4D-Fingerprints, Universal QSAR and QSPR Descriptors

Craig L. Senese,[†] J. Duca,[‡] D. Pan,[†] A. J. Hopfinger,[†] and Y. J. Tseng*[,†]

Laboratory of Molecular Modeling and Design (MC 781), College of Pharmacy, The University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231, and Schering Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, New Jersey 07033

Received March 24, 2004

An elusive goal in the field of chemoinformatics and molecular modeling has been the generation of a set of descriptors that, once calculated for a molecule, may be used in a wide variety of applications. Since such *universal descriptors* are generated free from external constraints, they are inherently independent of the data set in which they are employed. The realization of a set of universal descriptors would significantly streamline such chemoinformatics tasks as virtual high-throughout screening (VHTS) and toxicity profiling. The current study reports the derivation and validation of a potential set of universal descriptors, referred to as the 4D-fingerprints. The 4D-fingerprints are derived from the 4D-molecular similarity analysis. To evaluate the applicability of the 4D-fingerprints as universal descriptors, they are used to generate descriptive QSAR models for 5 independent training sets. Each of the training sets has been analyzed previously by several varying QSAR methods, and the results of the models generated using the 4D-fingerprints are compared to the results of the previous QSAR analyses. It was found that the models generated using the 4D-fingerprints are comparable in quality, based on statistical measures of fit and test set prediction, to the previously reported models for the other QSAR methods. This finding is particularly significant considering the 4D-fingerprints are generated independent of external constraints such as alignment, while the QSAR methods used for comparison all require an alignment analysis.

INTRODUCTION

High on the "wish list" of many workers in chemoinformatics and molecular modeling is a set of descriptors that once computed for a chemical structure can be repetitively used in a large range of applications. Perhaps the most progress toward realizing a set of *general* descriptors has come in the development of methods to explore molecular similarity and diversity. The Tanimoto Coefficients are one example of a simple set of descriptors that have wide applications in the characterization of molecular similarity.[1,2] Another example is that of Cramer's BC(DEF) parameters.[3,4] Cramer found that certain parameters (the aforementioned "BC" parameters), ultimately derived by factor analysis, could be used to construct a linear relationship to each of several physical properties of a diverse set of pure liquids.

Most *general* descriptors are 2D descriptors that do not capture the 3D chemical structure of a molecule, let alone the range of conformational freedom available to a molecule. Recently, we have developed a methodology called *4D-Molecular Similarity Analysis*, or simply *4D-MS Analysis*,[5] based on the 4D-QSAR paradigm pioneered in our laboratory.[6] The 4D-MS methodology permits the generation of sets of molecular fingerprints that embed the conformational information of the molecule as well as capture its size and chemical structure. Moreover, each molecular "finger" of the molecular fingerprint is specific to a particular atom/pharmacophore type present in a molecule. A unique set of molecular fingerprints can be constructed for each specific alignment assigned to the members of a training set or library. Alignment dependent molecular fingerprints permit molecular similarity measures to be developed as a function of, for example, the binding mode to a receptor site. However, another unique set of molecular fingerprints in this class can be developed for any molecule which are *independent of alignment* but do encompass the ensemble of conformational states available to the molecule.

A method that can assemble a unique set of molecular fingerprints which embeds the size, chemical structure, conformation, and pharmacophore information about a molecule and is independent of alignment is attractive for both chemoinformatics and molecular modeling applications. Such molecular fingerprints represent descriptors that contain all the salient information about a molecule. Moreover, since these descriptors are independent of alignment, once computed, they can be used in any type of molecular modeling and/or chemometrics application. That is, these 4D-fingerprints are **universal** descriptors. Still, questions and concerns arise regarding if these descriptors are both significant and reliable. One way to explore answering these concerns is to see if these 4D-MS descriptors can be used to build QSAR models of comparable quality and reliability to those developed using other QSAR approaches. This paper reports the results of performing five comparison QSAR studies. The five training sets span ranges in molecular flexibility, size, and structural heterogeneity. The types of biological endpoints, and ranges in activity measures, also vary significantly across the five training sets, although such variances were not an objective in the selection of the training sets.

* Corresponding author phone: (312)996-4816; fax: (312)413-3479; e-mail: ytseng2@uic.edu.
[†] The University of Illinois at Chicago.
[‡] Schering Plough Research Institute.

10.1021/ci049898s CCC: $27.50
Published on Web 08/12/2004

Table 1. Interaction Pharmacophore Elements, IPEs, Used in 4D-QSAR and 4D-Fingerprint QSAR Analyses

IPE description (abbreviation)	IPE code
all atoms in the molecule (any)	0
nonpolar atoms (np)	1
polar (+) atoms (p+)	2
polar (−) atoms (p-)	3
hydrogen bond acceptor atoms (hba)	4
hydrogen bond donor atoms (hbd)	5
aromatic atoms (a)	6
non-hydrogen atoms (hs)	7

Finally, to fully assess the validity of the universal molecular fingerprints introduced in this study, two independent data fitting methods were employed to generate the QSAR models. The data fitting techniques used, namely the genetic function approximation (GFA),[7] employing multidimensional linear regression (MLR) and partial least squares (PLS)[8,9] regression, are widely used in many popular QSAR paradigms[6,10−11] and therefore provide a convenient basis for comparison.

METHOD

The universal 4D-fingerprints are eigenvalues from the eigenvectors determined for a molecule from its absolute molecular similarity main distance-dependent matrix (MDDM). The theory and corresponding methodology are presented in detail in a previous publication[5] and are only summarized here.

Elements of the Main Distance-Dependent Matrices (MDDM). The elements of each of the MDDM capture the intrinsic size, shape, and conformational flexibility of a compound. To evaluate how shape and flexibility are distributed with respect to the atoms composing the compound, the molecule is divided into "functional pieces" referred to as *interaction pharmacophore elements* (IPE's). A unique MDDM is constructed for each of the distinct IPE pairs and for the whole molecule (all IPE's). The ability to spatially consider compounds as a whole, in addition to their functional pieces of interest, is unique to the 4D modeling paradigm. The current definitions of each of the IPE types are listed in Table 1. The elements of the MDDM are defined as follows:

$$E_{(\upsilon,dij)} = e^{(-\upsilon\langle dij\rangle)} \quad (1)$$

The constant υ in eq 1 was determined such that the difference in the sum of eigenvalues for any two arbitrary compounds which have the same number, *N*, of a particular IPE type is maximized. The value of υ set to 0.25, on average, maximizes the sum of the differences in the IPE eigenvalues of any two compounds. The inductive derivation, along with a complete description of assigning a value to this constant, can be found in ref 5. The term $\langle dij\rangle$ refers to the average distance between the atom pair *ij* of IPE types *u* and υ, such that

$$\langle dij\rangle = \Sigma_k\, dij(k)\, p(k) \quad (2)$$

In eq 2, p(k) refers to the thermodynamic probability of the kth conformer state sampled in the assessment of conformational flexibility, and d*ij*(k) is the corresponding distance between atom pair *i* and *j* of IPE types *u* and υ for the kth conformer state.

Similarity eigenvalues are derived by the diagonalization of the MDDM. For same-term IPE pairs, i.e., $u = \upsilon$, the MDDM are square upper/lower triangular. These matrices can be directly diagonalized. The resulting eigenvalues determined from the MDDM are normalized and ranked in numerically descending order in their eigenvector representation. To calculate the *n* normalized eigenvalues for IPE type *m* of compound α, $\epsilon_{mn}(\alpha)$, the nonscaled eigenvalues $\epsilon_{mn}'(\alpha)$ are scaled relative to the rank of the MDDM

$$\epsilon_{mn}(\alpha) = \epsilon_{mn}'(\alpha)/\mathrm{rank}(\alpha)_m \quad (3)$$

Thus, $\epsilon_{0,3}(2)$ would correspond to the third eigenvalue of the MDDM for IPE type 0 of compound 2.

Determination of eigenvalues of the MDDM for $u \neq \upsilon$, the so-called cross-terms for IPE pairs that are not the same, requires a different strategy since these matrices may, or may not, be square. In the case of rectangular MDDM ($u \neq \upsilon$), the following square MDDM are constructed

$$\mathrm{MDDM}(u,u) = \mathrm{MDDM}(n_u, n_\upsilon) \times \mathrm{MDDM}(n_u, n_\upsilon)^T \quad (4)$$

and for $[n_\upsilon \times n_u]$

$$\mathrm{MDDM}(\upsilon,\upsilon) = \mathrm{MDDM}(n_\upsilon, n_u) \times \mathrm{MDDM}(n_\upsilon, n_u)^T \quad (5)$$

Since MDDM(*u*,*u*) and MDDM(υ,υ) have the same rank and trace, both have the same set of eigenvalues. Hence, for each pair of IPE's for which $u \neq \upsilon$

$$\epsilon(\alpha)_{u,\upsilon} = \{[\epsilon(\alpha)]_{\mathrm{MDDM}(u,u)}\}^{1/2} \quad (6)$$

There are 36 possible molecular similarity eigenvectors from the MDDM for each compound α, corresponding to all possible combinations of the eight IPE types. Once the similarity eigenvectors have been calculated for the set of compounds, the estimation of molecular similarity for a pair of compounds α and β begins with a definition for molecular dissimilarity, given by

$$D_{\alpha\beta} = \Sigma_i|\epsilon(\alpha)_i - \epsilon(\beta)_i| \quad (7)$$

The normalized eigenvalues cause the value for $D_{\alpha\beta}$ to lie between 0 and 1. A value closer to 1 infers a higher degree of dissimilarity. Molecular similarity is described in an analogous manner

$$S_{\alpha\beta} = (1 - D_{\alpha\beta})\,(1 - \varphi) \quad (8)$$

where $\varphi = |\mathrm{rank}(\alpha) - \mathrm{rank}(\beta)|/(\mathrm{rank}(\alpha) + \mathrm{rank}(\beta))$. The rank of the matrices is essentially the number of atoms of a specific IPE type present. Therefore, the φ term in eq 8 serves to reincorporate molecular size information. Similar to the measure for dissimilarity, the similarity measure is a value between 1 and 0, where a value closer to 1 represents compounds that are more similar, and closer to 0 represents compounds that are more dissimilar.

