International Journal of
Molecular Sciences

MDPI

Article

In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9

Serena Nembri †, Francesca Grisoni †, Viviana Consonni and Roberto Todeschini *

Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milano, Italy; s.nembri@campus.unimib.it (S.N.); f.grisoni@campus.unimib.it (F.G.); viviana.consonni@campus.unimib.it (V.C.)

* Correspondence: roberto.todeschini@unimib.it; Tel.: +39-02-6448-2820

† These authors contributed equally to the work.

Academic Editor: Jesus Vicente De Julián Ortiz
Received: 16 May 2016; Accepted: 6 June 2016; Published: 9 June 2016

Abstract: Cytochromes P450 (CYP) are the main actors in the oxidation of xenobiotics and play a crucial role in drug safety, persistence, bioactivation, and drug-drug/food-drug interaction. This work aims to develop Quantitative Structure-Activity Relationship (QSAR) models to predict the drug interaction with two of the most important CYP isoforms, namely 2C9 and 3A4. The presented models are calibrated on 9122 drug-like compounds, using three different modelling approaches and two types of molecular description (classical molecular descriptors and binary fingerprints). For each isoform, three classification models are presented, based on a different approach and with different advantages: (1) a very simple and interpretable classification tree; (2) a local (*k*-Nearest Neighbor) model based classical descriptors and; (3) a model based on a recently proposed local classifier (*N*-Nearest Neighbor) on binary fingerprints. The salient features of the work are (1) the thorough model validation and the applicability domain assessment; (2) the descriptor interpretation, which highlighted the crucial aspects of P450-drug interaction; and (3) the *consensus* aggregation of models, which largely increased the prediction accuracy.

Keywords: cytochrome P450; QSAR; CYP2C9; CYP3A4; *in silico*; ADMET

1. Introduction

Cytochromes P450 (CYP) are a family of monooxygenase enzymes known for their crucial role in the metabolism of xenobiotics, as they are involved in the oxidation of the majority of compounds [1]. Despite the fact that the human genome encodes up to 57 different CYP genes, only six isoforms are mainly interested in drug metabolism [2,3]. Two of them account for 43% of the metabolism of known drugs: (1) the CYP3A4 isoform, which interacts with more than a half of all clinically used drugs, e.g., large and lipophilic molecules and; (2) the CYP2C9 isoform, which mainly metabolizes NSAIDs (Non-Steroidal Anti-Inflammatory Drugs) and weakly acidic molecules with a hydrogen bond acceptor [4]. These isoforms were the targets of the present work.

The evaluation of the interaction of CYP with chemicals constitutes a fundamental step for drug discovery/design, as well as for toxicity assessment [2,5–7]. For this reason, Cytochrome P450 isozymes have been the target of many modelling studies [8,9]. In this context, relevant contributions to the field have been given by Quantitative Structure-Activity Relationship (QSAR) studies, which link the molecular structure, encoded within the so-called molecular descriptors [10], to a biological activity of interest through a mathematical/statistical approach. Throughout the years, many QSAR models have been proposed [11], based on different statistical techniques, such as Support Vector Machine (SVM) [12–15], *k*-Nearest Neighbors (*k*-NN) [16] and Self Organizing Maps (SOM) [17]. However, the majority of proposed QSAR models have some limitations that can often hamper the applicability and

reliability of their predictions, especially for ADMET (adsorption, distribution, metabolism, excretion and toxicity) applications. In particular, the most notable drawbacks are (1) a limited applicability due to the small datasets they are often trained on or to the lack of an applicability domain assessment; and (2) the lack of biochemical interpretability due to the use of very complex modelling approaches and molecular descriptors.

The present work stems from these considerations and targets the development of novel, simple and easily interpretable QSAR systems to screen the drug binding to 3A4 and 2C9 isoform of CYP, which account for the metabolism of 30% and 13% of known drugs, respectively. Models were developed on High-Throughput Screening data and special attention was paid to model validation, applicability domain assessment and interpretation of the selected molecular descriptors.

2. Results and Discussion

2.1. Modelling Approach

The starting point of this study was the database of potency values for 17,143 drug-like compounds for five CYP isoforms developed by Veith *et al.* [18], determined through quantitative High-Throughput Screening with a bioluminescent assay, which recognizes both inhibitors and substrates. The database was retrieved from PubChem (AID: 1851) [19]. Our modelling approach followed six logical steps:

1. Data curation and splitting. CYP3A4 and CYP2C9 isoforms were curated separately by removing mismatching duplicates, inconsistent records and disconnected structures. The molecules were divided in (a) a Shared set, comprising the 9122 compounds with annotated activity for both the isoforms (Figure 1) and (b) an External set, having activity data for one isoform molecules (2996 and 2818 for CYP3A4 and CYP2C9, respectively). The Shared set molecules were randomly split into a training (70%, 6385 compounds) and a test set (30%, 2737 compounds), keeping the active/inactive proportion of both the isoforms (49:100 and 66:100 for 2C9 and 3A4, respectively). The training set served to select the variables, calibrate the models and perform the cross-validation (five-fold). The test set was used only in a later stage to validate the final pool of selected models. The external sets were used in the final stage to further validate the best models.
2. Molecular description. To allow for the mathematical treatment of molecules, they were described using the so-called molecular descriptors [10], that is, numbers encoding for the presence of particular structural features, fragments or chemical properties. Two types of descriptors were calculated: (a) 3763 classical Dragon 6 [20] molecular descriptors (MDs) from 0-dimensional to 2-dimensional molecular representation, from which only a set of 1472 non redundant MDs was finally retained (see Materials and Methods); and (b) two types of binary fingerprints (FPs), that is, the extended connectivity (ECFP) [21] and the path fingerprints (PFP) [22], which are 1024 bit strings encoding the presence of particular fragments/substructures of molecules. Three-dimensional descriptors were not considered, as in a preliminary phase they did not lead to an improvement in the predictions.
3. Variable selection and modelling. The Genetic Algorithms (GA) [23], a benchmark variable selection method characterized by an optimal trade-off between computational time and exploration/exploitation ability [24], were used to retain the most relevant subsets of variables. A refined two-step GA procedure (see Materials and Methods) was applied on the training set descriptors in combination with six classification techniques: (a) Classification and Regression Trees (CART) [25]; (b) *k*-Nearest Neighbor (*k*-NN) [26]; (c) *N*-Nearest Neighbors (N3) [27]; (d) Binned Nearest Neighbors (BNN) [27]; (e) Linear Discriminant Analysis (LDA) [28]; and (f) Partial Least Squares Analysis (PLSDA) [29]. Note that for FPs, no variable selection was performed, as they, unlike MDs, give a description of the molecule when used as a whole. To model FPs, only similarity-based classifiers (*k*-NN, N3 and BNN) were used. On both the isoforms, the best results were obtained using: (1) CART [25], based on binary splits of the data using one variable at time according to its optimized threshold values; (2) *k*-NN [26], in which

every molecule is classified according to the majority vote of its k more similar objects [14]; and (3) N3 [27], which uses all the available molecules as neighbors and, through an optimized α exponent, tunes their contribution as decreasing with decreasing their similarity to the new object. The model parameters (number of objects per leaf, k and α) were optimized in cross-validation as those giving the best classification performance.

4. Model selection and validation. From the pool of calculated models, the final models were chosen as the best compromise between classification performance in five-fold cross-validation (the higher the better) and number of variables (the smaller the better). Models with interpretable descriptors, if relevant, were preferred.
5. Applicability Domain Assessment. The selected models were evaluated for their chemical space of prediction reliability (Applicability Domain, AD). The AD assessment strongly depends on the nature of the modelling approach and the characteristics of the dataset [30], thus, it was calibrated it on a case-by-case basis, and rationalized according to the modeling approach (see Materials and Methods).
6. External validation. Models were selected according to the cross-validation results and the best models were screened on their performance on the test set. Finally, for each isoform, the external set molecules were used in order to test their robustness and predictivity towards real unknown data.

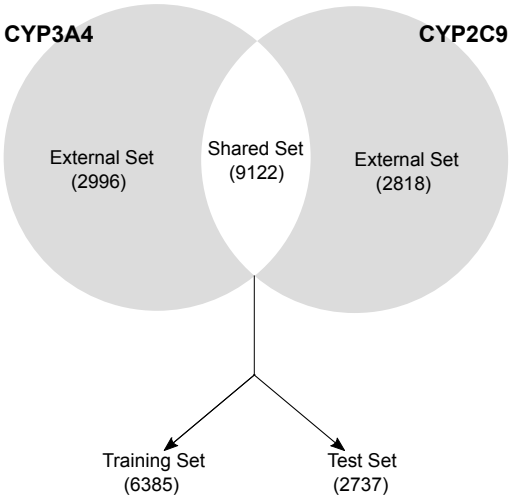


Figure 1. Scheme of the data splitting.

