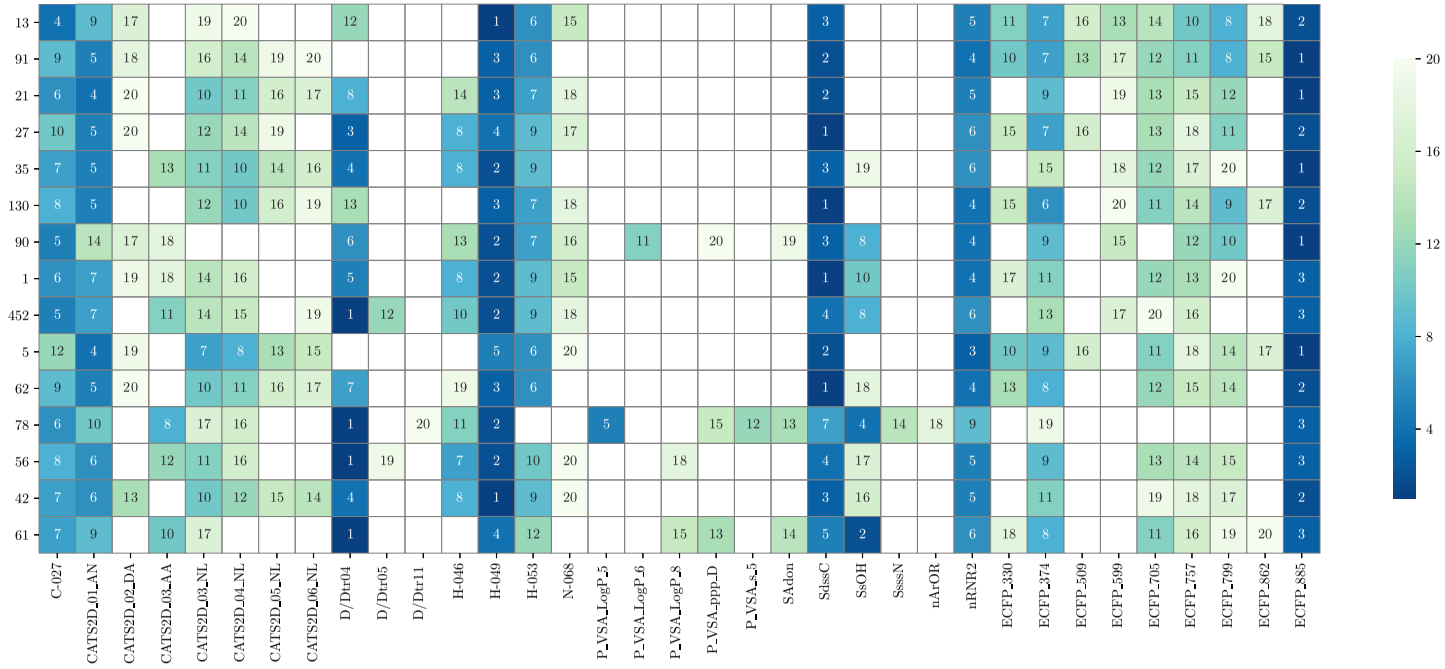


Figure 9: *Post hoc* feature analysis on CYP3A4. The rows represent different trials of our model sorted by performance, and the columns represent the 20 most relevant molecular descriptors. Both the color and the value in the cells represent the rank of a descriptor in terms of its relevance for a particular trial.

