Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates

Lisa Michielan,[†] Lothar Terfloth,[‡] Johann Gasteiger,*[,‡,§] and Stefano Moro*[,†]

Molecular Modeling Section (MMS), Dipartimento di Scienze Farmaceutiche, Università di Padova, via Marzolo 5, I-35131, Padova, Italy, Molecular Networks GmbH, Henkestrasse 91, D-91052, Erlangen, Germany, and Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052, Erlangen, Germany

Received August 10, 2009

Each drug can potentially be metabolized by different CYP450 isoforms. In the development of new drugs, the prediction of the metabolic fate is important to prevent drug−drug interactions. In the present study, a collection of 580 CYP450 substrates is deeply analyzed by applying multi- and single-label classification strategies, after the computation and selection of suitable molecular descriptors. Cross-training with support vector machine, multilabel *k*-nearest-neighbor and counterpropagation neural network modeling methods were used in the multilabel approach, which allows one to classify the compounds simultaneously in multiple classes. In the single-label models, automatic variable selection was combined with various cross-validation experiments and modeling techniques. Moreover, the reliability of both multi- and single-label models was assessed by the prediction of an external test set. Finally, the predicted results of the best models were compared to show that, even if the models present similar performances, the multilabel approach more coherently reflects the real metabolism information.

INTRODUCTION

In the drug discovery process, reliable chemoinformatic strategies can successfully contribute to resource optimization. In particular, the early detection of ADMET (absorption, distribution, metabolism, elimination, toxicity) properties of drugs under in vivo conditions is experimentally time-consuming and expensive; therefore, computational methods can profitably speed the collection of new data.[1−3] Currently, efficient in silico tools are increasingly employed in the development of drug candidates.

The metabolic profile of a drug candidate is an important aspect to be considered in the selection of a potential new drug. Several problems related to stability, toxicity of xenobiotics, and drug interactions might represent serious adverse effects. In fact, in the coadministration of drugs, the pharmacological profile of each drug might be modified by the presence of other drugs in the human body.[4,5] For example, if two drugs are coadministered and both are metabolized by the same enzyme, the competition for the binding site can result in the inhibition of the biotransformation of one or both drugs.

Focusing attention on metabolism in the ADMET process, a crucial role is played by the cytochrome P450 (CYP450) class of hemoprotein enzymes. The CYP450 superfamily of enzymes is abundantly expressed in the liver and, remarkably, in the small intestine, where it is responsible for the detoxification of xenobiotics.[6,7] In phase I metabolism, cytochrome P450 isoforms chemically modify a large variety of substrates mainly through oxidation reactions to make them more water-soluble and easier to eliminate.[8] However, this detoxification system is highly complex, as it includes many different cytochrome P450 isoforms characterized by multiple binding sites, polymorphism, and enzyme induction or modulation phenomena.[9] These aspects are involved in drug−drug interactions, which might lead to unpredictable blood concentrations of one or more xenobiotics, with consequent possible toxic effects or loss of activity. Cytochrome P450 enzymes are classified in several isoforms according to the similarity of their amino acidic sequences. In the present work, CYP450 1A2, 2C19, 2C8, 2C9, 2D6, 2E1, and 3A4 substrates were investigated because they cover almost all possible metabolism routes. Cytochrome P450 1A2 is ubiquitously expressed but shows a high concentration in the liver, where it metabolizes xenobiotics and endogenous compounds. In general, these substrates are planar molecules characterized by moderate volume and basicity.[6] The substrates of cytochrome P450 2C19, found in the liver and in the heart, are weakly acidic or neutral molecules and belong to several chemical classes,[6] whereas CYP450 2C8, mainly expressed in the liver, is responsible for the metabolism of some drugs with high volume (paclitaxel and statines). CYP1A2, CYP2C19, and CYP2C8 substrates are similar to CYP3A4 substrates, and no typical scaffold, unlike for the other class memberships, can be distinguished. CYP450 2C9 is abundantly located in the liver, where it is known to metabolize neutral or acidic and lipophilic molecules, particularly the sulfonylurea and NSAID (nonsteroidal anti-inflammatory drug) drug classes.[6] CYP450 2D6 shows polymorphism and is expressed in the liver, heart, brain, and mucosal enterocytes at low levels. Despite its low concentration, about 25% of all drugs are at least partial substrates of

* Corresponding author phone: +39 049 8275704; fax: +39 049 8275366; e-mail: stefano.moro@unipd.it.
[†] Università di Padova.
[‡] Molecular Networks GmbH.
[§] Universität Erlangen-Nürnberg.

10.1021/ci900299a CCC: $40.75
Published on Web 11/02/2009

this isoform. They show a hydrophilic character and have a basic nitrogen atom.[6] Mostly small and polar molecules, such as volatile anesthetics, are substrates of the CYP450 2E1 isoform, which is involved in many drug interactions.[6] The CYP450 3A4 isoform, ubiquitously found, is responsible for the metabolism of high-volume and lipophilic xenobiotics; indeed, almost 50% of all drugs are metabolized by this isoform.[6] For this reason, its activity is strongly affected by chemically different compounds, and CYP3A4 represents the most populated class in the data set investigated in this study. Moreover, different isoforms might be responsible for detoxification of the same drug.

A challenging problem in this field is the prediction of isoform specificity. The development of a model able to classify a compound according to the isoform by which it is metabolized and to anticipate the drug metabolic profile based on various molecular properties is a matter of wide interest. Several chemoinformatic tools have already been attempted for the prediction of CYP-related metabolism properties, as reported by Crivori et al. and reviewed by Li et al.[10,11]

Any classification model requires a mathematical representation of molecules through the computation of structure or property descriptors. In this work, molecular descriptors reflecting shape and acid−base properties, as well as the distribution of electrostatic properties on the 3D molecular surface, are used to encode the structure of CYPs substrates.

So far, several ligand-based approaches have been applied to classify CYP450 substrates according to their route of metabolism.[12−19] Among these, an attractive visualization system was published by Yamashita et al.,[12] but a small data set was used in the analysis. Block and Henry[13] built independent classifiers, whereas Yap and Chen[17] developed a classification model considering three nonoverlapping classes. Nevertheless, most solutions consider local models for each CYP450 isoform and do not approach the problem globally.

The traditional single-label classification approach deals exclusively with nonoverlapping classes. Following this approach, we recently developed a classification model to predict the isoform specificity for CYP3A4, CYP2D6, and CYP2C9 substrates considering nonoverlapping classes, that is, assuming each compound to be metabolized by a single, predominant CYP450 isoform.[20] Here, we aim to extend our previous model to cover other CYP450 isoforms and to find a strategy for predicting the substrates that are metabolized by more than one isoform. Such a multilabel classification analysis represents a different approach that can be applied regardlwhether our data set comprises elements assigned simultaneously to more than one class. The prediction of the metabolism profile of CYP450 substrates represents a novel application of this methodology.

A counterpropagation neural network (CPG-NN) is a powerful and widely applied technique for solving several classification problems.[21,22] More recently, two alternative multilabel classification methods, cross-training with support vector machines (ct-SVM) and multilabel *k*-nearest-neighbor (M_L-*k*NN) analysis, were reported.[23]

The present work aimed at the prediction of the isoform specificity from information on the substrates of these isoforms. The substrates were represented by different sets of structural and physicochemical descriptors. Then, various classification techniques—both multilabel and single-label approaches—were applied to derive models for the prediction of the isoform(s) responsible for the metabolism of CYP450 1A2, 2C19, 2C8, 2C9, 2D6, 2E1, and 3A4 substrates.

In our analysis, we followed three steps using different techniques and data sets to find a robust model. First, we applied a multilabel classification method to distinguish between substrates of seven CYP450 isoforms. In a second step, we reduced the number of isoforms to five and explored the prediction capability of the new multilabel model. Finally, we simplified the problem further and used only single-label compounds (i.e., compounds metabolized by a single isoform) to build a classifier. Variable selection and optimization of the models were performed in the present study. The procedure that we followed in the multilabel approach is illustrated in Figure 1. The single-label classification models were directly generated after a variable selection process, as summarized in Figure 2. The modeling results and metabolism profiles of the CYP450 substrates of the test sets predicted by the three different approaches were compared to verify the reliability of both single- and multilabel classification methods.

DATA SET PREPARATION

In the present study, a collection of 580 cytochrome P450 substrates with different chemical structures was used to derive our classification models. In particular, we considered the data set compiled in the recently published work by Block and Henry.[13] It includes 253 substrates metabolized by CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, and CYP3A4 isoforms. At least one isoform might be responsible for the metabolism of a single substrate. Because we considered all possible routes of metabolism, our classification problem deals with overlapping classes; that is, a compound might be metabolized by several isoforms, so that it belongs to several classes. Therefore, this classification problem is called multilabel. Further substrates were extracted from the work by Bonnabry et al.[24] (22 compounds metabolized by CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 isoforms) and from the list publicly available on the Web and published later by Pharmacist's Letter (14 substrates metabolized by CYP1A2, CYP2C9, CYP2D6, and CYP3A4 isoforms).[25,26] In addition, we considered 24 compounds used to build our previous model, extracted from the work published by Manga et al.[27] Each of these 24 compounds is a substrate of only CYP2C9, CYP2D6, or CYP3A4. The complete list of drugs and their classification by CYP450 isoform is provided as Supporting Information. Moreover, 267 additional compounds from the Metabolite reaction database were included in the data set in the following way: Metabolic reactions reported for the human body were exported from the Metabolic database.[28] Only the reactants metabolized by one CYP450 isoform, corresponding to a specific class and not present in the previously mentioned data sets, were extracted.

In some cases, we had to deal with inconsistent information, because some compounds were reported as substrates in a previously published data set and were not classified as substrates in the work by Block and Henry.[13] For this reason, these compounds were discarded from the final data set. In general, when inconsistencies were found in comparisons

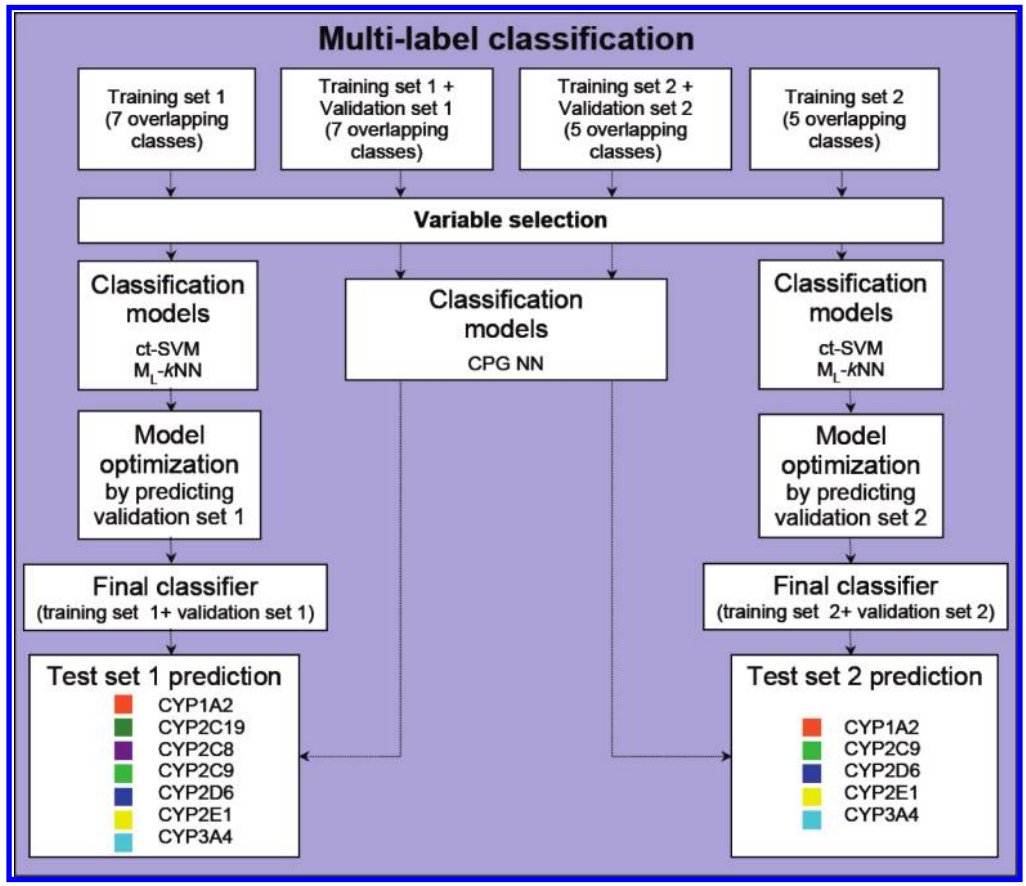

Figure 1. Flowchart of the multilabel approach for the prediction of CYP450 isoform specificity.