The model performance in recognizing active/inactive compounds was calculated through the Sensitivity (*Sn*), Specificity (*Sp*) and Non-Error Rate (*NER*), defined for two-class problems as follows:

$$Sn = \frac{TP}{TP+FN}$$
$$Sp = \frac{TN}{TN+FP} \quad (1)$$
$$NER = \frac{Sn+Sp}{2}$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives of each class, respectively. *Sn*, *Sp* and *NER* were calculated in fitting, cross-validation, and on the test/external sets.

2.2. Quantitative Structure-Activity Relationship (QSAR) Models

2.2.1. Isoform 3A4

The proposed QSAR models for 3A4 are collected in Table 1. For all the models, a similar performance on the training and test sets can be noted, indicating the robustness and reliability of the predictions towards unknown data. The CART model, which is based on three very simple molecular descriptors, showed a very good balance between *Sn* and *Sp*. The *k*-NN model has a slightly better prediction ability, especially for the inactive compounds (higher *Sp*). If the model is restricted to its applicability domain (AD), more balanced predictions (in terms of *Sn* and *Sp*) are obtained with the same *NER* value. Finally, the N3 model (based on ECFPs) is characterized by high *Sn* values, that is, it identifies well the active compounds.

Table 1. Model statistics for CYP3A4 isoform. Models are described according to the method and type of descriptors, the Applicability Domain (AD: yes/no (y/n)), number of variables (*p*) and classification parameters (parameter: object/leaf ratio for Classification and Regression Trees (CART), *k* for *k*-Nearest Neighbours (*k*-NN) and α for *N*-Nearest Neighbors (N3)). For each model, the Non-Error Rate (*NER*), the Sensitivity (*Sn*) and the Specificity (*Sp*) are reported in Fitting, Cross-Validation and on the test set. %out indicates the percentage of test set compounds outside of the AD. MD: molecular descriptors; ECFP: extended connectivity fingerprints.

Model	Descriptors	AD	*p*	Parameter	Fitting			Cross-Validation			Test Set			
					NER	*Sn*	*Sp*	*NER*	*Sn*	*Sp*	%out	*NER*	*Sn*	*Sp*
CART	MD	y	3	210	0.74	0.74	0.75	0.74	0.73	0.75	-	0.75	0.74	0.76
		n	3	210	0.74	0.74	0.75	0.74	0.73	0.75	0	0.75	0.74	0.76
k-NN	MD	y	6	14	0.76	0.73	0.79	0.76	0.73	0.78	-	0.77	0.75	0.79
		n	6	14	0.76	0.73	0.79	0.76	0.73	0.78	5	0.77	0.76	0.78
N3	ECFP	y	1024	1	0.79	0.88	0.71	0.79	0.87	0.70	-	0.78	0.86	0.71
		n	1024	1	0.79	0.88	0.71	0.79	0.87	0.70	1	0.78	0.86	0.71

The selected MDs are briefly presented in Table 2. In Figure 2a, their role in the CART classification is depicted. To visualize the effect of the descriptors on the *k*-NN model, a Principal Component Analysis (PCA) [31] was performed (Figure 2b,c). PCA is a well-known data visualization technique that allows to observe the objects in a few new variables (Principal Components, PC) space, according to their coordinates (scores, Figure 2b). The contribution of the variables in the PC space is given by the loadings (Figure 2c), the higher they are (in absolute value) the larger their role. PC1 and PC2 explain 2/3 of data variance. Information about PC3, which explains 18% of data variance, is reported in the Supplementary Material (Figure S1). PC3 explains 14% of data variance, and it leads to considerations similar to those derived from PC1/PC2. The PCA loadings (Figure 2c) allow for an understanding of the contribution of the molecular descriptors: the higher their PC1 loadings, the higher their value for inactive molecules and *vice versa* for active compounds. Active compounds group on the left side of the score plot (negative PC1 scores), while the inactive molecules distribute on the right side (positive PC1 scores). At the same time, active compounds have relatively lower PC2 values, indicating that high PC2 values relate to inactivity. The descriptors with high loadings on PC1 (e.g., *nROH*) and PC2 (e.g., *ATSC4p* and *ATSC6i*) influence at most the observed class separation. With the aid of the PCA and the CART classification scheme, a brief description and interpretation of the descriptors is given below.

1. The descriptor *nROH* represents the number of hydroxyl groups. It was independently selected in both the models based on MDs, underscoring its relevance in modelling the CYP3A4 activity. In particular, in both cases (Figure 2), high *nROH* values tend to correspond to inactive molecules. An increasing number of hydroxyl group generally increases the hydrophilicity, while molecules with no (or a few of) hydroxyl groups have a lipophilic nature. Molecular lipophilicity is known to facilitate the adsorption and to limit the excretion of the compounds. As a result, lipophilic

molecules are oxidated by CYP and converted into more hydrophilic compounds that can be easily eliminated [32].

2. *nBM* and *nBO* are the number of multiple bonds and bonds in the H-depleted molecular structure, respectively, which were selected within the CART model (Figure 2c). The descriptor *nBM* represents a measure of the unsaturation level and, therefore, gives information about the molecular interaction ability with CYP. In addition, this descriptor contains information about molecular size, flexibility, presence of heteroatoms (Table 2). In particular, small and flexible molecules with a few of the multiple bonds (*nBM* < 13, Figure 2c) tend to be inactive. The relevance of heteroatoms can be related to the P450-mediated oxygen addition to nitrogen, sulfur, phosphorus, and iodine atoms [33–36].
3. The descriptor *nBO* mainly encodes the information about molecular size. Molecules with no hydroxyl groups, less than 16 multiple bonds and with less than 25 non-hydrogen bonds tend to be inactive. These molecules probably have a small effective dimension within the receptor pocket, as they are either small or relatively small, but very flexible. In this branch also lie some hydrophilic compounds classified as inactive. They probably do not interact with CYP as they are easily dissolved in the aqueous body fluid and, consequently, are excreted from the body [37].
4. *C%* and *NaasC* are the percentage of C atoms and number of aromatic carbons bonded with non-H atoms, respectively (Table 2). They have a crucial role in identifying the active compounds of *k*-NN, as shown by their very high loadings on PC1. This means that relative large and/or aromatic molecules are generally classified as actives. The mechanisms of aromatic oxidation have been already elucidated [36,38].
5. *SIC5* is the structural information content of order 5 [39]. It encodes information about atom equivalence and represents a general measure of structural complexity, the higher, the larger *SIC5*. It has negative PC1/PC2 loadings, suggesting that relatively large, branched and/or polycyclic compounds tend to be active. The effect of dimension on activity was already suggested (e.g., [40]), as well-known CYP3A4 ligands are commonly relatively large molecules.
6. *ATSC4p* and *ATSC6i* are the Centred Broto-Moreau autocorrelations [10] of lag 4 and 6, respectively. The former is weighed on the atomic polarizability, while the latter on the atomic ionization potential. They lie very close to each other in the PC space and have positive loadings on PC2, indicating that inactive compounds tend to have high values of both these MDs. They increase when increasing the molecular dimension, the number of heteroatoms and the branching/cyclicity. Because of their weighting scheme type, *ATSC4p* and *ATSC6i* increase when increasing the atomic polarizability and the ionization potential, respectively. These features are known for their relevant contribution in receptor binding, especially when charged systems are taken into consideration [41,42].

In order to interpret the ECFP, the relevant molecular fragments (MFs) encoded by ECFP were generated by means of Dragon 7 [22], with the same settings used for the fingerprints calculation. A small set of 19 large fragments (length from 4 to 6) was retained from the original pool by: (a) removing those that were not selected for at least 100 molecules; and (b) considering only those that were characterized by a difference of frequency between the classes greater than or equal to 1%. The fragments are depicted in Figure 3. Only the cyclohexanediol derivate fragment (No. 1) occurs more frequently within the inactive compounds. All the other MFs have a higher frequency for the active molecules. In particular, the MFs quinazolinamine-like (No. 4 and 5) and benzamide-like (No. 14) MFs are mostly present within active compounds (frequency ranging from 3.6% to 5.1%), and their frequency value within the inactive molecules is lower than 0.5%. The butenamide- (No. 2) and fluorobenzene (No. 9)-derived MFs are less useful for the discrimination between the classes, since they have the lowest difference in the frequency. The remaining fragments are high for both the classes.

