QSAR
& Combinatorial Science

COREPA-M: A Multi-Dimensional Formulation of COREPA

Ovanes Mekenyan[a]*, Nina Nikolova[b], Patricia Schmieder[c] and Gilman Veith[d]

[a] Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", 8010 Bourgas, Bulgaria
[b] Central Laboratory of Parallel Processing, Bulgarian Academy of Sciences, "Acad. G.Bonchev" str. 25A, 1756 Sofia, Bulgaria
[c] U.S. Environmental Protection Agency, Mid-Continent Ecology Division, 6201 Congdon Blvd., Duluth, MN 55804, U.S.A.
[d] International QSAR Foundation for Reducing Animal Testing, Duluth, Minnesota, U.S.A.

Full Paper

Recently, the COmmon REactivity PAttern (COREPA) approach was developed as a probabilistic classification method which was formalized specifically to advance mechanistic QSAR development by addressing the impact of molecular flexibility on stereoelectronic properties of chemicals. In the initial version of COREPA, the probability distributions for only one stereoelectronic parameter at a time were analyzed for the series of chemicals under analysis. To go beyond considering probability distributions of one parameter at a time requires the capability of analyzing a suite of parameters simultaneously for each chemical. This work creates that capability for a multi-dimensional formulation of the COREPA which is expected to enhance the reliability of the method to discriminate complex patterns. Using probability distance measures such as Kullback-Leibler divergence and Hellinger distance, the set of parameters are defined that best discriminate activity. The COREPA-M system automatically identifies the parameters that best discriminates chemicals in groups defined by comparable reactivity endpoints. A detailed Bayesian decision tree is then used for classifying untested chemicals with measures of "goodness of fit" criteria. COREPA-M is illustrated using the example of modelling binding affinity of chemicals at the aryl hydrocarbon receptor.

1 Introduction

The evolution of QSAR approaches for chemical design and risk management involves the development of new methods for quantifying molecular structure, the identification of more mechanistic endpoints within biological pathways, and more objective approaches for discovering plausible mechanistic structure-activity relationships. The use of stereoelectronic parameters in quantifying structural variation in heterogeneous datasets and libraries requires a formal treatment of the flexibility of chemicals and the possibility that even a moderately flexible chemical can have many low-energy structures which are not adequately represented by minimum energy conformations. The lowest-energy conformer might have weak interactions with macromolecules or steric incompatibilities whereas other conformations within permitted energies boundaries may have strong interactions [1 – 5]. For example, QSARs for binding affinity to the aryl hydrocarbon receptor (AhR) using minimum energy conformations have generally failed, whereas QSARs using charge-transfer parameters computed for the most planar conformations were successful [6]. Even in the AhR model, nonetheless, the selection of the most planar conformation was only a reasonable assumption based on knowledge of the receptor which was imposed on the QSAR analysis instead of being derived directly from the data. For more complex QSAR explorations to be successful without a priori assumptions of geometry, a formal mathematical approach is needed to derive models for complex interactions.

The COmmon REactivity PAttern (COREPA) formalism treats this complex QSAR exploration as a classification task [2, 3]. Classification methods identify criteria which will classify an unknown object into predefined classes using a training set of objects from multiple classes. Probabilistic methods, discriminant analyses, nearest-neighbour classifiers, neural networks and decision trees are representative classification techniques. The COREPA formalism uses a Bayesian probabilistic method to identify common structural characteristics among chemicals that elicit similar biological activity, or class; but does so in a context that allows many possible conformations of individual chemicals and the probability distribution of molecular descriptor values instead of single parameter values for each chemical.

* To receive all correspondence

Key words: QSAR, chemical screening, drug design, Bayesian chemistry

The common structural characteristics can then be objectively encoded into a decision tree to screen large and structurally heterogeneous chemical libraries for the sought-after biological activity.

Our recent presentation of COREPA focused on mathematical and combinatorial problems of migrating from two dimensional (2D) chemical structures to three dimensional (3D) conformations and the identification of common patterns within single parameter distributions. Available algorithms for 2D-3D structure migration can introduce significant non-deterministic variation and ultimately affect the QSAR outcome if active conformers are designated by ad hoc selection [1, 4, 7]. The COREPA method examines the distribution of all energetically reasonable conformations for the structure, but selects active conformations on a case-by-case basis using the activity endpoints of specific studies. To achieve this capability, COREPA introduced a new approach to evaluate similarity between chemicals based on parameter distributions derived from the distribution of 3-D conformations. Typically, this requires aligning with the template structures that could yield ambiguous results when molecular flexibility is taken into account. COREPA circumvents the problem of structure alignment by defining and analyzing the common reactivity pattern. The common reactivity pattern consists of the normalized sum of conformational distributions of chemicals across the descriptor axis. The descriptors which best discriminating chemicals into classes of biological similarity are selected among those potentially associated with the specific biological endpoint. Thus, instead of aligning structural representatives of chemicals, their conformational distributions are compared.

In the original formulation of the method the common reactivity patterns have been determined across single parameter axis in terms of parameter ranges. While simple for interpretation, the one-dimensional formulation significantly limited the discrimination ability of the COREPA approach. The classification model, represented by a decision tree, consisted of multiple hierarchically ordered rules based on the parameter ranges that comprise common reactivity patterns. The decision tree within the original formulation, however, was built manually which imposed ambiguity and difficulties in statistical evaluation of the obtained results.

The present paper describes a new formulation of the COREPA method, where the efforts have been focused in two directions: developing multi-dimensional reactivity patterns (COREPA-M) and automated building of the decision trees under user defined constraints. The mathematical formalism of COREPA-M will be presented in details along with an example of deriving a COREPA decision tree for screening libraries

2 Methods

2.1 AhR Receptor Binding Data

The performance of the COREPA-M formulation was tested using a high quality database for the binding affinity of chemical to the aryl hydrocarbon receptor reported by Safe et al. [8]. The binding affinity, EC_{50}, is defined as the concentration of the test chemical necessary to reduce the specific binding of [3H]TCDD (2,3,7,8- tetrachlorodibenzo-p-dioxin) to 50% of the maximal value in the absence of the competitor. Ligands for this receptor include many polycyclic aromatic hydrocarbons and many adverse "dioxin-like" effects have been linked to the AhR binding as the molecular initiating event for these toxicity pathways. Consequently, AhR binding is one important endpoint in the evaluation of possible toxic effects of drugs, pesticides and industrial chemicals. The binding affinities for a broad range of polychlorinated biphenyls (PCBs), polychlorinated dibenzofurans (PCDFs) and polychlorinated dibenzo-p-dioxins (PCDDs) were compiled from the literature as described previously [8], and the compilation is presented in Table 1 to aid in the discussion of the COREPA-M models.

2.2 Ligand Conformational Analysis

The OASIS (Optimized Approach based on Structural Indices Set) software [9] was used to generate conformations of all ligands in the database [4, 7]. Although, the best approach for computing plausible conformations for large, highly flexible chemicals may be the OASIS genetic algorithm which avoids some combinatorial problems by minimizing similarity of conformers [4], a second option available for less flexible chemicals like PCBs is 3DGEN [7] which is a combinatorial approach to generate all conformers in the context of steric constraints on distances between non-bonded atoms, ring-closure limits, torsional resolution and expert rules for the likelihood of intramolecular hydrogen bonds, cis/trans and optical isomers. A unique aspect of the approach involves the initial propagation of a cyclic 3-D molecular skeleton prescribed by a topological ranking a recursive procedure based on the 3-D information of previously established bonds. This includes the atom type and hybridization of the atoms incident to the bond being constructed as well as the two atoms associated with the previously completed bond. Cyclic fragments incident to the bond being constructed are also retained. Bond lengths and valence angles are determined through a molecular mechanics parameterization. During the propagation of the acyclic components, cyclic character is gained through defined ring-closure constraints. Rotamers associated with all torsional angles that meet hybridization and specified geometric constraints are retained. Thus, the approach identifies flexibility in saturated cyclic and acyclic molecules where techniques involving only rotations around acyclic single bonds do not.

Strain within plausible conformations is corrected with a pseudo molecular mechanics (PMM) strain-relief procedure based on a truncated force field energy-like function, where the electrostatic terms are omitted [4, 7]. The basic form and parameterization of the interatomic interactions were taken from the Chem.-X force field [10, 11]. Geometry optimization was achieved using MOPAC 93 [12, 13] with the AM1 Hamiltonian. The conformers are also screened to eliminate those whose heat of formation (ΔH_o^{ff}) is greater from the ΔH_o^{ff} associated with the conformer with lowest energy by a threshold of 15 – 20 kcal/mol and eliminate conformational degeneracy [14, 15].