The descriptor set for α is all of the eigenvalues of all of the eigenvectors derived from all of the MDDM for compound α. Operationally, a threshold cutoff value is applied, and those normalized eigenvalues below the thresh-

old value are disregarded. For the data sets studied and reported in this paper, the threshold was set at 0.002.

When constructing an entire descriptor matrix for a training set of compounds, it is important to consider the regression technique that will be employed to generate the QSAR model. A GFA type analysis, using multiple linear regression (MLR), is a *variable selection* regression technique that ultimately eliminates those variables in the analysis that are not highly correlated with the variance of the dependent variable data. Therefore, in the case of a GFA regression (GFAR) analysis, to be able to treat an entire arbitrary training set as well as extract the most information, the maximum number of significant eigenvalues specific to that training set for a particular compound and a particular IPE type, m, is determined, $\epsilon_{m,\max}$. All the eigenvectors for IPE type, m, for each molecule across the training set are then assigned $\epsilon_{m,\max}$ eigenvalues for IPE type m. Eigenvectors that otherwise contain less than $\epsilon_{m,\max}$ elements have the "missing" eigenvalues set to zero. For example, if $\epsilon_{0,\max}$ is 10 and the eigenvector for IPE 0 of compound α has only nine significant eigenvalues, the tenth eigenvalue for IPE 0 of compound α is set to zero.

The total set of universal descriptors, ϵ_{total}, for a compound in the training set will be the sum of the 36 eigenvalues of $\epsilon_{m,\max}$ length, which can be a number that is quite large. This method of creating the universal descriptor data matrix introduces some degree of estimation. However, GFAR considers each column (descriptor) individually. If the estimation present in a column results in that variable being unfit to describe the variance in the dependent variable, that descriptor will simply not survive the evolution process to optimization.

The PLS regression (PLSR) technique, unlike the GFAR method, is not a variable selection technique. PLSR generates 'latent' variables that are, in short, linear combinations of *all* the independent variables in the original data set. Therefore, introducing estimation [the zeroes] into the universal descriptor matrix will result in a significant reduction in model quality as well as produce erroneous results. Consequently, the universal descriptor matrices used by the PLSR method are assembled slightly different than those of the GFA method. First, the minimum number of significant eigenvalues specific to the training set for a particular compound and a particular IPE type, m, is determined, $\epsilon_{m,\min}$. All the eigenvectors for IPE type, m, for each molecule across the training set are then assigned $\epsilon_{m,\min}$ eigenvalues for IPE type m. Eigenvectors that otherwise contain more than $\epsilon_{m,\min}$ elements are truncated to remove the "extra" eigenvalues. To guard against truncating significant values, the GFAR analysis is performed first. If a universal descriptor survives GFAR optimization, then it consequently is not eligible for truncation in the creation of the PLSR universal descriptor matrix.

As indicated in several reviews on the subject,[8,9] the quality of models produced in PLS type regression is highly dependent on the manner in which the data are preprocessed as well as the presence, or rather the absence, of outlier compounds. Therefore, the procedure employed to optimize models based on PLSR is as follows. First, the entire training set was used to generate the best model possible. The best model was identified by selecting the number of principle components yielding the highest cross-validated correlation coefficient. After the best model was determined for the raw data, the procedure was repeated for the entire training set with the corresponding data matrix mean-centered, i.e., the column means were subtracted from each element in the column resulting in a zero mean for each column. Finally, an optimum model was determined for the auto-scaled data matrix. Auto-scaling involves calibrating the column data to zero mean and unit variance by dividing each column by its standard deviation. Each of the preprocessing methods help to make the calculations and resulting interpretation more concise. For example, mean-centering allows variables to be evaluated as to how they vary from a common reference point. The optimum models resulting from the raw data as well as the two types of preprocessing are compared to determine which method is most appropriate for the particular training set. All data in the current study are either mean-centered or auto-scaled.

The second step in PLSR model optimization is to identify and remove outlier compounds, according to the PLS regression, from the training set. This is accomplished by examining the contribution of each compound to the cumulative cross-validated PRESS (*p*redicted *re*sidual *s*um-of-squares). Compounds that contribute a value of more than twice the standard deviation of the overall PRESS are identified as outliers. In all cases tested, removal of these compounds results in significantly improved model quality. It should be noted that the acceptable values of cross-validated correlation coefficients resulting from a PLSR model are typically lower (xv-$r^2 > 0.5$) than for models resulting from a GFAR model (xv-$r^2 > 0.7$). Also, because PLSR models are, essentially, linear combinations of the entire set of independent variables, or descriptors, they are not as readily interpretable, or as compact, as models from a variable selection and fitting technique. Therefore, the results provided from PLSR analyses consist only of the statistical quality of the models as a function of the number of model terms (principle components) and, when applicable, activity value predictions for a set of test compounds.

Inspection of the GFAR eigenvalue models, GFAR 4D-Fingerprint models, presented in this paper reveals that the eigenvalues that survive model optimization are present in varying positions, near the beginning, middle, or end, within their respective eigenvectors. For example, the best eigenvalue QSAR model, 4D-Fingerprint QSAR model, for the AHPBA study (see *Results*) contains the first eigenvalue from the (aro,all) eigenvector and the 60th eigenvalue from the (all,all) eigenvector. To establish whether these two descriptors are equivalent in significance, three diagnostic measures are determined. The first is how often the descriptor appears in the top models for a training set. A descriptor that repeatedly survives the GFAR model optimization is likely not due to chance occurrence. The second diagnostic is the statistical quality of models resulting from the GFAR analysis if the descriptor of interest is left out of the trial descriptor pool. An inferior model should be generated when a significant independent variable (descriptor) is not available. The third and final diagnostic is consideration of the variance vectors. Variance vectors are calculated by multiplying the regression coefficient by the corresponding descriptor value for each compound over the training set. The range in the descriptor variance vector reflects a descriptor's net contribution to the calculated dependent variable. Descriptors that

are similar in significance have similar ranges in their variance vectors. As an example, the determination of descriptor significance is presented completely in this paper for the AHPBA data set.

Random Scrambling Analysis. Random scrambling analysis is another means by which the significance of a model or descriptor set can be probed**.** When the descriptor pool is much larger than the number of observations (compounds) there is a significant chance of generating a chance correlation. Therefore, it is important to verify that the optimized QSAR models are not chance correlations. The likelihood of a chance correlation model can be explored and estimated using the method of random scrambling.

Random scrambling simply involves the dependent variables of the data set being assigned in a random manner to the independent variables. In the case of QSAR modeling, the random scrambling process reassigns the activity values to different compounds (and their corresponding descriptors) in the training set in a random manner. An optimized QSAR model is then built for this bogus training set and the statistics of fit (r^2, xv- r^2) are determined. This process is repeated a number of times to establish a random-fit profile. The statistical quality of the models resulting from the rearranged (scrambled) data should be significantly less than that of the parent model constructed from the nonscrambled data set.

There is an obvious statistical constraint when applying random scrambling. If the standard deviation of the set of dependent variables, i.e., the activities, of the data set is low, then the random scrambling analysis could be approximately equivalent to introducing a small amount of experimental error to the dependent variables. For example, in the case of the flavonoid analogues, given in this paper, the standard deviation for the dependent variables of the data set is 0.92, and the activity mean is 6.55. Therefore, random scrambling is equivalent to introducing an added experimental error of 14% in measuring the activities, which is often possible for many types of biological endpoint measurements. The results of the random scrambling analysis for data sets in which the standard deviation of the activity measure is small relative to the mean may, or may not, show a pronounced difference in statistical fits of the scrambled and nonscrambled data sets. Consequently, further diagnostic testing using a different method may be necessary.

RESULTS

Validation of the eigenvalue descriptors, or 4D-fingerprints, as universal descriptors requires the appropriate selection of sets of compounds to demonstrate the general applicability of the descriptors as well as the efficacy of the QSAR paradigm. The five data sets examined in this study explore a variety of structural, bioactivity, and statistical criteria deemed essential to validating the descriptors and the QSAR method employed. When possible, the QSAR models generated with the eigenvalue descriptors, the 4D-Fingerprint QSAR models (either GFA or PLS), are presented in parallel to the best models from previous QSAR analyses for a particular data set.

A. Glucose Inhibitors of Glycogen Phosphorylase. The set of glucose inhibitors of glycogen phosphorylase has been extensively studied by various QSAR methods.[12−14] For this study, the 4D-fingerprints are utilized to generate both PLSR and GFAR models that predict the free energy of binding, ΔG, of a set of 47 semiflexible glucose analogue inhibitors of glycogen phosphorylase *b*. The analogues are derived from the parent compound, glucose, depicted in Figure 1(a). The range in binding free energy for the training set is 1.77−6.65 kcal/mol. This ΔG range corresponds to a range in the binding constant, K_i, of 0.016 mM to 53.10 mM. Thus, these ligand inhibitors are not particularly potent and, as a consequence, can be difficult to model in QSAR studies. CEP's generated for each inhibitor are used to build both the 4D-Fingerprint QSAR models as well as a 4D-QSAR model.