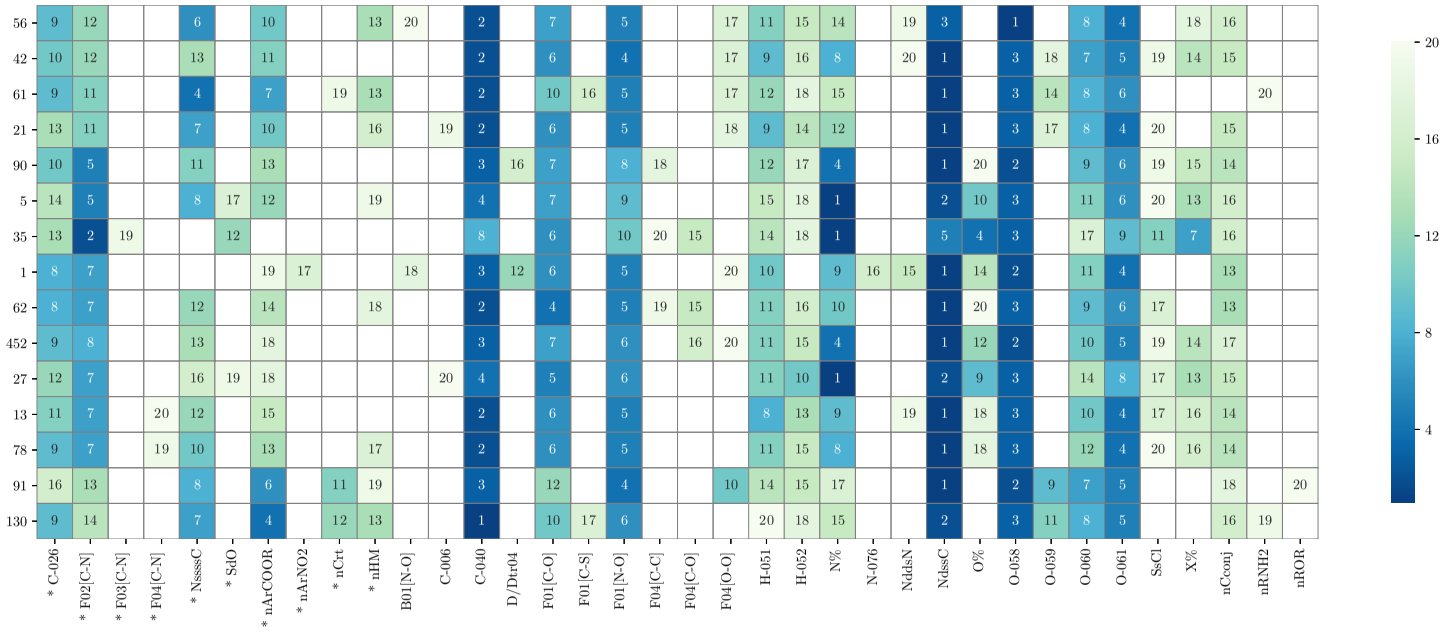


Figure 10: *Post hoc* feature analysis on Ready biodegradability (RB). The rows represent different trials of our model sorted by performance, and the columns represent the 20 most relevant molecular descriptors. Both the color and the value in the cells represent the rank of a descriptor in terms of its relevance for a particular trial.

the Original and Enriched versions of the dataset, while also showing balanced results between *Sn* and *Sp* average values. On the Enriched Held-out set, while the predictive performance slightly decreases compared to the results on the Validation set, balanced results between *Sn* and *Sp* values are obtained. This observation does not hold for the average results in the Original version of the dataset. It is worth noticing that average results were computed by taking into consideration all trials of the model, including those undertrained due to the model apparently getting caught on local minima. In a similar way as it happened for CYP2C9, *Best_E* obtained the highest *NER* for both the internal validation set and Held-out set—0.93 and 0.85, respectively—, which represents an increase of 0.05 over the same results for *Consensus 1*.

From Table 3 we can see that the predictive performance of our models improves the results using *Consensus 2* [46] for the RB dataset. The average *NER* of our models in the Validation set is higher than that of *Consensus 2* in both the Original and Enriched versions of the dataset, and at the same time showing more balanced *Sn* and *Sp* average values. The performance of the model on the Held-out set is higher in the Original dataset than in its Enriched version. Balance between *Sn* and *Sp* values is not observed for this partition, where *Sn* is consistently lower than *Sp* in all the experiments. It is noteworthy that this imbalance is also present in *Consensus 2*, which suggests an issue with the Held-out set data that makes prediction of inactive compounds to be inaccurate when compared to results for the Validation set. Nonetheless, the best trial on the Enriched version of RB, i.e., *Best_E*, attained the highest *NER* in both Validation set and Held-out set—0.94 and 0.89, respectively.

For both datasets CYP3A4 and CYP2C9, a high consistency is observed between the results got for both Validation and Held-out sets. These results show that the obtained models have strong generalization capabilities. Besides, for both Held-out sets no large disparity is observed between *Sn* and *Sp* values of *Best_E*, which in turn implies that the proposed models are able to classify active and inactive compounds unbiasedly. At a high-level analysis of the results, the results for the three datasets show that their Enriched versions lead to models

with higher predictive performance and more balanced Specificity and Sensitivity than just using the Original versions with the descriptors chosen by means of a feature selection approach. Attaining balanced performance in a binary classification problem is a desirable quality in a QSAR model [80]. The results of the Enriched models suggest that there is relevant data encoded on molecular descriptors that were not present in the Original versions of the datasets. Consequently, our experimental results imply that neural networks are able to learn in large dimensionality scenarios, and that performing a feature selection step could lead to valuable information loss, and hence to a decrease in the predictive performance. Furthermore, our work proves that neural-based QSAR models are capable of surpassing the benchmarked consensus-based models. Therefore, the use of neural networks constitutes a strong approach for QSAR modelling.