2.3 Molecular Descriptors

There are hundreds of molecular descriptors being used in QSAR and few have consistent mechanistic interpretation over large diverse sets of chemicals. When there is sufficient mechanistic evidence for the structure-activity relation of interest, an informal assessment to limit the molecular descriptors is prudent [16]. From preceding studies of AhR binding affinity of PCBs, PCDFs and PCDDs, available information suggests that charge-transfer interactions control binding energy and there are substantial steric constraints [6, 8]. Therefore, to demonstrate COREPA-M, we limited the list of molecular descriptors to LUMO energy (E_{LUMO}), HOMO energy (E_{HOMO}), HOMO-LUMO energy gap (E_{gap}), Electronegativity EN; dipole moment (μ), volume polarizability (VolP; defined as a sum of atomic self-polarizabilities), maximum donor (D_{max}) and acceptor (A_{max}) delocalizabilities, maximum (Q_{max}) and minimum (Q_{min}) charges, maximum (B_order_max) and minimum ((B_order_min) bond order, degree of stretching or compactness (quantified by the geometric analogue of Wiener index, GW, i.e., by the sum of interatomic steric distances), greatest interatomic distance (L_{max}), planarity (normalized sum of torsion angles in a molecule); effective cross-section diamater (Diameff), maximum diamater (Diammax), distance between wild card heteroatoms (O, N, F, Cl, Br, I). Physicochemical and volumetric indices – $\log K_{OW}$, Van der Waals volume (VAN_D_WAALS_VOL.), surface (VAN_D_WAALS_SUR.), solvent accessible surface (SAS1.5; assuming water as a solvent) calculated by making use of Connoly algorithm [17] and charged partial surface areas (CPSAs) as introduced in [18] by Stanton and Jurs.

3 The COREPA Method

Similarity is inherently a multi-dimensional problem. Classification methods seek to define n-dimensional patterns as well as decision boundaries between the groups in n-dimensional space. It is beyond the scope of this paper to review these methods, but common limitations in the measure of similarity between chemicals are the use of point estimates for chemical and distance measures as a measure of similarity. Euclidean distances can only be linear which limits application to classifications where active and inactive chemicals can be separated by a hyper-surface.

3.1 Probabilistic Approach to Quantification of Chemicals Structures

COREPA uses a probabilistic approach based on Bayes theorem and provides a theoretically optimal decision rule [19, 20]. The Bayesian minimum error decision rule guarantees lowest classification error if the class-conditional probability distributions are known. This requirement is met in COREPA by estimating the class-conditioned probability distribution through a series of mathematical approximations. While the number of plausible conformers for a flexible chemical is large, only a representative sample of the conformers are needed to estimate the resulting probability distributions for a each molecular descriptor for each chemical. The probability distribution for the conformers is approximated from a Boltzman energy distribution. The probability of forming a specific conformer is $p(x \mid C_{sj})$, where C_{sj} denotes the *j-th* conformer of the chemical *S*. The probability of that chemical having a specific molecular descriptor value is denoted as $p(x \mid S_i)$ in Eq. 1:

$$p(x|S_i) = \sum_{j=1}^{R_i} p(C_{ij}) p(x|C_{ij}) \tag{1}$$

where S_i is the *i*-th chemical in the data set, Ri is the number of conformers for the compound Si, and p(Cij) is the probability to have the j-th conformer of a i-th compound.

The Boltzman probability $p(C_{ij})$ can be estimated with Eq. 2:

$$p(C_{ij}) = \frac{e^{-\Delta E_j / k_B T}}{\sum_{m=1}^{N} e^{-\Delta E_m / k_B T}}, \tag{2}$$

where $\Delta E_j = E_j - E_{min}$ *and* E_{min} is the energy of the conformer with minimal energy.

The application of formula for the kernel density estimate to $p(x \mid Cij)$, gives

$$p(x|C_{ij}) = \frac{1}{N_{ij} h} \left[\sum_{k=1}^{N_{ij}} \varphi\left(\frac{x - x_{ijk}}{h}\right) \right] \tag{3}$$

where N_{ij} is the number of values of descriptor x for *j-th* conformer of *i-th* chemicals.

Substitution of Eq. 2 and Eq. 3 into Eq. 1 allows calculation of a *conformational distribution* of a compound across a descriptor x.

$$p(x|S_i) = \sum_{j=1}^{R_i} \frac{e^{-\Delta E_j / k_B T}}{\sum_{m=1}^{N} e^{\Delta E_m / k_B T}} \left[\frac{1}{N_{ij} h} \sum_{k=1}^{N_{ij}} \varphi\left(\frac{x - x_{ijk}}{h}\right) \right] \tag{4}$$

To create a probability distribution for each value of descriptor x a kernel density function [21] is superimposed on each individual data point, and these data density kernels are summed and normalized to give an overall probability distribution. The kernel density function, φ(x), provides a bounded symmetrical probability distribution function for estimation of the class-conditional probability distribution as shown in Eq. 5:

$$p(x) = \frac{1}{nh}\sum_{k=1}^{n}\varphi\left(\frac{x - x_k}{h}\right), \tag{5}$$

where h is a smoothing parameter. The smoothing function can be optimized through cross-validation; however, COREPA-M sets the initial smoothing as $h = 1.059\sigma\sqrt[5]{n}$, σ being the standard deviation of the data set and n being the number of data points [21].

Figure 1 illustrates the transform of discrete values of molecular descriptors for a specific conformation into a distribution around each value. In addition to varying the exploration the possible distribution of the molecular descriptors for a chemical by adjusting the smoothing function, COREPA-M provide options for Gaussian, Lorenz, Laplace, or Epanechnikov kernel density functions to make sure the QSAR outcome is not sensitive to selection of φ(x).

The "weight" of the kernel density is proportional to the probability of finding a conformation for any specific value of a molecular descriptor. The estimated probability distribution for the molecular descriptor is computed as the sum of the conformer kernel densities allong the axis as shown in Eq. 6.

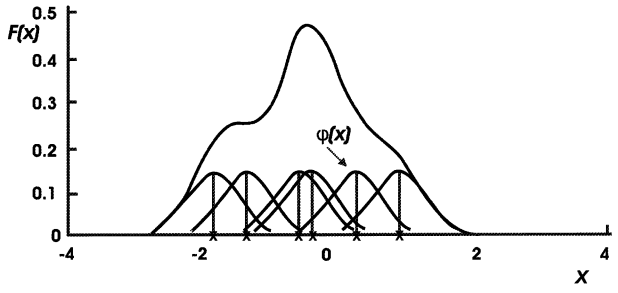


Figure 1. An illustration of the kernel density estimation

$$p(x|S_i) = \sum_{j=1}^{R_i}\sum_{k=1}^{N_{ij}}\frac{\alpha_{ij}}{h}\varphi\left(\frac{x - x_{ijk}}{h}\right),$$

$$\alpha_{ij} = \frac{e^{-\Delta E_j/k_B T}}{N_{ij}\sum_{m=1}^{N} e^{-\Delta E_m/k_B T}} \tag{6}$$

To illustrate this approach, conformations of Genestin and ICI_182780 were generated and the energies of the highest occupied molecular orbital E(HOMO) were computed as shown in Figure 2. These chemicals have significantly different E(HOMO) when computed for the lowest energy conformations; however, depending on which conformations were chosen to compare the two chemicals, different conclusions would be drawn about which chemical has greater E(HOMO). Figure 2 illustrates that some of the conformations have overlapping E(HOMO).

As a final stage in the quantification of chemical structure, COREPA generates the overall probability distribution for descriptor x for each chemical as shown in the composite graphs in Figure 3. Rather than using point estimates for a molecular descriptor from a single conformation of a chemical, the molecule is represented as probability distributions for all molecular descriptors in subsequent analysis of molecular similarity and differences with respect to classes of biologically active chemicals.