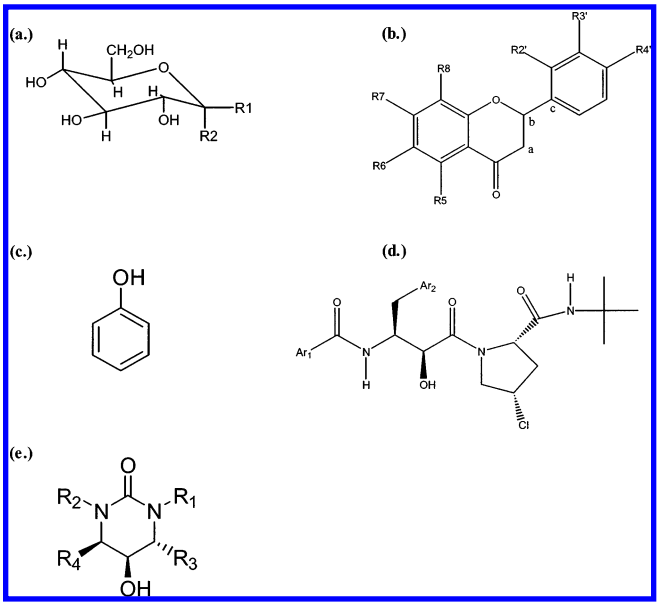

Figure 1. The core structures of the compounds used in the five data sets: (a) the glucose analogs, (b) the flavonoid analogs, (c) the propofol analogs, (d) the AHPBA analogs, and (e) the THP analogs.

4D-Fingerprint QSAR Models Using PLSR. The data matrix generated for PLSR for the set of glycogen phosphorylase inhibitors contains 182 eigenvalues [4D-fingerprints]. For this particular data set, preprocessing by autoscaling produces the highest quality models. Table 2(a) reports the r^2 and xv-r^2 (the statistical quality) of the 'best' models as a function of the number of descriptor terms when all 47 training set compounds are considered. The statistics of the overall 'best' model (4 model terms; $r^2 = 0.89$; xv-$r^2 = 0.3$) suggest that the model is *overfitting* the data. However, there are four compounds in the training data that fit the criteria of an outlier. Removal of these four compounds from the training set results in a significantly better model (4 model terms; $r^2 = 0.94$; xv-$r^2 = 0.59$), given in Table 2(b). There appears to be no obvious common characteristic(s) among the four compounds determined to be outliers. It is possible that there are slight differences in their conformational ensemble profiles (CEPs), generated from molecular dynamics simulation, that lead to a particular compound possessing a set of 4D-fingerprints that deviates from the norm of the remaining training set compounds. The optimum model with an xv-r^2 of 0.59 suggests that the universal 4D-fingerprint descriptors are significantly reflective of the salient 3-D as well as dynamic features of the training set compounds.

Table 2. Statistical Quality of the PLSR Models as a Function of the Number of Model Terms for the Set of Glucose Derived Inhibitors of Glycogen Phosphorylase[a]

no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2
				(a)				
1	0.37	0.08	4	0.89	0.30	7	0.97	0.17
2	0.67	0.18	5	0.92	0.25	8	0.98	0.15
3	0.82	0.28	6	0.95	0.20			
				(b)				
1	0.46	0.14	4	0.94	0.59	7	0.98	0.57
2	0.76	0.35	5	0.96	0.58	8	0.99	0.56
3	0.89	0.51	6	0.97	0.58			

(c) The Linear Correlation Matrix of the Residuals of Fit for the Pairs of Top 10 4D-Fingerprint GFAR QSAR Models of the Set of Glucose Analog Inhibitors of Glycogen Phosphorylase

model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.82	1.00								
3	0.81	0.82	1.00							
4	0.88	0.76	0.81	1.00						
5	0.79	0.94	0.87	0.82	1.00					
6	0.85	0.83	0.96	0.82	0.87	1.00				
7	0.87	0.88	0.89	0.85	0.90	0.93	1.00			
8	0.86	0.87	0.79	0.75	0.74	0.79	0.81	1.00		
9	0.74	0.87	0.80	0.73	0.93	0.82	0.83	0.69	1.00	
10	0.97	0.85	0.87	0.84	0.82	0.90	0.90	0.90	0.77	1.00

[a] Reported are r^2 and xv-r^2 for the models resulting when all 47 compounds are included (a) as well as when the four determined outlier compounds are removed (b).

4D-Fingerprint QSAR Models Using GFAR. The data matrix generated for GFAR model optimization for the set of glycogen phosphorylase inhibitors contains 255 nonscaled 4D-fingerprints. All of the resulting top-ten models are highly cross-correlated ($r > 0.69$), see Table 2c, suggesting that the best *overall* model can be represented by a single model, namely the model having the highest r^2 and xv-r^2 values. Because of the high correlation between the top-ten models, the GFAR model cross-correlation analysis was extended to include the top 30 models. The first 29 of the 30 models are highly correlated to one another, and the one remaining model ($r^2 = 0.68$, xv-$r^2 = 0.62$) is moderately correlated (0.57−0.7) to the other 29. The best GFAR 4D-Fingerprint QSAR model is

$$\Delta G = 25.54(\epsilon_4(\text{hba,hbd})) + 88.48(\epsilon_3(\text{all,p-})) + 34.55(\epsilon_1(\text{np,np})) + 73.07(\epsilon_3(\text{np,p+})) + 33.69(\epsilon_1(\text{all,p-})) - 59.98(\epsilon_4(\text{p-,p-})) - 12.76$$

$$r^2 = 0.75,\ \text{xv-}r^2 = 0.68 \qquad (9)$$

where, for example, ϵ_4(hba,hbd) represents the fourth largest eigenvalue from the MDDM of u = (hba) and v = (hbd). The eigenvectors of (hba,hbd) are fingerprints from which the relative molecular similarity of any pair of inhibitors with respect to both hydrogen bond donor and acceptor groups can be extracted. There is only one single highly correlated pair of fingerprints among the set of six terms in eq 9 with both descriptors derived from atoms belonging to the polar negative IPE type. Removal of either of the correlated descriptors from the parent descriptor pool results in a statistically inferior model compared to the original ($r^2 = 0.75$); $r^2 = 0.68$ when ϵ_1(all,p-) is removed, and $r^2 = 0.67$ when ϵ_4(p-,p-) is removed. This finding supports the need for both descriptors in generating a significant model for the given descriptor set.

Among the top-ten GFAR 4D-Fingerprint QSAR models, there are 15 unique descriptors, seven of which are defined as significant, that is, used more than once in each of the top-ten models. The standard deviation of the activity data for the inhibitor set is ± 1.34, and the activity mean is 3.68. These values suggest that if the descriptors are correlated quantitatively to the activity data, the results of a random scrambling analysis should yield models statistically inferior to that of the parent model. This is precisely the case, the parent 'best' model has a statistical measure of fit (xv-$r^2 = 0.68$) that is nearly double that of the average of the 'best' models constructed from the five sets of randomly scrambled data (xv-$r^2 = 0.36$).

Comparison to the 4D-QSAR Model. For comparison, the set of glucose inhibitors of glycogen phosphorylase was evaluated by the established 4D-QSAR method. The optimum model in a 4D-QSAR analysis results from the evaluation of several alignments. In the case of the glucose analogues, five unique alignments were considered. The "best" alignment is that which produces models with the highest r^2 and xv-r^2 values. The best model found from the 4D-QSAR analysis is

$$\Delta G = 14.7\text{GC1(any)} - 4.1\text{GC2(np)} + 4.0\text{GC3(hba)} + 4.1\text{GC4(any)} + 9.6\text{GC5(np)} + 6.1\text{GC6(p+)} + 2.05$$

$$r^2 = 0.84,\ \text{xv-}r^2 = 0.79 \qquad (10)$$

As mentioned previously, the GFAR- and PLSR 4D-Fingerprint QSAR models are generated independent of alignment. It is, therefore, possible that the 0.11 loss in xv-r^2 for the GFAR 4D-Fingerprint QSAR model, as compared to the 4D-QSAR model, is due to the absence of alignment information which is explicitly part of the 4D-QSAR model. The additional 0.09 loss in xv-r^2 for the PLSR 4D-Fingerprint QSAR model is most likely attributed directly to the data fitting technique.

The optimized GFAR 4D-Fingerprint QSAR and the 4D-QSAR models for the Gp*b* glucose inhibitors appear to share similar features with respect to the IPE types. In both cases, the nonpolar pharmacophore type is present, suggesting that this is a feature important to glucose inhibitor binding to glycogen phosphorylase. It is also evidence that the GFAR 4D-Fingerprint QSAR method is capable of identifying molecular features that likely contribute to the target biological endpoint and that the 4D-fingerprints are reflective of the conformational and structural features inherent to the training set.

B. Flavonoid Analogues as Ligands to $GABA_A$ Receptors. To test the validity of using the 4D-fingerprints as QSAR descriptors for large ligands, which have about the same conformational flexibility as the glucose inhibitors, a series of flavonoid analogues was considered. The flavonoids are of pharmacological interest due to their selective affinity for the benzodiazepine binding site on the $GABA_A$ receptor (BZR). The development of high-affinity ligands with

Table 3. Statistical Quality of the PLSR Models as a Function of the Number of Model Terms for the Set of Flavonoid Analogs[a]

no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2
		(a)			
1	0.67	0.47	6	0.97	0.59
2	0.79	0.55	7	0.98	0.61
3	0.88	0.61	8	0.98	0.63
4	0.92	0.59	9	0.98	0.63
5	0.96	0.59	10	0.99	0.65
		(b)			
1	0.68	0.48	6	0.98	0.80
2	0.81	0.61	7	0.99	0.80
3	0.92	0.73	8	0.99	0.81
4	0.95	0.77	9	0.99	0.81
5	0.97	0.80	10	0.99	0.82

[a] Reported are r^2 and xv-r^2 for the models resulting when all 38 compounds are included (a) as well as when the three determined outlier compounds are removed (b).

anxioselective properties would result in an improved therapeutic option for the treatment of anxiety related disorders. Previous QSAR studies utilized a training set of 38 flavonoids, all derived from the flavone structural template depicted in Figure 1(b), to develop CoMFA, CoMSIA, HQSAR,[15] and 4D-QSAR[16] models. Furthermore, four common test compounds were included in each study. Two of these test compounds have measured activity values outside the range ($5.1 < -\log K_i < 9.0$) of the training set. The training set of flavonoids represents a class of ligands having greater affinity for their target receptor as compared to the glucose inhibitors of glycogen phosphorylase.