5.2. Embedded Applicability Domain Technique

We proposed using an embedded AD model based on the class probabilities calculated by the output layer of our prediction model. This approach was applied on each QSAR model and evaluated on the Validation and Held-out sets by measuring the extent by which misclassification is correlated to the predicted class confidence.

The plots for CYP2C9, which are displayed in Figures 2 and 3, show that as the number of compounds in the AD increases, i.e., to the right on the horizontal axis, the values for all metrics tend to decay continuously, which implies that the predictive performance of the model is in fact correlated with our definition of AD.

It is fair to say that these curves are not smooth for the left-most compounds. Some peaks are observed in *NER* and *Acc* plots (Figures 2–a and 3–a), while slope fluctuations in the *Sn* and *Sp* curves are observed (Figures 2–b and 3–b). For the best trial *Best_E*, as it can be seen from Figure 3, a strong downward peak is observed on the *Sp* curve for the left-most compounds. This was caused by a few inactive compounds that were misclassified with a high output probability, which is clearly an unexpected result. It is worth noticing, however,

that the curve fluctuation stabilizes shortly after this peak is observed. Figure 2 shows that in average all trials exhibited similar behaviour to the best trial, *Best_E*. Moreover, irregularities on the slope of *Sn* and *Sp* for *Average_E* can be explained because of a few trials performing quite poorly compared to the majority of the trials. This issue can be examined in more detail in the Supplementary Material.

With regard to dataset CYP3A4, and similarly to what we observed for CYP2C9, Figures 4 and 5 show the effectiveness of our embedded AD approach. From the analysis of the Average performance on the Held-out set on Figure 4, *Sn* exhibits a fair variance for the left-most compounds, as depicted in its uneven curve for *NER*, while *Sp* is smooth along the whole set. In contrast, *Best_E* exhibits a fairly smooth *NER* curve. The irregularities on the slope of *Sn* for *Average_E* can be explained by a few trials that performed slightly worse than the majority of the trials.

Regarding RB (Figures 6 and 7), *Sn* on both the average performance plot and on the best trial plot are not smooth, as it presents abrupt decays for the left-most compounds. Some peaks can be observed on *NER* curves as well, which is caused by on the *Sn* curves (Figures 6–b and 7–b). These observations imply that the model is not able to predict active compounds with high certainty correctly. Even though the curves stabilize when fewer compounds are not assigned (%*na*), the behavior of the best trial appears to be unstable, even though this same model performed outstandingly well on the Validation set. This issue, along with the average results on the Held-out set from Figure 6, indicates that the model was not able to generalize well, since it exhibits a poor performance in terms of *Sn*. As discussed above, imbalanced *Sn* and *Sp* results are also observed for *Consensus 2*, as reported by Mansouri et al. [46].

One possible cause for this issue might be that the compounds in the Held-out set are significantly different from those in the Train and Validation sets. In order to test this hypothesis, we performed a similarity analysis between all partitions of the three datasets using two distances: standardized Euclidean and cosine. The results from such analysis are presented in the Supplementary Ma-

terial, and they show that there is indeed a large average difference between the compounds in the Held-out set of the RB dataset and those in the Validation set when measuring their respective distances to the RB Train set. These average distances are in turn considerably larger than those observed for CYP2C9 and CYP3A4 datasets. The difference between these sets of compounds would explain the generalization problems of *Best_E* on RB Held-out set.

This generalization problem due to data distribution differences gets exacerbated as the dimensionality of the data increases, so the models built upon the Enriched version are disfavoured in contrast to those models for the Original version.

From the figures discussed above it is clear that our models were able to reach the same *NER* values as reported by Nembri et al. [54] and Mansouri et al. [46] for the three datasets, yet dismissing many fewer compounds as not assigned (%*na*) than the models proposed therein. For CYP2C9, a *NER* value of 0.83 with 45%*na* is reported for *Consensus 1*, while *Best_E* attains the same *NER* value dismissing only 22% of compounds on the Held-out set. In the case of CYP3A4, a *NER* value of 0.8 with 42%*na* is reported for *Consensus 1*, while *Best_E* achieves the same *NER* value with only 20%*na* on the Held-out set. Finally, on the Held-out set of RB dataset, a *NER* value of 0.87 with 13%*na* is reported for *Consensus 2*, while *Best_E* attains the same *NER* value with only 5%*na*. A similar analysis could be performed by taking into consideration %*na* reported by Nembri et al. [54] and Mansouri et al. [46] for the three datasets, since our models systematically reached higher *NER* values for the same amount of discarded compounds than *Consensus 1* and *Consensus 2* in both Validation and Held-out sets.