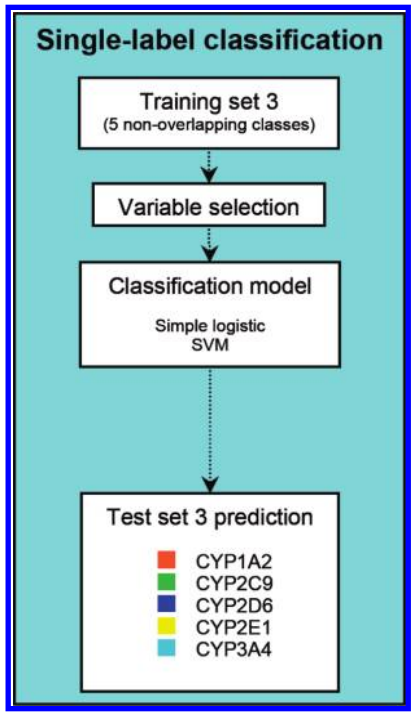

Figure 2. Flowchart of the single-label approach for the prediction of CYP450 isoform specificity.

with the other data sets, we considered as more reliable the information about a compound's metabolic fate published in the more recent data set by Block and Henry.[13]

In our final collection (580 compounds), 488 substrates are metabolized by one CYP450 isoform, and the remaining 92 compounds are metabolized by several CYPs (two or more, up to five different isoforms). Only 16% of all compounds in the data set are multilabeled data. Considering the 488 compounds metabolized by one isoform, 46 are CYP1A2 substrates (9%), 15 are CYP2C19 substrates (3%), 11 are CYP2C8 substrates (3%), 45 are CYP2C9 substrates (9%), 105 are CYP2D6 substrates (22%), 48 are CYP2E1 substrates (10%), and 218 are CYP3A4 substrates (44%). Not all isoforms have the same relevance in the xenobiotic metabolism; consequently, these classes are differently populated, and our data set is quite unbalanced. After the compound names and information about their metabolism had been extracted from the electronic version of the articles, the correct stereochemistry was assigned to the substrates. For this purpose, the PubChem and DrugBank databases were searched to confirm the 2D structures and stereochemistries.[29,30]

For our modeling studies, three different data sets were compiled. Two data sets were manually split into training, validation, and test sets, with similar distributions of the considered classes in the entire data set and the subsets. Most of the compounds from the Metabolite data set were used as the test set, even if some substrates were included in the training and validation sets. In particular, 173, 165, and 163 of the 267 compounds from the Metabolite data set were used in test set 1, test set 2, and test set 3, respectively. The last data set was simply split into a training and a validation set, by application of the same selection criterion (similar distribution of the substrates in the considered classes).

Data Set 1. The initial collection comprised 580 chemically different substrates, metabolized by seven CYP450 isoforms (CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 single- and multilabel substrates).

Table 1. Data Set 1 Used in This Analysis, Including the Distribution of the Substrates into Training, Validation, and Test Sets:[a] 580 Single- and Multilabel CP450 Substrates Classified in Seven Isoforms (CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, and CYP3A4)

	training set 1		validation set 1		test set 1	
isoform	single-label	multilabel	single-label	multilabel	single-label	multilabel
CYP1A2	26	17	6	5	14	9
CYP2C19	7	18	3	1	5	5
CYP2C8	6	12	2	2	3	0
CYP2C9	27	17	8	3	10	8
CYP2D6	45	25	7	8	53	11
CYP2E1	29	3	8	0	11	1
CYP3A4	95	47	21	10	102	14
total (580)	235	61	55	12	198	19

[a] Single-label substrates occur only once in each class; multilabel substrates belong to more than one class. Consequently, the sum of multilabel substrates for all the classes within a column is higher than the number of multilabel substrates for training set 1, validation set 1, or test set 1.

Table 2. Data Set 2 Used in This Analysis, Including the Distribution of the Substrates in Training, Validation, and Test Sets:[a] 554 Single- and Multilabel CP450 Substrates Classified in Five Isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1 and CYP3A4)

	training set 2		validation set 2		test set 2	
isoform	single-label	multilabel	single-label	multilabel	single-label	multilabel
CYP1A2	26	17	6	5	14	9
CYP2C9	31	13	8	3	11	7
CYP2D6	46	24	7	8	53	11
CYP2E1	30	2	8	0	11	1
CYP3A4	110	32	21	10	102	14
total (554)	243	40	50	12	191	18

[a] Single-label substrates occur only once in each class; multilabel substrates belong to more than one class. Consequently, the sum of multilabel substrates for all the classes within a column is higher than the number of multilabel substrates for training set 2, validation set 2, or test set 2. In comparison to data set 1, the distribution of single- and multilabel substrates is different, because some multilabel substrates changed into single-label ones.

This collection was split into training, validation, and test sets, in order to have similar distributions of the substrates for each class membership. Data set 1 was used to perform a multilabel classification analysis. The training, validation, and test sets selected for multilabel models using seven classes are reported in Table 1.

Data Set 2. All CYP2C19 and CYP2C8 single-label substrates were removed from data set 1. These classes are not very populated, and they were discarded for the following analysis. Thus, a new data set comprising 554 chemical structures metabolized by five different isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 single- and multilabel substrates) was obtained. Obviously, the exclusion of two classes from data set 2 changed the distribution of single- and multilabel substrates. In fact, some of the multilabel compounds turned into single-label ones, but the total number of substrates in the training, validation, and test sets remained the same for the five classes. A multilabel classification approach was applied to data set 2. The membership distribution in the training, validation, and test sets is summarized in Table 2.

Data Set 3. Finally, only the single-label substrates in data set 2 were selected to perform a single-label classification analysis. These 484 compounds, a subset of data set 2, were used as data set 3. In the new training set (training set 3), exactly the same single-label substrates collected in training set 2 and validation set 2 (a total of 293 compounds) were included, and in test set 3, only the single-label substrates in test set 2 (191 compounds) were considered. Data set 3 was utilized to develop a single-label classification model. The results of the splitting process for data set 3 can be inferred by considering the single-label columns for the training, validation, and test sets in Table 2.

COMPUTATIONAL METHODOLOGIES

All modeling studies in this work were carried out using the following software packages: Structure management and editing were performed with the CACTVS system.[31] ADRIANA.Code (version 2.2) was used for the calculation of molecular descriptors.[32] Cross-training SVM (ct-SVM) and multilabel *k*-nearest-neighbor (M_L-*k*NN) models were generated with R software (package e1071).[33,34] The counterpropagation neural network (CPG-NN) analysis was performed using SONNIA software.[35] Single-label classification models were built using Weka.[36,37]

Molecular Structure Building. Three-dimensional models of all substrates were obtained using the 3D structure generator CORINA,[38] also an integral part of the ADRIANA.Code suite. Conformer generation and best-conformer selection were performed using standard parameters of CORINA. Insufficient knowledge is available about the binding mode of the substrates within the substrate−CYP450 isoform complex, so the single, low-energy conformation produced by CORINA was selected for each compound. The chemical structures were assumed to be neutral, and it was verified that no charged fragments were present in our training, validation, and test sets.

Molecular Descriptor Calculation. Various combinations of molecular descriptors were selected to compute our

Table 3. List of Descriptors Used in the Analysis, Arranged by Class

no.	name	details	ref(s)
		Global	
1	MW	molecular weight	
2	HAccPot	highest hydrogen-bond acceptor potential	39
3	HDonPot	highest hydrogen-bond donor potential	39
4	HAcc	number of hydrogen-bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule	39
5	HDon	number of hydrogen-bonding donor derived from the sum of NH and OH groups in the molecule	39
6	TPSA	topological polar surface area	40
7	ASA	approximate surface area	41
8	α	mean molecular polarizability	42−45
9	μ	molecular dipole moment	46
		Topological	
10, 11	χ^0, χ^1	connectivity χ indices	47
12, 13	κ_1, κ_2	κ shape indices	47
14	W	Wiener path number	48
15	χ^R	Randic index	46
		Size/Shape	
16	D_3	diameter	49
17	R_3	radius	48
18	I_3	geometric shape coefficient	49, 50
19	r_2	radius perpendicular to D_3	
20	r_3	radius perpendicular to D_3 and r_2	
21−23	λ_1, λ_2, λ_3	principal moments of inertia	46
24	r_{gyr}	radius of gyration	51, 52
25	r_{span}	radius of the smallest sphere, centered at the center of mass which completely encloses all atoms in the molecule	52
26	ϵ	molecular eccentricity	46
27	Ω	molecular asphericity	46
		Functional-Group Counts	
28	n_{aliph_amino}	number of aliphatic amino groups	
29	n_{aro_amino}	number of aromatic amino groups	
30	n_{prim_amino}	number of primary aliphatic amino groups	
31	n_{sec_amino}	number of secondary aliphatic amino groups	
32	n_{tert_amino}	number of tertiary aliphatic amino groups	
33	$n_{prim\_sec\_amino}$	$n_{prim\_amino} + n_{sec\_amino}$	
34	$n_{aro\_hydroxy}$	number of aromatic hydroxy groups	
35	$n_{aliph\_hydroxy}$	number of aliphatic hydroxy groups	
36	$n_{guanidine}$	number of guanidine groups	
37	$n_{basic\_nitrogen}$	number of basic, nitrogen-containing functional groups	
38	n_{acidic_groups}	number of acidic functional groups	
39	$n_{acylsulfonamides}$	number of sulfonamide-C=O groups	
40	$n_{enolate\_groups}$	number of enolate groups	
		Vectorial	
41−51	2D-ACχ^{LP}	topological autocorrelation; property: lone-pair electronegativity χ^{LP}	
52−62	2D-ACχ_σ	topological autocorrelation; property: σ-electronegativity χ^σ	
63−73	2D-ACχ_π	topological autocorrelation; property: π-electronegativity χ^π	
74−84	2D-ACq_σ	topological autocorrelation; property: σ-charge q^σ	
85−95	2D-ACq_π	topological autocorrelation; property: π-charge q^π	
96−106	2D-ACq_{tot}	topological autocorrelation; property: total charge q_{tot}	
107−117	2D-AC$_\alpha$	topological autocorrelation; property: polarizability α	
118−245	3D-AC$_{identity}$	spatial autocorrelation; property: identity	
246	$\chi_\sigma\_1 = \sum\chi_\sigma^2$	property: σ-electronegativity χ^σ	
247	$\chi_\pi\_1 = \sum\chi_\pi^2$	property: π-electronegativity χ^π	
248	$q_\sigma\_1 = \sum q_\sigma^2$	property: σ-charge q^σ	
249	$q_\pi\_1 = \sum q_\pi^2$	property: π-charge q^π	
250−261	SurfACorr_ESP	spatial autocorrelation; property: molecular electrostatic potential	
262−273	SurfACorr_HBP	spatial autocorrelation; property: hydrogen-bonding potential	

classification models. Simple descriptors require knowledge of only the code of a molecule and consider the presence of a particular element. More complex molecular global properties or functional-group counts require the connection table to be computed. The descriptors reflecting molecular shape or the distribution of a property on the molecular surface require the previous computation of the 3D molecular structure. All molecular descriptors in our analysis were computed using the ADRIANA.Code suite for descriptor calculation.[32] All descriptors calculated in the present study are listed in Table 3.

These descriptors are, to a large extent, 2D and 3D molecular descriptors and reflect shape and reactivity properties. The capability for participating in hydrogen bonding is described directly by the number of hydrogen-bonding acceptors/donors or the hydrogen-bond acceptor/donor potential or indirectly by the number of basic nitrogen atoms and the number of acidic groups. Both 2D and 3D autocorrelation descriptors first require the computation of a particular property from the 3D structure of each molecule; then, the function of autocorrelation is applied to derive the autocorrelation vector.

The highest hydrogen-bonding acceptor potential (descriptor 2 in Table 3) is defined as the maximum lone-pair electronegativity on an atom considering all N, O, and F atoms in a compound. The highest hydrogen-bonding donor potential (descriptor 3 in Table 3) is defined as the most positive charge on the hydrogen atom in the functional groups −OH, −NH, and −SH in a compound.