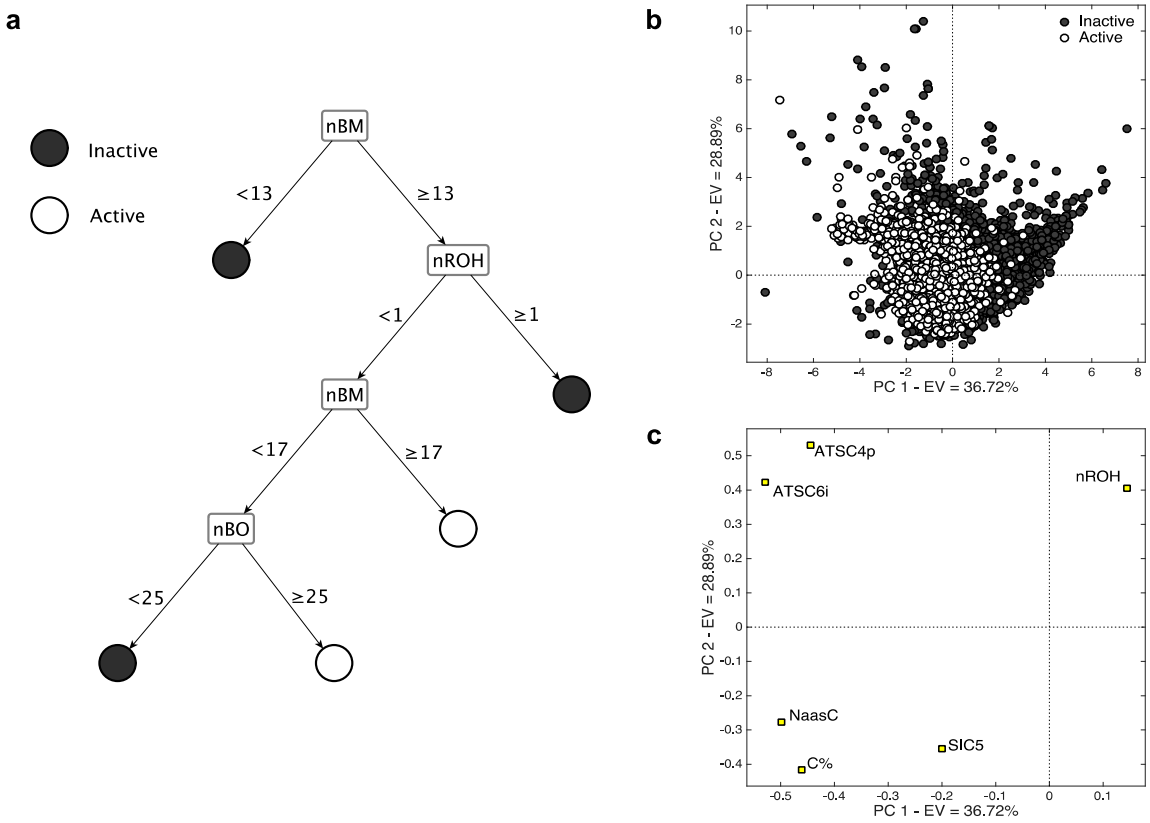


Figure 2. Representation of the molecular descriptor (MD)-based models for CYP3A4: (**a**) Classification and Regression Trees (CART) model; (**b**) Score plot of the training molecules described by the *k*-Nearest Neighbours (*k*-NN) descriptors, coloured according to their activity; (**c**) Loading plot of the *k*-NN descriptors.

Table 2. List and brief description of the classical molecular descriptors (MDs) selected for CYP3A4. Some examples of molecules with low and high MD values are also reported.

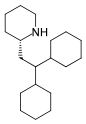
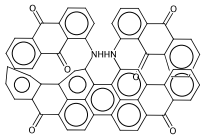
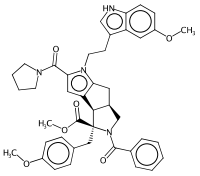
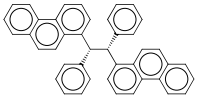
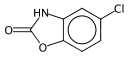
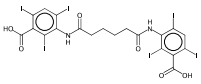
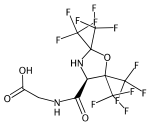
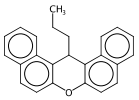
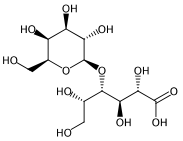
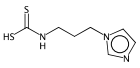
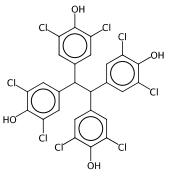
MD	Description	Reference	Model	Low Value		High Value	
nBM	Number of multiple bonds.	[10]	CART		0		66
nBO	Number of non-hydrogen bonds.	[10]	CART		3		59
C%	Percentage of C atoms.	[10]	*k*-NN		0		58.3

Table 2. *Cont.*

MD	Description	Reference	Model	Low Value		High Value	
SIC5	Structural Information Content—order 5.	[39]	*k*-NN		0.28		1.00
ATSC4p	Centred Broto-Moreau autocorrelations—lag 4 (weighted by atomic polarizability).	[20]	*k*-NN		0.21		46.6
ATSC6i	Centred Broto-Moreau autocorrelations—lag 6 (weighted by atomic ionization potential).	[20]	*k*-NN		0		4.1
nROH	Number of hydroxyl groups.	[10]	CART, *k*-NN		0		9
NaasC	Counts of the E-state atom types.	[43]	*k*-NN		0		16