3.2 Similarity Between Chemical Structures

One immediate application of the estimation of the probability distributions for the molecular descriptors of two chemicals would be to compare the distributions as a measure of molecular similarity [22]. The similarity between chemicals can be quantified using probability distance such as the Kullback-Leibler divergence or the Chernov, Bhattacharyya, Matusita, Hellinger, Mahalanobis, Patrick-Fisher distances [20, 23–26]. The Hellinger distance is used in COREPA-M software and is calculated with Eq. 7:

$$HD^2_{1,2} = HD(p_1(x), p_2(\mathrm{x})) = \int\left(\sqrt{p_1(x)} + \sqrt{p_2(x)}\right)^2 dx \tag{7}$$

The minimum value of the Hellinger distance is zero, and it is reached when two probability density functions are the

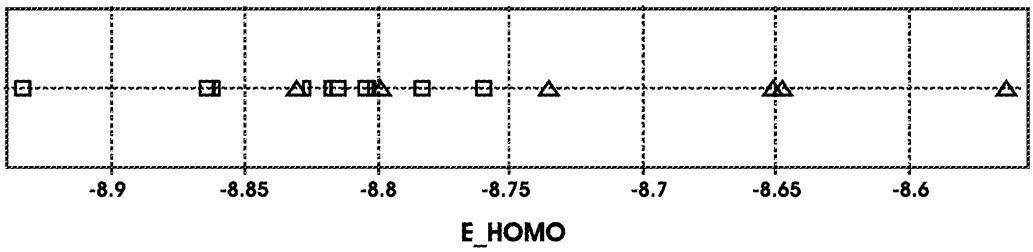


Figure 2. Representation of two chemicals with their discrete conformer distributions across E(HOMO) descriptor axis; Genestin with CAS = 129453618 and ICI 182780 with CAS = 529599.

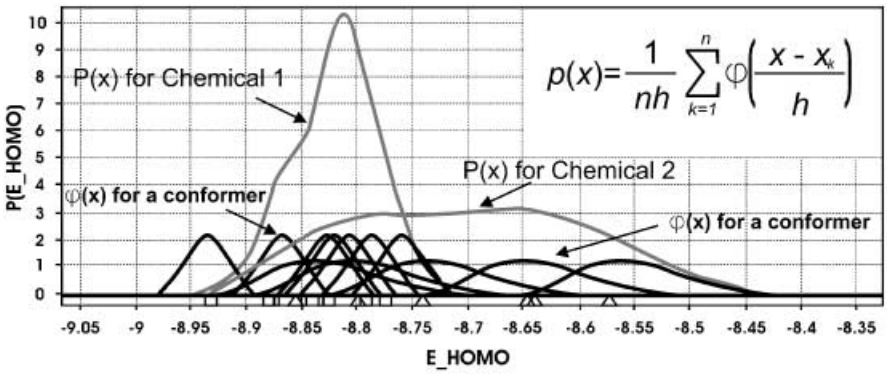


Figure 3. The conformer distribution of two chemicals

same. The maximum value of the Hellinger distance is two, and it is reached when two probability density functions are most distinct. Higher Hellinger distance values mean high dissimilarity between chemical probability distributions.

3.3 Similarity Between Classes of Chemical. Class-conditioned Probability Distributions

The larger objective of COREPA-M, however, is to develop a mathematical formalism for comparing the class-conditioned probability distributions of one class of chemicals with those of a second class as well as to assigning unclassified chemicals to one of the predefined classes based on their molecular descriptors. In the general form, the class-conditional probability is given by Eq. 8. The probability of having descriptor x with certain value, while being in *class m* can be calculated by:

$$p(x|class_m) = \sum_{i=1}^{M_m} P(S_i)p(x|S_i), \tag{8}$$

where M_m is the number of chemicals in $class_m$; $p(x \mid S_i)$ is defined by Eq. 6; and $P(S_i)$ is an *a priori* probability of a compound S_i to belong to $class_m$. If the *a priori* probabilities are approximately equal, the *class conditional* probability can be calculated by Eq. 9:

$$p(x|class_m) = \sum_{i=1}^{M_m} \frac{1}{M_m} p(x|S_i) \tag{9}$$

Substitution of Eq. 6 into Eq. 9 gives a formula for calculation of class-conditional probability density for a set of flexible chemicals with known descriptor x values:

$$p(x|class_m) = \frac{1}{M_m} \sum_{i=1}^{M_m} \sum_{j=1}^{R_i} \sum_{k=1}^{N_{ij}} \frac{\alpha_{ij}}{h} \varphi\left(\frac{x - x_{ijk}}{h}\right) \tag{10}$$

Using the kernel density estimation formula (Eq. 3) requires calculation of the kernel function for each chemical in the class. When N is large, the computational time becomes prohibitive and the probability density can be streamlined using preliminary binning and Fast Fourier transform methods [21].

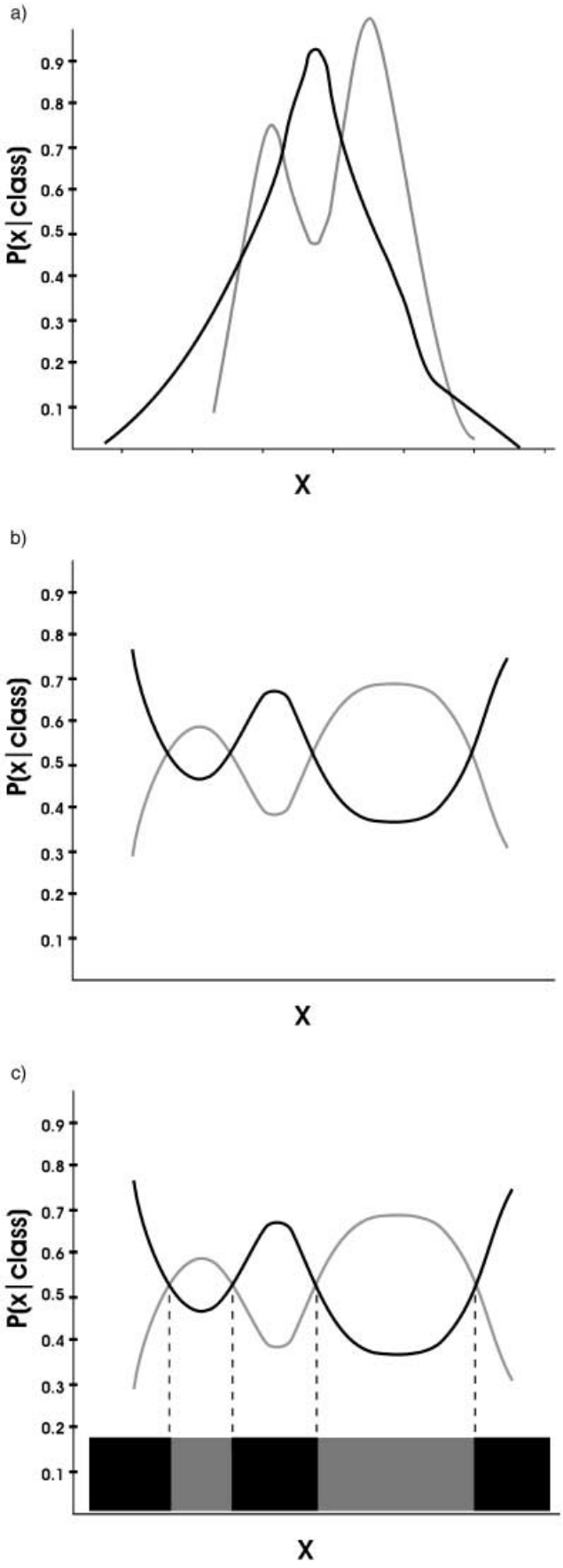
3.4 Decision Rules in Classification. A posteriori Probability Distributions

The class-conditioned probability distribution density can be interpreted as a reactivity pattern with respect to molecular descriptor x and for a class of chemicals with common biological activity. While $p(x \mid class_m)$ is a measure of similarity between chemicals in the class, only in the simplest cases would two or more classes of chemicals be delineated by a single molecular descriptor. Moreover, in addition to examining probabilitiy distributions in multiple demensions, successful classification requires the capability of having a chemicals with probabilities for belonging to more than one class. The decision rules for classification is summarized in Figure 4 where $p(x \mid class_m)$ is shown for two classes in 4(a). The Bayesian *a posteriori* probability distribution, $p(class_m \mid x)$, is computed in 4(b) and the decision rule is to classify a chemical with molecular descriptor x into the class with the greatest $p(class_m \mid x)$ as shown in 4(c).

3.5 Bayesian Multidimensional Classification

COREPA-M was formulated to facilitate the exploration of reactivity patterns for classes of chemicals which are defined using multiple molecular descriptors. Bayes formula has no restriction on the dimensionality of involved probability functions and class-conditioned probabilities using multiple molecular descriptors $x_1 \ldots x_n$.

$$P(class_i|(x_1, \ldots x_n)) = \frac{P(class_i)p((x_1, \ldots x_n)|class_i)}{\sum_{j=1}^{k} P(class_j)((x_1, \ldots x_n)|class_j)}, \tag{11}$$



◀ **Figure 4.** (a) Class-conditional probabilities; (b) *a posteriori* probabilities and classification for two classes of chemicals.

