Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation

José S. Duca† and A. J. Hopfinger*

Laboratory of Molecular Modeling and Design, Department of Medicinal Chemistry and Pharmacognosy M/C 781, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231

Received February 2, 2001

The 4D-QSAR paradigm has been used to develop a formalism to estimate molecular similarity measures as a function of conformation, alignment, and atom type. It is possible to estimate the molecular similarity of pairs of molecules based upon the entire ensemble of conformational states each molecule can adopt at a given temperature, normally room temperature. Molecular similarity can be measured in terms of the types of atoms composing each molecule leading to multiple measures of molecular similarity. Multiple measures of molecular similarity can also arise from using different alignment rules to perform relative molecular similarity, RMS, analysis. An alignment independent method of determining molecular similarity measures, referred to as absolute molecular similarity, AMS, analysis has been developed. Various sets and libraries of compounds, including the amino acids, are analyzed using 4D-QSAR molecular similarity analysis to demonstrate the features of the formalism. Exploration of molecular similarity as a function of chirality, identification of common embedded 3D pharmacophores in compounds, and elucidation of 3D-isosteric compounds from structurally diverse libraries are carried out in the application studies.

INTRODUCTION

A variety of approaches has been developed to estimate *molecular similarity* and correspondingly to characterize *chemical diversity*.[1−15] These approaches cover a relatively wide-range of "philosophies" regarding the *essence* of a molecule, but they share the following three common assumptions:

1. If a large number of properties and features can be determined for a molecule, the essence of the molecule will be captured to provide a basis for measuring molecular similarity and characterizing chemical diversity.

2. Regardless of the number of properties and features determined for a molecule, data reduction can be used to construct the minimum set of similarity/diversity descriptors that approximate the essence of a molecule.

3. The locations of the members of a set of molecules in data reduction descriptor space, DRDS, define both the relative similarities as well as the composite diversity of the set of compounds. Molecules close to one another in DRDS are quite similar, while compounds distant from one another are not similar or dissimilar. The distribution, and its range, of the set of molecules in DRDS defines the composite molecular diversity of the set.

Thus, the entire similarity/diversity characterization process begins with the selection of the properties and features to be used in the study. Correspondingly, the quality and significance of the similarity/diversity analysis will depend on the choice of the property and feature set. In most previous applications studies the focus has been to make the **number** of properties and features as large as possible, but with the restriction that these properties and features be easy and quick to compute. That is, the properties and features have been limited to atom composition, topological, and/or group additive molecular attributes [two-dimensional (2D) properties and features].

Molecular similarity algorithms have been proposed which use conformations of molecules [three-dimensional, 3D, information] which best fit a multiple point pharmacophore, or the molecular similarity measure is the average of a set of such measures determined for multiple pairs of conformations of two molecules.[12−14] In one of the other methods where 3D properties and features have been included in a similarity/diversity analysis, the "default" conformation used for each molecule, or parts of a molecule (topomers), in a set is a low-energy conformer state readily determined from the input structure file and software employed.[15] While this is not an unreasonable conformation to select, it may not reflect the **distribution** of conformer states available to a molecule that contribute to the overall similarity among molecules. *One goal of the methodology presented in this paper is to include the thermodynamic distribution of conformer states available to a molecule in constructing its set of similarity/diversity descriptors.* The methodology presented here **explicitly** incorporates the features and properties of **all** conformations of a molecule that are sampled as part of the construction of multiple molecular similarity measures. This approach is distinctly different from selecting a single conformation of a molecule to use in a molecular similarity estimation from some conformer set, or averaging over a set of molecular similarity measures realized from a set of pairs of conformations of two molecules being compared.

Molecular similarity can be measured on a *relative*, or an *absolute*, basis. The concept of relative and absolute mo-

* Corresponding author phone: (312)996-4816; fax: (312)413-3479; e-mail: hopfingr@uic.edu.

† Current address: Schering-Plough Research Institute, 2015 Galloping Hill Road, K-15-L-0300, Kenliworth, NJ 07033-1300.

10.1021/ci0100090 CCC: $20.00
Published on Web 07/27/2001

lecular similarities does not seem to have been explored, perhaps because this "dimension" of molecular similarity does not become prominent until 3D properties and features are considered. Relative similarity is dependent upon an alignment constraint (an external reference frame) while absolute similarity is alignment independent. By way of example, biologically active compounds (ligands) that bind to a common receptor are often characterized by a common 3D-pharmacophore[16] which is a set of points in space that define the ligand−receptor binding geometry. If the 3D-pharmacophore is used as an alignment constraint to establish molecular similarity, the resultant similarity measures will be different from those realized by not employing any type of constraint (absolute molecular similarity). *Another goal of the methodology presented in this paper is to be able to compute both absolute and relative molecular similarity measures from a common algorithm.*

Current approaches to estimating molecular similarity and molecular diversity yield measures that embody the **entire** molecule. It is usually not possible to break out similarity estimates of "parts" of molecules. Still, there are applications, particularly in pharmaceutical studies, where it might, for example, be useful to know how similar two molecules are with respect to their hydrogen bonding donor features when they are quite similar in overall steric shape. That is, the availability and use of a set of molecular similarity measures, each related to the *functional "pieces"* of a molecule, might be advantageous when compared to a single composite measure of molecular similarity. *The methodology reported here includes the capability of making molecular similarity measurements with respect to the whole molecule as well as functional pieces of the molecule.*

METHODS

A. Conformational Ensemble Sampling. The operational steps involved in the algorithm to compute both absolute and relative molecular similarity measures are schematically illustrated in Figure 1. Both absolute and relative molecular similarity calculations begin with, and are dependent upon, the estimation of the conformational energy profile, CEP, of the molecule. This is illustrated by graph 1 in Figure 1 and is indicated by the Boltzmann distribution plot of the number of conformers at energy ΔE, $N(\Delta E)$ of the molecule shown. The CEP can be estimated by a variety of methods including molecular dynamics simulations, MDS, Monte Carlo simulations, MCS, systematic conformational searching, and combinations of these individual methods. We are currently only employing MDS to estimate the CEP of a molecule. The MOLSIM MDS package[17] is used to estimate the CEP which consists of the MDS trajectory generated for each molecule. The major time-tradeoff inherent to this approach to estimating molecular similarity, and molecular diversity, is the extent of conformational sampling done per molecule versus the number of molecules considered in an analysis—the size of the test library. Typical CEP processing times on an average O^2 10K Silicon Graphics Workstation are 30−60 CPU seconds per compound for typical organic compounds. The CEP is generated from a MDS run of 50 ps generated at intervals of 0.001 ps time steps. Snapshots of the MDS are normally recorded every 0.2 ps and stored in a "trajectory" file as a collection of MDS "frames" constituting the CEP. These MDS conditions are the "standard conditions" used in studies reported throughout this paper, unless otherwise indicated.