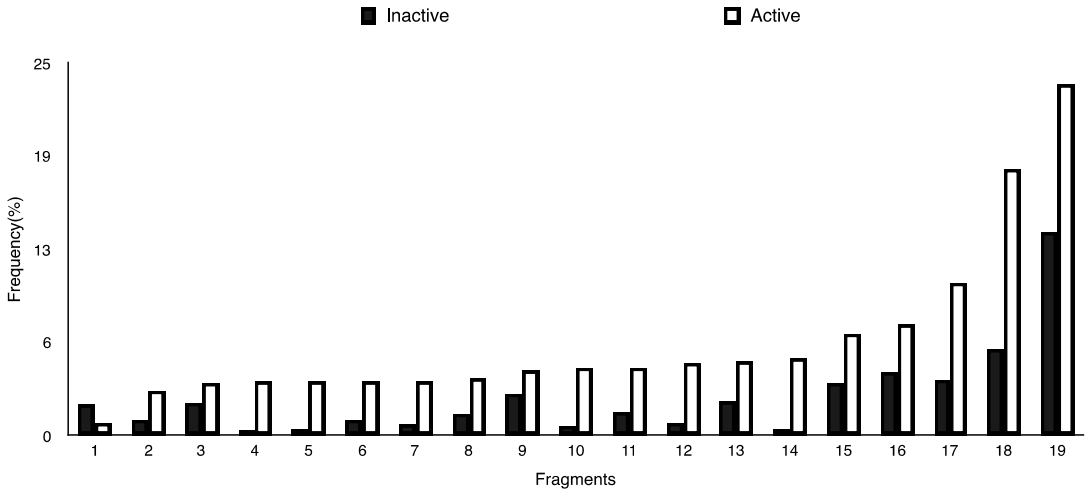


Figure 3. *Cont.*

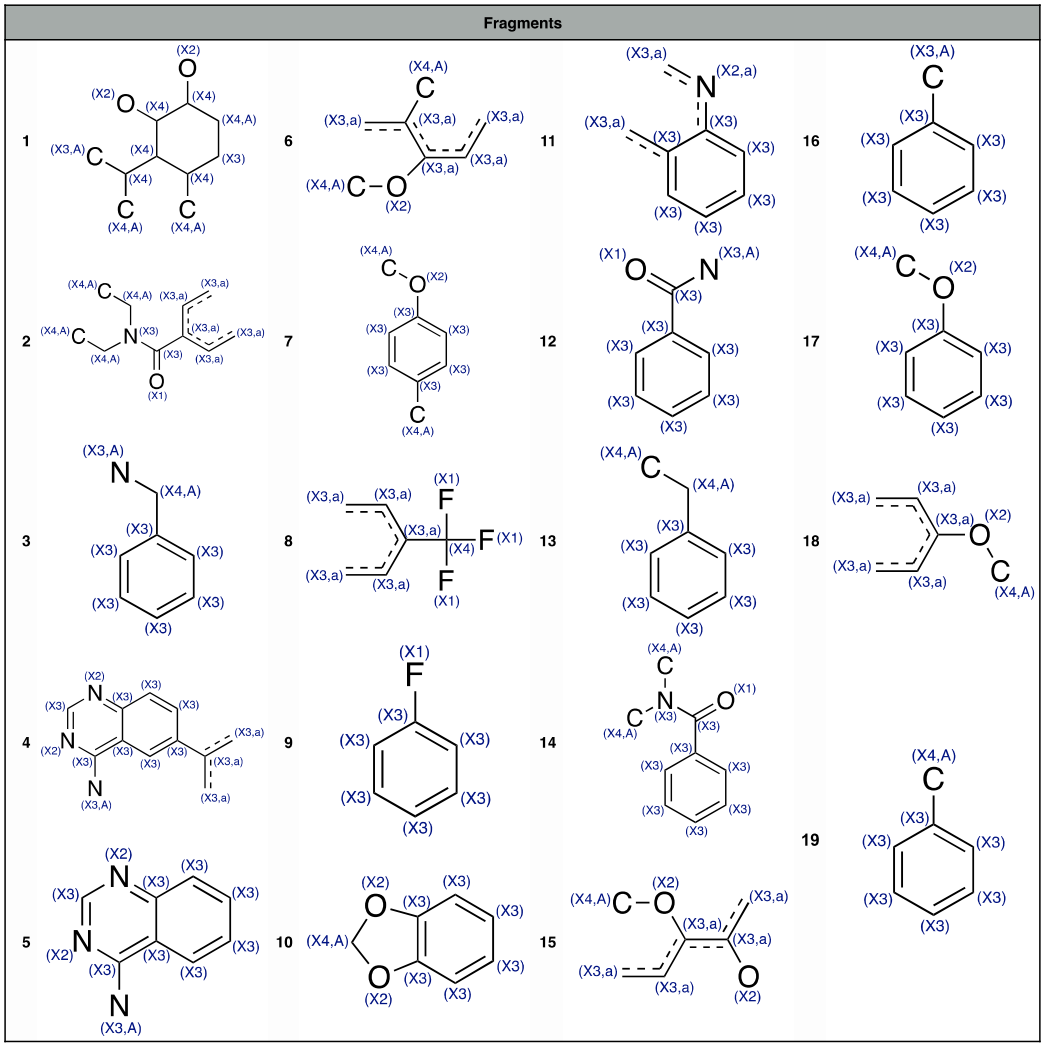


Figure 3. Occurrence frequency of the 19 selected fragments for CYP3A4 within the active/inactive compounds. Symbols associated with the fragments (according to SMARTS language) have the following meaning: *X* + number = number of total bonds in which the considered atom is involved; a = aromatic atom; A = aliphatic atom.

2.2.2. Isoform 2C9

The statistics of the selected classification models for the CYP2C9 are reported in Table 3. As for the 3A4 isoform, in this case the proposed models also have comparable performance on the training and test set. Furthermore, the CART model still appears to be the simplest and most balanced model, and only a few of the compounds are outside of its AD.

The best performance in recognizing inactive compounds was obtained with the *k*-NN approach (Table 3). On the contrary, N3 model, which is known for its ability to predict better the less represented class, is better in the actives recognition. The applicability domain characterization for the *k*-NN model slightly worsens the predictions on inactive compounds. For the N3 model, only a small fraction of compounds was excluded, not affecting the classification performance.

Table 3. Model statistics for CYP2C9 isoform. Models are described according to the method and type of descriptors, the Applicability Domain (AD: yes/no), number of variables (p) and classification parameters (parameter: object/leaf ratio for CART, *k* for kNN and α for N3). For each model, the Non-Error Rate (*NER*), the Sensitivity (*Sn*), and the Specificity (*Sp*) are reported in Fitting, Cross-Validation and on the test set. %out indicates the percentage of test set compounds outside the AD.

Model	Descriptors	AD	*p*	Parameter	Fitting			Cross-Validation			Test Set			
					NER	*Sn*	*Sp*	*NER*	*Sn*	*Sp*	%out	*NER*	*Sn*	*Sp*
CART	MD	y	4	210	0.75	0.75	0.75	0.75	0.75	0.75	-	0.75	0.75	0.74
		n	4	210	0.75	0.75	0.75	0.75	0.75	0.75	0	0.75	0.75	0.74
k-NN	MD	y	6	14	0.77	0.69	0.85	0.77	0.68	0.85	-	0.76	0.67	0.86
		n	6	14	0.77	0.69	0.85	0.77	0.68	0.85	5	0.76	0.67	0.84
N3	ECFP	y	1024	1	0.80	0.87	0.73	0.80	0.86	0.73	-	0.78	0.83	0.73
		n	1024	1	0.80	0.87	0.73	0.80	0.86	0.73	1	0.78	0.83	0.73

Table 4 collects all the selected MDs for CYP2C9. In analogy with the previous case, in Figure 4, the representation of the CART model (Figure 4a) and of the PCA on *k*-NN descriptors (Figure 4b,c) can be found. In this case, PC1 and PC2 explain 57% of data variance. PC3 explains 14% of data variance and leads to similar considerations than those derived from PC1/PC2 (see Supplementary Material). In analogy with CYP3A4, the active compounds are grouped on the left part, having a low score value on PC1 and PC2, while the inactive compounds mainly cover the right side (high PC1 and PC2 scores).

The selected MDs are briefly described and interpreted below:

- *nBM* (number of multiple bonds) resulted to be relevant also for this isoform. As for the 3A4 CART model, small flexible molecules with a few number of multiple bonds ($nBM < 13$, Figure 4c) are classified as inactive, while those with a high number of multiple bonds ($nBM > 17$, Figure 4c) are classified as active.
- The presence of pyrimidines (*nPyrimidines*), together with the high values of *Sp* (sum of atomic carbon-scaled polarizability, $Sp \geqslant 30.6$) characterize the active compounds. The inhibition ability of pyrimidine derivatives towards CYP2C9 was already suggested [44–47].
- *ARR* represents the aromatic ratio, that is, the ratio between the number of aromatic bonds and bonds in the H-depleted molecular structure. Molecules without pyrimidine rings but with an aromatic character ($ARR \geqslant 0.38$) and high atomic polarizability value ($Sp \geqslant 23.6$) are active. The presence of two MDs related to aromaticity (*i.e.*, *ARR* and *nPyrimidines*) suggests that this feature is fundamental for CYP2C9-drug interaction, in agreement with previous studies (e.g., [40]).
- *GATS2i* is the Geary autocorrelation of lag 2 weighted by the ionization potential [10]. It plays a crucial role in identifying the inactive compounds for the *k*-NN model, as denoted by its (high) positive loading on PC1. The inactive compounds, distributed on the right side of the score plot, are characterized by low values of this MD and have, therefore, low ionization potential. When increasing the ionization potential, the number of active compounds increases.
- *nRNR2* counts the number of aliphatic tertiary amines. It has high PC2 loading, meaning that it tends to be higher for inactive compounds. Tertiary aliphatic amines are, in fact, generally oxidized by Flavin monooxygenase [48] and are inactive on CYP.
- *F01[C–N]*, represents the frequency of bonded C and N atoms. It has a positive loading on PC2 and, in analogy with *nRNR2*, tends to be higher for inactive compounds, meaning that a high number of C-N could limit the interaction with the CYP2C9 isoform. In addition to the information overlap with *nRNR2*, *F01[C–N]* also takes into account the information about the presence of primary/secondary amines and of amides.
- *Eta_betaP_A*, is the average measure of π bonds and lone pairs of non-H atoms [49]. It has the lowest loading on PC2 and it increases when increasing the number of multiple bonds and lone pairs. Molecules with a high value of *Eta_betaP_A* tend to be active, confirming the

previous considerations, that is, the higher the unsaturation, the higher the probability of activity towards CYP.

- *MLOGP* is the Moriguchi octanol-water partition coefficient [50]. Drug lipophilicity is known for playing a crucial role in the CYP binding affinity [51], and a significant correlation was reported between the P450 binding affinity and the compound lipophilicity [51–54]. As noted from the PCA, the compounds with high *MLOGP* value are active.
- *HyWi_B(m)* is the hyper-Wiener-like index weighted by mass [10]. It contains information about molecular size, branching, cyclicity and presence of heavy atoms. *HyWi_B(m)* tends to be higher for relatively large and polycyclic compounds, in analogy with the 3A4 isoform. As this descriptor captures several types of chemical information and its contribution is not easily detectable from the PCA, it needs further investigation.

Table 4. List and brief description of the classical molecular descriptors (MDs) selected for CYP2C9. Some examples of molecules with low and high MD values are also reported.