Again, the decision rule for classification is to put a new chemical into the class with the greatest *a posteriori* probability. The only difference is that, if joint probability $p((x_1, \ldots x_n) \mid class_m)$ is to be estimated via kernel density technique, n-dimensional kernels should be placed on every data point and then summed. In the initial version of COREPA, the probability distributions are estimated only across single descriptor. COREPA-M allows multidimensional probability density estimation. For the sake of simplicity and faster calculation, descriptors are assumed to be independent. This allows calculation of joint probability as product of one-dimensional probabilities using Eq. 12:

$$p((x_1, \ldots x_n)|class_i) = \prod_{k=1}^{n} p(x_k|class_i) \tag{12}$$

Individual probability densities $p(x_k \mid class_i)$ are estimated through kernel density estimation as explained in section above. The assumption of othogonal descriptors is a severe requirement and generally cannot be met in most real data sets. Violation of the assumption of independence leads to lower classification quality. While the solution may be to use multivariate joint density estimates, sufficient data are seldom available to do so.

3.6 Bayesian Decision Networks

The final step in exploring structure-activity relationships is to objectively identify the molecular descriptors that best explain the variance in the data and provide mechanistic interpretations. COREPA-M uses class conditional probabilities to discriminate between classes of chemicals and identifies those molecular descriptors which show the least amount of overlap between class-conditional probabilities. The power of a molecular descriptors to distinguish chemical classes is measured in COREPA-M by Hellinger distance between classes. Mathematically, the delineation power of each descriptor is defined by Eq. 13:

$$DescriptorsQuality\ (x_1, \ldots, x_n, class_k) = HD(p(x_1, \ldots, x_n|class_k), p\left(x_1, \ldots, x_n \middle| \bigcup_{\substack{i=1 \\ i \neq k}}^{K} class_i\right) \tag{13}$$

To develop the classification rule-base which can be used for screening untested chemicals in heterogenous libraries, COREPA-M creates a binary decision network or tree from the probability distributions. The "best" set of descriptors is selected and chemicals are classified according to Bayesian

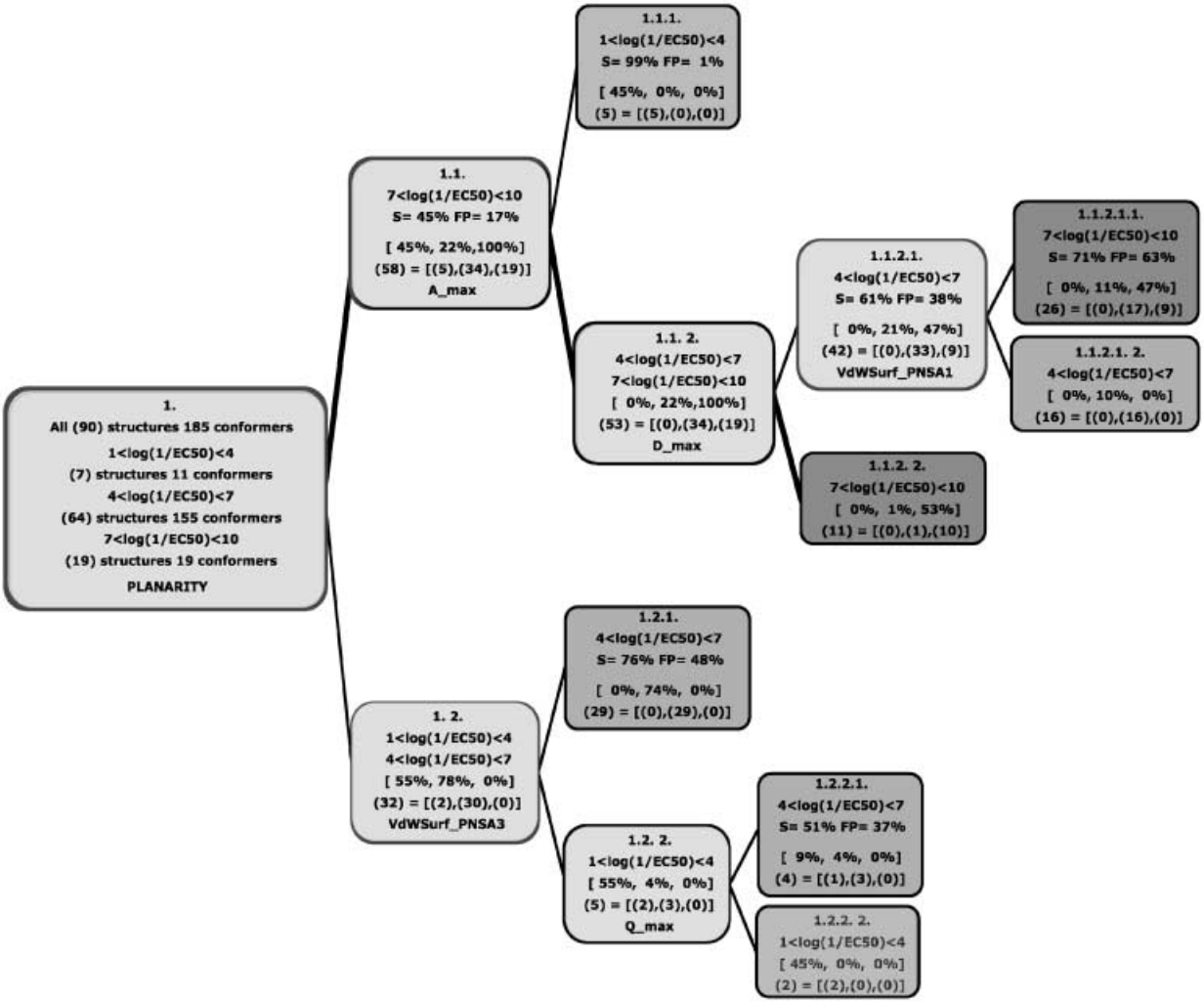


Figure 5. The COREPA-M decision tree for discriminating the three classes of chemicals according to their potency: $1 < \log(1/EC_{50}) \leq 4$ (class I), $4 < \log(1/EC_{50}) \leq 7$ (class II), and $7 < \log(1/EC_{50}) \leq 10$ (class III); generated by not applying "keep together" classification scheme, n = 1 (maximum two parameters at a node), P = 50%, (posterioir probability threshold, P%), Chi sq = on, and Hellinger Distance for selecting best parameters. Colors from the row model are not visualized in the graph.

minimum error decision rule. The training set is split according to Bayesian minimum decision rule with rejection. On creating a node, the class-conditional probability densities for the "best set" of parameters are evaluated and stored in the node. The a *posteriori* probabilities for all the conformers in the training set are evaluated as:

$$p(class_i|[x_1,\ldots,x_m]) = \frac{p([x_1,\ldots,x_m]class_i)}{\sum_{c=1}^{K} p([x_1,\ldots,x_m]class_c)} \qquad (14)$$

The conformers, which have maximum value of the probability to belong to "class to split" follow "Yes" branch. All other chemicals follow the "No" branch. The propagation of the decision tree is continued until stopping criteria such as 95% confidence limits, minimum numbers of conformers is a class, or maximum depth of the tree as set by the user are met.

3.7 Cross-validation

The COREPA-M decision tree could be cross-validated by deriving it with part of the data only and using the rest as and external test data. When *n*-fold cross-validation is applied, the initial data set is divided into *n* subsets; the decision tree is derived *n* times and for each of them one of the *n* subsets is eliminated from initial training set. The eliminated subsets of chemicals are used as external validation sets for decision trees derived using the rest of the initial training set. Eventually, each of the chemicals from initial training set is predicted within the respective external validation sets. The statistical estimates of the decision tree (sensitivity, S; rate of false positives, FP; concordance, C, etc.) derived on whole training set are compared with those (S_{cross}, FP_{cross}, and C_{cross}) associated with the models based on the reduced training sets; one should emphasize that cross-validated estimates

are based on the same number of chemicals as that of the initial training set (each of the chemicals from initial training set can be found on one of the *n* eliminated subsets). Similarly, one could compare the lists of the best-selected parameters of these models. The smaller the differences between goodness criteria of the models based on whole training set and part of it the more robust is the model and more reliable the predictions for unknown chemicals.