Two-dimensional topological autocorrelations derive atom properties from the connection table of the molecules. On the other hand, 3D spatial autocorrelation descriptors first require knowledge of the 3D structure of each molecule and have to consider atom identities or perform the computation of a property (σ-electronegativity, π-electronegativity, σ-charge, or π-charge). In both cases, the function of autocorrelation was applied to derive autocorrelation vectors. In this study, the molecular electrostatic potential (MEP) was derived by a classical point-charge model: The electrostatic potential for each molecule was obtained by moving a unit positive point charge across the molecular surface and thus calculating it at various points j on this surface, as follows

$$V_j = \frac{1}{4\pi\varepsilon_0} \sum_{i}^{\text{atoms}} \frac{q_i}{r_{ji}} \tag{1}$$

Here, q_i represents the partial charge of each atom i, and r_{ji} is the distance between a point j and an atom i. Starting from the 3D model of a molecule and its partial atomic charges, the electrostatic and hydrogen-bonding potentials are calculated for points on the molecular surface. Other properties (identity, σ-electronegativity, π-electronegativity, σ-charge, π-charge) are based on the atoms. Partial atomic electronegativity and charges were calculated by the PEOE (partial equalization of orbital electronegativity) method, and its extension to conjugated systems was implemented by the Petra (parameter estimation for the treatment of reactivity applications) module of the ADRIANA.Code suite. Once the autocorrelation function was applied, the autocorrelation vector for each property was obtained. An extensive presentation of the remaining descriptors in Table 3 was reported in our previous work.[20]

Autocorrelation Vectors. During the metabolism process, the substrates and CYP450 isoforms are in close contact, so the molecular properties have to be represented in order to explain the isoform specificity. The autocorrelation function transforms the constitution of a molecule into a fixed length representation. Starting from the first investigations of Moreau and Broto,[53,54] the introduction of this concept made possible the comparison of properties of molecules with a different size.[55,56] Topological autocorrelation considers the structure diagram, whereas spatial autocorrelation is based on the information encoded by the 3D molecular structure. Each component, $A(d)$, of the autocorrelation vector for the topological distance d is calculated as

$$A(d) = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} p_j p_i \delta(d_{t,ij}, d) \qquad \delta = \begin{cases} 1 \ \forall d_{t,ij} = d \\ 0 \ \forall d_{t,ij} \neq d \end{cases} \tag{2}$$

where $A(d)$ represents the autocorrelation coefficient referring to atom pairs i,j; N is the number of atoms in the molecule; p_i and p_j are the properties of atoms i and j, respectively; and $d_{t,ij}$ is the i,j topological distance (i.e., the number of bonds corresponding to the shortest path in the structure diagram). Equation 2 illustrates how a certain property p of an atom i is correlated with the corresponding property of atom j, and these products are summed over all atom pairs having a certain topological distance d. For the calculation of 2D topological autocorrelation coefficients, a maximum distance $d = 10$ bonds was selected, and 11 components per molecule were obtained.

In a similar way, 3D spatial autocorrelation vectors for the previously reported atomic or surface properties (see Table 3) were calculated separately for each compound by applying the following equation in which the spatial distance instead of the topological distance is used.

A set of randomly distributed points on the molecular surface had to be generated to compute the autocorrelation molecular electrostatic potential descriptor. In the computation of the 3D autocorrelation identity vectors, atomic positions and properties, instead of surface points and properties, were considered. Thus, all distances between the atoms or surface points were calculated and sorted into preset intervals

$$A(d_{\text{lower}}, d_{\text{upper}}) = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} p_j p_i \delta(d_{ij}, d_{\text{lower}}, d_{\text{upper}})$$
$$\delta = \begin{cases} 1 & \forall d_{\text{lower}} < d_{ij} \leq d_{\text{upper}} \\ 0 & \forall d_{ij} \leq d \vee d_{ij} > d_{\text{upper}} \end{cases} \tag{3}$$

where $A(d_{\text{lower}}, d_{\text{upper}})$ represents the component of the autocorrelation vector referring to the i,j distance d in the interval $d_{\text{lower}}-d_{\text{upper}}$ and is the sum of all of the products of the properties p_i and p_j for atoms i and j. The application of this concept made possible the comparison of different molecular properties, as this 3D descriptor represents a compressed expression of the distribution of property p on the molecular surface. The 3D autocorrelation vectors for property identity (descriptors 118−245 in Table 3) were computed by using $p_i = p_j = 1$ in eq 3. Default parameter values were $d_{\text{lower}} = 1$ Å and $d_{\text{upper}} = 13.8$ Å, with a resolution of 0.1 Å. Then, 128 components for this descriptor were obtained. The sums of the squares of the σ-electronegativity (descriptor 246 in Table 3), π-electronegativity (descriptor 247 in Table 3), σ-charge (descriptor 248 in Table 3), and π-charge (descriptor 249 in Table 3) were calculated to reflect the electronegativities and charge distributions in the aliphatic and conjugated systems. They can be obtained by calculating in ADRIANA.Code the first component of the autocorrelation vector while setting the distance to zero ($d_{\text{lower}} = 0$ Å, $d_{\text{upper}} = 0$ Å, and number of intervals = 1). For these descriptors, only one component ($d = 0$) of the autocorrelation coefficients was considered.

The parameters for the calculation of the molecular electrostatic potential (descriptors 250−261 in Table 3) and hydrogen-bonding potential (descriptors 262−273 in Table 3) autocorrelation coefficients were as follows: $d_{\text{lower}} = 1$ Å, $d_{\text{upper}} = 13$ Å, point density = 10 points/Å^2, and 12 autocorrelation coefficients were obtained.

The hydrogen atoms were included before the vectorial descriptors were computed. The autocorrelation vectors were computed by the ADRIANA.Code suite.

Variable Selection. In the multilabel classification approach, no automatic descriptors selection was performed,

but the molecular descriptors were selected or discarded after each single addition whether a better or worse classification model, respectively, was obtained. This process was repeated during the model optimization to find the best-performing descriptor set. The BestFirst automatic criterion implemented in Weka was applied to select the variables.[36,37] The descriptor space was explored to detect the subset that was likely to predict the classes best. The attribute evaluator CfsSubsetEval combined with the BestFirst search method was applied. In particular, the variable selection process was repeated for each fold during the model validation step.

Modeling Techniques. In the present work, several methods addressing multilabel and traditional single-label classification approaches were investigated in combination with the descriptors reported above. For the multilabel classifications, we applied cross-training with the recently developed support vector machine (ct-SVM) and multilabel *k*-nearest-neighbor (M_L-*k*NN) techniques, in addition to counterpropagation neural network (CPG-NN) analysis.[22,23,58] Simple logistic regression and SVM algorithms provided by Weka were selected as modeling methods for the single-label classification.[36,37]

Multilabel Classification. Cross-Training with Support Vector Machines (ct-SVM) Analysis. The concept of cross-training was introduced by Boutell and collaborators.[58] They turned from the too-simplistic or too-complex approaches and presented a new classification method suitable for multiclass and overlapping-classes problems. During the classification process, each multilabel example can be associated with more than one class. Moreover, the SVM technique represents a widely applied supervised learning method, recently developed by Vapnik.[59,60] SVM is applied to solve function approximation problems, where the data set is represented by pairs of samples $T = \{(x_1,y_1), (x_2,y_2), ..., (x_i,y_i)\}$, where x_i is the input data value and y_i is the target value. In a binary classification problem, usually $x_i \in \mathscr{R}^n$, and $y_i \in \{-1, +1\}$. The aim of the learning process is to find the optimal separating hyperplane in $\mathscr{R}^n$ space, corresponding to the hypothesis function $f(x)$, able to approximate the responses y_i with the minimum risk functional. Extensive literature is available on this subject.[59−61] ct-SVM represents a novel application of support vector machine analysis, because it was adapted to transform the real-valued scores of the starting independent binary classifiers into the labels of the final multilabel classifiers. Our models were developed using the R software and the e1071 package.[33,34] As reported in detail by Hristozov et al.,[23] in the present study, *n* binary classifiers, where *n* corresponds to the number of considered CYP450 isoforms ($n = 7$ and $n = 5$ for data sets 1 and 2, respectively), were built using a radial basis function (RBF) kernel. The parameters for each SVM classifier (C and γ) were automatically optimized on the training set during the learning process using a 10-fold cross validation and an external validation set.[23] After application of the cross-training approach, real-valued scores were obtained. Then, these values were turned into multilabel classifiers by applying different predefined testing criteria (P, T, and C).[58] The P criterion assigns a pattern to a particular class only if the SVM score is positive. The T criterion uses the closed-world assumption (CWA), according to which, if all of the SVM scores for a particular sample are negative, the pattern is assigned to the class corresponding to the less-negative score. The C criterion considers SVM scores without any sign, and the assignment of the patterns depends on the closeness of the SVM scores, which is determined via cross-validation. In this work, the same value (0.1) was used as an acceptable difference between two SVM scores. All of these different criteria were applied separately to transform the information given by the scores. Once the model had been optimized by predicting the validation set, the training and validation sets were merged, and a new model with the previously optimized parameters was computed. The final classifier was used for the prediction of a test set in order to evaluate its statistical robustness. Several parameters based on the scores were defined to evaluate the statistical quality of the models. In particular, accuracy was referred to the overall performance, whereas recall and precision were computed for each class. These parameters allowed for a comparison with the single-label classification results. On the other hand, the confusion matrix was calculated to compare the ct-SVM methodology with the CPG-NN technique.

Multilabel k-Nearest-Neighbors (M_L-kNN) Analysis. The multilabel *k*-nearest-neighbors (M_L-*k*NN) method represents an adaptation of a *k*NN learning algorithm to multilabel data.[62] In addition to the well-established *k*NN algorithm, it is based on the maximum a posteriori (MAP) principle, combining prior and posterior probabilities calculated from the training set. Generally, in a single-label classification problem, the *k*-nearest neighbors are found according to the Euclidian distance. In this way, the new samples are assigned to one of the classes. In the modified algorithm, the MAP principle both assigns the corresponding labels and ranks them. In the present work, the M_L-*k*NN technique was implemented in R. The parameters of the classifier, based on some selected descriptors, were optimized using a validation set as summarized in the previous subsection. After the optimization process, the training and validation sets were merged to build a new model with the optimal parameters. Then, it was used to predict an external test set. The same P, T, and C criteria described in the previous subsection were applied to the label rankings to obtain the final M_L-*k*NN classification labels.[23,58] The best values of the threshold (0.3) for each criterion and *k* (6) were selected during the analysis to achieve the best classification performance. The accuracy, recall, and precision parameters were computed to estimate the model robustness, as previously reported.

Counterpropagation Neural Network (CPG-NN). A counterpropagation neural network is a well-known extension of the Kohonen self-organizing map (SOM) analysis, where some output layers, corresponding to the classes (*Y* variables), are added to the Kohonen input layers, which represent the molecular descriptors (*X* variables). In the output layers, each vector component is 1 for positive examples or 0 for negative examples, according to the assigned classes. As an example in data set 1, haloperidol is a drug metabolized by three different CYP450 isoforms (CYP1A2, CYP2D6, and CYP3A4). Therefore, the value of 1 was set for the CYP1A2, CYP2D6, and CYP3A4 classes, and 0 was set for the other real labels before the computation of the CPG-NN model. During the learning process, only the input layers were considered to determine the winning neuron. Then, the weights for the output layers were used to predict the target values. Different topologies (rectangular and toroidal) and

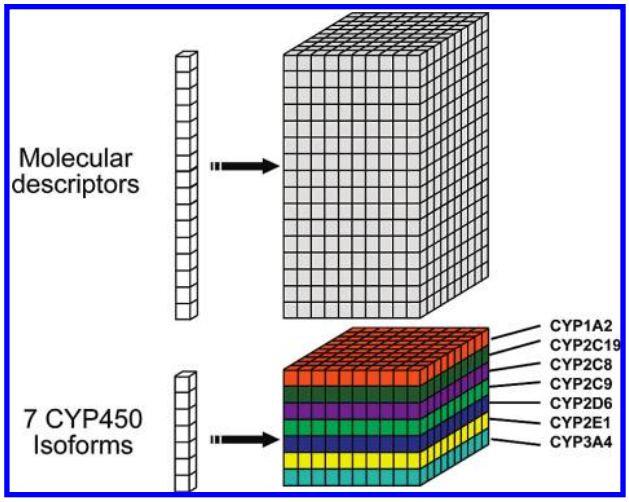

Figure 3. Flowchart of the CPG-NN analysis with seven output layers applied to our multilabel classification problem.

different map sizes were used in the CPG-NN analysis. In the present work, all CPG-NN models were derived using SONNIA software. The graphical representation of the CPG-NN technique applied to a problem with seven classes is shown in Figure 3.

Single-Label Classification. The same classification methods applied in the previous work were explored, using different combinations of descriptors.[20] The simple logistic regression and the SVM algorithms, which showed the best modeling power, are briefly described here.