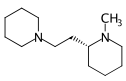
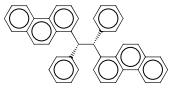
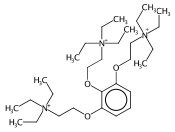
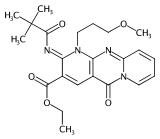
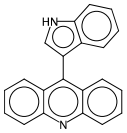
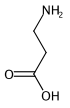
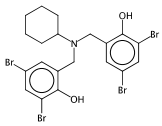
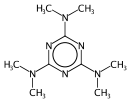
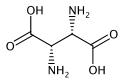
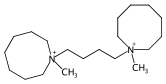
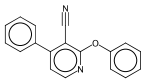
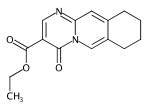
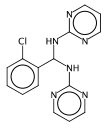
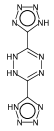
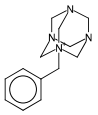
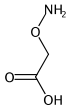
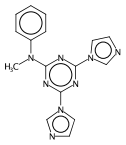
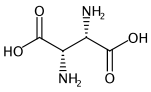
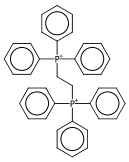
MD	Description	Reference	Model	Low Value		High Value	
nBM	Number of multiple bonds.	[10]	CART		0		44
Sp	Sum of atomic polarizabilities scaled on Carbon atom.	[10]	CART		5.0		56.1
ARR	Ratio between the number of aromatic bonds over the total number of non-H bonds.	[10]	CART		0		0.96
HyWi_B(m)	Hyper-Wiener-like index from Burden matrix weighted by mass.	[55]	*k*-NN		2.3		5.1
GATS2i	Geary autocorrelation of lag 2 weighted by ionization potential.	[56]	*k*-NN		0.09		1.93
Eta_betaP_A	Eta pi and lone pair average VEM count.	[49]	*k*-NN		0		1.02
nPyrimidines	Number of Pyrimidines.	[10]	CART		0		2
nRNR2	Number of aliphatic tertiary amines.	[10]	*k*-NN		0		3

Table 4. *Cont.*

MD	Description	Reference	Model	Low Value		High Value	
F01[C–N]	Frequency of C–N at topological distance 1.	[57]	*k*-NN		0		19
MLOGP	Moriguchi octanol-water partition coefficient.	[50]	*k*-NN		-6.3		9.6

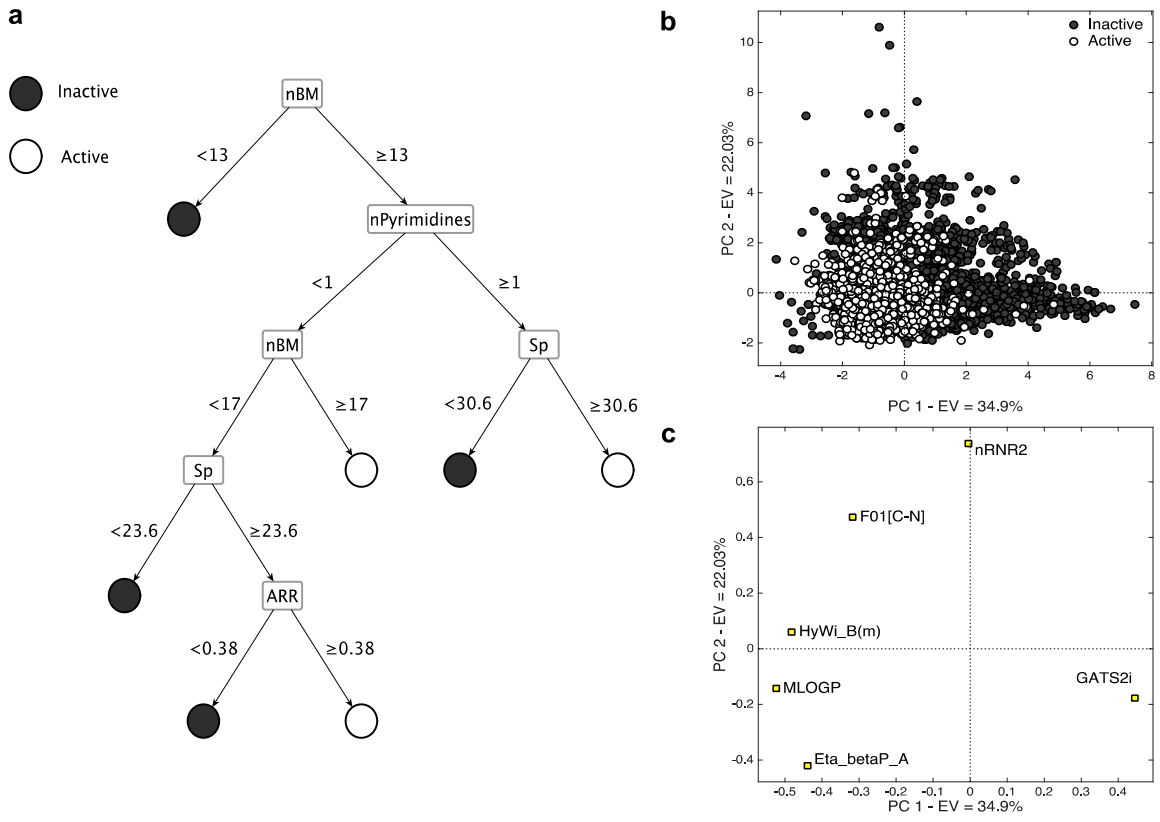


Figure 4. Representation of the MD-based models for CYP2C9: (**a**) CART model; (**b**) Score plot of the *k*-NN descriptors, colored according to their activity; (**c**) Loading plot of the *k*-NN descriptors.

In order to easily interpret information encoded by CYP2C9 fingerprints, the same approach as for 3A4 was used (Figure 5), obtaining in this case a subset of the 16 most relevant molecular fragments (MFs). Also for this isoform, the cyclohexanediol-like MF (No. 1) mainly characterizes the inactive compounds. The quinazolinamine-like MFs (No. 2 and 3) have a higher frequency within inactive structures. This is a relevant difference with respect to the isoform, in which the opposite situation was observed. The remaining MFs are more frequent within the actives class, and the same conclusion as for 3A4 isoform can be drawn.