4 Results and Discussion

The need for an objective method to explore structure-activity relationships can be demonstrated in a comparatively simple case of binding to the Ah receptor. Clearly, multivariate methods are adequate to explore relationships between activity and chemicals which have discrete molecular discriptors . However, when flexible chemicals are included and the molecular descriptors vary with conformation, objectivity is lost when a specific conformation is chosen for the exploration. For example, in modeling binding to the AhR, stereoelectronic descriptors did not predict binding unless we imposed the planarity constraint on the "active form" and used descriptors for the most planar conformation. COREPA-M was created to see if a formal methods could determine these patterns without investigator judgement.

The conformers for the PCBs, PCDFs and PCDDs were generated and optimized for the chemicals in the training set (Table 1), producing a total of 185 structures including 125 conformers of the 30 PCBs congeners and a single conformer for each of PCDFs and PCDDs. Each individual chemical (and all possible conformations thereof) was classified into one of three classes of reactivity based on the binding to the Ah receptor using the following thresholds: Class I-least active with $1 < \log(1/EC_{50}) \leq 4$ and seven

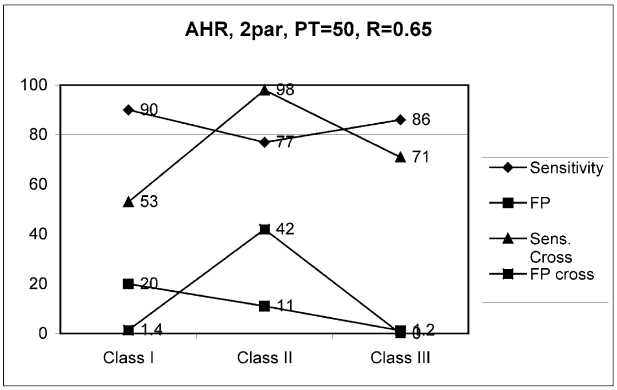
Table 1. Observed and predicted AhR binding affinities. The COREPA-M model from Figure 5 was used for predicting AhR binding affinity classes.

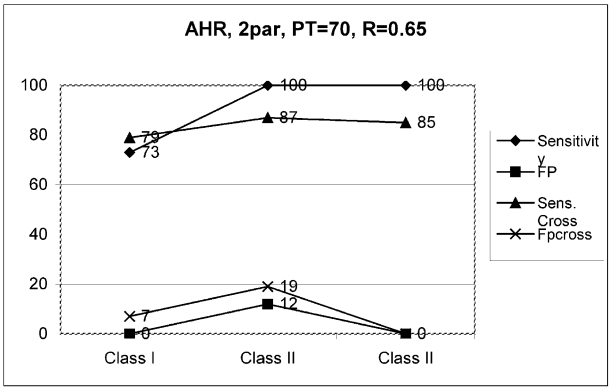
No.	Name	Observed		Predicted			
				Probabilities			Ultimate class
		$Log1/EC_{50}$	class	class III	class II	class I	
1	3,3′,4,4′-Tetrachlorobiphenyl	6,15	2	0.23	1	0	2
2	2,3,4,4′-Tetrachlorobiphenyl	4,55	2	0.23	1	0.1	2
3	3,3′,4,4′,5-Pentachlorobiphenyl	6,89	2	0.23	1	0	2
4	2′,3,4,4′,5-Pentachlorobiphenyl	4,85	2	0.23	1	0	2
5	2,3,3′,4,4′-Pentachlorobiphenyl	5,37	2	0.23	1	0	2
6	2,3′,4,4′,5-Pentachlorobiphenyl	5,04	2	0.23	1	0	2
7	2,3,4,4′,5-Pentachlorobiphenyl	5,39	2	0.23	1	0	2
8	2,3,3′,4,4′,5-Hexachlorobiphenyl	5,15	2	0.23	1	0	2
9	2,3′,4,4′,5,5′-Hexachlorobiphenyl	4,8	2	0.23	1	0	2
10	2,3,3′,4,4′,5′-Hexachlorobiphenyl	5,33	2	0.23	1	0	2
11	2,2′,4,4′-Tetrachlorobiphenyl	3,89	1	0.23	0	1	1
12	2,2′,4,4′,5,5′-Hexachlorobiphenyl	4,1	2	0.23	1	0	2
13	2,3,4,5-Tetrachlorobiphenyl	3,85	1	0.23	0.59	1	1
14	2,3′,4,4′,5′,6-Hexachlorobiphenyl	4	2	0.23	1	0	2
15	4′-Hydroxy-2,3,4,5-tetrachlorobiphenyl	4,05	2	0.23	1	0	2
16	4′-Methyl-2,3,4,5-tetrachlorobiphenyl	4,51	2	0	1	0	2
17	4′-Fluoro-2,3,4,5-tetrachlorobiphenyl	4,6	2	0.23	1	0	2
18	4′-Methoxy-2,3,4,5-tetrachlorobiphenyl	4,8	2	0	1	0	2
19	4′-Acetyl-2,3,4,5-tetrachlorobiphenyl	5,17	2	0	1	0	2
20	4′-Cyano-2,3,4,5-tetrachlorobiphenyl	5,27	2	0.23	1	0.39	2
21	4′-Ethyl-2,3,4,5-tetrachlorobiphenyl	5,46	2	0	1	0	2
22	4′-Bromo-2,3,4,5-tetrachlorobiphenyl	5,6	2	0.23	1	0	2
23	4′-Iodo-2,3,4,5-tetrachlorobiphenyl	5,82	2	0.23	1	0	2
24	4′-Isopropyl-2,3,4,5-tetrachlorobiphenyl	5,89	2	0	1	0	2
25	4′-Trifluormethyl-2,3,4,5-tetrachlorobiphenyl	6,43	2	0	1	0	2
26	3′-Nitro-2,3,4,5-tetrachlorobiphenyl	4,85	2	0.23	1	0	2
27	4′-N-Acetylamino-2,3,4,5-tetrachlorobiphenyl	5,09	2	0	1	0	2
28	4′-Phenyl-2,3,4,5-tetrachlorobiphenyl	5,18	2	0.23	1	0	2
29	4′-t-Butyl-2,3,4,5-tetrachlorobiphenyl	5,17	2	0	1	0	2
30	4′-n-Butyl-2,3,4,5-tetrachlorobiphenyl	5,13	2	0	1	0	2
31	2,3,7,8-Tetrachlorodibenzo-p-dioxin	8	3	0.79	0.21	0	3
32	1,2,3,7,8-Pentachlorodibenzo-p-dioxin	7,1	3	0.55	0.45	0	3
33	2,3,6,7-Tetrachlorodibenzo-p-dioxin	6,8	2	0.79	0.21	0	3
34	2,3,6-Trichlorodibenzo-p-dioxin	6,66	2	0.49	1	0	2

Table 1. (cont.)