5.3. Post hoc Feature Analysis

From the heat maps in Figures 8, 9 and 10, one interesting aspect in all three models is that we can pinpoint molecular descriptors that were highly influential to all trials of the same model. For instance, for CYP2C9, the ECFP fragment *ECFP_393* was the most relevant feature for eleven out of the fifteen

trials, being nine of those trials among the best performing ones. Molecular descriptors *H-046* and *nRNR2* were also frequently selected in the different trials; the latter descriptor is also present in the Original CYP2C9 dataset. In the case of CYP3A4, the fragment *ECFP_885* was a highly influential feature during the training phase of the model, as it was considered among the top three most relevant descriptors in all trials. Descriptors *SdssC* and *H-049* were also signaled as relevant to the majority of trials, according to our measure. Interestingly, the molecular descriptor *D/Dtr04* was identified as an important feature occupying the first place in five trials, although these trials were among the worst trials. For dataset RB, the molecular descriptor *NdssC* is chosen as the most relevant for most models, as it was considered the most relevant descriptor for eleven out of the fifteen trials. Descriptors *O-058* and *C-040* were also signaled as important features, occupying the top three positions for the majority of trials.

Another interesting aspect that can be observed from these heat maps is that similarly performing trials tend to choose the same descriptors and in similar order of relevance. For instance, in Figure 8 descriptors are: *N-071, NsssCH, C-034* and *H-051* were spotted as relevant only by the best performing trials and occupied similar positions on the relevance rankings of every trial. Similarly, *SsssN, T(N..N), H-047, MLOGP2, F01[C-N]* and the *P_VSA* family of descriptors were deemed as influential in low-performing trials. The ECFP fragments were mostly included by the best-performing trials. The same phenomenon is observed in Figure 9: *ECFP_509, ECFP_599* and *ECFP_862* were mostly signaled by our measure in the best trials, while descriptors *D/Dtr05, D/Dtr11, SAdon, SsOH, SsssN, nArOR* and the *P_VSA* family of descriptors were found to be somewhat relevant in the low-performing trials.

Among the molecular descriptors identified as relevant for each model by our technique, there are some descriptors that are also in the Original versions of the datasets. The largest number of descriptors shared between these two sets is observed for dataset RB, where 10 out of 36 of the features signaled as the most important were also present in the Original RB dataset. Out of these 10, only 4

descriptors were highly relevant to the majority of trials, yet occupying medium-to-low importance positions in all models. Figure 8 shows that only 3 out of the 36 descriptors are present in both the Original and Enriched CYP2C9 dataset; however, as mentioned before, the descriptor *nRNR2* was identified as one of the most relevant descriptors for all of the trials by our technique. Lastly, in Figure 9 it is shown that no molecular descriptors present in the Original CYP3A4 dataset were marked as relevant for the Enriched model. It is worth noticing that both models *Consensus 1* developed for CYP2C9 and CYP3A4 datasets by Nembri et al. [54] take into account ECFP as inputs to one of their base models, hence all of the ECFP fragments are considered to be present in the Original version of these two datasets.

Taking into consideration that all of our models outperformed the reference models reported by Nembri et al. [54] and Mansouri et al. [46], while at the same time identifying relevant molecular descriptors not included in the Original datasets, it is possible to conclude that meaningful information for the model might be encoded in such molecular descriptors, and hence that relevant data could have been lost in the Original feature selection process. Furthermore, by means of this technique it was possible to identify molecular descriptors that were relevant to our models, and to find interesting relationships among them. Therefore, the proposed technique for *post hoc* feature analysis represents a way of providing interpretability to our neural-based models. One observation of the heat map visualizations is that they are practically limited by the maximum number of compounds that can be visually analyzed at the same time. Yet we note that an analysis on the top-20 or 30 features can be carried out with no problems as it was described previously.

6. Conclusions

QSAR modeling has become a key stage in the complex drug discovery process throughout the years. Upon the recent increase in the volume and quality of accessible datasets as well as computational power, more complex machine

learning algorithms have established as current state of the art in QSAR modeling. Consensus approaches have consistently proven their efficacy for bioactivity prediction, but tend to lack interpretability and suffer from the limitations of their base classifiers. DNNs have not been widely adopted as a standard for QSAR modeling yet, although their effectiveness in solving high-dimensional problems make them a suitable technique for this area.