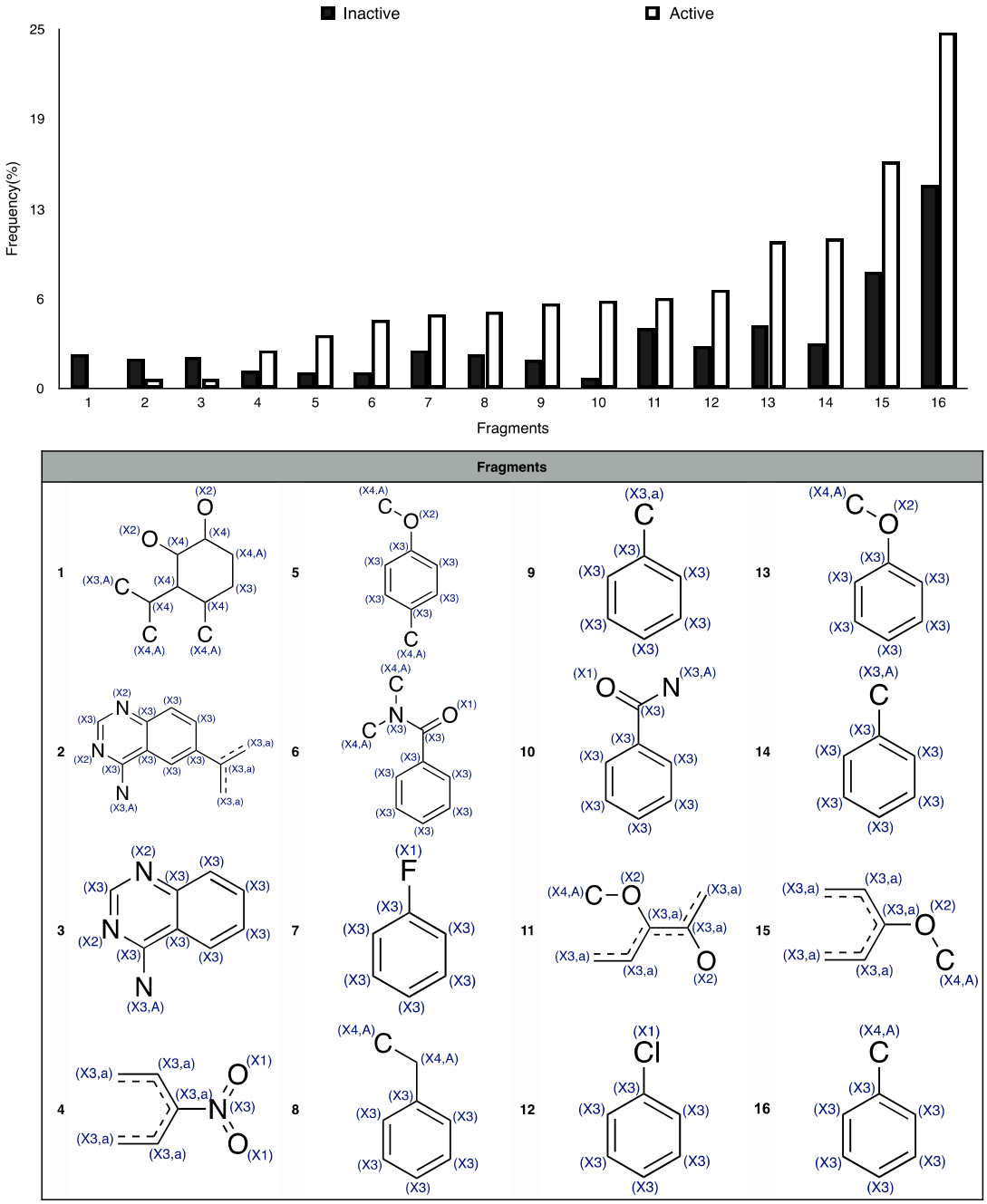


Figure 5. Occurrence frequency of the 16 selected fragments for CYP2C9 within the active/inactive compounds. Symbols associated with the fragments (according to SMARTS language) have the following meaning: *X* + number = number of total bonds in which the considered atom is involved; a = aromatic atom; A = aliphatic atom.

2.3. Consensus Modelling

The consensus approach is well-known in the field of QSAR [58–61] and consists in combining the predictions of several models to maximize their advantages and minimize their drawbacks, aiming to increase the global prediction accuracy. In this work, the selected classification models, which are based on very different descriptors and classification techniques, were merged in two types of *consensus* approaches (Table 5):

1. *Consensus* 1: a molecule was classified if and only if the following conditions were met: (1) all the model predictions agreed in its predicted class; (2) the molecule was inside the AD of all of the models.
2. *Consensus* 2: is based on the majority vote approach, *i.e.*, the compound is classified according to the most frequently predicted class. In this case, the AD of the models used for the prediction was considered.

Table 5. *Consensus* models (*cons*) 3A4 and 2C9. Non-Error Rate (*NER*), Sensitivity (*Sn*), and Specificity (*Sp*) are reported in Fitting, Cross-Validation and on the test set.

CYP	Type	Fitting				Cross-Validation				Test set			
		%na	*NER*	*Sn*	*Sp*	%na	*NER*	*Sn*	*Sp*	%na	*NER*	*Sn*	*Sp*
3A4	*cons 1*	33	0.88	0.92	0.84	33	0.88	0.92	0.84	36	0.88	0.92	0.83
	cons 2	-	0.79	0.80	0.78	-	0.78	0.79	0.77	6	0.80	0.81	0.80
2C9	*cons 1*	33	0.89	0.90	0.88	34	0.89	0.90	0.88	40	0.89	0.89	0.88
	cons 2	-	0.81	0.80	0.82	-	0.81	0.80	0.82	1	0.79	0.77	0.81

For both the isoforms, the same outcome of the consensus approach was reached. Consensus 1 provided increased predictions reliability, especially for the active compounds (*Sn* ranging from 0.89 to 0.92 on the test set), at the expense of a large number of excluded molecules (up to 40% when both the AD and the model disagreement are considered). Consensus 2 showed a *NER* comparable with the single models, but with more balanced *Sn* and *Sp* values and the advantage of providing a prediction for each model.

2.4. External Validation

After recalibrating the models including the test set compounds within the AD domain, their predictivity was further tested on each isoform's external set (Table 6). The single model performances tend to decrease on the external set. In particular, for CYP3A4 individual models, the *Sp* decreases substantially, while *Sn* values are stable; this means that the considered models are more reliable in recognizing the actives. This represents a prominent feature in the field of drug-drug interaction prediction and virtual screening [62,63]. On the contrary, on CYP2C9 single models, a comparable and moderately high decrease of *Sn* and *Sp* values can be noted, with the exception of N3, which has higher *NER* and *Sn* values. The best results are reached by the consensus 1 for both the isoforms, confirming the reliability of these models. For what concerns the consensus 2, they are slightly less accurate than the former case, but the advantage is that the predictions are provided for all the molecules, with a higher predictivity than the single models.

Table 6. Classification results of the models on the external set in terms of *Sn*, *Sp*, *NER* and percentage of not assigned/out of the AD compounds (%na).

CYP	Mod.	Fitting [a]				Cross-Validation [a]				External Set			
		%na	*NER*	*Sn*	*Sp*	%na	*NER*	*Sn*	*Sp*	%na	*NER*	*Sn*	*Sp*
3A4	CART	-	0.75	0.74	0.75	-	0.74	0.73	0.76	-	0.66	0.68	0.63
	k-NN	-	0.76	0.73	0.79	-	0.76	0.73	0.79	1	0.70	0.70	0.69
	N3	-	0.80	0.87	0.73	-	0.79	0.87	0.71	1	0.72	0.85	0.59
	cons 1	32	0.88	0.91	0.85	33	0.88	0.91	0.84	42	0.80	0.89	0.70
	cons 2	-	0.80	0.80	0.79	-	0.79	0.80	0.78	1	0.71	0.76	0.67
2C9	CART	-	0.75	0.77	0.74	-	0.74	0.73	0.76	-	0.66	0.66	0.66
	k-NN	-	0.77	0.69	0.85	-	0.77	0.68	0.85	1	0.69	0.58	0.81
	N3	-	0.80	0.86	0.74	-	0.79	0.85	0.74	-	0.75	0.83	0.68
	cons 1	34	0.89	0.91	0.88	35	0.89	0.90	0.88	45	0.83	0.85	0.82
	cons 2	-	0.81	0.81	0.82	-	0.80	0.79	0.82	1	0.73	0.71	0.75

[a] Performance on the models recalibrated on all the shared set (9122 molecules), *i.e.*, on training and test set compounds.

3. Materials and Methods

3.1. Data Curation

This study is based on the publicly available CYP bioactivity database developed by Veith *et al.* [18] of potency values for 17,143 drug-like compounds for five Cytochrome P450 isoforms (3A4, 2D6, 2C9, 2C19, 1A2), screened using a quantitative High-Throughput Screening with a bioluminescent assay, which recognizes both inhibitors and substrates. The dataset was retrieved from PubChem (AID: 1851), which provides the class of activity (active/inactive) for each compound, identified by a SMILES (Simplified Molecular Input Line Entry System). For each isoform of interest (3A4, 2C9), data were curated by: (a) removing the records without SMILES and/or activity class; (b) removing duplicate structures with mismatching class activity; (c) removing disconnected structures. The comparison between the isoform datasets led to the development of two types of sets (Table 7):

1. Shared set (9122 molecules), comprising the molecule available for both the isoforms. It served to calibrate, validate and choose the optimal model(s) for each isoform.
2. External sets (2996 and 2818 molecules for CYP3A4 and CYP2C9, respectively). These datasets comprise the molecules with annotated activity values for one isoform only. They were used as an external validation tool to further evaluate the model predictivity on unknown molecules.

Table 7. Characteristics of the shared and external sets for each isoform: *n* = number of molecules, %act = percentage of active compounds.

CYP Isoform	Shared Set		External Set	
	n	**%act**	***n***	**%act**
3A4	9122	40	2996	49
2C9	9122	33	2818	36

The shared set of molecules was randomly split into a training set (70%, 6385 compounds) and a test set (30%, 2737 compounds), keeping the active/inactive proportion of both the isoforms (49:100 and 66:100 for 2C9 and 3A4, respectively). The training set was used to select the variables, calibrate the models and cross-validate them (five-fold). The test set served in a later stage to validate the final pool of selected models. The external set was used as a further validation tool only.