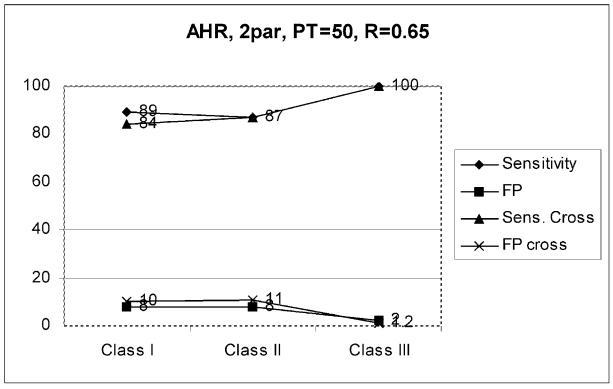
No.	Name	Observed		Predicted			
				Probabilities			Ultimate class
		Log1/EC_{50}	class	class III	class II	class I	
35	1,2,3,4,7,8-Hexachlorodibenzo-p-dioxin	6,55	2	0.47	1	0	2
36	1,3,7,8-Tetrachlorodibenzo-p-dioxin	6,1	2	0.66	0.34	0	3
37	1,2,4,7,8-Pentachlorodibenzo-p-dioxin	5,96	2	0.4	0.6	0	2
38	1,2,3,4-Tetrachlorodibenzo-p-dioxin	5,89	2	0.32	0.68	0	2
39	2,3,7-Trichlorodibenzo-p-dioxin	7,15	3	0.71	0.29	0	3
40	2,8-Dichlorodibenzo-p-dioxin	5,5	2	0	1	0	2
41	1,2,3,4,7-Pentachlorodibenzo-p-dioxin	5,19	2	0.54	0.46	0	3
42	1,2,4-Trichlorodibenzo-p-dioxin	4,89	2	0.79	0.21	0	3
43	1,2,3,4,6,7,8,9-Octachlorodibenzo-p-dioxin	5	2	0	1	0	2
44	1-Chlorodibenzo-p-dioxin	4	2	0.14	0.86	0	2
45	2,3,7,8-Tetrabromodibenzo-p-dioxin	8,82	3	1	0	0	3
46	2,3-Dibromo-7,8-dichlorodibenzo-p-dioxin	8,83	3	1	0	0	3
47	2,8-Dibromo-3,7-dichlorodibenzo-p-	9,35	3	1	0	0	3
48	2-Bromo-3,7,8-trichlorodibenzo-p-dioxin	7,94	3	1	0	0	3
49	1,3,7,8,9-Pentabromodibenzo-p-dioxin	7,03	3	1	0	0	3
50	1,3,7,8-Tetrabromodibenzo-p-dioxin	8,7	3	1	0	0	3
51	1,2,4,7,8-Pentabromodibenzo-p-dioxin	7,77	3	1	0	0	3
52	1,2,3,7,8-Pentabromodibenzo-p-dioxin	8,18	3	1	0	0	3
53	2,3,7-Tribromodibenzo-p-dioxin	8,93	3	1	0	0	3
54	2,7-Dibromodibenzo-p-dioxin	7,81	3	0.94	0.06	0	3
55	2-Bromodibenzo-p-dioxin	6,53	2	0.74	0.26	0	3
56	2-Chlorodibenzofuran	3,55	1	0	0	1	1
57	3-Chlorodibenzofuran	4,38	2	0	1	0.3	2
58	4-Chlorodibenzofuran	3	1	0	0	1	1
59	2,3-Dichlorodibenzofuran	5,33	2	0	1	0	2
60	2,6-Dichlorodibenzofuran	3,61	1	0	0	1	1
61	2,8-Dichlorodibenzofuran	3,59	1	0	0	1	1
62	1,3,6-Trichlorodibenzofuran	5,36	2	0.66	0.34	0	3
63	1,3,8-Trichlorodibenzofuran	4,07	2	0	1	0	2
64	2,3,4-Trichlorodibenzofuran	4,72	2	0.65	0.35	0	3
65	2,3,8-Trichlorodibenzofuran	6	2	0	1	0	2
66	2,6,7-Thrichlorodibenzofuran	6,35	2	0	1	0	2
67	2,3,4,6-Tetrachlorodibenzofuran	6,46	2	0.7	0.3	0	3
68	2,3,4,8-Tetrachlorodibenzofuran	6,7	2	0	1	0	2
69	1,3,6,8-Tetrachlorodibenzofuran	6,66	2	0.56	0.44	0	3
70	2,3,7,8-Tetrachlorodibenzofuran	7,39	3	0.78	0.22	0	3
71	1,2,4,8-Tetrachlorodibenzofuran	5	2	0	1	0	2
72	1,2,4,6,7-Pentachlorodibenzofuran	7,17	3	0.79	0.21	0	3
73	1,2,4,7,9-Pentachlorodibenzofuran	4,7	2	0.38	0.62	0	2
74	1,2,3,4,8-Pentachlorodibenzofuran	6,92	2	0	1	0	2
75	1,2,3,7,8-Pentachlorodibenzofuran	7,13	3	0.77	0.23	0	3
76	1,2,4,7,8-Pentachlorodibenzofuran	5,89	2	0.78	0.22	0	3
77	2,3,4,7,8-Pentachlorodibenzofuran	7,82	3	0.77	0.23	0	3
78	1,2,3,4,7,8-Hexachlorodibenzofuran	6,64	2	0.79	0.21	0	3
79	1,2,3,6,7,8-Hexachlorodibenzofuran	6,57	2	0.56	0.44	0	3
80	1,2,4,6,7,8-Hexachlorodibenzofuran	5,08	2	0.36	0.64	0	2
81	2,3,4,6,7,8-Hexachlorodibenzofuran	7,33	3	0.56	0.44	0	3
82	2,3,6,8-Tetrachlorodibenzofuran	6,66	2	0.77	0.23	0	3
83	1,2,3,6-Tetrachlorodibenzofuran	6,46	2	0.65	0.35	0	3
84	1,2,3,7-Tetrachlorodibenzofuran	6,96	2	0.77	0.23	0	3
85	1,3,4,7,8-Pentachlorodibenzofuran	6,7	2	0.77	0.23	0	3
86	2,3,4,7,9-Pentachlorodibenzofuran	6,7	2	0.54	0.46	0	3
87	1,2,3,7,9-Pentachlorodibenzofuran	6,4	2	0.32	0.68	0	2
88	2,3,4,7-Tetrachlorodibenzofuran	7,6	3	0.64	0.36	0	3
89	1,2,4,6,8-Pentachlorodibenzofuran	5,51	2	0.66	0.34	0	3
90	dibenzofuran	3	1	0	0	1	1

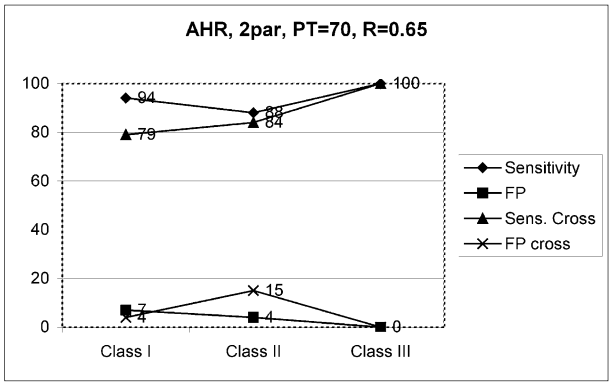
a)





b)





◀ **Figure 6.** Comparison between sensitivity and rate of false positive predictions of models with n = 2 derived when all chemicals have been included in correlation sample (S and FP, respectively) and after 4-fold cross validation (Scross, FPcross): with "keep together" classification scheme and P = 50% (a) and P = 70% (b) and without "keep together" classification scheme and P = 50% (c) and P = 70% (d).

chemicals, 11 conformers; Class II-with $4 < \log(1/EC_{50}) \leq 7$ and 64 chemicals, 155 conformers; Class III-most active with $7 < \log(1/EC_{50}) \leq 10$ and 19 chemicals- 19 conformers. A typical COREPA-M decision tree for discriminating these three classes of chemicals is illustrated in Figure 5.

The results of the COREPA-M classification can be interpreted as a potential model for statistical classification of chemicals. The predicted activity classes for the set chemicals using the decision tree from Figure 5 are listed in Table 1. The statistics of the COREPA models obtained without and with the "keep conformaters together" classification schemes are listed in Tables 2a and 2b, respectively. Models are evaluated by confusion matrix classification errors for chemicals and conformers. Goodness criteria, such as sensitivity and false positives for three classes of chemicals are presented as well when all chemicals are used for deriving models (S and FP, respectively) and after 4-fold cross validation (Scross, FPcross, respectively). S and FP are based on the classification of conformers of the chemicals according to the highest class-conditional probabilities, $p(x \mid \text{class}_i)$ Hence, this classification of chemicals is based on their most active conformers wherein the chemical is asigned to the class with highest activity reached by all conformers.

The increase in number of parameters in nodes appears to slightly increase the model accuracy. Thus, %Corr/Incorr predictions for chemicals are 80/20% and 89/11% for n = 1and n = 2, respectively (same holds for other statistical estimates, as seen in Table 2b). Increasing the posteriori probability thresholds from P = 50% to P = 70% was expected to enhance the rate of false negative identifications at the cost of reduced rates of false positives. However, while the false positives were reduced, the rates of false negative identifications were not always increased. In general, the "keep together" classification scheme demonstrated better performance. This clearly can be seen from the summary classification errors listed in the confusion matrices. The same holds for goodness criteria (S and FP). Moreover, "keep together" models are more stable than the alternative classification scheme with respect to 4-fold cross-validation. As demonstrated in Figure 6 (and Table 2), the difference in S and Scross , and FP and FPcross are smaller for COREPA-M models derived with "keep together" classification scheme. In general, cross-validation analysis showed that all derived COREPA models are stable especially at higher thresholds for posteriori probabilities (P%).