Preferred alignment searching in 3D-QSAR analysis becomes increasingly challenging as the molecules in the training set become larger. It is often difficult to identify the actual pharmacophore from the multitude of candidates for a particular ligand analogue set, arising from the multitude of corresponding alignments. To approach the alignment problem in the CoMFA analysis of the flavonoids, a general three-point alignment rule was assumed which corresponds to the virtual focal points of a proposed pharmacophore. One alignment point is located at the center of the phenyl ring of the chromone moiety, a second at the center of the phenyl ring, and the final alignment point is at the carbonyl oxygen of the chromone moiety, thereby defining a triangle that traverses the entire molecule. This triangular pattern serves as the basis for the conformational superimposition. The alignment, as used in CoMFA, has also served as the alignment in the CoMSIA analysis. The statistical results for the 4D-QSAR, CoMFA, CoMSIA, and HQSAR analyses are utilized for comparison to the 4D-Fingerprint QSAR models.

4D-Fingerprint QSAR Models Using PLSR. The data matrix created for the PLSR study of flavonoid binding to the benzodiazepine $GABA_A$ receptor contains 192 'universal' 4D-fingerprints. The optimum preprocessing method was determined to be auto-scaling for the flavonoid data set. Table 3(a) contains the statistical quality of the PLSR models generated as a function of the number of model terms. The 'best' model shown in Table 3(a) (3 model terms; $r^2 = 0.88$; xv-$r^2 = 0.61$) possesses statistical characteristics that are consistent with a high-quality PLSR model. However, three outliers were identified from the 38 compound training set. Removal of the outliers results in a model with superior statistical significance (5 model terms; $r^2 = 0.97$; xv-$r^2 = 0.80$) as shown in Table 3(b). There appears to be no common characteristic among the outlier compounds. The excellent statistical characteristics of the PLSR 4D-Fingerprint QSAR models suggests that the 4D-fingerprints are more than adequate for producing predictive models for the flavonoid compounds, capturing 3-D and conformational information while, in turn, being devoid of any alignment-dependent limitations. As will also be demonstrated, the PLSR 4D-Fingerprint QSAR models for the flavonoid analogues are also successful at predicting activities for a number of test compounds.

4D-Fingerprint QSAR Models Using GFAR. The data matrix created for GFAR model optimization for the set of flavonoid analogues contains 330 nonscaled 4D-fingerprints. Unlike the glucose analogues, among the top-ten GFAR 4D-Fingerprint QSAR models for the flavonoid analogues there are two unique sets of models. This fact is clear from the linear cross-correlation matrix of the residuals of fit given in Table 4. The r-values in bold ($r < 0.5$) indicate the presence of multiple unique models. The existence of manifold models in QSAR techniques has been clearly established, yet for the sake of brevity, only the best model from each distinct set will be considered in this analysis. The optimum GFAR 4D-fingerprint QSAR model is

$$-\log K_i = 2126.58(\epsilon_{17}(\text{all,all})) + 832.86(\epsilon_{10}(\text{all,all})) - 214.08(\epsilon_{6}(\text{all,all})) - 489.1(\epsilon_{8}(\text{np,np})) - 124.15(\epsilon_{17}(\text{hs,np})) - 9.11$$

$$r^2 = 0.93,\ \text{xv-}r^2 = 0.89 \qquad (11)$$

where K_i is the binding constant of the flavonoid to the BZR. The first obvious feature of the 'best' GFAR 4D-Fingerprint QSAR model, eq 11, is the excellent r^2 and xv-r^2 values, suggesting a highly predictive model. The total number of unique descriptors among the top-ten 4D-Fingerprint QSAR models is seventeen, with nine descriptors used more than once in each of the top-ten models, and thus considered significant descriptors. This finding is further support that there is a distinct relationship between the 4D-fingerprints for the flavonoid ligands that survive the GFAR model optimization and the corresponding activity measures.

It is important to note that as a result of the eigenvector normalization process, shorter eigenvectors possess greater variance between each element within the eigenvector. Simply put, a shorter eigenvector will possess a greater disparity regarding the percent of variance explained by each of its elements. Consequently, when a larger data matrix is created to accommodate larger compounds, there is the possibility that eigenvalues [4D-fingerprints] from the middle or end of the size-sorted eigenvector will survive the GFAR model optimization. This behavior is seen by larger regression coefficients present in eq 11 as compared to eq 9. The larger size of the flavonoids, as compared to the glucose analogues, is included in eq 11 by the presence of 4D-fingerprints such as (ϵ_{17}(all,all)).

To rule out the possibility of chance correlations, random scrambling was performed for the GFAR 4D-Fingerprint QSAR model. The standard deviation of the activity data is $\pm$ 0.92, and the mean is 6.55. The xv-r^2 for the parent model

Table 4. Linear Cross-Correlation Matrix of the Residuals of Fit for the Pairs of Top 10 4D-Fingerprint GFAR QSAR Models of the Set of Flavonoid Inhibitors of $GABA_A$

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8	model 9	model 10
model 1	1.00									
model 2	0.90	1.00								
model 3	0.66	0.63	1.00							
model 4	0.82	0.87	0.63	1.00						
model 5	0.82	0.84	0.78	0.78	1.00					
model 6	0.40	0.47	0.63	0.55	0.61	1.00				
model 7	0.81	0.87	0.68	0.78	0.92	0.58	1.00			
model 8	0.56	0.59	0.85	0.71	0.75	0.67	0.60	1.00		
model 9	0.49	0.61	0.59	0.61	0.60	0.80	0.65	0.62	1.00	
model 10	0.78	0.72	0.83	0.70	0.84	0.52	0.69	0.70	0.37	1.00

Table 5. Comparison of the Statistical Measures of Fit of Each Best Model from the QSAR Methodologies Used To Model the Set of Flavonoid Analogs

QSAR methodology	no. of model terms	r^2	xv-r^2
PLS-EV-QSAR[a]	5	0.97	0.80
GFA-EV-QSAR	5	0.93	0.89
CoMFA	5	0.97	0.75
CoMS IA	5	0.95	0.73
HQSAR	NA	0.97	0.75

[a] 35 compounds used in training set.

is 0.87, while the average of the xv-r^2 for the scrambled models is 0.45. This again suggests that the high correlation coefficient observed in the parent model is not a result of a chance relationship between the independent and dependent variables but rather illustrates that a significant quantitative relationship exists between the descriptors and the activity data, thus contributing to the validity of the 4D-fingerprints as universal descriptors.

Comparison of the 4D-Fingerprint QSAR Models to Other Reported QSAR Models. The CoMFA and 4D-QSAR studies suggests that steric intermolecular interactions are prominent contributors to the activity profile of the flavonoid training set. The steric-interaction evidence is present in the optimum 4D-QSAR equation determined for this data set, and reported previously, shown in eq 12.

$$\text{-log } K_i = -12.38\text{GC1(any)} + 13.26\text{GC2(p-)} - 7.56\text{GC3(any)} + 1.76\text{GC4(p-)} - 6.78\text{GC5(p+)} + 6.89$$

$$r^2 = 0.89,\ \text{xv-}r^2 = 0.84 \qquad (12)$$

In the 4D-QSAR model, descriptors of IPE type (any) with a negative regression coefficient indicate steric repulsions. This finding is seemingly mirrored in the GFAR 4D-Fingerprint QSAR model by the presence of three descriptors in eq 11 of IPE type (all-all), one with a negative regression coefficient, and also by the fact that there are two descriptors used in each of the top 10 models that are of type (all-all).

A summary of the statistical properties of the best QSAR models from each of the methodologies used to model the flavonoid training set is given in Table 5. Based on the statistical fits, it is clear that the 4D-Fingerprint QSAR models are as significant in modeling flavonoid binding to BZR as any of the other QSAR methodologies employed. The 4D-Fingerprint QSAR models are also independent of alignment.

Table 6. Predicted -log K_i Values for the Four Test Set Compounds by the 4D-QSAR, PLSR-EV-QSAR, and GFAR-EV-QSAR Models for the Set of Flavonoid Analogs[a]

compound	obs. -log K_i	predicted -log K_i 4D-QSAR	PLS-EV-QSAR[c]	GFA-EV-QSAR
Baicalin[b]	4.11	5.80(+1.69)	7.98(+3.87)	5.5(+1.39)
Baicalein	5.25	5.89(+0.64)	5.69(+0.44)	5.14(−0.11)
Scutellarein[b]	4.92	5.92(+1.00)	5.17(+0.25)	5.77(+0.85)
Wogonin	5.69	5.32(−0.37)	5.64(−0.05)	6.44(0.75)

[a] The residual -log K_i for each prediction is given in parentheses. [b] Compounds with -log K_i values outside the range of the training set [c] The PLSR-EV-QSAR model generated after removal of the three determined outlier compounds.

Finally, to evaluate the predictive power of the models created from the 'universal' 4D-fingerprint descriptors, the activities of four test set compounds were determined using the models as virtual screens. The results are given in Table 6. The models generated from the 'universal' 4D-fingerprint descriptors perform well in making predictions for the test set compounds, with the exception of the predictions made for baicalin. However, this compound possesses the activity measure, -log K_i (4.11), most outside the range of the training set compounds ($5.1 <$ -log $K_i < 9.0$). Therefore, it is likely that these models, including the 4D-QSAR model, are inadequate for estimating the low binding activity of this compound, perhaps due to baicalin having additional adverse binding sites when compared to the training set compounds.