While DNN-based models attain higher predictive performance than other established techniques, they have their own challenges, such as low interpretability and proneness to overfitting. In this work we developed three neural-based QSAR models, which outperformed the state-of-the-art results for the three properties under study. At the same time we address the interpretability drawback without the need for performing feature selection. In addition, in this work we posed a strategy for analyzing the applicability domain of a neural-based QSAR model based on network output probabilities, which was shown to be correlated to the likelihood of correct classification. We also provided a technique based on an aggregation of the network weights for identifying the most relevant molecular descriptors and fingerprint fragments in a *post hoc* manner, which provides a sense of interpretability to our models. As future work we plan to investigate the impact of multi-task training, as a way of improving the performance of neural-based QSAR models.

Acknowledgements

This work is kindly supported by CONICET, grant PIP 112-2012-0100471 and UNS, grant PGI 24/N042. Authors thank MinCyT for its grant "PIDRI / PRH-2017-0007". Authors also thank Dr. Gustavo Vazquez for his help in the calculation of molecular descriptors.

Conflict of Interest

The authors declare no competing conflict of interest.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] Vinicius M. Alves, Alexander Golbraikh, Stephen J. Capuzzi, Kammy Liu, Wai In Lam, Daniel Robert Korn, Diane Pozefsky, Carolina Horta Andrade, Eugene N. Muratov, and Alexander Tropsha. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. *Journal of Chemical Information and Modeling*, 58(6):1214–1223, jun 2018. doi: 10.1021/acs.jcim.8b00124.

[3] Sorin Avram, Alina Bora, Liliana Halip, and Ramona Curpăn. Modeling Kinase Inhibition Using Highly Confident Data Sets. *Journal of Chemical Information and Modeling*, 58(5):957–967, may 2018. doi: 10.1021/acs.jcim.7b00729.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, jul 2015. doi: 10.1371/journal.pone.0130140.

[5] Davide Ballabio, Fabrizio Biganzoli, Roberto Todeschini, and Viviana Consonni. Qualitative consensus of QSAR ready biodegradability predictions. *Toxicological & Environmental Chemistry*, pages 1–24, dec 2016. doi: 10.1080/02772248.2016.1260133.

[6] Davide Ballabio, Francesca Grisoni, Viviana Consonni, and Roberto Todeschini. Integrated QSAR Models to Predict Acute Oral Systemic Toxicity. *Molecular Informatics*, page minf.201800124, dec 2018. doi: 10.1002/minf.201800124.

[7] Igor I. Baskin, Vladimir A. Palyulin, and Nikolai S. Zefirov. Neural Networks in Building QSAR Models. In *Artificial Neural Networks*, pages 133–154. Humana Press, 2006. doi: 10.1007/978-1-60327-101-1_8.

[8] Igor I. Baskin, David Winkler, and Igor V. Tetko. A renaissance of neural networks in drug discovery. *Expert Opinion on Drug Discovery*, 11(8): 785–795, aug 2016. doi: 10.1080/17460441.2016.1201262.

[9] Francois Berenger and Yoshihiro Yamanishi. A Distance-Based Boolean Applicability Domain for Classification of High Throughput Screening Data. *Journal of Chemical Information and Modeling*, page acs.jcim.8b00499, jan 2019. doi: 10.1021/acs.jcim.8b00499.

[10] Audrey Cayot, Davy Laroche, Anne Disson-Dautriche, Anaïs Arbault, Jean-Francis Maillefert, and Paul Ornetti. Cytochrome P450 interactions and clinical implication in rheumatology. *Clinical Rheumatology*, 33(9): 1231–1238, sep 2014. doi: 10.1007/s10067-014-2710-3.

[11] Lidia Ceriani, Ester Papa, Simona Kovarich, Robert Boethling, and Paola Gramatica. Modeling ready biodegradability of fragrance materials. *Environmental Toxicology and Chemistry*, 34(6):1224–1231, jun 2015. doi: 10.1002/etc.2926.

[12] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Dis-*

covery Today, 23(6):1241–1250, jun 2018. doi: 10.1016/J.DRUDIS.2018.01.039.

[13] Feixiong Cheng, Yue Yu, Jie Shen, Lei Yang, Weihua Li, Guixia Liu, Philip W. Lee, and Yun Tang. Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *Journal of Chemical Information and Modeling*, 51(5):996–1011, may 2011. doi: 10.1021/ci200028n.

[14] R. Vasundhara Devi, S. Siva Sathya, and Mohane Selvaraj Coumar. Evolutionary algorithms for de novo drug design – A survey. *Applied Soft Computing*, 27:543–552, feb 2015. doi: 10.1016/J.ASOC.2014.09.042.

[15] Dimitar Dobchev and Mati Karelson. Have artificial neural networks met expectations in drug discovery as implemented in QSAR framework? *Expert Opinion on Drug Discovery*, 11(7):627–639, jul 2016. doi: 10.1080/17460441.2016.1186876.