B. Construction of the Main Distance-Dependent Matrices (MDDM). The second operation in the estimation of the molecular similarity measures is the construction of a main distance-dependent matrix, MDDM, for each pair of *interaction pharmacophore elements*, IPEs[18] of the molecule. The IPEs are the "functional pieces" of a molecule and constitute a set of specific and independent groups to use in constructing measures of molecular similarity. The IPEs currently being used are given in Table 1 and can be seen to correspond to the major atom types composing any molecule. A new IPE type, IPE-7, has also been defined in order to incorporate information about the molecular "skeleton" and molecular shape. IPE-7 considers all atoms but hydrogens, and is coded HS, which stands for hydrogen-suppressed.

A unique MDDM is constructed for *each unique (u,υ) IPE pair* for each molecule.

The "(all-atom)−(all-atom)" IPE pair [$u = \upsilon =$ (all-atom)] corresponds to the *whole molecule* measure of molecular similarity. Other measures of molecular similarity that will be considered are: self-measures [$u = \upsilon =$ any IPE but (all-atom)] and cross-term measures [$u \neq \upsilon$]. The transformation of conformational geometry to a MDDM of a typical IPE is illustrated by graphs 3 and 4 in Figure 1.

C. Absolute Molecular Similarity MDDM Elements. The algorithm for the estimation of molecular similarity diverges for absolute and relative molecular similarity in the estimation of the MDDM elements, $f(\nu,d_{ij})$. The absolute molecular similarity MDDM matrix elements are defined as

$$f(\nu,d_{ij}) = \exp(-\nu \langle d_{ij} \rangle) \tag{1}$$

In eq 1 ν is a "universal constant" which maximizes the ranges in the molecular similarity measures, and its determination is discussed below. The average distance, $\langle d_{ij} \rangle$, between atom pair i and j of IPE types u and υ, respectively, is computed as

$$\langle d_{ij} \rangle = \Sigma_k d_{ij}(k) p(k) \tag{2}$$

The thermodynamic probability of conformer state k of the CEP is p(k), and $d_{ij}(k)$ is the distance between atom pair i and j of IPE types u and υ, respectively, for the kth conformer state. The p(k) are computed from the conformational energies determined in the MDS and recorded in the corresponding trajectory file. The temperature of the system is normally assumed to be 300 K in both the MDS and in estimating the p(k). The geometries of the CEP of a randomly selected molecule are illustrated in graph 2a in Figure 1. Only interatomic distances associated with atom pairs (i,j) of the appropriate IPE pair types *(u,υ)* are included in constructing the MDDM of that IPE pair type. The $d_{ij}(k)$ of an IPE pair *(u,υ)* for the molecule are illustrated in graph 3a of Figure 1.

D. Relative Molecular Similarity MDDM Elements. Determination of the MDDM elements, $f(\nu,d_{ij})$, for relative molecular similarity measures requires performing a *partial* 4D-QSAR analysis.[19−21] The grid cell space must be defined, the alignment rule selected, and the grid cell occupancy descriptors, GCODs, determined for each molecule from its CEP. A GCOD is the probability that a given IPE type will occupy a grid cell for a given molecule. The set of GCODs

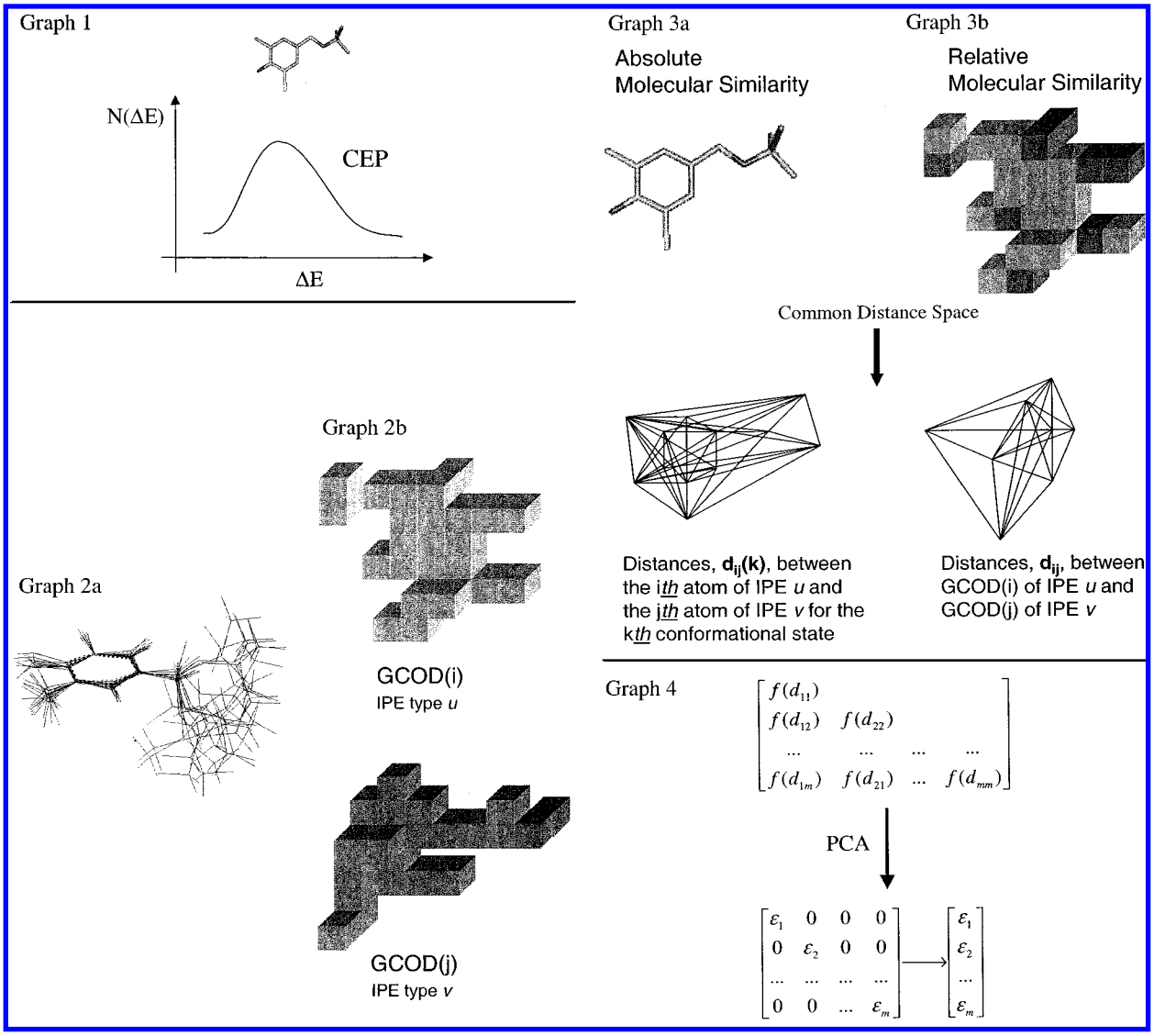


Figure 1. The operational steps for computing absolute and relative molecular similarity, AMS and RMS, respectively. Graph 1: schematic representation of the conformational energy profile, CEP, of a molecule represented as a plot of the number of conformations, N(Δ*E*), having conformational energy, Δ*E*. Graph 2a: molecular stick model illustration of the distribution of conformer states used in absolute molecular similarity analysis. Graph 2b: schematic illustration of the grid cell occupancy distribution of IPE types *u* and *v* for the CEP used in relative molecular similarity analysis. Graph 3a: schematic representation of generating the $d_{ij}(k)$ for the kth conformer state of the CEP for absolute molecular similarity analysis. Graph 3b: schematic illustration of generating the d_{ij} between GCODs of *u* and *v* for relative molecular similarity analysis. Graph 4: schematic illustration of the data reduction of a MDDM by PCA to its eigenvalue, $\{\epsilon_i\}$ representation. The subscript "m" of the matrix elements indicates the number of atoms of the molecule being a particular IPE type.