C. Propofol Analogues as General Anesthetic Agents. Often the size of a training set is small, but a corresponding QSAR model is nevertheless desired. Generating reliable quantitative models using a limited size training set SAR may not always be possible, depending both on the QSAR method employed and the training set SAR. The number of descriptor terms in the QSAR models of small data sets must also be limited to avoid *overfitting* of the data. An example of a small training set for which QSAR models have been successfully constructed[17,18] is the set of 14 propofol analogues given in Table 7, and the parent structure is given in Figure 1(c). Propofol and its analogues are of clinical interest as general anesthetic agents.[19] The activity measure for the propofol analogue set, EC_{50}, corresponds to the loss of righting reflex in *Xenopus laevis* tadpoles. The propofol analogues are relatively rigid structures compared to both the flavonoids and the glucose inhibitors. The propofol training set can be effectively modeled by alignment dependent 3D-QSAR techniques due to the relatively small size and rigid nature of this set of compounds. However, for

Table 7. Training Set of Propofol Analogs[a]

compound	-log EC50
1. phenol	2.8
2. 2,6-dimethylphenol	4.2
3. 2-isopropylphenol	3.9
4. 2,6-diethylphenol	4.7
5. 2-*tert*-butyl-6-methylphenol	4.5
6. 2,6-diethylphenyl bromide	3.8
7. 2,6-diethylphenyl isocyanate	3.8
8. 2,6-diisopropylphenol (propofol)	5.7
9. 3,5-diisopropylcatecol	4.3
10. 3,5-di-*tert*-butylphenol	4.0
11. 4-iodo-2,6-diisopropylphenol	5.6
12. 2,6-di-*sec*-butylphenol	6.4
13. 2,6-diethylphenyl isothiocyanate	3.8
14. 2,4-di-*sec*-butylphenol	5.0

[a] -log EC50 is the loss of righting activity evaluating in tadpoles, a measure of anesthetic potency.

Table 8. Statistical Quality of the PLSR Models as a Function of the Number of Model Terms for the Training Set of 14 Propofol Analogs

no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2
1	0.50	0.33	4	0.98	0.68
2	0.83	0.42	5	0.98	0.69
3	0.96	0.59			

validation of the QSAR models, and the application of the models to seek novel compounds (virtual screening), it is important to evaluate the role of alignment on model quality and significance.

The 4D-fingerprints-QSAR models, utilizing the 4D-fingerprint descriptors, incorporate the contributions of conformational flexibility and molecular shape into a QSAR model. The MDDM, and thus the resulting eigenvectors, contain the information about molecular flexibility. The rigid substructure on which the propofol analogues are based, Figure 1, minimizes the role of conformational flexibility in the resulting eigenvectors making these analogues appear similar with respect to shape and conformation. Overall, compounds such as the propofol analogues probe to what extent the models generated from the 4D-fingerprints are able to incorporate conformation-independent properties of a training set, such as atom bonding topology and molecular size, into QSAR model building.

The CEP of each propofol analogue was generated in a fashion similar to those of the previous data sets, and the PLSR and GFAR 4D-fingerprintQSAR methodologies were applied.

4D-Fingerprint QSAR Models Using PLSR Regression. The relatively limited size of the propofol analogue structures as well as the small number of compounds in the training set results in a relatively small data matrix containing only 80 eigenvalue descriptors. Unlike both the glucose and flavonoid data sets, mean-centering was determined to be the optimum variable preprocessing method for the PLSR. The statistical quality of the models constructed as a function of the number of principle components, i.e., model terms, is given in Table 8. Due to the limited number of compounds in the data set, the three-term model was chosen as the 'best' model, although the four- and five-term models appear to have slightly higher xv-r^2 values. This behavior in statistical fit is a common occurrence with small training sets, and choosing the model with the least possible number of model terms is the strategy of choice for model selection to avoid overfitting the data. The statistics of the three-term model are consistent with a quality fit ($r^2 = 0.96$; xv-$r^2 = 0.59$).

There are also no outliers for the three-term PLSR 4D-Fingerprint QSAR model, further suggesting the model is stable and robust.

4D-Fingerprint QSAR Models Using the GFAR Method. The data matrix generated for GFAR model optimization for the set of propofol analogues contains 224 nonscaled 4D-fingerprints. Table 9(a) lists the statistical measures (quality) of the top-ten models. All of the top-ten models possess good model statistics and are either two or three-term models, an appropriate size relative to the number of training set compounds. Table 9(b) is the linear cross-correlation matrix of the residuals of fit among pairs of the top-ten models. It is evident from Table 9(b) that there are several unique models present. To fully capture the information contained in the data set, a manifold model consisting of several of the top-ten models would be necessary. However, solely for the sake of comparison, the model having the highest cross-validated correlation coefficient, given in eq 13, was chosen to represent the 'best' of the GFAR 4D-fingerprint-QSAR models.

$$-\log EC_{50} = -12.51(\epsilon_1(\text{all,hs})) - 246.56(\epsilon_8(\text{all,all})) + 15.95$$

$$r^2 = 0.87,\ \text{xv-}r^2 = 0.79 \qquad (13)$$

Although the model exhibits satisfactory model statistics, its limited size reduces the information that may be extracted from the descriptors. It appears the model is suggesting that steric factors play a role in the anesthetic properties of the propofol analogues. The good model statistics of eq 13 as well for the PLSR 4D-Fingerprint QSAR model suggest that the eigenvectors of the various IPE MDDM's of a training set not only embed conformational information but also are a compilation of atom bonding topology and molecular size.

The results of random scrambling also reflect the quality of the information contained in the MDDM eigenvectors of the training set. The standard deviation of the activity data for the propofol analogue set is $\pm$ 0.95, and the average activity is 4.46. The model constructed from the non-scrambled data has an xv-$r^2 = 0.79$, while the average xv-r^2 for the five models constructed from scrambling is 0.56.

Comparison to the 4D-QSAR Models. An optimized two-term 4D-QSAR has also been developed for the same set of propofol analogues and is given in eq 14.

$$-\log EC_{50} = 16.71\text{GC1(hbd)} + 5.96\text{GC2(all)} + 3.70$$

$$r^2 = 0.87,\ \text{xv-}r^2 = 0.84 \qquad (14)$$

The GFAR 4D-Fingerprint QSAR model as well as the 4D-QSAR model share a similarity in their representation of the 'all' IPE type. Generally, this is an indication of steric factors, or at least a particular region of the receptor, contributing to the activity being modeled. Therefore, an interesting point to consider for this data set is whether the GFAR 4D-Fingerprint QSAR model and the 4D-QSAR model are representing the same or possibly similar information. The linear cross-correlation of the residuals of fit

Table 9. (a) Statistical Quality of the Top-Ten GFAR-EV-QSAR Models for the Training Set of 14 Propofol Analogs and (b) the Linear Cross-Correlation Matrix for the Top-Ten Models

(a)

model no.	no. of descriptor terms	r^2	xv-r^2	model no.	no. of descriptor terms	r^2	xv-r^2
1	2	0.87	0.79	6	3	0.92	0.82
2	3	0.93	0.82	7	3	0.92	0.77
3	2	0.84	0.77	8	2	0.81	0.72
4	3	0.92	0.84	9	3	0.92	0.81
5	2	0.82	0.66	10	2	0.81	0.74

(b)

model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.53	1.00								
3	0.57	0.34	1.00							
4	0.63	0.66	0.38	1.00						
5	0.69	0.73	0.61	0.68	1.00					
6	0.70	0.43	0.16	0.54	0.48	1.00				
7	0.70	0.50	0.72	0.33	0.66	0.30	1.00			
8	0.60	0.20	0.93	0.26	0.58	0.23	0.72	1.00		
9	0.29	0.29	−0.19	0.22	−0.06	0.64	0.01	−0.20	1.00	
10	0.47	0.21	0.97	0.23	0.52	0.06	0.63	0.96	−0.27	1.00

between the 4D-QSAR model with the GFA 4D-Fingerprint QSAR model is 0.16. Thus, the 4D-QSAR and 4D-Fingerprint QSAR models are quite different and fitting this training set data in distinctly different ways.

D. AHPBA Inhibitors of HIV-1 Protease. Another set of analogues employed for the exploration and validation of the MDDM eigenvalues as "universal" QSAR descriptors is the set of 3(S)-amino-2(S)-hydroxyl-4-phenylbutanoic acids (AHPBAs). These compounds are inhibitors of HIV-1 protease (HIVPR). The APHBAs are relatively large, structurally diverse, possess significant conformational freedom, and contain multiple chiral centers. A 3D-QSAR study in which CoMFA, CoMSIA, and HQSAR[20] models were generated from a training set of 27 analogues whose parent structure is shown in Figure 1(d). Additionally, a 4D-QSAR study[21] has also been performed for this data set thereby providing a convenient means for comparison. The activity measurements for this data set, expressed as -log IC_{50}, correspond to the inhibition of HIVPR by the AHPBA derivatives. The training set has a small range in activity ($7.3 <$ -log $IC_{50} < 9.1$). The conformational flexibility of these compounds is well suited to descriptors that can embed molecular motion, i.e., the 4D-fingerprints, while the limited activity range tests the ability of the 4D-Fingerprint QSAR method to develop a model with high resolution (distinguish small activity differences). Seven test compounds, not included in the training set, are also available to test the predictivity of the 4D-Fingerprint QSAR models.

4D-Fingerprint QSAR Models Using PLSR Regression. The AHPBA analogues have the largest common scaffold of the training sets used to evaluate the 4D-fingerprints. Therefore, the data matrix of independent variables for PLSR is significantly larger, containing 445 4D-fingerprints. The optimum preprocessing method was determined to be autoscaling for the set of AHPBA analogues.

Table 10 contains the statistical quality measures of the models produced as a function of the number of model terms. The models generated when including all the training set analogues in the analysis, Table 10(a), are obviously unstable as evidenced by the high value for r^2 and corresponding low, and in some cases negative, value for xv-r^2. Among the 27 compounds in the training set, five were determined to be outliers in the PLSR analysis. Removal of these five compounds results in significantly increased model stability, as shown in Table 10(b). It is possible that the large number of trial descriptors increases the probability of creating outlier compounds in the training set. As with the other data sets containing outliers in the PLSR analysis, there appears to be no common trait among the outliers. The best PLSR 4D-Fingerprint QSAR for the reduced data set for the AHPBA analogues contains 3 terms and possesses acceptable model statistics ($r^2 = 0.94$; xv-$r^2 = 0.54$). The PLSR model is also successful at predicting the -log IC_{50} for the test set compounds, as will be shown in the QSAR model comparisons.