Behind the classification results in the COREPA approach, however, are the estimated class-conditional prob-

Table 2. Classification error for the COREPA-M decision trees for AhR binding affinity as listed in the confusion matrices for chemicals and conformers. The COREPA models are obtained without (a, b) and with (c, d) applying the "keep conformaters together" classification schemes. Sensitivity and False positive predictability of models (S, Scross and FP, FPcross, respectively) are listed when all chemicals are used in correlation sample (Tables 2a and 2c, respectively) and after 4-fold cross validation (Tables 2b and 2d, respectively). Models are derived with n-parameters at a node, posteriori probability threshold, P%, Chi sq = on, Hellinger Distance for selecting best parameters and without (a) and with (b) using "keep together" classification schemes.

a

n	P%	Chemicals (Confision Matrix)				Conformers (Confision Matrix)				Number
		Class I Error%	Class II Error%	ClassIII Error%	Summary %Corr/Incorr	Class I Error%	Class II Error%	ClassIII Error%	Summary %Corr/Incorr	nodes/leafs
1	50	0	27	12	80/20	0	12	9	90/10	6/7
	70	0	27	12	80/20	0	12	9	90/10	6/7
2	50	10	20	0	82/18	10	11	0	89/11	8/9
	70	26	0	0	94/6	26	0	0	96/4	9/10
3	50	5.2	5.2	0	95/5	5.2	3	0	97/3	7/8
	70	16	5	0	92/8	16	3	0	95/5	9/10

b

n	P%	Chemicals (Probabilities)			Chemicals (Probabilities)Cross-validation		
		Class I S/FP	ClassII S/FP	ClassIII S/FP	Class I S/FP	ClassII S/FP	ClassIII S/FP
1	50	100/25	72/0	100/0	68/20	78/23	100/0
	70	100/25	72/0	100/0	90/20	78/12	86/0
2	50	90/20	77/11	86/1	53/1	98/42	71/0
	70	74/0	100/11	100/0	68/6	88/15	86/2.4
3	50	95/4	95/4	100/0	79/3	95/19	86/1.2
	70	84/1.4	97/8	100/0	79/4	95/19	71/0

c

n	P%	Chemicals (Confision Matrix)				Conformers (Confision Matrix)				Number
		Class I Error%	Class II Error%	ClassIII Error%	Summary %Corr/Incorr	Class I Error%	Class II Error%	ClassIII Error%	Summary %Corr/Incorr	nodes/leafs
1	50	0	28	0	80/20	0	12	0	90/10	6/7
	70	5	9	0	92/8	5	4	0	96/4	11/12
2	50	10	12	0	89/11	10	5	0	95/5	10/11
	70	5	12	0	90/10	5	5	0	95/5	10/11
3	50	0	19	0	87/13	0	7	0	94/6	8/9
	70	42	3	0	89/11	42	1	0	95/5	10/11

d

n	P%	Chemicals (Probabilities)			Chemicals (Probabilities)Cross-validation		
		Class I S/FP	ClassII S/FP	ClassIII S/FP	Class I S/FP	ClassII S/FP	ClassIII S/FP
1	50	100/25	72/0	100/0	79/14	84/15	100/0
	70	95/4	91/0	100/0	90/18	80/4	86/0
2	50	89/8	88/8	100/2	84/10	88/11	100/1.2
	70	95/7	88/4	100/0	79/4	84/15	100/0
3	50	100/17	81/8	71/0	79/0	100/19	86/0
	70	58/0	97/15	86/0	89/13	86/12	71/0

abilities which may be more instructive than the classification, itself. For example, chemical #13 (2,3,4,5-tetrachlorobiphenyl) has three conformers, and two of them were predicted to belong to class I with a class-conditional probabilities of 1.0, whereas the third conformers had probabilities of 0.41, 0.59, 0.23 for belonging to classes I, II, and III, respectively. According to maximum $p(x \mid class_i)$ rule, this conformer was assigned to class II. Hence, two of the three conformers of the chemical were classified to belong to class I, and one – to class II. Eventually, the

Table 3. Descriptors in COREPA-M model for AhR binding affinity listed with their importance for discriminating three classes of chemicals as assessed by Hellinger distance; i.e., *DescriptorsQuality* as defined by Eq. 13.

	Class III	Class II	Class I
Descriptors	$7 < \log(1/EC50) < 10$	$4 < \log(1/EC50) < 7$	$1 < \log(1/EC50) < 4$
PLANARITY	1.1	0.54	0.39
E GAP	0.92	0.16	0.34
VdWSurf PNSA1	0.7	0.2	0.32
D max	0.67	0.38	0.56
B ord max	0.61	0.28	0.44
VOLUME POLARIZAB.	0.6	0.39	0.8
B ord min	0.5	0.27	0.56
A max	0.39	0.24	0.61
Log(Kow)	0.37	0.14	0.71
VdWSurf PPSA1	0.37	0.19	0.68
DIPOLE MOMENT	0.34	0.21	0.26
VdWSurf PPSA3	0.34	0.23	0.58
Energy LUMO	0.28	0.08	0.43
Diameff	0.25	0.32	0.28
VdWSurf PNSA3	0.21	0.33	0.38
Q max	0.21	0.7	0.29
E HOMO	0.15	0.18	0.1
ELECTRONEGATIVITY	0.08	0.11	0.36

chemical was classified to belong to class I because has a conformer whose class-conditional probability to belong to class I ($p(x \mid classI) = 1$) was higher than the probability of the other conformer to belong to class II ($p(x \mid classII) = 0.59$). Thus, the probabilistic nature of the COREPA method allows introducing the notion of continuous behaviour of chemicals. A chemical may be classified predominantly in one of the activity classes, and at the same time have finite or nearly equal probabilities of belonging to other classes of biologically active chemicals.

Returning now to the use of COREPA-M decision trees to provide insight into the mechanism involved, Figure 7 illustrates a progression of descriptors that discriminate the assigned abitrary classes. Planarity, Amax and Dmax were the most important discriminators and the one-dimensional COREPA patterns for these descriptors at the subsequent nodes are presented in Figure 6a – 6c, respectively. The data show that the most active chemicals (class III) have maximum class-conditional probability at the lowest range of planarity index (i.e., highest planarity, Figure 7a), and highest delocalizabilities – A_max (Figure 7b) and D_max (Figure 7c). This derived result from COREPA-M is consistent with the literature that the AhR is flat and with our earlier finding that superimposing the "planar conformation" constraint on the calculation of stereoelectronic descriptors in developing the QSAR for AhR. Moreover, the COREPA analysis placed a descriptor of molecular shape as the first order parameter and then found realistic electronic descriptors that decriminated at lower levels in the tree. If this finding hold for other biological receptors, it will provide an objective steric filter for chemicals that are flexible enough to conform to the steric requirements of a binding site.

The parameters with highest importance for AhR activity are planarity, E_gap, D_max and CPSAs (VdWSurf_PNSA1). The parameters selected by the system are in accordance with the experimentally established stacking type of interaction of ligands with AhR. This interaction is conditioned by a charge transfer process, which could be determined by the energy of frontier orbitals or their difference (E_gap), electron delocalizability (D_max) and charged surfaces (partially negative surface area, VdWSurf_PNSA1). Logically, these parameters are less important (i.e., have lower *DescriptorsQualityIv* values) for classes II and III that consist of chemicals with lower binding affinity to AhR (Table 3).

To summarize the information derived from COREPA-M, the program provides: 1) a decision tree wherein each node illustrates the distribution pattern by visualizing the class-conditional probabilities, $p(x \mid class_i)$ in the descriptor space formed by the best parameters. The user has access to all chemicals and their conformers classified at certain node or leaf of the tree which can be saved as a binary file and used for chemical screening purposes; 2) a confusion matrix for all chemicals with information about actual and predicted classifications for the classification; 3) a list of most discriminating molecular descriptors, each of which is evaluated according to the calculated Hellinger distances for discrimination of all classes (see Table 3); 4) a summary table with class-conditional probabilities, $p(x \mid class_i)$, for each conformers across classes, and ultimate classification of conformers based on calculated posteriori probability $p(class_i \mid x)$ and Bayes classification rule; 5) a summary table with distribution of conformers of the chemicals in classes according to calculated posteriori probability $p(class_i \mid x)$ and Bayes classification rule; and 6) a summary table with