Simple Logistic Regression. In the algorithm implemented in Weka, a linear regression technique is used to assign the correct class to each sample. Instead of the values of 0 and 1 as targets, the class probabilities are considered. Then, a linear model based on the transformed target variable is built to approximate the target function. The LogitBoost algorithm was applied with simple logistic functions in the learning process to perform automatic variable selection.[63]

Support Vector Machines (SVM). SVM is a well-known nonlinear learning method, which selects a hypothesis to optimally approximate the target function. In a binary classification problem, a hyperplane in the feature space is built to separate samples into two different classes. The support vectors are defined as the points located close to the decision boundary and separated by it. In Weka, several binary classifiers are generated according to the number of the classes and by using the algorithm developed by Platt.[63,64]

Validation Techniques. A large variety of validation methods were applied to assess the statistical robustness of the developed models. In the ct-SVM and M_L-*k*NN methodologies, a small validation set (see Tables 1 and 2) was selected to optimize the parameters of each binary classifier within the same model. The best parameters were detected by using 10-fold cross-validation and by applying separately the P, T, and C criteria. The confusion matrix was extracted from the validation set predictions. All validation processes were performed by using the R software.[33] Finally, the training and validation sets were used together to build new ct-SVM and M_L-*k*NN models with the optimized parameters. In the CPG-NN analysis, leave-one-out (LOO) and 5-fold cross-validation were performed. The confusion matrix resulted from the cross-validated predictions. CPG-NN validation analysis was carried out using the SONNIA package.[35] An extensive *n*-fold validation procedure was applied to check the prediction capability of single-label models developed with Weka. In particular, LOO, 10-, 5-, 3-, and 2-fold cross-validations were performed for model validation, and automatic variable selection was applied before each step. In *n*-fold cross-validation procedure, the data set is divided randomly into *n* subsets with a similar number of samples and class distribution, according to a stratified methodology. In the first step, one partition *n* is considered as test set, while the other $(n - 1)$ partitions are used to fit the model and then to predict the test set. The process is repeated *n* times, until all of the partitions are considered as the test set. The average, standard deviation, minimum, and maximum rates were collected for *n*-fold cross-validation methods. The package Weka was used to perform LOO and *n*-fold cross-validations.[36,37]

Model Evaluation. We introduced several parameters to assess the robustness of our models in order to compare the strategies applied using the same approach (multilabel or single-label) and, in a following step, the multilabel with the single-label approaches. The true positive rate (TP rate), the false positive rate (FP rate), the true negative rate (TN rate), the false negative rate (FN rate), the recall, and the precision were calculated for each class after the comparison between the predicted and real-valued labels for the validation set in the ct-SVM and M_L-*k*NN techniques. The values of the rates are included in the interval between 0 and 1. Low values of FP and FN rates and high values of TP and TN rates correspond to good modeling performances. Moreover, we calculated the Matthews correlation coefficients (MCCs) to better compare the prediction results of multilabel and single-label models on the test sets. The Matthews correlation coefficient falls in the range $-1 \leq \mathrm{MCC} \leq 1$. A value of $\mathrm{MCC} = 1$ indicates perfect agreement between predicted and experimental classes for each binary classifier, whereas a Matthews correlation coefficient of $\mathrm{MCC} = -1$ indicates the worst possible prediction.

Further performance measures were defined to evaluate the ranking process, which provides a ranking function to order the labels for each sample and to assign scores to the samples. One-error represents the ratio of the number of substrates not present in the top-ranked class to the total number of compounds. It can take on values between 0 and 1, so values close to 0 indicate a good performance. Coverage is the number of classes assigned, on average, to each substrate in order to correctly predict the real-valued classes. The coverage interval has a value between 1 and the number of the classes; therefore, the best performance corresponds to a value of 1. Average precision reflects the effectiveness of the label ranking and indicates the frequency of the top ranking for the experimental classes. The extreme values are 0 and 1. Perfect performance is achieved when the average precision is equal to 1. These parameters were mathematically defined in the work of Hristozov et al.,[23] as previously proposed by Boutell and co-workers.[58]

The CPG-NN models and single-label models developed using Weka were evaluated by computing the confusion matrix after the LOO cross-validation process and extracting the TP, FP, TN, and FN rates for an easier comparison of their performances with the first multilabel results. The

Table 4. Statistical Parameters Used to Determine the Model Performance

name	calculation details
true positive rate	TP/(FN + TP)
false positive rate	FP/(TN + FP)
true negative rate	TN/(TN + FP)
false negative rate	FN/(FN + TP)
recall	TP rate
precision	TP/(FP + TP)
% correct predictions	[(TP + TN)/total number of compounds] × 100
Matthews correlation coefficient	$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FN} \cdot \mathrm{FP}}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})}}$
	where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives in the confusion matrix

statistical parameters used in this work to assess the model prediction capability were calculated as summarized in Table 4.

RESULTS

The prediction of isoform specificity represents a multilabel classification problem, characterized by high complexity of the feature space. Here, we built a model to simultaneously classify a collection of substrates metabolized by seven CYP450 isoforms. Moreover, a simplified model was derived using five classes to improve predictivity. Then, the methods used in our previous publication[20] were applied to generate a single-label model with five classes, in order to compare the results with the multilabel strategy. In particular, various descriptors and data analysis techniques were combined to predict the isoform specificity using three different data sets, which correspond to (1) multilabel classification models with seven isoforms (CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4), (2) multilabel classification models with five isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1, CYP3A4), and (3) single-label classification models with five isoforms (CYP1A2, CYP2C9, CYP2D6, CYP2E1, CYP3A4). Only the best-performing experiments are reported in the present work.

Modeling of Data Set 1: Multilabel Classification Using Seven Isoforms. In a first investigation, data set 1 was considered (see Table 1). As a starting point, we used the descriptors selected in our previous publication.[20] See the Supporting Information for details.

This set of 12 descriptors, including some vectorial properties, functional-group counts, and shape descriptors, were found to be sufficient for building novel classification models with more than three classes. In fact, the corresponding ct-SVM model is not able to predict well the CYP1A2, CYP2C19, and CYP2C8 substrates, confirming that this descriptor set is not suitable for a good multilabel classification approach to such an unbalanced data set.

Moreover, we analyzed by visual inspection the interval of values of the descriptors (global, shape, topological, and functional-group counts) for each class in data set 1 (see the section Molecular Descriptor Calculation). Some descriptors displayed different values for a specific class membership, and this information was taken into consideration in the subsequent descriptor selection process. Because the information about the reaction site of the substrates was not reported, several models were derived combining global, topological, shape, and functional-group-count descriptors with 2D topological or 3D spatial autocorrelation vector components. We also included autocorrelation vectors of the molecular electrostatic potential (descriptors 250−261 in Table 3) and the hydrogen-bonding potential (descriptors 262−273 in Table 3) vectors in our set to achieve a possible improvement in the model performance. To this end, several global, topological, shape, functional-group-count, and vectorial descriptors were computed for each compound from the 3D molecular structure (descriptors 1−40 and 246−273 in Table 3).

A manual descriptor selection process was combined with the ct-SVM multilabel modeling method in the optimization procedure using training set 1. The best subset of global, topological, shape, and functional-group-count descriptors was selected according to the model performance after each single-descriptor addition step. In more detail, a descriptor was included in the best subset if the model predictivity results on validation set 1 improved. Furthermore, some descriptors encoding electronegativity and charge properties (descriptors 246−249 in Table 3) were added to this first subset, and the new descriptor set was optimized. The resulting subset was further extended by autocorrelation molecular electrostatic potential vectors (descriptors 250−261 in Table 3) and autocorrelation hydrogen-bonding potential vectors (descriptors 262−273 in Table 3), separately. Two different descriptor sets were obtained to generate the ct-SVM models in parallel. The best results corresponded to the model based on autocorrelation molecular electrostatic potential vectors. The variable selection procedure yielded a final set of 27 descriptors as reported in Table 5.

The information encoded by the autocorrelation molecular electrostatic potential descriptors improved the predictivity of the model in comparison to the model computed without these 12 variables or the model including autocorrelation hydrogen-bonding potential vectors. It supports the concept that the distribution of electrostatic properties on the molecular surface represents an important determinant in the prediction of isoform specificity.

In the ct-SVM modeling methods, we applied all of the P, T, and C criteria, but the best results were achieved using the T criterion to transform the real-valued scores assigned by the corresponding classifier into labels. The same analysis

Table 5. Twenty-Seven Descriptors Selected for the Training Set in Multilabel Classification Models with Data Sets 1 and 2

no.	name	details
1	MW	molecular weight
2	HAccPot	hydrogen-bond acceptor potential
6	TPSA	topological polar surface area
7	ASA	approximate surface area
16	D_3	diameter
17	R_3	radius
18	I_3	geometric shape coefficient
19	r_2	radius perpendicular to D_3
20	r_3	radius perpendicular to D_3 and R_2
29	n_{aro_amino}	number of aromatic amino groups
32	n_{tert_amino}	number of tertiary aliphatic amino groups
33	$n_{prim\_sec\_amino}$	$n_{prim\_amino} + n_{sec\_amino}$
37	$n_{basic\_nitrogen}$	number of basic, nitrogen-containing functional groups
38	n_{acidic_groups}	number of acidic functional groups
249	$q_{\pi\_}1 = \sum q_{\pi}^2$	property: π-charge q^{π}
250−261	SurfACorr_ESP	surface autocorrelation; property: molecular electrostatic potential

Table 6. Multiclassification ct-SVM/*7classes* Model: Statistical Parameters for Each Class after Prediction on Validation Set 1 (67 Substrates) for the Optimization of the Descriptors Set

class	TP rate	TN rate	recall	precision	% correct predictions
CYP1A2	0.64	0.96	0.64	0.78	91.0
CYP2C19	0.00	0.98	0.00	0.00	92.5
CYP2C8	0.00	0.98	0.00	0.00	92.5
CYP2C9	0.82	0.91	0.82	0.64	89.5
CYP2D6	0.87	1.00	0.87	1.00	97.0
CYP2E1	0.88	1.00	0.88	1.00	98.5
CYP3A4	0.90	0.78	0.90	0.78	83.6

Table 7. Multiclassification ct-SVM/*7classes* Model: Performance Measures after Prediction on Validation Set 1 (67 Substrates) and Test Set 1 (217 Substrates)

model prediction	$accuracy_{ML(\alpha=1)}$	one-error	coverage	average precision
validation set 1	0.76	0.13	2.06	0.87
test set 1	0.66	0.29	1.77	0.82

was repeated using the M_L-*k*NN method. All of the models that we generated showed a low predictivity; consequently, they were not considered in our study. The statistical parameters of the ct-SVM (ct-SVM/*7classes*) model to classify validation set 1 in seven CYP450 isoforms using the T criterion are summarized in Table 6, and the measures of ranking performances are shown in Table 7.

In validation set 1, there were four CYP2C19 and four CYP2C8 substrates, but ct-SVM/*7classes* model was not able to correctly classify them. As seen in Table 6, a good predictivity was achieved for all of the other classes if one considers the values of recall and precision. In more detail, a perfect prediction of the negative samples characterizes the CYP2D6 and CYP2E1 classes. After the optimization process, a new model using training set 1 and validation set 1 was carried out to predict test set 1. In Table 7, the performance measures of the final classifier are reported. The values of accuracy, precision, and one-error decrease compared to those of the first model. In fact, the ranking process is affected by the increase of the chemical diversity of test set 1 in comparison to the validation set 1.

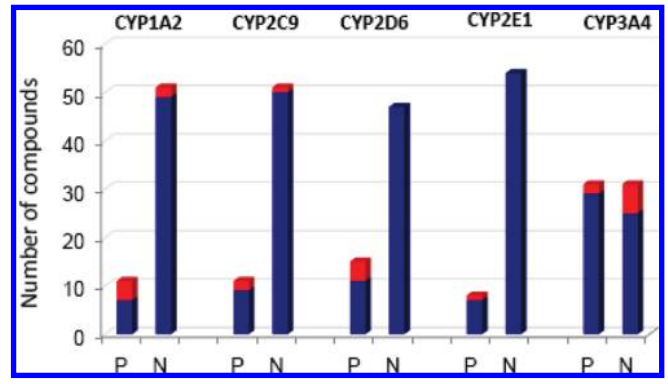

Figure 4. ct-SVM/*5classes* model: Graphical representation of the confusion matrix after the prediction of the validation set during model optimization. In blue, the number of true positives (TP) out of the total number of hits (P) and the true negatives (TN) out of the total number of nonhits (N) are indicated for each class. In red, the number of false negatives (FN) out of the total number of hits (P) and the false positives (FP) out of the total number of nonhits (N) are shown for each class.

Table 8. Multiclassification ct-SVM/*5classes* Model: Statistical Parameters for Each Class after Prediction on Validation Set 2 (62 Substrates)

classes	TP rate	TN rate	recall	precision	% correct predictions
CYP1A2	0.64	0.96	0.64	0.78	90.3
CYP2C9	0.82	0.95	0.82	0.75	95.2
CYP2D6	0.87	1.00	0.87	1.00	93.5
CYP2E1	0.88	1.00	0.88	1.00	98.4
CYP3A4	0.87	0.75	0.87	0.75	87.1

Table 9. Multiclassification ct-SVM/*5classes* Model: Performance Measures after Prediction on Validation Set 2 (62 Substrates) and Test Set 2 (209 Compounds)

model prediction	$accuracy_{ML(\alpha=1)}$	one-error	coverage	average precision
validation set 2	0.84	0.10	1.53	0.93
test set 2	0.70	0.25	1.52	0.85

We combined the same set of 27 descriptors to perform the CPG-NN analysis. Similar results after LOO cross-validation were obtained, with minimum predictivity for the CYP2C19 and CYP2C8 classes.