Table 10. Statistical Quality of the PLSR Models as a Function of the Number of Model Terms for the Set of AHPBA Derived Inhibitors of HIV-1 Protease[a]

no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2
		(a)			
1	0.38	0.02	5	0.96	−0.02
2	0.65	−0.06	6	0.98	0.06
3	0.83	−0.06	7	0.99	0.02
4	0.93	0.02	8	0.99	−0.02
		(b)			
1	0.54	0.30	5	0.99	0.41
2	0.88	0.43	6	0.99	0.45
3	0.94	0.54	7	0.99	0.45
4	0.98	0.46	8	0.99	0.46

[a] Reported are the r^2 and xv-r^2 for the models resulting when all 27 compounds are included (a) as well as when the five determined outlier compounds are removed (b).

4D-Fingerprint QSAR Models Using the GFAR. The data matrix created for GFA model optimization for the set of AHPBA analogues contains 660 nonscaled 4D-fingerprints. All of the top-ten models are correlated ($r > 0.5$), indicating that the best *overall* model is the model having the highest

Table 11. Eigenvectors and Variance Vectors for the Five Eigenvalue Descriptors Present in the GFAR-EV-QSAR Model, Eq 15, for the Set of AHPBA Analogs

eigenvectors					AHPBA	variance vectors[a]				
ϵ_{14}(all,all)	ϵ_{41}(all,all)	ϵ_{60}(all,all)	ϵ_{12}(hs,hs)	ϵ_{1}(aro,all)	analog	ϵ_{14}(all,all)	ϵ_{41}(all,all)	ϵ_{60}(all,all)	ϵ_{12} (hs,hs)	ϵ_{1} (aro,all)
0.01230	0.00434	0.00250	0.01670	0.38050	1	2.86628	11.32401	17.03284	7.57686	1.22520
0.01261	0.00419	0.00247	0.01526	0.37476	2	2.93946	10.93524	16.84169	6.92225	1.20674
0.01280	0.00425	0.00250	0.01508	0.36302	3	2.98257	11.09962	17.08063	6.84195	1.16892
0.01273	0.00413	0.00249	0.01524	0.37004	4	2.96649	10.78391	16.99188	6.91363	1.19153
0.01212	0.00425	0.00258	0.01428	0.43606	5	2.82480	11.09962	17.61994	6.47903	1.40410
0.01250	0.00419	0.00253	0.01391	0.41981	6	2.91243	10.92741	17.24447	6.31027	1.35179
0.01193	0.00418	0.00249	0.01382	0.41146	7	2.78075	10.90393	16.99871	6.26717	1.32491
0.01187	0.00418	0.00258	0.01397	0.37530	8	2.76654	10.90915	17.59264	6.33522	1.20846
0.01162	0.00426	0.00258	0.01491	0.37890	9	2.70781	11.12310	17.61994	6.76211	1.22006
0.01230	0.00416	0.00257	0.01353	0.37061	10	2.86535	10.85957	17.55850	6.13879	1.19336
0.01183	0.00431	0.00248	0.01533	0.41022	11	2.75721	11.25618	16.95775	6.95445	1.32090
0.01265	0.00423	0.00249	0.01490	0.39532	12	2.94762	11.03961	17.01919	6.75757	1.27293
0.01338	0.00415	0.00257	0.01635	0.36492	13	3.11821	10.81783	17.53120	7.41672	1.17503
0.01252	0.00420	0.00258	0.01574	0.37210	14	2.91685	10.95872	17.57898	7.13954	1.19815
0.01260	0.00418	0.00257	0.01476	0.37187	15	2.93573	10.90393	17.52437	6.69769	1.19743
0.01180	0.00416	0.00256	0.01542	0.42450	16	2.75046	10.86218	17.45610	6.99347	1.36689
0.01171	0.00408	0.00256	0.01499	0.42062	17	2.72925	10.65345	17.44928	6.79976	1.35440
0.01222	0.00421	0.00248	0.01441	0.41064	18	2.84740	10.97177	16.95092	6.53483	1.32226
0.01214	0.00414	0.00251	0.01473	0.31381	19	2.82923	10.81000	17.10111	6.68045	1.01046
0.01285	0.00402	0.00250	0.01384	0.32650	20	2.99516	10.48646	17.03284	6.27716	1.05133
0.01183	0.00395	0.00251	0.01369	0.33033	21	2.75582	10.31164	17.10794	6.20820	1.06365
0.01358	0.00411	0.00251	0.01397	0.34155	22	3.16365	10.72128	17.12159	6.33840	1.09980
0.01318	0.00430	0.00249	0.01556	0.36027	23	3.07230	11.21443	17.00553	7.05925	1.16006
0.01255	0.00415	0.00261	0.01501	0.40044	24	2.92361	10.81783	17.81110	6.80929	1.28943
0.01293	0.00418	0.00261	0.01456	0.39416	25	3.01287	10.90393	17.79744	6.60650	1.26920
0.01223	0.00416	0.00261	0.01482	0.39811	26	2.84974	10.86479	17.81110	6.72128	1.28191
0.01252	0.00425	0.00248	0.01539	0.41221	27	2.91755	11.09701	16.94409	6.98349	1.32733
AHPBA EV-QSAR models					min	2.70781	10.31164	16.84169	6.13879	1.01046
					max	3.16365	11.32401	17.81110	7.57686	1.40410
					range	0.45585	1.01238	0.96940	1.43807	0.39364

[a] Absolute values for the regression coefficients were utilized in calculating the variance vectors.

values of r^2 and xv-r^2. The best GFAR 4D-Fingerprint QSAR model is given in eq 15.

$$\text{-log IC}_{50} = 233.05(\epsilon_{14}(\text{all,all})) + 2609.22(\epsilon_{41}(\text{all,all})) - 6826.79(\epsilon_{60}(\text{all,all})) - 453.65(\epsilon_{12}(\text{hs,hs})) - 3.22(\epsilon_{1}(\text{aro,all})) + 19.82$$

$$r^2 = 0.88,\ \text{xv-}r^2 = 0.80 \qquad (15)$$

A feature of immediate interest from eq 15 is that due to the extended length of the eigenvectors (owing to the large size of the molecules), there are eigenvalue descriptors near the eigenvector "tails", and, consequently, these descriptors possess much larger regression coefficients. Second, the optimal 4D-Fingerprint QSAR model contains the descriptor ϵ_1(aro,all), the largest eigenvalue for the aromatic-all atom eigenvector. This descriptor may reflect the importance of the lipophilic aromatic ring system of the AHPBA's, a factor that has been previously determined to be an integral factor in modulating HIVPR inhibition potency. This aromatic ring system fits in the S_1 hydrophobic pocket of HIVPR. The 4D-Fingerprint QSAR model describes inhibition activity based, in part, on an aromatic descriptor of the inhibitors. In turn, eq 15 suggests that the 4D-MS based method can identify important types of ligand−receptor interactions and also illustrates the ability of the methodology to relate the local environments of the receptor to the intrinsic properties of the ligands in the training set.

The standard deviation of the observed -log IC_{50} values for the AHPBA data set is ± 0.51, and the mean is 8.39. In this case, the standard deviation is merely 6% of the mean, not a good situation for performing a QSAR or random scrambling analysis. Although the average of the xv-r^2 for the models from random scrambling is 0.71, which is less than the xv-r^2 (0.88) of the optimized nonscrambled model, the numerical difference between the two is less than seen in the previous three application studies. For this reason, the variance vectors were explored to help establish descriptor significance. All five descriptors show a relatively similar range in their variance vectors, shown in Table 11, suggesting that each descriptor is of nearly equal importance in determining -log IC_{50}. The top-ten models contain thirteen unique descriptors, seven of which are used in more than one of the top-ten models. There are three descriptors used in each of the top-ten models (ϵ_{12}(hs,hs), ϵ_{60}(all,all), ϵ_{41}(all,all)). These findings help verify that the descriptor correlations to the dependent variable are genuine and not attributed to a chance occurrence.

Comparison to Previous QSAR Studies. A 4D-QSAR analysis was performed for the AHPBA data set.[21] The 4D-QSAR analysis suggests that steric as well as nonpolar interactions are critical to determining inhibition potency of the AHPBA derivatives. This finding is also supported by

Table 12. Comparison of the Statistical Measures of Fit for Best Models from Each of the QSAR Methodologies Used To Model the Set of AHPBA Analogs

QSAR method	no. of model terms	r^2	xv-r^2
PLS-EV-QSAR[a]	3	0.94	0.54
GFA-EV-QSAR	5	0.88	0.8
4D-QSAR	5	0.93	0.9
CoMFA	5	0.98	0.61
CoMS IA	6	0.97	0.53
HQSAR		0.95	0.72

[a] Model generated with the reduced data set of 22 compounds.

the presence of three descriptors of type "all-all" as well as a descriptor of type "hs-hs" in the GFAR 4D-Fingerprint-QSAR model, eq 15. For comparison sake, one of the five 4D-QSAR manifold models is

$$\text{-log IC}_{50} = 3.99\text{GC1(any)} + 3.41\text{GC2(np)} - 9.29\text{GC3 (any)} + 9.38\text{GC4(any)} - 10.84\text{GC5(any)} + 8.05 \quad (13)$$

$$r^2 = 0.93,\ \text{xv-}r^2 = 0.90 \quad (16)$$

For the CoMFA, CoMSIA, and HQSAR analyses the crystal structure of HIVPR has been indirectly used. Both the conformation and alignment utilized in each analysis were based on a molecular docking experiment performed at the active site of the HIVPR. The *active* conformation assigned to each inhibitor was ultimately decided upon by considering the binding free energy associated with the AHPBA-HIVPR complex. Therefore, these particular CoMFA related studies inherently include receptor information as well. Interestingly, the results of the CoMFA study also suggest that aromatic ligand−receptor interactions contribute to the inhibition potency of the AHPBA analogues. A comparison of the statistics of fit for the best models from the three 3D-QSAR methods, the 4D-QSAR model, as well as the PLSR- and GFAR 4D-Fingerprint QSAR models is summarized in Table 12. The 4D-Fingerprint QSAR models, *generated independent of any receptor structure or alignment information*, exhibit excellent statistics of fit. This is strong evidence that the spatial structure−activity information from the training set are captured by the models generated from the 'universal' 4D-fingerprint descriptors.