Table 1. Interaction Pharmacophore Elements, IPEs, Used in Both 4D-QSAR Absolute and Relative Molecular Similarity Analyses

IPE code	symbol	definitions
0	ALL	all atoms in the molecule
1	NP	nonpolar atoms
2	P+	polar atoms with positive charge
3	P−	polar atoms with negative charge
4	HBA	hydrogen bond acceptor atoms
5	HBD	hydrogen bond donor atoms
6	ARO	aromatic atoms
7	HS	non-hydrogen atoms (hydrogen-suppressed)

for a particular IPE pair, *u* and *v*, in light and dark gray colors, respectively, is schematically illustrated in graph 2b of Figure 1. The distances d_{ij} between the centers of GCOD(i) of IPE type *u* and GCOD(j) of IPE type *v* are determined. These distances are represented by the set of line-segments in graph 3b of Figure 1 which is also used to represent the $d_{ij}(k)$ in the absolute molecular diversity formulation. The $f(\nu,d_{ij})$ are then defined as

$$f(\nu,d_{ij}) = \mathrm{GCOD(i)GCOD(j)}\exp(-\nu\, d_{ij}) \qquad (3)$$

The product of the grid cell occupancy descriptors corresponds to the joint probability of an IPE of type *u* being at grid cell i and an IPE of type *v* being at grid cell j. Thus, the product of the GCODs in eq 3 is the equivalent of summing over the p(k) in eq 2. The estimation of ν in eqs 1 and 3 is discussed below.

E. Data Reduction of the MDDM. 1. Whole and Self-Molecular Similarity Representations. The MDDM for IPE pairs that are the same, $u = v$, are square upper/lower triangular matrices which can be readily diagonalized. Thus, principal component analysis, PCA, is the data reduction process. The general Jacobi transformation algorithm[22] is

being used to perform the PCA on any molecule α. The set of eigenvalues obtained from the PCA of the MDDM is then normalized (weighed between zero and one) and numerically sorted in descending size. Normalization is performed using the rank of the MDDM matrix, rank(α), as a weighting factor. That is, nonscaled eigenvalues, $\{\epsilon_i'(\alpha)\}_{u,u}$, are transformed by the relationship $\{\epsilon_i(\alpha)\}_{u,u} = \{\epsilon_i'(\alpha)\}_{u,u}/\text{rank}(\alpha)$.

Overall, the set of normalized eigenvalues, $\{\epsilon_i(\alpha)\}_{u,v}$, are defined as the **essential** molecular similarity measures for molecule α with respect to the IPE of types u and v. There is one unique "eigenvector", $\boldsymbol{\epsilon(u,u,\alpha)} = \{\epsilon_i(\alpha)\}_{u,u}$, for each pair of identical IPE (u,v) for α. The diagonalized representation of the MDDM is schematically illustrated in graph 4 of Figure 1.

2. Cross-Term Molecular Similarity Representations. The MDDM for IPE pairs that are not the same, $u \neq v$, are not necessarily square matrices because the number of u and number of v IPE elements can be different ($n_u \neq n_v$). The following square matrices are defined for these particular cases of rectangular MDDM matrices.

In the case $[n_u \times n_v]$, MDDM(u,u) is defined as

$$\text{MDDM}(u,u) = \text{MDDM}(n_u,n_v) \times \text{MDDM}(n_u,n_v)^T \quad (4a)$$

and for $[n_v \times n_u]$, MDDM(v,v) is defined by

$$\text{MDDM}(v,v) = \text{MDDM}(n_v,n_u) \times \text{MDDM}(n_v,n_u)^T \quad (4b)$$

Because both matrices MDDM(u,u) and MDDM(v,v) defined in eqs 4a and 4b have the same rank and trace, both have the same set of eigenvalues, $\{\epsilon_i(\alpha)\}_{\text{MDDM}(u,u)} = \{\epsilon_i(\alpha)\}_{\text{MDDM}(v,v)}$, from the PCA.

In composite, the unique eigenvector, $\boldsymbol{\epsilon(u,v,\alpha)} = \{\epsilon(\alpha)\}_{u,v} = [\{\epsilon_i(\alpha)\}_{\text{MDDM}(u,u)}]^{1/2}$, is defined for each pair of IPEs (u,v) for compound α.

F. Determination of ν. The value of the parameter ν can be used to influence the values of the $\{\epsilon(\alpha)\}_{u,v}$. In particular, ν can be chosen to maximize the sum of the eigenvalue differences between arbitrary compounds α and β which have the same number, N, of atoms of a particular IPE type. Obviously, the choice of ν which yields this maximum difference will be different for each set of {α, β, N, IPE}. The value of ν being sought is a global average value that will yield an "average maximum difference" in eigenvalues over all compounds. To find this "universal" value of ν, the following function is defined for the eigenvalues of any IPE pair, (u,v)

$$D_{\alpha\beta}(\nu) = \Sigma_i|\epsilon(\alpha,\nu)_i - \epsilon(\beta,\nu)_i| \quad (5)$$

The goal is to maximize the value of $D_{\alpha\beta}(\nu)$ for any, and all, choices of α and β. The general solution to eq 5 for arbitrary α and β is seemingly impossible. However, the inference of a general solution to eq 5 from the solutions for simple choices in α and β is possible. If two general diatomic molecular systems, $\alpha = 1$ and $\beta = 2$, are selected, see Figure 2a, then for the IPE pair $u = v =$ "any atom" the function $D_{\alpha\beta}(\nu)$ is

$$D_{12}(\nu) = 2[\exp(-\nu d_1) - \exp(-\nu d_2)] \quad (6)$$

In eq 6 d_1 and d_2 are the distances between the two atoms of molecules 1 and 2, respectively. Correspondingly, the value of ν which optimizes $D_{12}(\nu)$ is