3.2. Molecular Descriptors

Two types of molecular descriptors were considered:

1. Classical molecular descriptors (MDs) were calculated using Dragon 6 [20]. From the obtained 3763 MDs, we filtered out those that: (a) were constant or near-constant; (b) had at least one missing value; (c) had a pairwise correlation larger than 0.95. Eventually, 1472 descriptors were retained.
2. Two types of binary fingerprints (FPs) were calculated, namely: (1) extended connectivity (ECFP) [21], circular/topological FPs that encode for several molecular features including stereochemical information; and (2) path connectivity (PFP) [64] based on the presence of particular molecular fragments without accounting for the stereochemical information. A total of 1024 bit FPs were calculated using Dragon 7 [22]. The detailed settings can be found in the Supplementary Material. PFP were considered in the preliminary phase, but their results were not reported as they were outperformed by ECFP.

3.3. Variable Selection

The most relevant MDs were selected using Genetic Algorithms [23], in the version proposed by Leardi *et al.* [65]. For each method, the following strategy was always followed: (1) GA were run on

the pool of MDs (1472); (2) the most frequent MDs of phase 1 (up to a maximum of 200) were subjected to a second selection phase; and (3) the most relevant MDs of phase 2 (up to 15) were tested for all of their possible combinations (All Subset Modelling Strategy, e.g., [66]).

3.4. Applicability Domain (AD) Assessment

The AD approach was rationalized according to the classifier characteristics, as explained below.

1. As CART is based on univariate splits of data, a "bounding box" approach was applied, by excluding the molecules with descriptor values outside the training set ranges (reported as Supplementary Material) [58].
2. For *k*-NN, every compound too far from its *k* neighbors was considered out of the AD, using the approach proposed by Sahigara *et al.* [67]. The same distance metric of the best *k*-NN models (Euclidean distance) was used.
3. For what concerns N3, as the largest contribute to the classification is given by the closest compound, a query was considered inside the AD if its similarity with its nearest neighbor was larger than or equal to 0.2. The same similarity metric of best models (Jaccard-Tanimoto) was used.

3.5. Software and Code

Data curation was performed using KNIME [68] v 2.11.3 workflows. Training/test splitting, variable selection, cross-validation and model calibration were performed in MATLAB [69] environment, using routines written by Milano Chemometrics and QSAR Research Group [70].

4. Conclusions

The overall goal of this work was to obtain reliable QSAR screening systems for two of the most relevant isoforms of Cytochrome P450, namely CYP2C9 and CYP3A4.

The study was based on the CYP bioactivity database developed by the National Institutes of Health Chemical Genomics Center (NCGC), from which activity data for 14,936 molecules were obtained and used to train/validate the models.

Different classification approaches were tested on both classical molecular descriptors and binary fingerprints. For each isoform, three models led to the optimal results: (1) a classification tree, characterized by high interpretability and prediction balance on the classes; (2) a *k*-Nearest Neighbor model, with high complexity but also high predictivity towards inactive compounds; and (3) a *N*-Nearest Neighbor model, which had the highest performance towards the actives.

The interpretation of the selected molecular descriptors made it possible to gather insights into the structural features that determine the activity on Cytochrome P450. In particular, for both the isoforms, molecular dimension and flexibility ultimately influenced the activity towards CYP, in that small and flexible molecules tend to be inactive, while lipophilic compounds tend to be active. As for CYP3A4, distinctive features of the active molecules turned out to be the presence of hydroxyl groups, and/or a low polarizability/ionization potential. Regarding the CYP2C9, the analysis underscored the fact that active molecules tend to have a large number of aromatic bonds, along with a high polarizability, a high unsaturation degree and often the presence of pyrimidines. Moreover, the interaction with CYP2C9 could be limited by a high number of bonded C and N atoms in the molecule.

As a final classification step, the models were aggregated in a *consensus* manner. This allowed us to obtain a system with a good predictive ability towards both the activity classes and to exploit the advantages of the single models. Finally, the models were validated using more than 2000 molecules for each isoform as an external set. The external validation confirmed the model reliability and stability. A future perspective will be the development of analogous approaches for the remaining CYP isoforms of relevance.

Supplementary Materials: The curated datasets along with the calculated molecular descriptors and other supplementary materials can be found at http://www.mdpi.com/1422-0067/17/6/914/s1. The dataset can be also freely downloaded from Milano Chemometrics website (http://michem.disat.unimib.it/chm/).

Author Contributions: Serena Nembri and Francesca Grisoni conceived the study. Serena Nembri curated the dataset and performed the calculations. Serena Nembri and Francesca Grisoni analyzed and interpreted the results, and wrote the paper. Viviana Consonni and Roberto Todeschini supervised the study and reviewed the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AD	Applicability Domain
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
CART	Classification and Regression Trees
CYP	Cytochrome P450
FP	binary Fingerprints
ECFP	Extended Connectivity Fingerprints
k-NN	*k*-Nearest Neighbors
MD	classical Molecular Descriptors
MF	Molecular Fragment
N3	*N*-Nearest neighbors
NER	Non-Error Rate
QSAR	Quantitative Structure-Activity Relationship
SMILES	Simplified Molecular Input Line Entry System
Sn	Sensitivity
Sp	Specificity

References

1. Munro, A.W.; Girvan, H.M.; Mason, A.E.; Dunford, A.J.; McLean, K.J. What makes a P450 tick? *Trends Biochem. Sci.* **2013**, *38*, 140–150. [CrossRef] [PubMed]
2. Yan, Z.; Caldwell, G.W. Metabolism Profiling, and Cytochrome P450 inhibition & induction in drug discovery. *Curr. Top. Med. Chem.* **2001**, *1*, 403–425. [PubMed]
3. Singh, D.; Kashyap, A.; Pandey, R.V.; Saini, K.S. Novel advances in cytochrome P450 research. *Drug Discov. Today* **2011**, *16*, 793–799. [CrossRef] [PubMed]
4. Zanger, U.M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **2013**, *138*, 103–141. [CrossRef] [PubMed]
5. Pb, W. Role of cytochromes P450 in drug metabolism and hepatotoxicity. *Semin. Liver Dis.* **1990**, *10*, 235–250.
6. Gonzalez, F.J.; Gelboin, H.V. Role of human cytochromes P450 in the metabolic activation of chemical carcinogens and toxins. *Drug Metab. Rev.* **1994**, *26*, 165–183. [CrossRef] [PubMed]
7. Gonzalez, F.J. Role of cytochromes P450 in chemical toxicity and oxidative stress: Studies with CYP2E1. *Mutat. Res. Mol. Mech. Mutagen.* **2005**, *569*, 101–110. [CrossRef] [PubMed]
8. Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Deliv. Rev.* **2002**, *54*, 407–415. [CrossRef]
9. Kirton, S.B.; Baxter, C.A.; Sutcliffe, M.J. Comparative modelling of cytochromes P450. *Adv. Drug Deliv. Rev.* **2002**, *54*, 385–406. [CrossRef]
10. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 41.
11. Li, H.; Sun, J.; Fan, X.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput. Aided Mol. Des.* **2008**, *22*, 843–855. [CrossRef] [PubMed]
12. Yap, C.W.; Chen, Y.Z. Prediction of Cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992. [CrossRef] [PubMed]
13. Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: Prediction of cytochromes P450 inhibition. *Bioinformatics* **2013**, *29*, 2051–2052. [CrossRef] [PubMed]
14. Pan, X.; Chao, L.; Qu, S.; Huang, S.; Yang, L.; Mei, H. An improved large-scale prediction model of CYP1A2 inhibitors by using combined fragment descriptors. *RSC Adv.* **2015**, *5*, 84232–84237. [CrossRef]