[16] Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling*, 54(3):837–843, mar 2014. doi: 10.1021/ci400573c.

[17] Jiansong Fang, Ranyao Yang, Li Gao, Shengqian Yang, Xiaocong Pang, Chao Li, Yangyang He, Ai-Lin Liu, and Guan-Hua Du. Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. *Molecular Diversity*, 19(1):149–162, feb 2015. doi: 10.1007/s11030-014-9561-3.

[18] Alberto Fernández, Robert Rallo, and Francesc Giralt. Prioritization of in silico models and molecular descriptors for the assessment of ready biodegradability. *Environmental Research*, 142:161–168, oct 2015. doi: 10.1016/J.ENVRES.2015.06.031.

[19] Erik Gawehn, Jan A. Hiss, and Gisbert Schneider. Deep Learning in Drug

Discovery. *Molecular Informatics*, 35(1):3–14, jan 2016. doi: 10.1002/minf.201501008.

[20] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks, mar 2010. URL `http://proceedings.mlr.press/v9/glorot10a.html`.

[21] Garrett B. Goh, Khusheemn Sakloth, Charles Siegel, Abhinav Vishnu, and Jim Pfaendtner. Multimodal Deep Neural Networks using Both Engineered and Learned Representations for Biodegradability Prediction. *arXiv preprint arXiv:1808.04456*, aug 2018. doi: arXiv:1808.04456v2.

[22] Mohammad Goodarzi, Bieke Dejaegher, and Yvan Vander Heyden. Feature selection methods in QSAR studies. *Journal of AOAC International*, 95 (3):636–51, 2012.

[23] M.A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, nov 2003. doi: 10.1109/TKDE.2003.1245283.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[25] Dan Hendrycks and Kevin Gimpel. Generalizing and Improving Weight Initialization. *undefined*, 2016.

[26] Gerhard Hessler, Karl-Heinz Baringhaus, Gerhard Hessler, and Karl-Heinz Baringhaus. Artificial Intelligence in Drug Design. *Molecules*, 23(10):2520, oct 2018. doi: 10.3390/molecules23102520.

[27] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, feb 1970. doi: 10.1080/00401706.1970.10488634.

[28] JP Hughes, S Rees, SB Kalindjian, and KL Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, mar 2011. doi: 10.1111/j.1476-5381.2010.01127.x.

[29] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *undefined*, 2015.

[30] Joanna Jaworska, Nina Nikolova-Jeliazkova, and Tom Aldenberg. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Alternatives to laboratory animals : ATLA*, 33(5):445–59, oct 2005.

[31] Berith F. Jensen, Christian Vind, Søren B. Padkjær, Per B. Brockhoff, and Hanne H. F. Refsgaard. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. *Journal of medicinal chemistry*, 2007. doi: 10.1021/JM060333S.

[32] Supratik Kar, Kunal Roy, and Jerzy Leszczynski. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. In *Methods in molecular biology (Clifton, N.J.)*, volume 1800, pages 141–169. Springer, 2018. doi: 10.1007/978-1-4939-7899-1_6.

[33] Seyran Khademi, Xiangwei Shi, Tino Mager, Ronald Siebes, Carola Hein, Victor de Boer, and Jan van Gemert. Sight-Seeing in the Eyes of Deep Neural Networks. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 407–408. IEEE, oct 2018. ISBN 978-1-5386-9156-4. doi: 10.1109/eScience.2018.00125.

[34] Kabiruddin Khan, Emilio Benfenati, and Kunal Roy. Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms:

Ranking and prioritization of the DrugBank database compounds. *Ecotoxicology and Environmental Safety*, 168:287–297, jan 2019. doi: 10.1016/J.ECOENV.2018.10.060.

[35] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, dec 2014.

[36] Waldemar Klingspohn, Miriam Mathea, Antonius ter Laak, Nikolaus Heinrich, and Knut Baumann. Efficiency of different measures for defining the applicability domain of classification models. *Journal of Cheminformatics*, 9(1):44, dec 2017. doi: 10.1186/s13321-017-0230-2.

[37] Alexios Koutsoukas, Keith J. Monaghan, Xiaoli Li, and Jun Huan. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, 9(1):42, dec 2017. doi: 10.1186/s13321-017-0226-y.

[38] Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20(3):318–331, mar 2015. doi: 10.1016/J.DRUDIS.2014.10.012.

[39] Tailong Lei, Youyong Li, Yunlong Song, Dan Li, Huiyong Sun, and Tingjun Hou. ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *Journal of Cheminformatics*, 8(1):6, dec 2016. doi: 10.1186/s13321-016-0117-7.