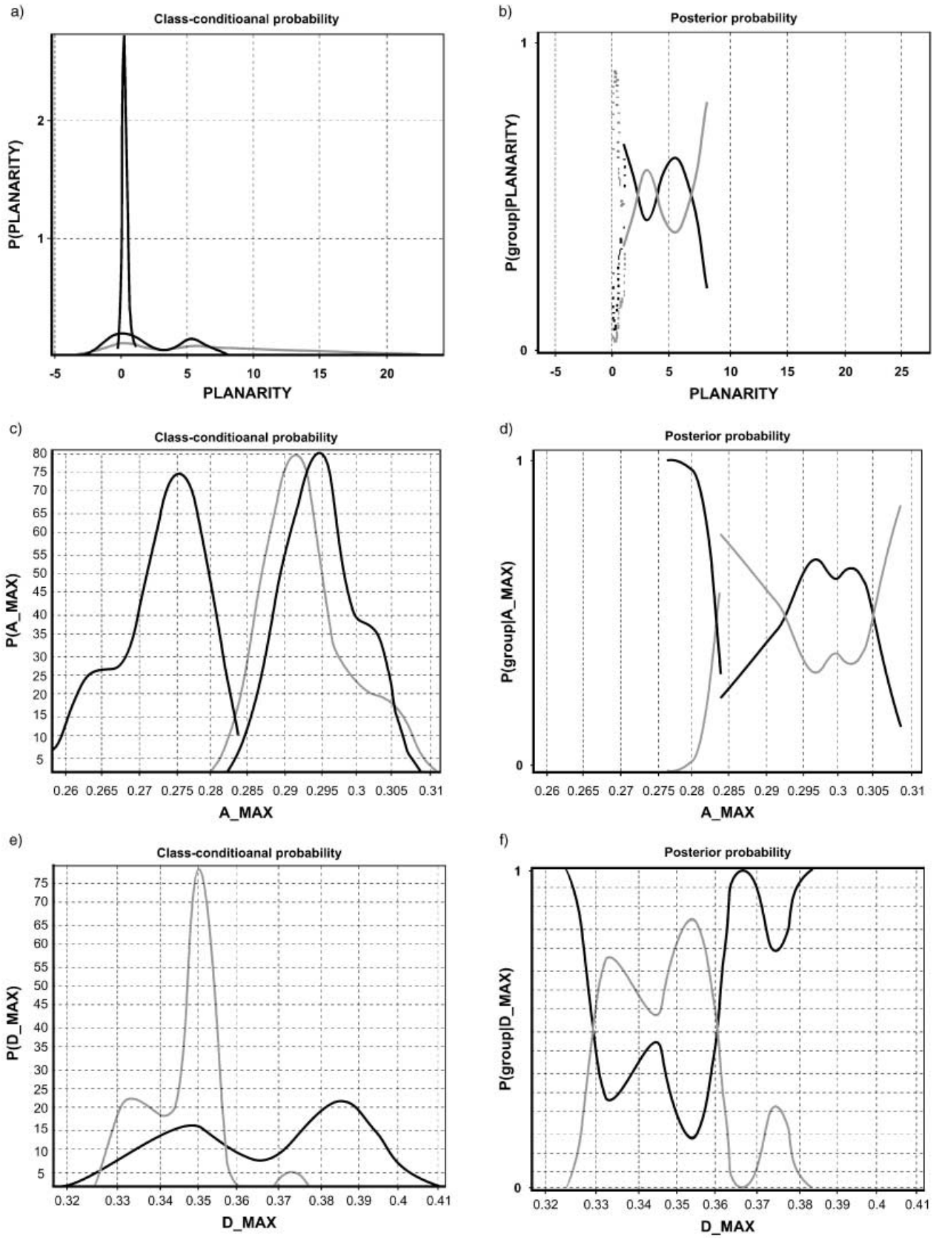


Figure 7. One-dimensional COREPA patterns across the path of the decision tree (the upper highlighted branch, in Fig. 5) providing classification of two most active classes (II and III). The patterns are described as distribution of class-conditional probabilities, $p(x|class_i)$, and posteriori probabilities $p(class_i|x)$ across Planarity (a), Amax (b) and Dmax (c), associated with the subsequent nodes.

highest class-conditional probabilities, $p(x \mid class_i)$, which conformers of the chemicals have reached for each class and classification of chemicals according to these maximal $p(x \mid class_i)$ (see Table 1).

5 Conclusions

The evolution of methods to quantify chemical structure includes the inclusion of many conformations for individual chemicals and the need for computerized systems to manage the distributions of molecular descriptors associated with the conformers. Identifying common reactivity patterns (COREPA) has been originally introduced for one molecular descriptor, this paper introduces a multi-dimensional formulation of the COREPA method. The method was evaluated using binding data for a receptor of known shape and for which charge-transfer interaction are important. The new formulation increased discrimination power of the method and allowed automated building of a decision tree using Bayesian decision rules for classification of biologically similar chemicals. The set of best discriminating parameters and class to split at a node are defined by evaluating similarity between conformer distributions of chemicals. Planarity at the AhR was found to be the most important descriptor for binding to this flat receptor. Stereoelectronic parameters related to charge-transfer interactions were predicters of binding with classes. The automated building of decision tree for classification of chemicals provided opportunity for a detailed statistical evaluation of derived model. N-fold cross validation is used for that purpose. The probabilistic nature of the COREPA classification supports the concept of a more continuous behaviour of chemicals because some conformers of the same chemical can be different classes.

Acknowledgements

This paper has been subjected to review by the USEPA National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. Research associated with this paper was funded in part through an EPA cooperative research agreement (CR828826).

References

[1] J. M. Ivanov, O. G. Mekenyan, S. P. Bradbury, G. Schuurmann, *Quant. Struct.-Act. Relat.* **1998**, *17*, 437 – 449.
[2] O. G. Mekenyan, J. M. Ivanov, S. Karabunarliev, S. Bradbury, G. Ankley, W. Karcher, *Environ. Sci. Technol.* **1997**, *31*, 3702 – 3711.
[3] O. G. Mekenyan, N. Nikolova, S. Karabunarliev, S. Bradbury, G. Ankley, B. Hansen, *Quant. Struct.-Act. Relat.* **1999**, *18*, 139 – 153.
[4] O. G. Mekenyan, D. Dimitrov, N. Nikolova, and S. Karabunarliev, *Chem. Inf. Comput. Sci.* **1999**, *39*, 997 – 1016.
[5] O. G. Mekenyan, *Curr. Pharm. Des.* **2002**, *8,* 1605 – 1624.
[6] O. G. Mekenyan, G. D. Veith, D. J. Call, G. T. Ankley, *Environ. Health Perspect.* **1996**, *104*, 1302 – 1310.
[7] J. Ivanov, S. Karabunarliev, O. Mekenyan, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 234 – 243.
[8] S. Safe, S. Bandiera, T. Sawyer, B. Zmudzka, G. Mason, M. Romkes, M. Dehomme, J. Sparling, A. Okey, T. Fujita, *Environ. Health Perspect*. **1985**, *61*, 21 – 23.
[9] O. Mekenyan, St. Karabunarliev, J. Ivanov, D. Dimitrov, *Comput. Chem.* **1994**, *18*, 173 – 187.
[10] E. K. Davies, N. W. Murrall, *Comput. Chem.* **1989**, *13*, 149 – 156.
[11] D. N. J. White, *Spec. Rep. Chem. Soc*. **1987**, *6*, 38 – 63.
[12] J. J. P. Stewart**,** *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1 – 105.
[13] J. J. P. Stewart, MOPAC 93, Fujitsu Limited, Chiba-city, Chiba 261, Japan, and Stewart Computational Chemistry, Colorado Springs, CO, **1993.**
[14] T. Wiese, S. C. Brooks, *J Steroid Biochem Molec Biol*. **1994**, *50*, 61 – 72.
[15] G. M. Anstead, K. E. Carlson, J. A. Katzenellenbogen, *Steroids* **1997**, *62*, 268 – 303.
[16] O. Mekenyan, N. Nikolova, P. Schmieder, *J. Mol. Structure* (*THEOCHEM*) **2003**, *622*, 147 – 165.
[17] M. L. Connoly, *J. Appl. Crystallogr.* **1983**, *16*, 548 – 558.
[18] D. T. Stanton, P. C. Jurs, *Anal. Chem*. **1990**, *62*, 2323 – 2329.
[19] R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, **1973.**
[20] R. O. Duda, P. E. Hart, D. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, **2000.**
[21] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, **1986.**
[22] O. G. Mekenyan, J. M. Ivanov, S. H. Karabunarliev, B. Hansen, G. T. Ankley, S. P. Bradbury, A new approach for estimating 3-D similarity that incorporates molecular flexibility, in F. Chen, G. Schuurmann (Eds.), *Quantitative Structure Activity Relationships in Environmental Sciences – VII*, SETAC Press, Pensacola, FL, USA, **1998**, pp. 39 – 57.
[23] L. Devroye, L. Gyorfi, G. Lugosi, *A probabilistic Theory of Pattern Recognition*, Springer, **1996.**
[24] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, **1998.**
[25] G. McLachlan, *Discriminant Analisys and Statistical Pattern Recognition*, John Wiley and Sons, **1992.**
[26] D. W. Scott, *Density Estimation*, Rice University, Houston, TX (http://rice.edu)

Received on October 14, 2003; Accepted on November 3, 2003