As a final test of model fitness and descriptor validity for the AHPBA analogues, the PLSR and GFAR 4D-Fingerprint QSAR models were used to predict the -log IC_{50} for the seven test set inhibitors. The predicted -log IC_{50} values for these test compounds using the two 4D-Fingerprint QSAR models as well as the 4D-QSAR model are listed in Table 13. Overall, the models perform well in accurately predicting the potencies of the test set compounds and are taken as further proof that the 4D-fingerprints are not only universal but also are a genuine representation of the 3-dimensional and conformational properties of the compounds.

E. THP Inhibitors of HIV-1 Protease. The final application preformed in this study for descriptor validation also involves ligands that target the aspartic protease of the HIV-1 virus (HIVPR). The training set analogues are a set of cyclic urea derived compounds, the tetrahydropyrimidine-2-ones (THPs). The core structure is given in Figure 1(e). The THP analogues are potent inhibitors of HIVPR, and several comprehensive QSAR studies[22] have been performed to aid in fully exploiting this class of inhibitors. This training set is composed of 49 relatively large and flexible analogues. The activity measurements are expressed as -log K_i values, reflecting the binding of the inhibitors to the HIVPR enzyme. The range in inhibition potency is quite large, spanning 5 orders of magnitude ($6.01 <$ -log $K_i < 11.00$). Eleven test set compounds are also included in the analysis to evaluate the predictive power of the resulting models. The large potency range, and relatively large sizes of the training and test sets, make this a good data set to evaluate any QSAR method's ability to capture subtle as well as substantial features that lead to the vast differentiation in inhibition potency.

Table 13. Predicted -log IC50 Values for the Seven Test Set Compounds by the 4D-QSAR, PLSR-EV-QSAR, and GFAR-EV-QSAR Models for the Set of AHPBA Analogs[a]

compd	obs. -log IC50	predicted -log IC50		
		4D-QSAR	PLS-EV-QSAR[b]	GFA-EV-QSAR
1	7.41	7.91(+0.50)	8.61(+1.21)	8.38(+0.97)
2	8.05	7.70(−0.35)	8.50(+0.45)	8.01(−0.04)
3	8.82	7.90(−0.92)	8.73(−0.09)	8.29(−0.53)
4	8.47	8.51(0.04)	8.69(+0.22)	8.71(+0.24)
5	8.89	8.23(−0.66)	8.98(+0.09)	8.12(−0.77)
6	8.27	8.57(0.30)	9.22(+0.95)	8.85(+0.58)
7	7.85	8.19(0.34)	8.00(+0.15)	7.42(−0.43)
	⟨Δ(-log IC50)⟩[c]	0.44	0.45	0.51

[a] The residual − log IC50 for each prediction is given in parentheses as well as the absolute value for the residuals from each method. [b] The PLSR-EV-QSAR model generated after removal of the five determined outlier compounds. [c] Absolute values for the residuals used to calculate residual averages.

Table 14. Statistical Quality of the PLSR Regression Models as a Function of the Number of Model Terms for the Set of THP Derived Inhibitors of HIVPR[a]

no. of model terms	r^2	xv-r^2	no. of model terms	r^2	xv-r^2
		(a)			
1	0.55	0.51	5	0.90	0.70
2	0.77	0.70	6	0.91	0.66
3	0.81	0.69	7	0.93	0.62
4	0.87	0.70	8	0.95	0.59
		(b)			
1	0.55	0.51	5	0.90	0.65
2	0.77	0.70	6	0.93	0.54
3	0.84	0.70	7	0.94	0.54
4	0.86	0.71	8	0.95	0.51

[a] Reported are r^2 and xv-r^2 for the models resulting when all 49 compounds are included (a) as well as when the one determined outlier compounds is removed (b).

4D-Fingerprint QSAR Models Using PLSR. The THP molecules are large, leading to a data matrix containing 300 descriptors. The optimum preprocessing method for the PLSR analysis of the THP analogues was determined to be mean-centering. Table 14 contains the statistical measures of fit as a function of the number of model terms. The best model generated for all analogues in the analysis, the two-term model in Table 14(a), shows good model statistics ($r^2 = 0.77$; xv-$r^2 = 0.70$). However, a single outlier analogue was identified. Removal of the single outlier leads to marginally better models, shown in Table 14(b). The four-

term model from the reduced data set ($r^2 = 0.86$; xv-$r^2 = 0.71$) was chosen as the best model, as it possesses the highest value for xv-r^2. Based on statistical quality, the PLSR 4D-Fingerprint QSAR model generated from the 4D-fingerprints performs quite as well in spanning the range of potencies present in the training set. Moreover, as will be demonstrated, the model is also adequate in predicting the potencies of the test set compounds.

4D-Fingerprint QSAR Models Using the GFAR Method. The data matrix generated for GFAR model optimization for the set of THP analogues is the largest of the five training sets reported in this paper, containing 738 nonscaled 4D-fingerprints. The top-ten models are all very highly correlated ($r > 0.89$), suggesting that the top-ten models are nearly identical. Therefore, the 'best' model is chosen solely based on statistical measures of fit and given in eq 17.

$$-\log K_i = 1108.24(\epsilon_{18}(\text{all,all})) + 339.16(\epsilon_{48}(\text{all,all})) - 15.67(\epsilon_{5}(\text{hs,p-})) - 13.84(\epsilon_{1}(\text{p-,hba})) - 445.29(\epsilon_{19}(\text{hs,hs})) + 8.81$$

$$r^2 = 0.92, \text{ xv-}r^2 = 0.88 \quad (17)$$

The excellent r^2 and xv-r^2 of the 'best' GFAR 4D-Fingerprint QSAR model, eq 17, are indicative of a highly predictive model. There are a total of 14 unique descriptors among the top-ten models, with three of them being used in each of the top-ten models. The descriptors that survive model optimization, namely (ϵ_{18}(all,all)), (ϵ_{48}(all,all)), and (ϵ_{19}(hs,hs)), indicate the importance of the molecular shape and size in modulating the potency of these compounds. The size feature that has also been identified in previous QSAR studies. As discussed for the AHPBA analogues, larger data matrices indirectly lead to larger regression coefficients, that is again illustrated in the GFAR 4D-Fingerprint QSAR model for the THP analogues.

The results of the random scrambling analysis further establish the authenticity of the 4D-fingerprints, while also ruling out the possibility that the corresponding models arise from chance correlation. The standard deviation of the -log K_i values is $\pm$ 1.26, and the mean is 8.72. The r^2 and xv-r^2 for the parent model, 0.92 and 0.88, respectively, is much higher than that of the corresponding average values of the models generated from random scrambling analysis, namely $r^2 = 0.56$ and xv-$r^2 = 0.4$.

Comparison to Previous QSAR Studies. The 4D-QSAR study performed for this analogue set is also an evaluation of a novel clustering method to improve model selection and predictivity. It is interesting to note that the optimal clustering method determined from the study involved grouping compounds based *on their similarity eigenvectors*. These are the same eigenvectors being employed as 'universal' descriptors in the current study. The results of the 4D-QSAR study as well as the evidence presented in this paper is strong support for the authenticity of the 4D-fingerprint descriptors. The 4D-QSAR model with the optimal test set predictivity was chosen for comparison and is given by eq 18.

Table 15. Comparison of the Statistical Measures of Fit for the Best Models from Each of the QSAR Methodologies Used To Model the Set of THP Analogs

QSAR method	no. of model terms	r^2	xv-r^2
PLS-EV-QSA R	4	0.86	0.71
GFA-EV-QSA R	5	0.92	0.88
4D-QSAR	7	0.91	0.86
CoMFA	2	0.97	0.80

$$-\log K_i = 6.17 + 12.91\text{GC1 (any)} + 13.45\text{GC2 (np)} - 22.42\text{GC3 (np)} - 28.75\text{GC4 (any)} - 8.19\text{GC5 (a)} + 45.78\text{GC6 (any)} + 18.70\text{GC7 (np)}$$

$$r^2 = 0.91, q^2 = 0.86 \quad (18)$$

It is again apparent that the descriptor IPE types in the GFAR 4D-Fingerprint QSAR model agree are consistent with those found in the optimized 4D-QSAR model. Although the underlying paradigms are different, the conclusions reached by each methodology are much the same. The 4D-QSAR model also exhibits excellent statistical measures of fit, another feature shared with the GFAR 4D-Fingerprint QSAR model.

The CoMFA study performed for this set of inhibitors focuses on enhancing model quality by including non-CoMFA descriptors to account for such binding features as solvation energy and buried surface area (BSA) of the ligands. The study reports seventeen different CoMFA models. The best model, chosen as the one possessing the highest statistical measures of fit, was selected for comparison. A summary of the statistical measures of fit for the best models from each method is given in Table 15. Overall, the models generated using the 4D-fingerprints have statistics of fit as significant as those of the *enhanced* CoMFA model as well as the 4D-QSAR model. This finding is of particular significance since the 4D-fingerprint descriptors contain no alignment or receptor information.

The test set compounds for the THP training set were selected to span the large range of potencies present in the training set. The results of the inhibition potency predictions by the QSAR methods used to model the set of THP analogues is given in Table 16. It is clear that while none of the methods used yield exceptional predictions, which is not too surprising given the wide range in potency, the models using the universal 4D-fingerprints perform as well as the methods requiring alignments and/or additional receptor/ligand data. Table 16 illustrates the exceptional ability of a model derived from the universal descriptors to be employed as an effective virtual screening tool.

Model generation utilizing the universal descriptors is relatively straightforward and efficient since no alignment analysis or descriptor calculation is required. The result is models that predict for unknown compounds at least as well as the more time-consuming 3D-QSAR methods.