[40] Eelke B. Lenselink, Niels ten Dijke, Brandon Bongers, George Papadatos, Herman W. T. van Vlijmen, Wojtek Kowalczyk, Adriaan P. IJzerman, and Gerard J. P. van Westen. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics*, 9(1):45, dec 2017. doi: 10.1186/s13321-017-0232-0.

[41] Hao Lin, Hui Ding, Feng-Biao Guo, and Jian Huang. Prediction of subcellular location of mycobacterial protein using feature selection techniques. *Molecular Diversity*, 14(4):667–671, nov 2010. doi: 10.1007/s11030-009-9205-1.

[42] Ruifeng Liu, Hao Wang, Kyle P. Glover, Michael G. Feasel, and Anders Wallqvist. Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure–Activity Relationship Models Based on Deep Neural Networks? *Journal of Chemical Information and Modeling*, page acs.jcim.8b00348, nov 2018. doi: 10.1021/acs.jcim.8b00348.

[43] Junshui Ma, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274, feb 2015. doi: 10.1021/ci500747n.

[44] Stephani Joy Y. Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, 38(9):1686–1701, sep 2015. doi: 10.1007/s12272-015-0640-5.

[45] Christopher D. Manning, Prabhakar. Raghavan, and Hinrich. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN 9780521865715.

[46] Kamel Mansouri, Tine Ringsted, Davide Ballabio, Roberto Todeschini, and Viviana Consonni. Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals. *Journal of Chemical Information and Modeling*, 53(4):867–878, apr 2013. doi: 10.1021/ci4000213.

[47] Richard L. Marchese Robinson, Anna Palczewska, Jan Palczewski, and Nathan Kidley. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets.

Journal of Chemical Information and Modeling, 57(8):1773–1792, aug 2017. doi: 10.1021/acs.jcim.6b00753.

[48] Alfonso E. Márquez-Chamorro, Gualberto Asencio-Cortés, Cosme E. Santiesteban-Toca, and Jesús S. Aguilar-Ruiz. Soft computing methods for the prediction of protein tertiary structures: A survey. *Applied Soft Computing*, 35:398–410, oct 2015. doi: 10.1016/J.ASOC.2015.06.024.

[49] María Jimena Martínez, Ignacio Ponzoni, Mónica F Díaz, Gustavo E Vazquez, and Axel J Soto. Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods. *Journal of Cheminformatics*, 7(1):39, dec 2015. doi: 10.1186/s13321-015-0092-4.

[50] María Jimena Martínez, Julieta Sol Dussaut, and Ignacio Ponzoni. Biclustering as Strategy for Improving Feature Selection in Consensus QSAR Modeling. *Electronic Notes in Discrete Mathematics*, 69:117–124, aug 2018. doi: 10.1016/J.ENDM.2018.07.016.

[51] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, nov 2015.

[52] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, feb 2018. doi: 10.1016/J.DSP.2017.10.011.

[53] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. *undefined*, 2010.

[54] Serena Nembri, Francesca Grisoni, Viviana Consonni, and Roberto Todeschini. In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. *International Journal of Molecular Sciences*, 17(6):914, jun 2016. doi: 10.3390/ijms17060914.

[55] Ester Papa, Simona Kovarich, and Paola Gramatica. Development, Validation and Inspection of the Applicability Domain of QSPR Models for

Physicochemical Properties of Polybrominated Diphenyl Ethers. *QSAR & Combinatorial Science*, 28(8):790–796, aug 2009. doi: 10.1002/qsar.200860183.

[56] Carlos J S Passos and Donna Mergler. Human mercury exposure and adverse health effects in the Amazon: a review. *Cadernos de saude publica*, 24 Suppl 4:s503–20, 2008.

[57] M. E. Pina, P. Coimbra, P. Ferreira, P. Alves, A. I. Figueiredo, and M. H. Gil. Polymeric Materials in Ocular Drug Delivery Systems. In *Handbook of Polymers for Pharmaceutical Technologies*, pages 439–458. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2015. doi: 10.1002/9781119041412.ch16.

[58] Pavel Polishchuk. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *Journal of Chemical Information and Modeling*, 57(11):2618–2639, nov 2017. doi: 10.1021/acs.jcim.7b00274.

[59] Ignacio Ponzoni, Víctor Sebastián-Pérez, Carlos Requena-Triguero, Carlos Roca, María J Martínez, Fiorella Cravero, Mónica F Díaz, Juan A Páez, Ramón Gómez Arrayás, Javier Adrio, and Nuria E Campillo. Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Scientific reports*, 7(1):2403, 2017. doi: 10.1038/s41598-017-02114-3.

[60] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1135–1144, New York, New York, USA, 2016. ACM Press. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

[61] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, may 2010. doi: 10.1021/ci100050t.