These results suggested to simplify our classification problem. To this end, the substrates metabolized by CYP2C19 and CYP2C8 isoforms were removed from data set 1, and we reduced the number of analyzed classes to five. In this way data set 2 was obtained.

Modeling of Data Set 2: Multilabel Classification Using Five Isoforms. The models generated using data set 2 represent a simplification of the approach reported in the preceding section with seven classes. The same descriptor set (Table 5) was used to build a first ct-SVM/*5classes* classifier with training set 2. As previously described, the model parameters were optimized by predicting validation set 2. The TP, FP, TN, and FN rates were calculated from the confusion matrix, which is graphically summarized in Figure 4. The prediction accuracies and performance measures of the ct-SVM/*5classes* model are reported in Tables 8 and 9, respectively.

In comparison to the ct-SVM/*7classes* model, the overall accuracy improved. However, the TP, FP, TN, and FN rates had almost the same values for the five classes as previously,

Table 10. Multiclassification CPG-NN/*5classes* Model: Statistical Parameters for Each Class after LOO Cross-Validation and 5-Fold Cross-Validation (345 Substrates)

	TP rate		TN rate		recall		precision	
class	LOO	5-fold	LOO	5-fold	LOO	5-fold	LOO	5-fold
CYP1A2	0.33	0.48	0.90	0.90	0.33	0.48	0.39	0.47
CYP2C9	0.60	0.58	0.93	0.93	0.60	0.58	0.62	0.60
CYP2D6	0.79	0.79	0.91	0.93	0.79	0.79	0.74	0.80
CYP2E1	0.87	0.85	0.99	0.98	0.87	0.85	0.95	0.77
CYP3A4	0.77	0.77	0.76	0.78	0.77	0.77	0.76	0.78

when these five classes were part of the seven-class data set. As in the previous study with data set 1, training set 2 and validation set 2 were merged to derive a new classifier with the optimized parameters. The performance measures of the final classifier are shown in Table 9. Again, all of the modeling parameters decrease after the prediction of test set 2 in comparison to the first classifier, regardless of whether the values are higher than the results of the ct-SVM/*7classes* model after the prediction of test set 1.

Satisfactory results for the multilabel approach were also achieved by the application of CPG-NN technique with the set of 27 descriptors in Table 5. Different map sizes, topologies, and numbers of epochs were selected for the learning process, and the best model with a rectangular topology is discussed here. After LOO and 5-fold cross-validation procedures, the best CPG-NN/*5classes* model was selected. The model performances are reported in Table 10.

Good values of precision and recall were obtained for classes CYP2C9, CYP2D6, CYP2E1, and CYP3A4, upon application of the best CPG-NN/*5classes* model. A low predictivity of the CPG-NN/*5classes* model can be observed for the CYP1A2 class in terms of a higher TP rate or, alternatively, lower FN rate. In general, a model with a minimum predictivity for one class might not be able to detect substrates metabolized by this particular isoform, regardless of whether it was applied to an external test set. However, at least 75% of the compounds for each class were correctly classified. In Figure 5, the five output layers of the CPG-NN/*5classes* network are shown.

In each layer, the CYP450 substrates tend to cluster. This tendency is not particularly evident for class CYP1A2, where the occupied neurons are spread out in the corresponding layer. Few conflict neurons are present in the maps, and most of them are caused by CYP3A4 substrates conflicting with the substrates metabolized by other classes. In fact, CYP3A4 represents the major class and the most chemically heterogeneous class. In some cases, the descriptor set we selected was not able to correctly classify substrates with similar structural features and different class memberships.

Modeling of Data Set 3: Single-Label Classification Using Five Isoforms. Data set 3 includes only single-label substrates extracted from data set 2, and it is therefore suitable for the following data analysis. In the single-label approach, a systematic variable selection procedure was performed. Different sets of descriptors were considered in our study. The global, topological, shape, and functional-group-count descriptors (descriptors 1−40 in Table 3) were added separately to 2D topological autocorrelation descriptor vectors (descriptors 41−117 in Table 3) and to 128 spatial autocorrelation vectors (descriptors 118−245 in Table 3).

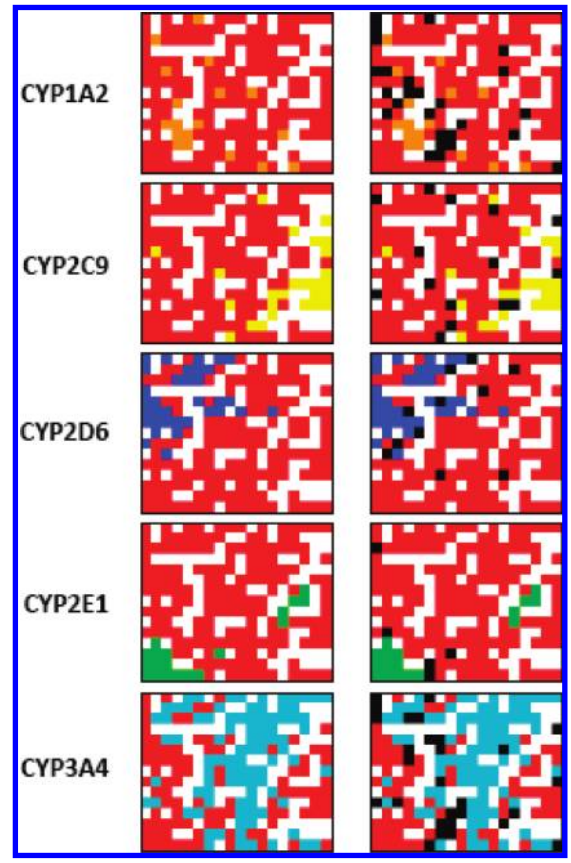

Figure 5. CPG-NN/*5classes* model: Projection of the CYP450 substrates into five maps, corresponding to CYP1A2, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 classes, as reported on the left. CYP1A2 substrates in orange, CYP2C9 substrates in yellow, CYP2D6 substrates in blue, CYP2E1 substrates in green, and CYP3A4 substrates in light blue are indicated (according to the most frequent output). For each map, neurons containing substrates that do not belong to the corresponding class are in red. In the maps on the right side, conflicting neurons in black are also shown. Black neurons contain compounds of different CYP450 classes. White squares represent empty neurons.

The best models we achieved are based on the second descriptor set as the input matrix, which comprises several global, topological, shape, functional-group-count, and 3D autocorrelation descriptors, giving a total of 168 variables. The 3D autocorrelation identity descriptors reflect the distribution of the interatomic distances in the 3D molecular structure and complete the information given by the first selected subset. Automatic variable selection using the BestFirst method was combined with different modeling techniques by applying Weka algorithms. Nineteen descriptors were selected for training set 3 during the selection procedure, as summarized in Table 11.

The selected nine components of 3D autocorrelation identity vectors correspond to particular atom distances and show an important contribution in the model building process. They seem to positively substitute the information encoded by the 2D topological autocorrelation used in the previous work to classify CYP450 into three classes. The numbers of aliphatic amino groups; acidic groups; and basic, nitrogen-containing functional groups were also selected in our published classification model.[20]

The best models were generated by combining the automatic variable selection with simple logistic regression

Table 11. Descriptors Resulting from Automatic Variable Selection by Applying the BestFirst Method for the Training Set in Single-Label Classification Models with Data Set 3

no.	name	details
3	HDonPot	hydrogen-bond donor potential
6	TPSA	topological polar surface area
7	ASA	approximate surface area
9	μ	molecular dipole moment
19	r_2	radius perpendicular to D_3
20	r_3	radius perpendicular to D_3 and r_2
28	$n_{\text{aliph\_amino}}$	number of aliphatic amino groups
33	$n_{\text{prim\_sec\_amino}}$	$n_{\text{prim\_amino}} + n_{\text{sec\_amino}}$
37	$n_{\text{basic\_nitrogen}}$	number of basic, nitrogen-containing functional groups
38	$n_{\text{acidic\_groups}}$	number of acidic functional groups
120	3D-AC_{identity} (1.2−1.3 Å)	spatial autocorrelation; property: identity (1.2−1.3 Å)
121	3D-AC_{identity} (1.3−1.4 Å)	spatial autocorrelation; property: identity (1.3−1.4 Å)
122	3D-AC_{identity} (1.4−1.5 Å)	spatial autocorrelation; property: identity (1.4−1.5 Å)
126	3D-AC_{identity} (1.7−1.8 Å)	spatial autocorrelation; property: identity (1.7−1.8 Å)
132	3D-AC_{identity} (2.3−2.4 Å)	spatial autocorrelation; property: identity (2.3−2.4 Å)
136	3D-AC_{identity} (2.7−2.8 Å)	spatial autocorrelation; property: identity (2.7−2.8 Å)
140	3D-AC_{identity} (3.1−3.2 Å)	spatial autocorrelation; property: identity (3.1−3.2 Å)
151	3D-AC_{identity} (4.2−4.3 Å)	spatial autocorrelation; property: identity (4.2−4.3 Å)
162	3D-AC_{identity} (5.3−5.4 Å)	spatial autocorrelation; property: identity (5.3−5.4 Å)

Table 12. Single-Label Classification SimLog/*5classes* Model: Statistical Parameters Using 19 Descriptors for Training Set 3

			% correct predictions			
partition	CV	no. of runs	mean	stdev	min	max
training set 3		1	87.0	−	87.0	87.0
training set 3	LOO	1	77.1	−	77.1	77.1
	10-fold	10	78.4	1.9	76.1	81.9
	5-fold	20	78.9	1.4	76.4	82.6
	3-fold	33	78.5	1.5	74.7	80.5
	2-fold	50	76.8	2.1	70.3	80.9
test set 3		1	77.5	−	77.5	77.5

Table 13. Single-Label Classification SVM/*5classes* Model: Statistical Parameters Using 19 Descriptors for Training Set 3

			% correct predictions			
partition	CV	no. of runs	mean	stdev	min	max
training set 3		1	85.7	−	85.7	85.7
training set 3	LOO	1	75.8	−	75.8	75.8
	10-fold	10	76.3	1.3	74.7	78.2
	5-fold	20	76.0	1.2	73.3	78.2
	3-fold	33	75.3	1.7	71.0	78.1
	2-fold	50	75.1	2.2	69.3	80.2
test set 3		1	78.0	−	78.0	78.0

(SimLog/*5classes* model) and SVM (SVM/*5classes* model) methods. The results are reported in Tables 12 and 13.

The standard parameters suggested in Weka software were selected for the computation. In the SVM/*5classes* model, a polynomial kernel with exponent equal to 3 was used. Both models are quite stable if one considers the profile of the standard deviation values for each *n*-fold cross-validation and the predictivity in Tables 12 and 13. In fact, in the SimLog/*5classes* model, training set 3 is predicted with an accuracy of 77.1% after LOO cross-validation. The minimum value of predictivity (76.8%) is achieved if a 2-fold cross-validation procedure is applied. Interestingly, 77.5% of the substrates in test set 3 are correctly predicted. Regarding the SVM/*5classes* model, in the LOO cross-validation, a predictivity of 75.8% was obtained. Again, the percentage of correct predictions is lower for the other *n*-fold cross-validation procedures, with a difference of 10.6% between training set 3 predictivity and the average prediction accuracy in 2-fold cross-validation. SimLog/*5classes* model seems to have quite better performances than SVM/*5classes* model, but test set 3 is predicted with an accuracy of 78%. The prediction performances are similar, and both models show a correspondence in the trends of the percentage of correct predictions for each *n*-fold cross-validation. Tables 14 and 15 show the TP, FP, TN and FN rates of training set 3 in LOO cross-validation for the SimLog/*5classes* and SVM/*5classes* models, respectively.

Table 14. Single-Label Classification SimLog/*5classes* Model: Predictivity Results after LOO Cross-Validation for Training Set 3

class	TP rate	TN rate	recall	precision
CYP1A2	0.44	0.95	0.44	0.54
CYP2C9	0.67	0.95	0.67	0.68
CYP2D6	0.83	0.96	0.83	0.81
CYP2E1	0.84	0.97	0.84	0.78
CYP3A4	0.84	0.85	0.84	0.82

Table 15. Single-Label Classification SVM/*5classes* Model: Predictivity Results after LOO Cross-Validation for Training Set 3

class	TP rate	TN rate	recall	precision
CYP1A2	0.34	0.99	0.34	0.73
CYP2C9	0.72	0.96	0.72	0.74
CYP2D6	0.77	0.95	0.77	0.77
CYP2E1	0.89	0.94	0.89	0.68
CYP3A4	0.82	0.82	0.82	0.79

By analyzing the single classes, it is noticeable that, in the SVM/*5classes* model, the recall for classes CYP1A2 and CYP2C9 decreases in comparison with the SimLog/*5classes*, whereas better results in the prediction of CYP2D6 and CYP2E1 substrates were obtained, if one compares the corresponding TP and TN rates. For these reasons, even if the SVM/*5classes* model shows the highest predictivity value for the external test set, we considered both models reliable in the final comparison of the test set predictions.