$$\nu = [\ln(d_2/d_1)]/(d_2 - d_1) \quad (7)$$

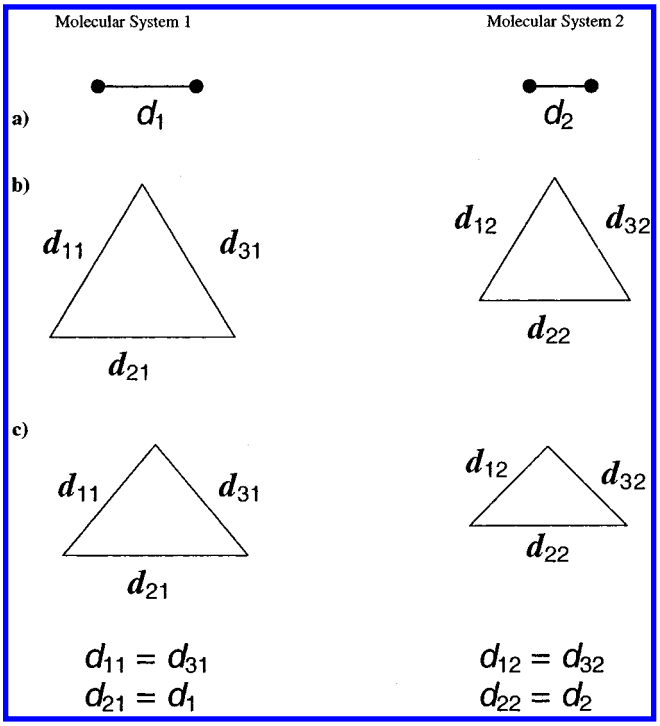


Figure 2. Schematic representation of symmetry conditions applied in the estimation of the "universal constant" ν for (a) two general diatomic molecular systems, (b) two triatomic molecular systems where complete symmetry has been applied, and (c) two triatomic molecular systems where partial symmetry has been applied.

If two triatomic molecules [A triatomic molecule can also refer to three atoms belonging to a specific IPE, for example, polar-negative atoms in the case of 1,3,5-trifluorobenzene. Hence, the term "molecular system", as opposed to simply "molecule", is used in this section.] α and β are next considered, symmetry conditions must be imposed to permit an analytical solution. The most simple symmetry condition that can be selected for $\alpha = 1$ and $\beta = 2$ to yield an analytic solution is, using Figure 2b for reference, $[d_{11} = d_{21} = d_{31} = d_1]$ and $[d_{12} = d_{22} = d_{32} = d_2]$. In d_{ij} i refers to molecule α or β and j refers to the atom number. The value of ν which optimizes $D_{\alpha\beta}(\nu)$ for this case of the triatomic molecular system is also given by eq 7. If only two of the three bonds in both α and β are the same (as pictured in Figure 2c), then the only values that appear in eq 7 are the corresponding nonsymmetric distances, that is, $d_{21} = d_1$ and $d_{22} = d_2$.

Overall, the assumption is made that the solution to $D_{\alpha\beta}(\nu)$ for two arbitrary molecules, α and β, will be a function similar to eq 6, and maximization of this sum of differences for ν will be of the form given by eq 7. In essence, two "average" distances representative of the sizes of α and β, respectively, replace the specific two distances in eqs 6 and 7. Thus, maximizing the sum of the differences between the corresponding eigenvalues of α and β in eq 5 ultimately becomes a function of the "average" distances, d_1 and d_2, of α and β, respectively. In Figure 3 the plot of ν as a function of d_1 and d_2 is shown. The functional dependence of ν on the two distances is modest when the distances are both greater than three ångstroms. In so far as the "average"

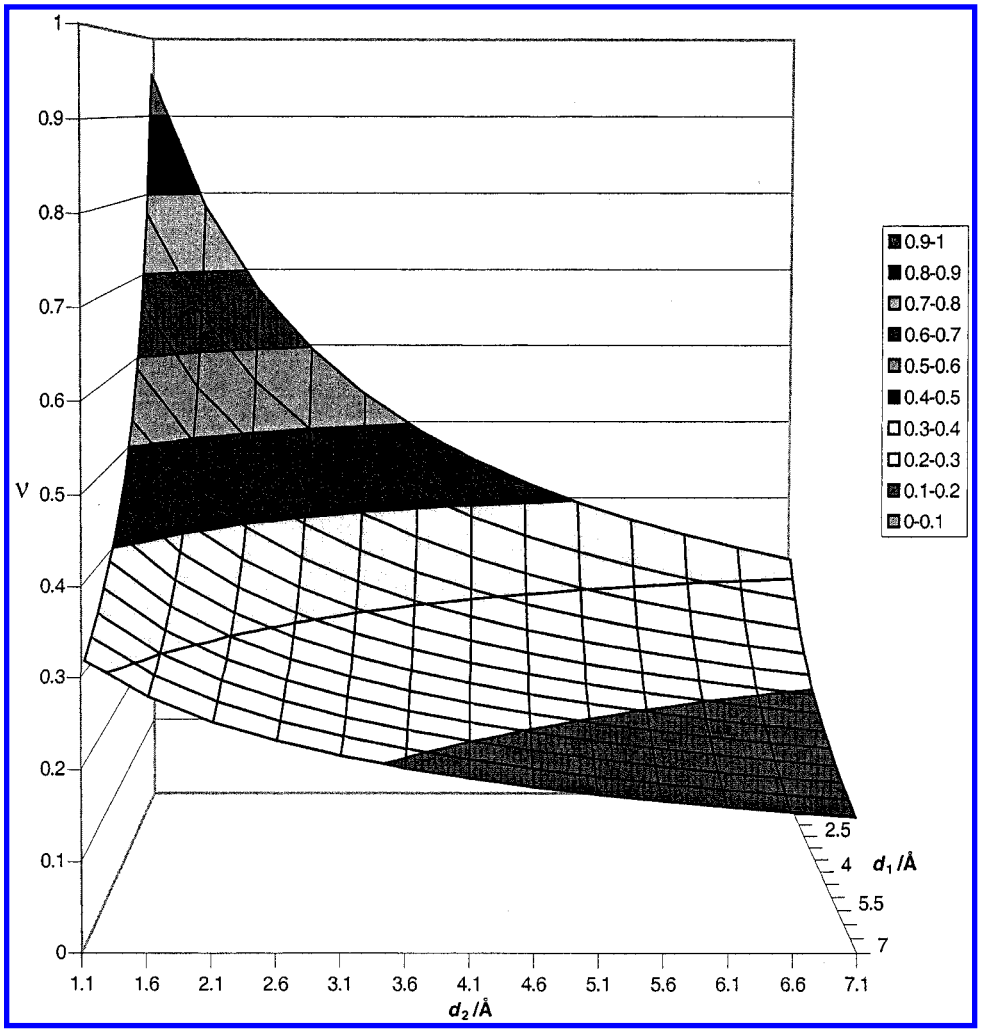


Figure 3. The 3D-plot of the functional relationship between ν and both d_1 and d_2.