[62] David Rosner and Gerald Markowitz. Persistent pollutants: A brief history of the discovery of the widespread toxicity of chlorinated hydrocarbons. *Environmental Research*, 120:126–133, jan 2013. doi: 10.1016/J.ENVRES.2012.08.011.

[63] Kunal Roy and Asim Sattwa Mandal. Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 23(6):980–995, jan 2008. doi: 10.1080/14756360701811379.

[64] Kunal Roy, Supratik Kar, and Pravin Ambure. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 145:22–29, jul 2015. doi: 10.1016/j.chemolab.2015.04.013.

[65] Kunal Roy, Pravin Ambure, and Rahul B. Aher. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? *Chemometrics and Intelligent Laboratory Systems*, 162:44–54, mar 2017. doi: 10.1016/J.CHEMOLAB.2017.01.010.

[66] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986. doi: 10.1038/323533a0.

[67] Daniel P. Russo, Kimberley M. Zorn, Alex M. Clark, Hao Zhu, and Sean Ekins. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Molecular Pharmaceutics*, 15(10):4361–4370, oct 2018. doi: 10.1021/acs.molpharmaceut.8b00546.

[68] Pranav Shah, Alexey Zakharov, R. Scott Obach, Anton Simeonov, Cornelis Hop, Dac-Trung Guyen, Eric Gonzalez, Hongmao Sun, and Xin Xu. Development of a multitask deep learning QSAR model using data from individual cytochrome P450 isozymes. *Drug Metabolism and Pharmacokinetics*, 33(1):S35–S36, jan 2018. doi: 10.1016/J.DMPK.2017.11.131.

[69] Watshara Shoombuatong, Philip Prathipati, Wiwat Owasirikul, Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen, Jarl E. S. Wikberg, and Chanin Nantasenamat. Towards the Revival of Interpretable QSAR Models. In *Advances in QSAR Modeling*, pages 3–55. Springer, 2017. doi: 10.1007/978-3-319-56850-8_1.

[70] Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*, mar 2017.

[71] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[72] Kode srl. Dragon (software for molecular descriptor calculation). Pisa, Italy, 2016. URL `https://chm.kode-solutions.net/products_dragon.php`. Version 7.0.

[73] Sheng Tian, Junmei Wang, Youyong Li, Dan Li, and Lei Xu. The application of in silico drug-likeness predictions in pharmaceutical research. *Advanced Drug Delivery Reviews*, 86:2–10, jun 2015. doi: 10.1016/J.ADDR.2015.01.009.

[74] Roberto Todeschini, Davide Ballabio, Matteo Cassotti, and Viviana Consonni. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *Journal of Chemical Information and Modeling*, 55(11):2365–2374, nov 2015. doi: 10.1021/acs.jcim.5b00326.

[75] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting Statistical Interactions from Neural Network Weights. *arXiv preprint arXiv:1705.04977*, may 2017.

[76] Alfredo Vellido, José David Martín-Guerrero, and Paulo J. G. Lisboa. Making machine learning models interpretable. *ESANN*, 2012.

[77] Yulan Wang, Jing Xing, Yuan Xu, Nannan Zhou, Jianlong Peng, Zhaoping Xiong, Xian Liu, Xiaomin Luo, Cheng Luo, Kaixian Chen, Mingyue Zheng, and Hualiang Jiang. In silico ADME/T modelling for rational drug design. *Quarterly Reviews of Biophysics*, 48(04):488–515, nov 2015. doi: 10.1017/S0033583515000190.

[78] Shane Weaver and M. Paul Gleeson. The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling*, 26(8):1315–1326, jun 2008. doi: 10.1016/j.jmgm.2008.01.002.

[79] David A. Winkler. Neural Networks as Robust Tools in Drug Lead Discovery and Development. *Molecular Biotechnology*, 27(2):139–168, jun 2004. doi: 10.1385/MB:27:2:139.

[80] Alexey V Zakharov, Megan L Peach, Markus Sitzmann, and Marc C Nicklaus. QSAR modeling of imbalanced high-throughput screening data in PubChem. *Journal of chemical information and modeling*, 54(3):705–12, mar 2014. doi: 10.1021/ci400737s.

[81] Yadi Zhou, Suntara Cahya, Steven A. Combs, Christos A. Nicolaou, Jibo Wang, Prashant V. Desai, and Jie Shen. Exploring Tunable Hyperparameters for Deep Neural Networks with Industrial ADME Data Sets. *Journal of Chemical Information and Modeling*, page acs.jcim.8b00671, jan 2019. doi: 10.1021/acs.jcim.8b00671.