Validation of the Models with an External Test Set. In our analysis, different test sets were studied, according to the data set used to build up the classification models, as seen in Tables 1 and 2. There is a good correspondence between the chemical space of the training sets and the test sets, which include similar molecular structures. The isoform specificity was predicted for each test set by applying the multilabel (test sets 1 and 2) or the single-label (test set 3) classification models.

Test Set 1. A test set comprising 217 compounds metabolized by seven different CYP450 isoforms was used to validate our multilabel approach. The prediction results for test set 1 after the application of the ct-SVM/*7classes* model are summarized in Table 16.

The performances of the ct-SVM/*7classes* model are statistically acceptable, but again, as resulted during the

Table 16. Multilabel Classification ct-SVM/*7classes* Model: Predicted Results of the Model for Test Set 1

classes	TP rate	TN rate	recall	precision	% correct predictions
CYP1A2	0.57	0.89	0.57	0.38	87.5
CYP2C19	0.00	0.98	0.00	0.00	93.5
CYP2C8	0.33	0.98	0.33	0.17	97.7
CYP2C9	0.50	0.98	0.50	0.64	93.5
CYP2D6	0.69	0.87	0.69	0.69	81.6
CYP2E1	0.75	0.98	0.75	0.64	96.3
CYP3A4	0.81	0.72	0.81	0.77	76.9

optimization procedure, the predictivity for the CYP2C19 class is equal to 0, and the TP rate for CYP2C8 class is very low. Better results were achieved for the other classes, but the value of the FN rate is surprisingly low for CYP2C9 class in comparison to the FN rate after the prediction of validation set 1. Then, the prediction performance after the training and validation sets 1 were merged decreases. This is clearly evident if one compares the recall and precision for the single CYP450 isoforms.

Test Set 2. Test set 2, comprising 209 substrates, was analyzed by both the ct-SVM/*5classes* and CPG-NN/*5classes* isoform predictors. The prediction results on test set 2 obtained using the ct-SVM and CPG-NN techniques are summarized in Tables 17 and 18, respectively.

In the ct-SVM/*5classes* model predictions the FN rate is still remarkable. The prediction results are interesting, and the TP rate for class CYP1A2 is similar to the value after the prediction of validation set 2 and higher than the predictivity using the ct-SVM/*7classes* model. If the test set 2 predictions (Table 17) are compared with the validation set 2 predictions (Table 8), the former values of TP rate for CYP2C9 and CYP2D6 classes are lower than the latter results. Considering that the classifier is based on five classes, the model prediction capability is quite accurate. The performance measures are slightly lower than the previous values for the model built using training set 2 (Table 9). In more detail, the accuracy and the average precision drop from 0.84 to 0.70 and from 0.93 to 0.85, respectively.

The same test set was used to assess the performances of the CPG-NN/*5classes* model. Only for the CYP2E1 class were surprisingly good results were achieved, with a recall of 0.92. Statistically acceptable values of the TP rate corresponded to the CYP2D6 (0.70) and CYP3A4 (0.72) isoforms. Regarding CYP1A2 and CYP2C9, the CPG-NN/*5classes* model showed low predictivity, with recall values of 0.52 and 0.61, respectively, and decreasing precision. These results reflected the profile of the statistical parameters obtained during model validation using validation set 2.

Test Set 3. In the single-label models, test set 3 (191 CYP450 substrates) was used to assess their predictivity. As seen in Tables 12 and 13, the percentage of correct predictions for test set 3 was found to be 77.5% for the SimLog/*5classes* model and 78.0% for the SVM/*5classes* model. These values are quite close to or slightly higher than the corresponding values of correct prediction after 2-fold cross-validation. In Tables 19 and 20, the prediction rates of the same single-classification models for each class of test set 3 are reported.

The prediction performances of the SimLog/*5classes* and SVM/*5classes* models show almost perfect correspondence for all of the classes, if one compares the corresponding recall and precision values in Tables 19 and 20. Slight differences can be seen for the CYP2C9, CYP2D6, CYP2E1, and CYP3A4 isoforms. In both cases, all of the CYP2E1 substrates are correctly predicted (the TP rate is equal to 1), and the values of TP rate for the remaining classes are included in the interval 60−80%. Moreover, both the SimLog/*5classes* and SVM/*5classes* models were applied to predict the 18 multilabel substrates in test set 2 that are not included in test set 3. An arbitrary experimental class was

Table 18. Multilabel Classification CPG-NN/*5classes* Model: Predicted Results for Test Set 2

class	TP rate	TN rate	recall	precision	% correct predictions	Matthews correlation coefficient
CYP1A2	0.52	0.85	0.52	0.30	81.3	0.29
CYP2C9	0.61	0.93	0.61	0.44	89.9	0.46
CYP2D6	0.70	0.88	0.70	0.72	82.8	0.58
CYP2E1	0.92	0.97	0.92	0.69	97.1	0.78
CYP3A4	0.72	0.79	0.72	0.81	75.6	0.52

Table 19. Single-Label Classification SimLog/*5classes* Model: Predicted Results for Test Set 3

class	TP rate	TN rate	recall	precision
CYP1A2	0.64	0.96	0.64	0.53
CYP2C9	0.73	0.98	0.73	0.67
CYP2D6	0.77	0.91	0.77	0.77
CYP2E1	1.00	0.98	1.00	0.73
CYP3A4	0.77	0.83	0.77	0.84

Table 20. Single-Label Classification SVM/*5classes* Model: Predicted Results for Test Set 3

class	TP rate[a]	TN rate	recall	precision	Matthews correlation coefficient
CYP1A2	0.64 (9/14)	0.97	0.64	0.60	0.59
CYP2C9	0.73 (8/11)	0.98	0.73	0.67	0.68
CYP2D6	0.77 (41/53)	0.91	0.77	0.76	0.67
CYP2E1	1.00 (11/11)	0.97	1.00	0.65	0.79
CYP3A4	0.78 (80/102)	0.85	0.78	0.85	0.64

[a] Also indicates number of TPs in the confusion matrix for each class.

Table 17. Multilabel Classification ct-SVM/*5classes* Model: Predicted Results of the Model for Test Set 2

class	TP rate[a]	TN rate	recall	precision	% correct predictions	Matthews correlation coefficient
CYP1A2	0.65 (15/23)	0.90	0.65	0.44	87.1	0.47
CYP2C9	0.44 (8/18)	0.96	0.44	0.53	91.9	0.44
CYP2D6	0.59 (38/64)	0.92	0.59	0.78	82.3	0.58
CYP2E1	0.75 (9/12)	0.98	0.75	0.69	96.6	0.70
CYP3A4	0.84 (97/116)	0.70	0.84	0.78	77.5	0.56

[a] Also indicates number of TPs in the confusion matrix for each class.

assigned to these compounds, but they were not used in the model performance evaluation. The prediction results were considered in the final comparison with the other multilabel classification methods.

DISCUSSION

Aspects Related to the Data Set. The classification of multilabel data represents a challenging problem. So far, many different strategies have been explored to classify drugs metabolized by a single isoform. However, this approach does not reflect the real scenario, in which the route of metabolism might involve several enzymes for the biotransformation.

A multilabel classification model able to simulate what really happens under in vivo conditions is still not available, because the collection of a robust and reliable data set represents the most important requirement as well as a difficult starting point.

CYP450 enzymes are not selective, and in fact, some xenobiotics are known to be metabolized by at least one CYP450 isoform. The same molecular structure can be recognized by different isoforms; on the other hand, the same CYP450 isoform might metabolize chemically diverse compounds, and this is particularly true for the CYP3A4 isoform. Moreover, in the metabolic process, the CYP450 enzymes show a different role and relevance. As a consequence, the data set is unbalanced, with few compounds classified as CYP1A2, CP2C19, CYP2C8, CYP2C9, and CYP2E1 substrates and more represented by classes CYP2D6 and CYP3A4.

The correct experimental determination of the metabolic profile represents a difficult task. In our analysis, we dealt with information coming from different sources, and in some cases, it was inconsistent or incomplete. After a comparison of the data reported in several articles, it became apparent that it is impossible to unequivocally determine the class memberships for all the substrates. The uncertainty of information led us to choose a compromise by considering the most recent source to be most reliable.

Considerations on the Selected Descriptors. In the metabolic process, a recognition mechanism is responsible for the complementary interaction between the substrates and the cytochrome P450 isoforms. Therefore, the chemical nature of the substrates and especially the distribution of particular properties on their surface are involved in the determination of their metabolic fate.

The specificity of the interactions is driven by several molecular properties. In this study, we collected 273 different molecular descriptors. Moreover, the function of autocorrelation is a useful strategy to overcome the dependence on the spatial rotation and translation of the molecules. In fact, the autocorrelation concept is able to describe the distribution of a particular property on the molecular surface and to represent molecules of different sizes with a vector of fixed length.

It was shown that the molecular shape and particular functional-group-count descriptors are crucial properties to describe CYP2C9, CYP2D6, and CYP3A4 substrates in a single-label classification approach.[20] The importance of these descriptors in the multilabel analysis was uncertain. As automatic variable selection is not available in multilabel classification problems, the same set of 12 descriptors as used in a previous study[20] was considered to classify the substrates in seven classes. See the Supporting Information for details. After the manual selection procedure, some descriptors were confirmed to be relevant (Table 5). The radius and the numbers of basic, nitrogen-containing and acidic functional groups were selected again, and other global, shape, functional-group-count, and 3D autocorrelation descriptors were included. As reported in Table 16 for the ct-SVM/*5classes* model, the prediction results considerably improved for CYP1A2, whereas slightly higher values were obtained for CYP2C9 and CYP3A4 isoforms. The TP rates for CYP2D6 and CYP2E1 are lower in comparison to the previous model. The CPG-NN/*5classes* model is based on the same descriptor set. The clearly distinguishable clusters in the maps corresponding to each layer/class further support the good choice of the variables. Still, in the published model, the meaning of the 3D autocorrelation identity components was not clear, and in the present study, the multilabel techniques do not provide an automatic variable selection method. It is significant to notice that the manual descriptor selection led to a model with similar performances for most of the classes, in comparison to the published model. Therefore, the autocorrelation molecular electrostatic potential descriptors can easily substitute the vectorial properties used in the first work.

The 19 descriptors in the single-label classification in the SimLog/*5classes* and SVM/*5classes* models resulted after an automatic attribute selection process by applying the Weka algorithms. The selection was repeated for each fold during the LOO cross-validation procedure. Some descriptors are in common with the descriptor set of the previous work (descriptors 28, 37, and 38) and with the manually selected variables in the multilabel classification models (descriptors 6, 7, 19, 20, 33, 37, and 38). Various shape- and size-related descriptors were recognized as important in the single-label analysis, whereas specific acid−base properties (descriptors 37 and 38) were selected in all of the models, confirming that these descriptors are crucial in the prediction of isoform specificity, whatever the followed approach (multi- or single-label). In any case, we achieved a quite good predictivity using a limited number of descriptors in the model-building process.

Model Comparison in the Multilabel Classification Approach. *ct-SVM/5classes and ct-SVM/7classes.* The same descriptor set (Table 5) was used to build both the ct-SVM/*5classes* and ct-SVM/*7classes* models. Concerning the optimization procedure using validation sets 1 and 2, the overall accuracy of the first model shows a higher value (0.84) than the same parameter measure in the second model (0.76), as shown in Tables 7 and 9. Also, the performance measures support the increased predictivity, with the average precision rising from 0.87 (ct-SVM/*7classes* model) to 0.93 (ct-SVM/*5classes* model). Test sets 1 and 2 were used to assess the statistical quality of the ct-SVM/*7classes* and the ct-SVM/*5classes* models, respectively. The values of the prediction accuracies are similar (Tables 7 and 9); however, different performances are displayed in the single classes if one considers Tables 16 and 17. The TP rate drops for classes CYP2C9 and CYP2D6 in test set 2; on the other hand, the TN rate increases for class CYP2D6. This means that, in test set 2, only 8−18 CYP2C9 substrates and 38−64

CYP2D6 substrates are correctly assigned to the corresponding class, whereas for test set 1, the numbers of TPs are 9 and 44, respectively. A higher recall resulted for CYP1A2 (from 0.57 to 0.65) and CYP3A4 (from 0.81 to 0.84) isoforms by comparing the values in test sets 1 and 2. Furthermore, ct-SVM/*5classes* is able to predict 15 of 23 CYP1A2 substrates, 9 of 12 CYP2E1 substrates, and 97 of 116 CYP3A4 substrates, improving the predictivity for the above-mentioned classes in the ct-SVM/*7classes* model.