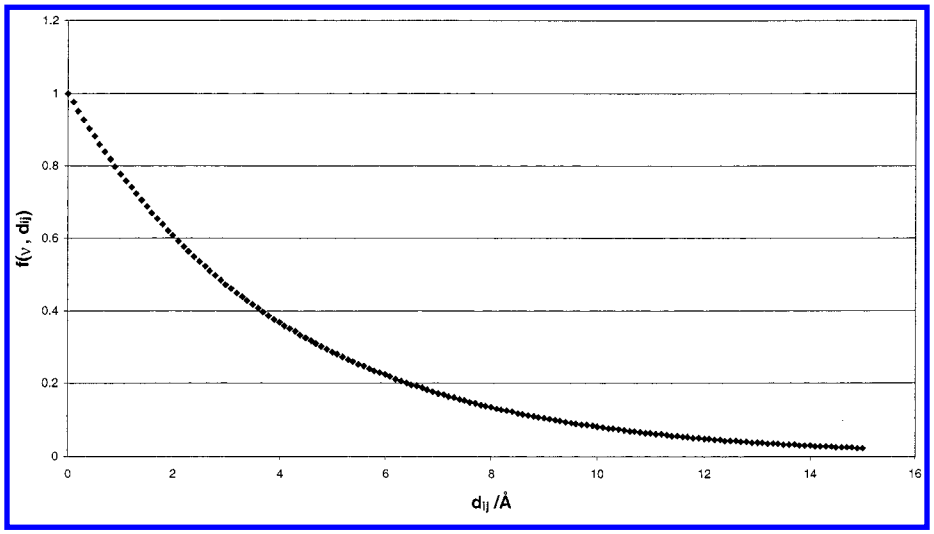


Figure 4. The dependence of f (ν, d_{ij}) [eq 2] on d_{ij} for the chosen value of $\nu = 0.25$.

distance reflects the size of the molecule, most "average" distance values for "typical" organic compounds can be expected to be greater than this value. The range of ν over the "average" distances between 2.5 and 7 Å in Figure 3, the $d_1 = d_2$ pathway, was monitored over the $\nu(d_1,d_2)$ function to determine a representative value of ν. A value of 0.25 is a good choice to maximize eigenvalue differences for organic compounds and has been selected as the "universal" constant value of ν. Figure 4 illustrates the dependence of f(ν, d_{ij}), as defined by eq 3, on the distance element d_{ij} when ν is set as equal to 0.25.

G. Definition of Molecular Dissimilarity. A variety of functions can be defined to "measure" molecular similarity. None will be absolutely "correct". In fact, several representations are discussed in the literature,[1−10] but the overwhelming majority are related to binary bitstrings used as basis sets,

Table 2. Absolute Molecular Similarity, AMS, for the Three Analog Compounds Benzene, Naphthalene, and Anthracene[a]

IPE(0,0)			
anthracene	1.00		
benzene	0.45	1.00	
naphthalene	0.75	0.64	1.00
IPE(7,7)			
anthracene	1.00		
benzene	0.34	1.00	
naphthalene	0.66	0.55	1.00

[a] AMS $S_{\alpha\beta}$ measures are given for *ALL*, IPE(0,0), and *HS*, IPE(7,7).

or descriptors, during the similarity analysis. In the present study the "eigenvectors", $\epsilon(u,v,\alpha)$, have numbers spanning large ranges of values as elements for constructing similarity measures.

The approach taken here to define molecular similarity is to begin by defining *molecular dissimilarity* in terms of eq 5. Once v has been assigned (0.25), then eq 5 measures the composite difference between corresponding eigenvalues of molecules α and β. This composite difference in the eigenvalues is defined as the *molecular dissimilarity* between α and β

$$D_{\alpha\beta} = \Sigma_i|\epsilon(\alpha)_i - \epsilon(\beta)_i| \tag{8}$$

Since the eigenvalues are normalized, $D_{\alpha\beta}$ will vary between 0 and 1. In the same way that molecular dissimilarity is defined by eq 8, the *molecular similarity* for a pair of compounds α and β is defined as

$$S_{\alpha\beta} = (1 - D_{\alpha\beta})\{1 - \Phi\} \tag{9}$$

where $\Phi = |\text{rank}(\alpha) - \text{rank}(\beta)|/[\text{rank}(\alpha) + \text{rank}(\beta)]$ represents the possible differences in the actual eigenvalues' sizes. In other words, it provides an accounting for different molecular sizes. The terms, "rank(α)" and "rank(β)", correspond to the dimensionality of the MDDM matrices for compounds α and β. The presence of factor $\{1 - \Phi\}$ represents the reincorporation, in a realistic manner, of molecular size information deleted during the normalization process of the eigenvalues. In eq 9 the dependence of $S_{\alpha\beta}$ on Φ is assumed to be linear based upon chemical "intuition". A simple experiment was designed in order to validate the performance of eq 9 for predicting the chemical similarity of molecular systems having different sizes. Absolute molecular similarity was computed for benzene, naphthalene, and anthracene considering only the aromatic carbons, that is using the hydrogen suppressed IPE. Taking anthracene as reference, the chemical similarity values obtained using eq 9 for benzene and naphthalene are 0.34 and 0.66, respectively. These results, listed in Table 2, show a behavior that can be called "correct" in terms of chemical intuition. Moreover, this example demonstrates the sensitivity desired from a methodology designed to discern geometrical properties among molecular systems having a common embedded moiety (benzene in this case).

The eigenvalues of the similarity eigenvectors of all molecular systems are sorted in descending size so that they can be compared in terms of *molecular dissimilarity* as defined by eq 8. However, the manner in which the eigenvalues are sorted and compared cannot be viewed as trivial, random, or even "intuitive". The eigenvalues can be viewed as the axes of an N-dimensional space defined by their eigenvector. As such, sorting of the eigenvalues to compute $S_{\alpha\beta}$ can be interpreted as employing a general transformation, τ, that is needed to project the different eigenvalues of any pair of molecules, α and β, onto a common collection of orthogonal axes so that

$$\tau\ \epsilon(\alpha)_i \Rightarrow \epsilon(\alpha)_i' \text{ and}$$

$$\tau\ \epsilon(\beta)_i \Rightarrow \epsilon(\beta)_i' \text{ such that } |\epsilon(\alpha)_i' - \epsilon(\beta)_i'| \text{ is minimum} \tag{10}$$

The new common collection of orthogonal axes to which $\epsilon(\alpha)_i$ and $\epsilon(\beta)_i$ are mapped must possess the property of reorienting each sorted pair of the N-dimensional components for compounds α and β as being parallel to one another. Therefore, the minimum difference $|\epsilon(\alpha)_i' - \epsilon(\beta)_i'|$ required in eq 10 guarantees that the eigenvalue comparison defined by eq 8 leads to a maximum measure of the magnitude of the molecular similarity, $S_{\alpha\beta}$, for a given pair of eigenvectors.

The eigenvalues of an eigenvector can also be employed as "component dissimilarity areas". If each eigenvalue is assigned a common arbitrary differential unity "width", dx, then the eigenvector $\epsilon(u,v,\alpha)$ can be interpreted as a scalar representing an effective area that incorporates the **essential** properties of compound α for IPE(u,v). This representation is depicted in Figure 5. Under this representation, eq 8 symbolizes the differences in effective areas for compounds α and β once integration over the entire width dimension (the number of eigenvalues) in similarity space is carried out. Equation 8 is a particular case, called the Hamming distance, of a general metric known as Minkowski distances[7a] which are defined by

$$D_{\alpha\beta} = [\Sigma_i(|\epsilon(\alpha)_i - \epsilon(\beta)_i|)^t]^{1/t} \tag{11}$$

where $t = 1$ for the Hamming distance and $t = 2$ for the Euclidean distance definition. This latter representation is one of the most popular metrics used in comparison of binary bitstrings.[13a]

RESULTS

Selection of sets of compounds to use in the demonstration, elucidation, and validation of a molecular similarity paradigm is difficult, somewhat arbitrary, and always incomplete. The examples given below have been chosen to meaningfully demonstrate the role of chirality, conformation, chemical structure, relative molecular size, and alignment on measures of molecular similarity using the 4D-QSAR molecular similarity paradigm. Nevertheless, other compounds sets may possibly suggest themselves to the reader as being more appropriate.