Table 16. Predicted -log K_i Values for the Eleven Test Set Compounds by the 4D-QSAR, CoMFA, PLSR-EV-QSAR, and GFAR-EV-QSAR Models for the Set of THP Analogs[a]

predicted -log K_i compd	obs. -log K_i	4D-QSAR	CoMFA	PLS-EV-QSAR[b]	GFA-EV-QSAR
1	10.70	8.22(−2.48)	10.65(−0.05)	9.96(−0.74)	10.33(−0.37)
2	6.34	7.77(+1.43)	7.39(+1.05)	7.77(+1.43)	7.98(+1.64)
3	7.00	6.95(−0.05)	7.69(+0.69)	7.40(+0.40)	5.97(−1.03)
4	7.85	9.83(+1.98)	7.61(−0.24)	7.18(−0.67)	7.45(−0.40)
5	9.60	9.88(+0.28)	8.39(−1.21)	8.69(−0.91)	7.93(−1.67)
6	10.10	10.22(+0.12)	12.19(+2.09)	11.00(+0.90)	11.70(+1.60)
7	10.52	11.71(+1.19)	12.28(+1.76)	10.55(+0.03)	10.15(−0.37)
8	7.36	7.06(−0.30)	8.90(+1.54)	9.36(+2.00)	9.41(+2.05)
9	10.10	9.79(−0.31)	9.29(−0.81)	9.91(−0.19)	9.39(−0.71)
10	10.22	9.29(−0.93)	8.94(−1.28)	9.58(−0.64)	9.53(−0.69)
11	10.70	9.32(−1.38)	9.15(−1.55)	10.27(−0.43)	9.77(−0.93)
	$\langle\Delta(\text{-log } K_i)\rangle$*	0.95	1.12	0.76	1.04

[a] The residual -log K_i for each prediction is given in parentheses. [b] Absolute values for the residuals used to calculate residual averages.

DISCUSSION

The study reported here provides strong support that the eigenvalues, or 4D-fingerprints, derived from the main distance-dependent matrices (MDDM) of the absolute similarity analysis of the 4D-MS method can be used to develop descriptive QSAR models for a training set. The purpose of developing such models, as far as this work is concerned, is to establish the validity of a novel set of descriptors. It is certainly intriguing that a set of 'universal' descriptors, derived independently of any alignment considerations, can be used to produce descriptive QSAR models that are of similar quality to the much more time-consuming and computationally intense QSAR methods that are the mainstay tools of drug design today. It is even more peculiar that when considering the amount of effort being afforded to QSAR in recent years, very little has been allocated to *simplifying* analyses. From a drug cost-to-market perspective, the upside of developing a highly efficient screening tool, such as one involving 'universal' descriptors, is undeniable. The models developed in this study represent but a small fraction of the possibilities that stem from such a useful tool. A set of descriptors that include 3D and conformational information, and yet are independent of any external constraints, may be used for, but not limited to, ADME property predictions, similarity clustering, and developing toxicity profile libraries, to name only a few important applications.

A limitation, or tradeoff, of the 4D-fingerprints is that they are quite abstract when compared to other descriptors that are used to develop QSAR models. Although the 4D-fingerprints are derived from the 3-dimensional structures of molecules, visualizing them in Euclidean space is not possible. Still, PLS principal components are also difficult to interpret for the same reasons as the 4D-fingerprints, but PLS[8] QSAR and QSPR models are readily constructed, accepted, and used in current applications. Moreover, topological descriptors, like the Kier-Hall connectivity indices,[23] do not directly admit to structural and/or mechanistic interpretation but are readily used throughout the QSAR and QSPR community.

Actually, the 4D-fingerprints can be interpreted in a manner, and to an extent, that permits mechanistic interpretation and also provides design constraints in designing new compounds. For example, eq 17, restated below, for the THP analogues can be interpreted as follows:

$$\text{-log } K_i = 1108.24(\epsilon_{18}(\text{all,all})) + 339.16(\epsilon_{48}(\text{all,all})) - 15.67(\epsilon_{5}(\text{hs,p-})) - 13.84(\epsilon_{1}(\text{p-,hba})) - 445.29(\epsilon_{19}(\text{hs,hs})) + 8.81$$

$$r^2 = 0.92,\ \text{xv-}r^2 = 0.88 \qquad (19)$$

The three 4D-fingerprints ϵ_{18}(all,all), ϵ_{48}(all,all), and ϵ_{19}(hs,hs) all work together to specify the steric size and shape requirements of the ligands necessary for binding potency. Of all of the individual (all, all) and (hs,hs) 4D-fingerprints, the three found in eq 17 best account for the steric size and shape requirements of the ligands across the training set. The 4D-fingerprint ϵ_5(hs,p-) indicates that the number and/or spatial distribution of polar negative groups across the ligand structure, in general, diminish binding potency. Finally, ϵ_1(p-,hba) suggests that the relative numbers and spatial distributions of polar negative groups and hydrogen-bonding acceptors across a ligand can diminish binding potency. More generally, this 4D-fingerprint suggests that sites of negative partial charge across the ligand are not conducive to high binding affinity.

Overall, the presence of a same IPE pair eigenvalue term in a QSAR model is indicative of that IPE type of functionality being specifically important in expressing activity. Moreover, the (all,all) same IPE pair strongly suggests that the size and shape of the molecule is crucial to activity. The different, or cross, IPE pairs generally reflect that the *joint distribution* of the two IPE types over the size and shape of the molecule is significantly related to the particular biological endpoint. The significance and meaning of the actual particular eigenvalue [1,2,3, ...] that is found in a QSAR model from the set of eignevalues for a given pair of IPE types are very difficult to assign. One possible interpretation is that the selected eigenvalue reflects the distribution of that pair of IPE types over the training set. The smaller the eigenvalue number, the better distributed over the training set are the corresponding IPE[s] features, at least with respect to the observed biological endpoint measures.

It may be true that in order to simplify descriptors to the point of universality, there are dimensions that must be sacrificed in the process. However, when used in a proper

sense, the 'universal' descriptors, regardless of their abstract nature, are an extremely powerful tool in the drug design process. A possible application would be to develop screening models using the 4D-fingerprints to filter out nonapplicable compounds in a study and use the results to develop highly descriptive visual models using a technique such as 4D-QSAR.

The 4D-fingerprint descriptors presented here represent a step forward from "universal descriptors" that, for the most part, capture only two-dimensional qualities of a molecule. Incorporation of the thermodynamically accessible conformer states into a "universal descriptor" captures the salient 2- and 3-dimensional characteristics of a compound and should provide near maximum information about the compound.

ACKNOWLEDGMENT

Partial funding for this study was provided by National Institutes of Health Grant P01-GM 62195. Resources of the Laboratory of Molecular Modeling and Design at UIC and The ChemBats21 Group, Inc. were used in performing this work. An unrestricted financial gift from the Procter and Gamble Company is also gratefully acknowledged.

REFERENCES AND NOTES

(1))Xin, C.; Reynolds, C. H. Performance of similarity measures in 2D-fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.

(2))Holliday, J. D.; Salim, N.; Martin, W.; Willett, P. Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.

(3))Cramer III, R. D. BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.* **1980**, *102*, 1837−1849.

(4))Cramer III, R. D. BC(DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties. *J. Am. Chem. Soc.* **1980**, *102*, 1849−1859.

(5))Duca, J. S.; Hopfinger, A. J. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(6))Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, G. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(7))Rogers, D.; Hopfinger, A. J. Application of the genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(8))Glen, W. G.; Dunn, W. J., III.; Scott, D. R. Principle component analysis and partial least squares regression. *Tetrahedron Comput. Method.* **1989**, *2* (6), 349−376.

(9))Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(10))Martin, Y. C. 3D-QSAR: current state, scope, and limitations. *Perspect. Drug Discovery Des.* **1998**, *12/13/14*, 3−23.

(11))Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(12))Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, *43*.

(13))Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding free energy by 4D-QSAR analysis: application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150.

(14))Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J.; Wang, S. Construction of a virtual high-throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.

(15))Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.; Wong, B. L.; Pei, G.; Jiang, H.; Chen, K. 3D-QSAR model of flavonoids binding at benzodiazepine site in $GABA_A$ receptors. *J. Med. Chem.* **2001**, *44*, 1883−1891.

(16))Hong, X.; Hopfinger, A. J. 3D-pharmacophores of flavonoid binding at the benzodiazepine $GABA_A$ receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.

(17))Krasowski, M. D.; Hong, X.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the $GABA_A$ receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.

(18))Krasowski, M. D.; Jenkins, A.; Flood, P.; Kung, A. Y.; Hopfinger, A. J.; Harrison, N. L. The general anesthetic properties of a series of propofol analogues correlate with potency for potentiation of GABA current at the $GABA_A$ receptor but not with lipid solubility. *J. Pharmacol. Exp. Ther.* **2001**, *297*, 338−351.

(19))James, R.; Glen, J. B. Synthesis, biological evaluation, and preliminary structure−activity considerations of a series of alkyl phenols as intravenous anesthetic agents. *J. Med. Chem.* **1980**, *23*, 1350−1357.

(20))Huang, X.; Xu, L.; Luo, X.; Fan, K.; Ji, R.; Pei, G.; Chen, K.; Jiang, H. Elucidating the inhibiting mode of AHPBA derivatives against HIV protease and building predictive 3D-QSAR models. *J. Med. Chem.* **2002**, *45*, 333−343.

(21))Senese, C. L.; Hopfinger, A. J. Receptor independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297−1307.

(22)) Nair, A. C.; Jayatilleke, P.; Wang, X.; Miertus, S.; Welsh, W. J. Computational studies of tetrahydropyrimidine-2-one HIV-1 protease inhibitors: improving three-dimensional quantitative structure−activity relationship comparative molecular field analysis models by inclusion of calculated inhibitor and receptor-based properties. *J. Med. Chem.* **2002**, *45*, 973−983.

(23)) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; J. Wiley & Sons: New York, 1986.

CI049898S