In conclusion, although the ct-SVM/*5classes* model presents better performances than the ct-SVM/7*classes* model on the respective validation sets, a difference after the comparison of the predictivity on the test sets is not evident. We selected the ct-SVM/*5classes* model, because the CYP2C19 and CYP2C8 classes contribute to unbalance our data set. Moreover, these classes belong to the main CYP2C class and were distinguished from the CYP2C9 isoform only in a second moment.

ct-SVM/5classes and CPG-NN/5classes. The prediction results for test set 2 were analyzed to compare the performances of the ct-SVM/*5classes* and CPG-NN/*5classes* models. The predictivity of ct-SVM/*5classes* for classes CYP1A2 and CYP3A4 is higher than that of the other multilabel model. On the other hand, if one considers the remaining isoforms, the values of recall underline the better performances of the CPG-NN/*5classes* model. After the comparison of the percentages of correct predictions, the values are similar for the corresponding isoforms.

Comparison of Multi- and Single-Label Classification Models. The single-label approach applied to a data set including seven isoforms did not give an appreciable predictivity for both the CYP2C19 and CYP2C8 classes. Therefore, we used the same methods to generate a model for the substrates metabolized by five CYP450 isoforms. The performances on test set 3 of SimLog/*5classes* and the SVM/*5classes* classifiers are very similar. Therefore, we compared the prediction results of the ct-SVM/*5classes* model on test set 2 (209 substrates) and the predictivity of the SVM/*5classes* model for test set 3 (191 substrates) to verify whether the multilabel approach might be a valid alternative to the single-label methodology, by applying the same algorithm as modeling method. In this analysis, we had to consider that the test sets comprise a different number of compounds, because in test set 2, 18 compounds are multilabel.

On first analysis, the recall shows similar performances of the models for class CYP1A2 and an increase of predictivity for class CYP3A4 in the multilabel model. Regarding classes CYP2C9, CYP2D6, and CYP2E1, the recall decreases in the single-label classifier. On the other hand, the profile of the precision values reflects better performances of the ct-SVM/*5classes* model if the CYP2D6 and CYP2E1 classes are considered, with the precision values of 0.78 and 0.69, respectively. In the single-label model, the values of recall are 0.60 (CYP1A2 class), 0.67 (CYP2C9 class), and 0.85 (CYP3A4 class), higher than the corresponding values in the multilabel classifier. If one considers the number of TPs in the single-label model, 9 of 14 CYP1A2 substrates, 8 of 11 CYP2C9 substrates, 41 of 53 CYP2D6 substrates, 11 of 11 CYP2E1 substrates, and 80 of 102 CYP3A4 substrates resulted. The number of compounds predicted correctly using the single-label approach is very close in comparison to the number of TPs in the multilabel results reported in Table 17. In fact, 15, 8, 38, 9, and 97 substrates are the TPs for classes CYP1A2, CYP2C9, CYP2D6, CYP2E1, and CYP3A4, respectively. In particular, for CYP2C9, CYP2D6, and CYP2E1 isoforms, the values of TP are very similar in the ct-SVM/*5classes* and the SVM/*5classes* models, in contrast to the significant differences detected by observing the other statistical parameters.

It seems clear that the single-label model is not able to provide a complete picture of the metabolism information. In fact, the ct-SVM/*5classes* model performances are at least comparable to the SVM/*5classes* model ones. However, in the single-label approach, inevitably, important details about isoform specificity are lost, because each substrate is implicitly supposed to be metabolized by a unique CYP450 isoform.

Analysis of Some Classified Compounds. We compared the prediction results of the ct-SVM/*5classes*, CPG-NN/*5classes*, and SVM/*5classes* models on test set 2, which includes multi- and single-label substrates. The SVM/*5classes* model instead of the SimLog/*5classes* classifier was selected in order to compare the predictivity between similar modeling techniques, which apply the support vector machines (SVM) algorithm. In the following analysis, the multilabel compounds in test set 2 were also predicted by the single-label model. The CYP450 substrates not extracted from the Metabolite database were analyzed.[28] The experimental and predicted classes for these compounds are reported in Table 21.

Nine of 44 compounds (two of which are multilabel) are incorrectly predicted by both the multilabel ct-SVM/*5classes* and CPG-NN/*5classes* models. The single-label SVM/*5classes* model assigned a wrong class to 10 compounds and correctly assigned all of the multilabel compounds to one of the experimental classes. Deprenyl (no. 175) is metabolized by the CYP2D6 isoform, but it is predicted to be a CYP3A4 substrate by the ct-SVM/*5classes* multilabel model and a CYP2E1 substrate by the SVM/*5classes* single-label model. It was only partially predicted by the CPG-NN/*5classes* model, which classified deprenyl as a multilabel substrate. The drug phenylbutazone (no. 195) is incorrectly predicted to be metabolized by CYP1A2 and CYP3A4 by the multi- and single-label models, respectively. It was also reported as a CYP2C9 inhibitor,[24] and the pyrazolidine-3,5-dione cycle in its structure recalls other CYP1A2 or CYP3A4 substrates in the training sets (ropirinole, alfentanil, and ethosuximide). In addition, it was wrongly predicted also by in our previous work considering three CYP450 isoforms.[20] However, the tautomer of phenylbutazone is correctly predicted as a CYP2C9 substrate by the single-label classifier (and as a CYP3A4 substrate by the multilabel model). Quercetine (no. 176) is reported to be a CYP3A4 substrate, wrongly predicted in the CYP1A2 class membership by both ct-SVM/*5classes* and CPG-NN/*5classes* models and by the SVM/*5classes* model. In fact, this polar compound presents several hydroxilic and ketone groups in its three-rings scaffold, similarly to different CYP1A2 substrates in the selected training sets. Most of the CYP2D6 substrates are small molecules with a basic amino group incorporated in the structure. The antipsychotic drug remoxipride (no. 198) has this profile, but it is predicted to be metabolized by the CYP3A4 isoform by all the models. Many compounds are

Table 21. Some Experimental and Predicted Isoforms after Application of the ct-SVM/*5classes*, CPG-NN/*5classes*, and SVM/*5classes* Models[a]

			predicted class(es)		
no.	name[b]	experimental class(es)	ct-SVM/*5classes*	CPG-NN/*5classes*	SVM/*5classes*
166	acetaminophen	CYP1A2	CYP1A2	CYP1A2	CYP1A2
167	alpidem	CYP3A4	CYP3A4	CYP3A4	CYP3A4
168	amiflamine	CYP2D6	CYP1A2	CYP2D6	CYP2D6
169	**aripiprazole**	CYP2D6, CYP3A4	CYP3A4	CYP3A4	CYP3A4
170	azatadine	CYP3A4	CYP3A4	CYP3A4	CYP3A4
171	bufuralol	CYP2D6	CYP2D6	CYP2D6	CYP2D6
172	cinnarizine	CYP2D6	CYP2D6	CYP2D6	CYP2D6
173	**clomipramine**	CYP1A2, CYP2C9, CYP2D6, CYP3A4	CYP1A2, CYP2D6, CYP3A4	CYP2D6	CYP2D6
174	**clopidogrel**	CYP1A2, CYP3A4	CYP2D6	CYP2D6	CYP3A4
175	deprenyl	CYP2D6	CYP3A4	CYP2D6, CYP3A4	CYP2E1
176	**desipramine**	CYP1A2, CYP2D6	CYP2D6	CYP2D6	CYP2D6
177	**dihydrocodeine**	CYP2D6, CYP3A4	CYP2D6, CYP3A4	CYP2D6	CYP2D6
178	ebastine	CYP3A4	CYP3A4	CYP3A4	CYP3A4
179	enalapril	CYP3A4	CYP3A4	CYP2C9	CYP1A2
180	fluconazole	CYP3A4	CYP2C9, CYP3A4	CYP3A4	CYP1A2
181	flunarizine	CYP2D6	CYP2D6	CYP2D6	CYP2D6
182	**formoterol**	CYP2C9, CYP2D6	CYP2D6	CYP2D6	CYP2D6
183	indomethacin	CYP2C9	CYP3A4	CYP2C9	CYP2C9
184	levonorgestrel	CYP3A4	CYP3A4	CYP3A4	CYP3A4
185	**lidocaine**	CYP2D6, CYP3A4	CYP1A2	CYP2D6, CYP3A4	CYP2D6
186	lisuride	CYP3A4	CYP1A2, CYP3A4	CYP3A4	CYP3A4
187	lobeline	CYP2D6	CYP2D6, CYP3A4	CYP1A2, CYP2D6, CYP3A4	CYP3A4
188	lornoxicam	CYP2C9	CYP2C9	CYP2C9	CYP2C9
189	**methadone**	CYP1A2, CYP2D6, CYP3A4	CYP2D6, CYP3A4	CYP2D6, CYP3A4	CYP3A4
190	methoxyphenamine	CYP2D6	CYP2D6	CYP2D6	CYP2D6
191	**mexiletine**	CYP1A2, CYP2D6	CYP2D6	CYP2D6	CYP2D6
192	**montelukast**	CYP2C9, CYP3A4	CYP3A4	CYP3A4	CYP3A4
193	**omeprazole**	CYP2C9, CYP3A4	CYP3A4	CYP1A2	CYP3A4
194	**ondansetron**	CYP1A2, CYP2D6, CYP3A4	CYP1A2	CYP3A4	CYP3A4
195	phenylbutazone	CYP2C9	CYP1A2	CYP1A2	CYP3A4
196	quercetin	CYP3A4	CYP1A2	CYP1A2	CYP1A2
197	**ramelteon**	CYP1A2, CYP2C9, CYP3A4	CYP3A4	CYP2C9	CYP3A4
198	remoxipride	CYP2D6	CYP3A4	CYP3A4	CYP3A4
199	sertindole	CYP3A4	CYP3A4	CYP3A4	CYP3A4
200	sparteine	CYP2D6	CYP2D6	CYP2D6	CYP2D6
201	sulfamethizole	CYP2C9	CYP2C9	CYP2C9	CYP3A4
202	sulfidimidine	CYP3A4	CYP3A4	CYP2C9	CYP3A4
203	**tamsulosin**	CYP2D6, CYP3A4	CYP3A4	CYP3A4	CYP3A4
204	**theophylline**	CYP1A2, CYP2E1	CYP1A2, CYP3A4	CYP1A2	CYP1A2
205	**tolterodine**	CYP2D6, CYP3A4	CYP2D6, CYP3A4	CYP2D6, CYP3A4	CYP2D6
206	trimethoprim	CYP2C9	CYP3A4	CYP3A4	CYP3A4
207	**valdecoxib**	CYP2C9, CYP3A4	CYP2C9	CYP2C9	CYP3A4
208	zidovudine	CYP3A4	CYP3A4	CYP3A4	CYP1A2
209	**zileuton**	CYP1A2, CYP2C9, CYP3A4	CYP3A4	CYP2E1	CYP1A2

[a] The ct-SVM/*5classes* and CPG-NN/*5classes* models are multilabel; the SVM/*5classes* model was applied by using the single-label approach. [b] Multilabel substrates in bold.

metabolized by both CYP2D6 and CYP3A4, the main isoforms responsible in drug metabolism, so it is difficult to correctly assign the class. Our models are not always able to strictly distinguish between CYP2D6 and CYP3A4 substrates. Similar considerations can be made for trimethoprim (no. 206). Experimentally, this drug is a CYP2C9 substrate, but it is classified as a CYP3A4 substrate by both the multi- and the single-label models. This predicted profile corresponds to the results of the recently published models.[20] It is reported to be a neutral molecule, according to the chemical profile of CYP2C9 substrates, highly similar to the drug trimetrexate, which is metabolized by the CYP3A4 isoform and included in the training sets. Regarding the multilabel compounds in Table 21, most of the ct-SVM/*5classes* predictions are correct, even if partial. Dyhydrocodeine (no. 177) and tolterodine (no. 205) are correctly assigned to both CYP2D6 and CYP3A4 classes by the ct-SVM/*5classes* model, whereas a partial classification of dyhydrocodeine as a CYP2D6 substrate is obtained by the CPG-NN/*5classes* model. The drug clomipramine (no. 173) is metabolized by four different CYP450 isoforms, and the ct-SVM/*5classes* model correctly predicted three of them (CYP1A2, CYP2D6, and CYP3A4), whereas the CPG-NN/*5classes* classifier and single-label approach can assign only one class. Similarly, methadone (no. 189) is predicted by the multilabel models to be metabolized by the CYP2D6 and CYP3A4 isoforms, showing a good correspondence with the experimental metabolic profile. These examples confirm that the multilabel approach is able to perform an extensive investigation of the drug metabolism, whereas the single-label results are limited to the prediction of a single class. Considering the remaining multilabel compounds in Table 21, all of the models are able to correctly assign one of the experimental classes for each compound. In more detail,

aripiprazole (no. 169), desipramine (no. 176), formoterol (no. 182), mexiletine (no. 191), montelukast (no. 192), omeprazole (no. 193), ondansetron (no. 194), ramelteon (no. 197), tamsulosin (no. 203), valdecoxib (no. 207), and zileuton (no. 209) are predicted to be metabolized by one isoform, coherently with the experimental classes. Omeprazole was reported to be an inductor of CYP1A2: the prediction of the CPG-NN/*5classes* model is not totally incorrect, because this drug interacts with CY1A2 without being metabolized. In these cases, the performances of the multi- and single-label models are very close. Some single-label compounds were found to be multilabel after application of the ct-SVM/*5classes* classifier. The drug lisuride (no. 186) is a CYP3A4 substrate, correctly predicted by both the multilabel CPG-NN/*5classes* and SVM/*5classes* models and assigned to both CYP1A2 and CYP3A4 classes by the ct-SVM/*5classes* model. On the other hand, lobeline (no. 187) is reported to be metabolized by CYP2D6, but it is classified as a CYP3A4 substrate and as a CYP1A2, CYP2D6, and CYP3A4 substrate by the single- and multilabel models, respectively. In the first case, the ct-SVM/*5classes* model prediction is partially correct, but one of the results is confirmed by both the CPG-NN/*5classes* and SVM/*5classes* models. On the contrary, the SVM/*5classes* model prediction of lobeline is incorrect; nevertheless, both multilabel models assign it to one of the correct classes. So, for these specific compounds, the multilabel methodology represents a valid alternative. Regarding the drug theophylline (no. 204), both multi- and single-label approaches are able to predict one of the experimental isoforms responsible for its metabolism. This analysis confirms that the ct-SVM/*5classes* and CPG-NN/*5classes* models have similar performances. If compared to the single-label model, they provide a more truthful description of the metabolism.