A. The Amino Acids. The L and D forms of each of the natural amino acids (in the neutral state) and two different monoprotonated tautomers of histidine were analyzed using both the relative and absolute molecular similarity methodologies. The notation utilized for representation of all 46 compounds is presented in Table 3.

The goals of this study were to (a) establish the absolute molecular similarity measures for the set of amino acids, (b) demonstrate that L and D isomers of a given amino acid cannot be distinguished from one another using absolute

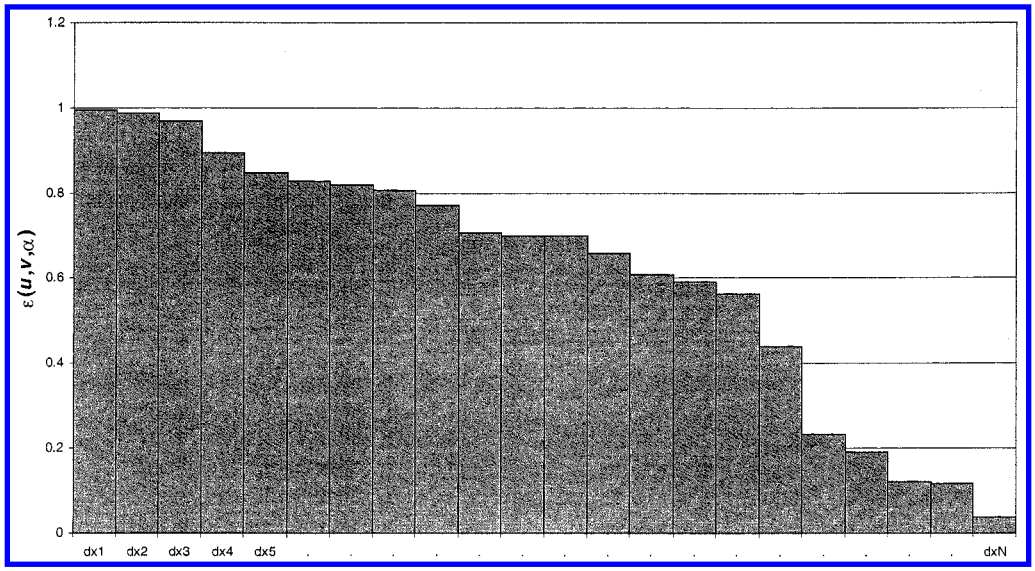
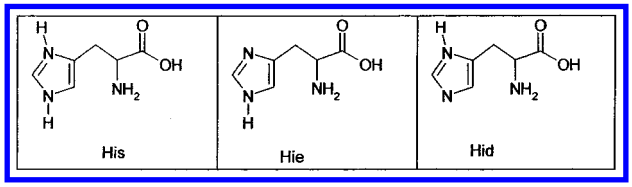


Figure 5. A scalar representation for an arbitrary set of eigenvalues $\{\epsilon_i\}$, with their index values, i, assigned arbitrary differential widths all of which have been set equal to one, $dx = 1$.

Table 3. Nomenclature for the Set of Amino Acids, AAs[a]



code	amino acid	code	amino acid
D/L-Ala	alanine	D/L-Thr	threonine
D/L-Val	valine	D/L-Arg	arginine
D/L-Asp	aspartic acid	D/L-Phe	phenylalanine
D/L-Gly	glycine	D/L-Cys	cysteine
D/L-Asn	asparragine	D/L-Leu	leucine
D/L-His	histidine	D/L-Hid	histidine
D/L-Lys	lysine	D/L-Tyr	tyrosine
D/L-Ser	serine	D/L-Ile	isoleucine
D/L-Gln	glutamine	D/L-Hie	histidine
D/L-Trp	triptophane	D/L-Glu	glutamic acid
D/L-Pro	proline	D/L-Met	methionine

[a] All amino acids are modeled in the neutral state.

similarity measures but can be distinguished using relative similarity measures, and (c) demonstrate that relative molecular similarity measures are alignment dependent.

1. Absolute Molecular Similarity Measures. Absolute molecular similarity, AMS, analysis was performed for the set of L and D amino acids, and the results were interpreted using the AMS matrices for all possible IPE representations defined in *Methods: Section B*. Each AMS matrix is presented as a lower triangular matrix because of symmetry. Each element of each of the AMS matrices corresponds to the absolute molecular similarity measured for a particular pair of compounds α and β, as defined by eq 9. Table 4 lists the AMS matrix for IPE(0,0) of all the amino acids. The corresponding absolute similarity measures for pairs of D and L isomers of each amino acid are boldfaced so that they can be readily identified.

Table 5 contains the AMS self-IPE measures for only the set of D and L amino acid isomer pairs. It can be seen from Table 5 that AMS $S_{\alpha\beta}$ cannot distinguish between D and L isomers for a given amino acid. $S_{\alpha\beta}$ values are higher than 0.94 and with a few exceptions in the range between 0.91 and 0.94 in the cases of D/L Ser [for IPE(2,2) and IPE(5,5)] and D/L Gln [for IPE(2,2), IPE(3,3), and IPE(5,5)]. These slight deviations from a $S_{\alpha\beta}$ value of 1.0 (complete inability to distinguish D/L isomers) are due to "noise" introduced by the MDS during the generation of the CEP. It is readily demonstrated that the AMS values will be closer to 1.0 if the number of snapshots (trajectory conformations) recorded during the MDS increases as well.

In a more general sense, AMS cannot distinguish between isomers of a compound having a *single* chiral center. AMS will identify $S_{\alpha\beta}$ differences between isomers of a compound having two, or more, chiral centers which arise because of differences in conformational distributions. However, differences in the *absolute* molecular shapes of the isomers arising from chirality will not be detected by AMS unless there are differences in the sets of atom pair distances of the two chiral dependent molecular shapes.

The only two-chiral center AA in the library of AAs studied (see Table 3) is Ile which contains a beta carbon chiral center in its side chain. A comparison of the *ALL* and *self* AMS $S_{\alpha\beta}$ values of Ile to the corresponding $S_{\alpha\beta}$ of other AAs, given in Table 5, indicates no significant differences among these measures.

Hence, there does not appear to be major differences in the overall conformational profiles of L and D Ile, otherwise there would be major differences in these AMS molecular similarity measures.

However, the (0,0) − *ALL*, (1,1) − *NP* and (7,7) − *HS* AMS $S_{\alpha\beta}$ measures are less than the other IPE $S_{\alpha\beta}$ measures for Ile. The Ile side chain is nonpolar so one would expect *ALL, NP, and HS* molecular similarity measures to be most sensitive to chiral conformational differences between the D and L isomers. While this seems to be the case, the magnitude of the differences in $S_{\alpha\beta}$ is small.

2. Relative Molecular Similarity Measures. Relative molecular similarity, RMS, measures can be helpful to distinguish different spatial atom locations with respect to a common reference frame (a receptor site, for example), previously defined as part of an alignment rule. Two distinct alignments were used for the set of D/L amino acids and are defined in Table 6.