15. Sun, H.; Veith, H.; Xia, M.; Austin, C.P.; Huang, R. Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51*, 2474–2481. [CrossRef] [PubMed]
16. Jensen, B.F.; Vind, C.; Brockhoff, P.B.; Refsgaard, H.H.F. *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. *J. Med. Chem.* **2007**, *50*, 501–511. [CrossRef] [PubMed]
17. Balakin, K.V.; Ekins, S.; Bugrim, A.; Ivanenkov, Y.A.; Korolev, D.; Nikolsky, Y.V.; Skorenko, A.V.; Ivashchenko, A.A.; Savchuk, N.P.; Nikolskaya, T. Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metab. Dispos.* **2004**, *32*, 1183–1189. [CrossRef] [PubMed]
18. Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C.P.; Lloyd, D.G.; *et al.* Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.* **2009**, *27*, 1050–1055. [CrossRef] [PubMed]
19. NCBI The PubChem Project. Available online: http://pubchem.ncbi.nlm.nih.gov/ (accessed on 1 March 2016).
20. Talete srl. Dragon (Software for Molecular Descriptor Calculation). Version 6.0. 2012. Available online: http://www.talete.mi.it/(accessed on 8 June 2016).
21. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef] [PubMed]
22. Kode srl. *Dragon (Software for Molecular Descriptor Calculation)*, Version 7.0. Pisa, Italy, 2016.
23. Goldberg, D.E.; Holland, J.H. Genetic algorithms and machine learning. *Mach. Learn.* **1988**, *3*, 95–99. [CrossRef]
24. Grisoni, F.; Cassotti, M.; Todeschini, R. Reshaped sequential replacement for variable selection in QSPR: Comparison with other reference methods. *J. Chemom.* **2014**, *28*, 249–259. [CrossRef]
25. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
26. Kowalski, B.R.; Bender, C.F. *K*-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.* **1972**, *44*, 1405–1411. [CrossRef]
27. Todeschini, R.; Ballabio, D.; Cassotti, M.; Consonni, V. N3 and BNN: Two new similarity based classification methods in comparison with other classifiers. *J. Chem. Inf. Model.* **2015**, *55*, 2365–2374. [CrossRef] [PubMed]
28. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley: Hoboken, NJ, USA, 1992.
29. Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196. [CrossRef]
30. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791–4810. [CrossRef] [PubMed]
31. Jolliffe, I. Principal component analysis. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Boca Raton, FL, USA, 2014.
32. Smith, D.A.; Jones, B.C.; Walker, D.K. Design of drugs involving the concepts and theories of drug metabolism and pharmacokinetics. *Med. Res. Rev.* **1996**, *16*, 243–266. [CrossRef]
33. De Montellano, P.R.O. *Cytochrome P450: Structure, Mechanism, and Biochemistry*; Springer Science & Business Media: Berlin, Germany; Heidelberg, Germany, 2005.
34. Guengerich, F.P. Mechanisms of cytochrome P450 substrate oxidation: MiniReview. *J. Biochem. Mol. Toxicol.* **2007**, *21*, 163–168. [CrossRef] [PubMed]
35. Guengerich, F.P. Oxidation of halogenated compounds by cytochrome P-450, peroxidases, and model metalloporphyrins. *J. Biol. Chem.* **1989**, *264*, 17198–17205.
36. Meunier, B.; Visser, S.P.D. Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. *Chem. Rev.* **2004**, *104*, 3947–3980. [CrossRef] [PubMed]
37. Mutschler, E.; Derendorf, H. *Drug Actions: Basic Principles and Theraputic Aspects*; CRC Press: Boca Raton, FL, USA, 1995.
38. Guroff, G.; Daly, J.W.; Jerina, D.M.; Renson, J.; Witkop, B.; Udenfriend, S. Hydroxylation-induced migration: The NIH shift. Recent experiments reveal an unexpected and general result of enzymatic hydroxylation of aromatic compounds. *Science* **1967**, *157*, 1524–1530. [CrossRef] [PubMed]

39. Magnuson, V.R.; Harriss, D.K.; Basak, S.C. Topological indices based on neighborhood symmetry: Chemical and biological applications. *Chem. Appl. Topol. Graph Theory* **1983**, 178–191.
40. Raunio, H.; Kuusisto, M.; Juvonen, R.O.; Pentikäinen, O.T. Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes. *Front. Pharmacol.* **2015**, *6*. [CrossRef] [PubMed]
41. Jiao, D.; Golubkov, P.A.; Darden, T.A.; Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6290–6295. [CrossRef] [PubMed]
42. Stauffer, D.A.; Karlin, A. Electrostatic potential of the acetylcholine binding sites in the nicotinic receptor probed by reactions of binding-site cysteines with charged methanethiosulfonates. *Biochemistry* **1994**, *33*, 6840–6849. [CrossRef] [PubMed]
43. Butina, D. Performance of Kier-hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules. *Molecules* **2004**, *9*, 1004–1009. [CrossRef] [PubMed]
44. Gunes, A.; Coskun, U.; Boruban, C.; Gunel, N.; Babaoglu, M.O.; Sencan, O.; Bozkurt, A.; Rane, A.; Hassan, M.; Zengil, H.; *et al.* Inhibitory effect of 5-fluorouracil on cytochrome P450 2C9 activity in cancer patients. *Basic Clin. Pharmacol. Toxicol.* **2006**, *98*, 197–200. [CrossRef] [PubMed]
45. Gilbar, P.J.; Brodribb, T.R. Phenytoin and fluorouracil interaction. *Ann. Pharmacother.* **2001**, *35*, 1367–1370. [CrossRef] [PubMed]
46. Brown, M.C. An adverse interaction between warfarin and 5-fluorouracil: A case report and review of the literature. *Chemotherapy* **1999**, *45*, 392–395. [CrossRef] [PubMed]
47. Stiborová, M.; Bieler, C.A.; Wiessler, M.; Frei, E. The anticancer agent ellipticine on activation by cytochrome P450 forms covalent DNA adducts. *Biochem. Pharmacol.* **2001**, *62*, 1675–1684. [CrossRef]
48. Beedham, C. The role of non-P450 enzymes in drug oxidation. *Pharm. World Sci.* **1997**, *19*, 255–263. [CrossRef] [PubMed]
49. Roy, K.; Saha, A. QSPR with TAU indices: Water solubility of diverse functional acyclic compounds. *Intern. Electron. J. Mol. Des.* **2003**, *2*, 475–491.
50. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127–130. [CrossRef]
51. Lewis, D.F.V.; Jacobs, M.N.; Dickins, M. Compound lipophilicity for substrate binding to human P450s in drug metabolism. *Drug Discov. Today* **2004**, *9*, 530–537. [CrossRef]
52. Hansch, C.; Zhang, L. Quantitative structure-activity relationships of cytochrome P-450. *Drug Metab. Rev.* **1993**, *25*, 1–48. [CrossRef] [PubMed]
53. Al-Gailany, K.A.S.; Houston, J.B.; Bridges, J.W. The role of substrate lipophilicity in determining type 1 microsomal P450 binding characteristics. *Biochem. Pharmacol.* **1978**, *27*, 783–788. [CrossRef]
54. Ramos-Nino, M.E.; Clifford, M.N.; Adams, M.R. Quantitative structure activity relationship for the effect of benzoic acids, cinnamic acids and benzaldehydes on Listeria monocytogenes. *J. Appl. Bacteriol.* **1996**, *80*, 303–310. [CrossRef] [PubMed]
55. Consonni, V.; Todeschini, R. New spectral indices for molecule description. *MATCH* **2008**, *1*, 2.
56. Geary, R.C. The contiguity ratio and statistical mapping. *Inc. Stat.* **1954**, *5*, 115–146. [CrossRef]
57. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73. [CrossRef]
58. Grisoni, F.; Consonni, V.; Vighi, M.; Villa, S.; Todeschini, R. Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environ. Int.* **2016**, *88*, 198–205. [CrossRef] [PubMed]
59. Cassotti, M.; Consonni, V.; Mauri, A.; Ballabio, D. Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards Daphnia magna. *SAR QSAR Environ. Res.* **2014**, *25*, 1013–1036. [CrossRef] [PubMed]
60. Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A.M.; Grulke, C.M.; *et al.* CERAPP: Collaborative estrogen receptor activity prediction project. *Environ. Health Perspect.* **2016**. [CrossRef] [PubMed]
61. Gissi, A.; Nicolotti, O.; Carotti, A.; Gadaleta, D.; Lombardo, A.; Benfenati, E. Integration of QSAR models for bioconcentration suitable for REACH. *Sci. Total Environ.* **2013**, *456–457*, 325–332. [CrossRef] [PubMed]
62. Fp, G. Role of cytochrome P450 enzymes in drug-drug interactions. *Adv. Pharmacol.* **1996**, *43*, 7–35.
63. Walters, W.P.; Stahl, M.T.; Murcko, M.A. Virtual screening—An overview. *Drug Discov. Today* **1998**, *3*, 160–178. [CrossRef]

64. James, C.A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, USA, 1995; 3951p.
65. Leardi, R.; Lupiáñez González, A. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195–207. [CrossRef]
66. Cassotti, M.; Grisoni, F. Variable selection methods: An introduction. Available online: http://www.moleculardescriptors.eu (accessed on 2 February 2016).
67. Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a novel *k*-Nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J. Cheminform.* **2013**, *5*, 27. [CrossRef] [PubMed]
68. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The konstanz information miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, P.D.H., Schmidt-Thieme, P.D.L., Decker, P.D.R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer: Berlin, Germany; Heidelberg, Germany, 2008; pp. 319–326.
69. *MATLAB, 2015. R2015a*; The MathWorks Inc.: Natick, MA, USA, 2015.
70. Milano Chemometrics and QSAR Research Group. Available online: http://michem.disat.unimib.it/chm/download/datasets.htm (accessed on 3 March 2016).