CONCLUSIONS AND PERSPECTIVES

In the present work, we present several classification strategies to predict the isoform specificity of known CYP1A2, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 substrates. The multilabel approach was applied to different data sets, including seven and five classes, using the ct-SVM, M_L-*k*NN, and CPG-NN methods.

The best model (ct-SVM/*5classes*) was derived after the selection of 27 descriptors and yielded 77.5−96.6% correct predictions for the five classes of the corresponding test set. Similarly, the CPG-NN/*5classes* model achieved 75.6−97.1% correct predictions. A five-class data set was used to perform an extensive single-label classification analysis, in combination with automatic variable selection. The highest predictivity on the corresponding test set was achieved by the SVM/*5classes* model based on 19 descriptors, with 78% correct predictions. All of the presented models show acceptable performances; nevertheless, the multilabel prediction results more coherently reflect the real metabolic fate of the drugs.

In comparison to the previous work published by Terfloth et al.,[20] we increased the number of the classes (SVM/*5classes* model) and extended the analysis to substrates metabolized by more than one CYP450 isoform.

In conclusion, our results underline the high complexity of this classification problem and suggest the application of the multilabel approach to predict isoform specificity. The advantage of the CPG-NN technique is the graphical visualization of the results. Both the ct-SVM and the CPG-NN strategies might be extended to quantitative data. The multilabel methodology might be used to explore the metabolic profile of new chemical entities, and its prediction capability might be improved by collecting other multilabel substrates in the database.

ACKNOWLEDGMENT

We gratefully thank Elsevier MDL Inc. for providing the Metabolite reaction database. Our work was assisted by Molecular Networks GmbH with their C++ chemoinformatics toolkit MOSES and the package ADRIANA.Code for descriptor calculation, as well as SONNIA for the development of the CPG-NN models. J.G. and L.T. gratefully acknowledge BMBF for financial support in the funding initiative "Systems of Life−Systems Biology" (Grant 0313080). The molecular modeling work coordinated by S.M. was carried out with financial support from the University of Padova, Italy, and the Italian Ministry for University and Research (MIUR), Rome, Italy.

Supporting Information Available: Complete data set used in the analysis (Table 1); the descriptors selected (Table 2), performance measures (Tables 3 and 4), and predicted results for test set 1 (Table 5) in the multilabel classification model applied using the descriptor set from the publication of Terfloth et al.;[20] the model parameters (Tables 6−9), together with the CPG-NN/*5classes* model percentage (%) of correct predictions (Table 10). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) de Groot, M. J.; Kirton, S. B.; Sutcliffe, M. J. In Silico Methods for Predicting Ligand Binding Determinants of Cytochrome P450. *Curr. Top. Med. Chem.* **2004**, *4*, 1803–1824.
(2) Lewis, D. F. V.; Modi, S.; Dickins, M. Structure−Activity Relationship for Human Cytochrome P450 Substrates and Inhibitors. *Drug Metab. Rev.* **2002**, *34*, 69–82.
(3) de Groot, M. J. Designing Better Drugs: Predicting Cytochrome P450 Metabolism. *Drug Discovery Today* **2006**, *11*, 601–606.
(4) Mann, H. J. Drug-Associated Disease: Cytochrome P450 Interactions. *Crit. Care Clin.* **2006**, *22*, 329–345.
(5) Lynch, T.; Price, A. The Effect of Cytochrome P450 Metabolism on Drug Response, Interactions, and Adverse Effects. *Am. Fam. Physician* **2007**, *76*, 391–396.
(6) Anzenbacher, P.; Anzenbacherová, E. Cytochrome P450 and Metabolism of Xenobiotics. *Cell. Mol. Life Sci.* **2001**, *58*, 737–747.
(7) Kalra, B. S. Cytochrome P450 Enzyme Isoforms and Their Terapeutic Implications: An Update. *Indian J. Med. Sci.* **2007**, *61*, 102–116.
(8) Brown, C. M.; Reisfeld, B.; Mayeno, A. N. Cytochrome P450: A Structure-Based Summary of Biotransformations Using Representative Substrates. *Drug Metab. Rev.* **2008**, *40*, 1–100.
(9) Ingelman-Sundberg, M. Human Drug Metabolising Cytochrome P450 Enzymes: Properties and Polymorphisms. *Biomed. Life Sci.* **2003**, *369*, 89–104.
(10) Crivori, P.; Poggesi, I. Computational Approaches for Predicting CYP-related Metabolism Properties in the Screening of New Drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795–808.
(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.
(12) Yamashita, F.; Hara, H.; Ito, T.; Hashida, M. Novel Hierarchical Classification and Visualization Method for Multiobjective Optimization of Drug Properties: Application to Structure−Activity Relationship Analysis of Cytochrome P450 Metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 364–369.
(13) Block, J. H.; Henry, D. R. Evaluation of Descriptors and Classification Schemes to Predict Cythchrome Substrates in Terms of Chemical Information. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 385–392.

(14) Vermeulen, N. P. Prediction of Drug Metabolism: The Case of Cytochrome P450 2D6. *Curr. Top. Med. Chem.* **2003**, *3*, 1227–1239.
(15) de Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. Cytochrome P450 in Silico: An Integrative Modeling Approach. *J. Med. Chem.* **2005**, *48*, 2725–2755.
(16) Fox, T.; Kriegel, J. M. Machine Learning Techniques for in Silico Modeling of Drug Metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579–1591.
(17) 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.
(18) Arimoto, R. Computational Models for Predicting Interactions with Cytochrome P450 Enzyme. *Curr. Top. Med. Chem.* **2006**, *6*, 1609–1618.
(19) Yap, C. W.; Xue, Y.; Chen, Y. Z. Application of Support Vector Machines to in Silico Prediction of Cytochrome P450 Enzyme Substrates and Inhibitors. *Curr. Top. Med. Chem.* **2006**, *6*, 1593–1607.
(20) Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-Based Models for the Isoform Specificity of Cytochrome P450 3A4, 2D6 and 2C9 Substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.
(21) Spycher, S.; Pellegrini, E.; Gasteiger, J. Use of Structure Descriptor to Discriminate between Modes of Toxic Action of Phenols. *J. Chem. Inf. Model.* **2004**, *45*, 200–208.
(22) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VHC: Weinheim, Germany, 1999.
(23) Hristozov, D.; Gasteiger, J.; Da Costa, F. B. Multilabeled Classification Approach to Find a Plant Source for Terpenoids. *J. Chem. Inf. Model.* **2008**, *48*, 56–67.
(24) Bonnabry, P.; Sievering, J.; Leemann, T.; Dayer, P. Quantitative Drug Interactions Prediction System (Q-DIPS). A Dynamic Computer-Based Method to Assist in the Choice of Clinically Relevant in Vivo Studies. *Clin. Pharmacokinet.* **2001**, *40*, 631–640.
(25) P450 Drug Interaction Table. http://medicine.iupui.edu/clinpharm/ddis/table.asp (accessed Feb 10, 2008).
(26) Detail-Document #220233, Pharmacist's Letter, 2006. http://www.pharmacistsletter.com (accessed Feb 10, 2008).
(27) Manga, N.; Duffy, J. C.; Rowe, P. H.; Cronin, M. T. Structure-Based Methods for the Prediction of the Dominant P450 Enzyme in Human Drug Biotransformation: Consideration of CYP3A4, CYP2C9, CYP2D6. *QSAR Environ. Res.* **2005**, *16*, 43–61.
(28) Metabolite Database; MDL Inc. http://www.mdl.com/products/predictive/metabolite/index.jsp (accessed May 3, 2008).
(29) The PubChem Project, National Library of Medicine, National Institutes of Health. http://pubchem.ncbinlm.nih.gov/ (accessed Feb 27, 2008).
(30) Drug Bank. http://www.drugbank.ca (accessed Feb 12, 2008).
(31) *CACTVS Chemoinformatics Toolkit*; Xemistry GmbH: Königstein, Germany. http://www.xemistry.com (accessed Feb 3, 2008).
(32) *ADRIANA.Code*; Molecular Networks GmbH: Erlangen, Germany. http://www.molecular-networks.com (accessed February 10, 2008).
(33) R Development Core Team. *R: A Language and Environment for Statistical Computing*, version 2.2.1, 2005. http://www.r-project.org (accessed Jun 2006).
(34) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. *e1071: Misc functions of the Department of Statistics (e1071)*; TU Wien: Vienna, Austria, 2005.
(35) *SONNIA*; Molecular Networks GmbH: Erlangen, Germany. http://www.molecular-networks.com (accessed Apr 2008).
(36) *Weka: Waikato Environment for Knowledge Analysis*; University of Waikato: Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/ (accessed May 27, 2008).
(37) Witten, I. H.; Frank, E. *Data Mining: Pratical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.
(38) *CORINA*; Molecular Networks GmbH: Erlangen, Germany. http://www.molecular-networks.com (accessed Feb 2008).
(39) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug. Delivery Res.* **1997**, *23*, 3–25.
(40) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
(41) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
(42) Gasteiger, J.; Hutchings, M. G. New Empirical Models of Substituent Polarisability and Their Application to Stabilisation Effects in Positively Charged Species. *Tetrahedron Lett.* **1983**, *24*, 2537–2540.
(43) Gasteiger, J.; Hutchings, M. G. Quantitative Models of Gas-Phase Proton-Transfer Reaction Involving Alcohols, Ethers, and Their Thio Analogues. Correlation Analysis Based on Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489–6495.
(44) Kang, Y. K.; Jhon, M. S. Additivity of Atomic Polarisabilities and Dispersion Coefficients. *Theor. Chim. Acta* **1982**, *61*, 41–48.
(45) Miller, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.
(46) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000; Vol. 11, pp 1−667.
(47) Hall, L. H.; Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling*; VHC: New York, 1991; Vol. 2, pp 367−422.
(48) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
(49) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.
(50) Bath, P. A.; Poirette, A. R.; Willet, P.; Allen, F. H. The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 714–716.
(51) Tanford, C. *Physical Chemistry of Macromolecules*; Wiley: New York, 1961.
(52) Volkenstein, M. V. *Configurational Statistics of Polymeric Chains*; Wiley-Interscience: New York, 1963; pp 1−562.
(53) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures, Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757–764.
(54) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359–360.
(55) Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J. Representation of Molecular Electrostatic Potential by Topological Feature Maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608–4620.
(56) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
(57) Boutell, M. R.; Luo, J.; Shen, X.; Brown, C. M. C. Learning Multi-Label Scene Classification. *Pattern Recogn.* **2004**, *37*, 1757–1771.
(58) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(59) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.
(60) Smola, A. J.; Scholkopft, B. Learning with Kernels. *Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.
(61) Zhang, M. L.; Zhou, Z. H. A *k*-Nearest Neighbor Based Algorithm for Multi-Label Classification. In *IEEE International Conference on Granular Computing*; IEEE Press: Picataway, NJ, 2005; Vol. 2, pp 718−721.
(62) Landwehr, N.; Hall, M.; Frank, E. Logistic Model Trees. *Machine Learning* **2005**, *59*, 161–205.
(63) Platt, J. *Fast Training of Support Vector Machine using Sequential Minimal Optimization*; MIT Press: Cambridge, MA, 1999; pp 185−208.
(64) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.* **2001**, *13*, 637–649.

CI900299A