Table 4. IPE(0,0) AMS $S_{\alpha\beta}$ Partial Matrix for the Set of D/L Amino Acids[a]

D-Ala	**1**	1.00																					
L-Ala	**2**	**0.99**	1.00																				
D-Val	**3**	0.65	0.65	1.00																			
L-Val	**4**	0.66	0.67	**0.96**	1.00																		
D-Asp	**5**	0.83	0.83	0.77	0.79	1.00																	
L-Asp	**6**	0.82	0.82	0.77	0.79	**0.99**	1.00																
D-Gly	**7**	0.72	0.73	0.48	0.49	0.63	0.64	1.00															
L-Gly	**8**	0.72	0.73	0.48	0.49	0.63	0.64	**1.00**	1.00														
D-Asn	**9**	0.70	0.71	0.87	0.86	0.85	0.85	0.56	0.56	1.00													
L-Asn	**10**	0.70	0.70	0.86	0.87	0.85	0.85	0.56	0.56	**0.99**	1.00												
D-His	**11**	0.54	0.54	0.81	0.81	0.66	0.67	0.41	0.41	0.77	0.78	1.00											
L-His	**12**	0.54	0.54	0.83	0.83	0.67	0.67	0.42	0.42	0.79	0.79	**0.97**	1.00										
D-Lys	**13**	0.44	0.44	0.67	0.68	0.54	0.54	0.33	0.33	0.63	0.64	0.80	0.79	1.00									
L-Lys	**14**	0.43	0.43	0.66	0.67	0.53	0.53	0.32	0.32	0.63	0.63	0.79	0.78	**0.98**	1.00								
D-Ser	**15**	0.91	0.91	0.72	0.72	0.89	0.89	0.68	0.68	0.78	0.77	0.60	0.61	0.49	0.48	1.00							
L-Ser	**16**	0.89	0.90	0.71	0.73	0.91	0.91	0.70	0.70	0.78	0.78	0.61	0.61	0.50	0.49	**0.97**	1.00						
D-Gln	**17**	0.57	0.57	0.86	0.86	0.70	0.70	0.44	0.44	0.82	0.82	0.93	0.93	0.78	0.77	0.64	0.64	1.00					
L-Gln	**18**	0.59	0.59	0.89	0.89	0.72	0.73	0.46	0.46	0.85	0.85	0.91	0.92	0.75	0.74	0.66	0.67	**0.96**	1.00				
D-Trp	**19**	0.42	0.42	0.65	0.66	0.52	0.53	0.31	0.31	0.62	0.62	0.77	0.76	0.90	0.90	0.47	0.48	0.74	0.73	1.00			
L-Trp	**20**	0.43	0.43	0.66	0.67	0.54	0.54	0.32	0.32	0.62	0.63	0.77	0.76	0.90	0.89	0.48	0.49	0.74	0.74	**0.97**	1.00		
D-Pro	**21**	0.75	0.76	0.87	0.88	0.89	0.89	0.56	0.56	0.94	0.93	0.73	0.74	0.61	0.60	0.82	0.82	0.77	0.81	0.59	0.60	1.00	
L-Pro	**22**	0.75	0.76	0.86	0.88	0.89	0.89	0.57	0.57	0.94	0.93	0.73	0.74	0.61	0.60	0.82	0.83	0.77	0.80	0.59	0.60	**0.99**	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**

D-Thr	**1**	1.00																					
L-Thr	**2**	**0.97**	1.00																				
D-Arg	**3**	0.58	0.58	1.00																			
L-Arg	**4**	0.58	0.59	**0.99**	1.00																		
D-Phe	**5**	0.68	0.67	0.84	0.85	1.00																	
L-Phe	**6**	0.68	0.67	0.84	0.85	**0.99**	1.00																
D-Cys	**7**	0.81	0.82	0.47	0.47	0.55	0.55	1.00															
L-Cys	**8**	0.82	0.83	0.47	0.47	0.55	0.55	**0.99**	1.00														
D-Leu	**9**	0.75	0.73	0.77	0.77	0.90	0.90	0.60	0.61	1.00													
L-Leu	**10**	0.75	0.74	0.77	0.78	0.90	0.90	0.61	0.61	**0.99**	1.00												
D-Hid	**11**	0.79	0.78	0.72	0.73	0.85	0.85	0.65	0.65	0.89	0.89	1.00											
L-Hid	**12**	0.81	0.79	0.71	0.71	0.84	0.84	0.66	0.66	0.90	0.90	**0.97**	1.00										
D-Tyr	**13**	0.64	0.63	0.88	0.88	0.94	0.94	0.51	0.52	0.85	0.85	0.81	0.80	1.00									
L-Tyr	**14**	0.64	0.63	0.88	0.88	0.94	0.94	0.51	0.52	0.85	0.85	0.81	0.80	**1.00**	1.00								
D-Ile	**15**	0.75	0.76	0.76	0.77	0.87	0.87	0.62	0.62	0.96	0.96	0.87	0.88	0.82	0.82	1.00							
L-Ile	**16**	0.76	0.75	0.76	0.76	0.89	0.89	0.61	0.62	0.98	0.98	0.88	0.90	0.84	0.84	**0.97**	1.00						
D-Hie	**17**	0.78	0.76	0.74	0.74	0.86	0.86	0.64	0.64	0.87	0.88	0.97	0.95	0.82	0.82	0.85	0.86	1.00					
L-Hie	**18**	0.80	0.79	0.73	0.73	0.84	0.85	0.65	0.66	0.89	0.90	0.97	0.97	0.81	0.81	0.87	0.89	**0.96**	1.00				
D-Glu	**19**	0.89	0.89	0.66	0.66	0.75	0.75	0.74	0.75	0.81	0.82	0.87	0.88	0.71	0.71	0.81	0.81	0.85	0.87	1.00			
L-Glu	**20**	0.89	0.89	0.65	0.66	0.75	0.75	0.74	0.75	0.81	0.82	0.87	0.88	0.71	0.71	0.81	0.81	0.85	0.87	**1.00**	1.00		
D-Met	**21**	0.76	0.75	0.74	0.75	0.85	0.85	0.61	0.62	0.86	0.86	0.93	0.92	0.82	0.82	0.84	0.85	0.95	0.94	0.83	0.83	1.00	
L-Met	**22**	0.77	0.76	0.73	0.74	0.85	0.85	0.62	0.62	0.87	0.87	0.94	0.93	0.82	0.82	0.85	0.86	0.96	0.95	0.84	0.84	**0.99**	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**

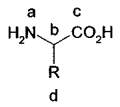
[a] The IPE(0,0) AMS $S_{\alpha\beta}$ values for the individual AA D/L isomer pairs are indicated in bold.

Table 5. *ALL* and *Self* AMS $S_{\alpha\beta}$ Measures for D and L Isomer Pairs of the Amino Acids

IPE	D/L-Ala	D/L-Val	D/L-Asp	D/L-Gly	D/L-Asn	D/L-His	D/L-Lys	D/L-Ser	D/L-Gln	D/L-Trp	D/L-Pro
(0,0)	0.989	0.958	0.991	1.000	0.986	0.971	0.977	0.967	0.958	0.971	0.990
(1,1)	0.987	0.959	0.999	1.000	0.994	0.969	0.965	0.983	0.987	0.966	0.993
(2,2)	1.000	1.000	1.000	1.000	0.984	0.986	0.986	0.923	0.970	0.976	1.000
(3,3)	0.987	0.980	0.996	1.000	0.997	0.944	0.990	0.975	0.917	0.942	0.999
(4,4)	0.987	0.980	0.996	1.000	0.997	0.991	0.996	0.975	0.917	0.966	0.999
(5,5)	1.000	1.000	1.000	1.000	0.984	0.986	0.986	0.923	0.970	0.976	1.000
(6,6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(7,7)	0.992	0.980	0.998	1.000	0.997	0.958	0.981	0.981	0.938	0.964	0.998

IPE	D/L-Thr	D/L-Arg	D/L-Phe	D/L-Cys	D/L-Leu	D/L-Hid	D/L-Tyr	D/L-Ile	D/L-Hie	D/L-Glu	D/L-Met
(0,0)	0.972	0.992	0.994	0.990	0.990	0.967	0.998	0.966	0.961	0.997	0.985
(1,1)	0.976	0.997	0.998	0.990	0.990	0.975	0.997	0.971	0.980	0.995	0.974
(2,2)	0.986	0.997	0.999	1.000	0.999	0.987	0.999	1.000	0.977	0.999	1.000
(3,3)	0.964	0.990	0.991	0.986	0.983	0.941	0.996	0.997	0.946	0.994	0.989
(4,4)	0.964	0.994	0.991	0.987	0.983	0.954	0.996	0.997	0.951	0.994	0.984
(5,5)	0.986	0.997	0.999	1.000	0.999	0.987	0.999	1.000	0.977	0.999	1.000
(6,6)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
(7,7)	0.980	0.992	0.998	0.968	0.995	0.957	0.998	0.979	0.965	0.997	0.987

Table 6. Alignments Selected for the Computation of the RMS $S_{\alpha\beta}$ Measures of D/L Isomer Pairs of Amino Acids[a]



alignment 1	a b c
alignment 2	a c d

[a] Atom d refers to the atom of the *alpha* side chain farthest from C^{α} or alignment atom b.

Table 7 contains the results of the RMS calculations for the set of D/L amino acids for both alignments. The RMS measures given in Table 7 demonstrate that relative similarity measures are more useful than the AMS measures to determine molecular structure differences associated with chirality. Moreover, since RMS is alignment-dependent, the capability to explore different alignments enhances the amount of molecular similarity information that can be gleaned from a library relative to a specific pharmacophore and/or binding site. Alignment 1 is restricted to a common amino acid moiety in that it is independent of alpha side chain atoms. Thus, this alignment should have lower chiral "resolving power" than the second alignment, which employs an alpha chain atom. Table 7a,b shows this general trend in chiral resolving power due to the fact that lower value RMS measures are obtained for alignment 2 in the cases of IPEs (0,0) and (1,1), which are the IPEs that most fully encompass the entire molecular structure of each of the amino acid isomers.

To exemplify the different RMS measures obtained for the two alignments, a set of the smallest amino acids (Gly, Ala, Ser, and Val) is chosen. The resolution power shown by both RMS alignments for this set is not large, which is expected due to the small and similar sizes of the side chains of these amino acids. However, even these small compounds exhibit a revealing trend in their RMS measures. Basically, molecular size ordering is followed in terms of D/L $S_{\alpha\beta}$ resolution for both alignments. Moreover, the amount of chiral resolution as measured by $S_{\alpha\beta}$ changes with IPE. IPE(3,3) and IPE(4,4) give the highest discrimination among isomers followed by IPE(2,2) and IPE(5,5), IPE(0,0) and IPE(1,1). This discriminating order is the same for both alignments, but alignment 2 presents a range of $S_{\alpha\beta}$ values of 0.69−0.92 for IPEs (3,3) and (4,4) and 0.83−0.92 for IPEs (2,2) and (5,5). Alignment 1 presents smaller ranges in $S_{\alpha\beta}$ for chiral discrimination [0.85−0.90 and 0.86−0.94 for IPEs (3,3) and (4,4) and IPEs (2,2) and (5,5), respectively].

Three types of protonated isomers of histidine (see Table 3) were also studied. From an inspection of Table 7 it is clear that the two different alignments used in RMS measures not only differentiate between stereoisomers but also dis-

Table 7. *ALL* and *Self* RMS $S_{\alpha\beta}$ Measures for D and L Isomer Pairs of Amino Acids

Part a[a]

IPE	D/L-Ala	D/L-Val	D/L-Asp	D/L-Gly	D/L-Asn	D/L-His	D/L-Lys	D/L-Ser	D/L-Gln	D/L-Trp	D/L-Pro
(0,0)	0.975	0.933	0.988	1.000	0.947	0.937	0.887	0.962	0.848	0.930	0.950
(1,1)	0.992	0.946	0.967	1.000	0.939	0.901	0.896	0.989	0.818	0.925	0.938
(2,2)	0.941	0.864	0.902	1.000	0.705	0.911	0.654	0.856	0.673	0.889	0.852
(3,3)	0.899	0.852	0.921	1.000	0.860	0.781	0.834	0.853	0.829	0.847	0.900
(4,4)	0.899	0.852	0.921	1.000	0.860	0.866	0.934	0.853	0.784	0.916	0.900
(5,5)	0.941	0.864	0.902	1.000	0.705	0.911	0.654	0.856	0.673	0.889	0.852
(6,6)	1.000	1.000	0.985	1.000	1.000	0.814	1.000	1.000	1.000	0.806	1.000

IPE	D/L-Thr	D/L-Arg	D/L-Phe	D/L-Cys	D/L-Leu	D/L-Hid	D/L-Tyr	D/L-Ile	D/L-Hie	D/L-Glu	D/L-Met
(0,0)	0.902	0.928	0.935	0.966	0.922	0.939	0.893	0.944	0.924	0.967	0.847
(1,1)	0.974	0.959	0.981	0.978	0.922	0.971	0.856	0.939	0.927	0.941	0.803
(2,2)	0.853	0.759	0.815	0.902	0.785	0.925	0.775	0.782	0.807	0.945	0.901
(3,3)	0.707	0.716	0.891	0.951	0.845	0.782	0.725	0.844	0.828	0.960	0.921
(4,4)	0.707	0.873	0.891	0.987	0.845	0.674	0.866	0.844	0.860	0.936	0.936
(5,5)	0.853	0.759	0.815	0.902	0.785	0.925	0.775	0.782	0.807	0.945	0.901
(6,6)	1.000	0.919	0.836	1.000	1.000	0.819	0.695	1.000	0.868	0.828	1.000

Part b[b]

IPE	D/L-Ala	D/L-Val	D/L-Asp	D/L-Gly	D/L-Asn	D/L-His	D/L-Lys	D/L-Ser	D/L-Gln	D/L-Trp	D/L-Pro
(0,0)	0.965	0.923	0.981	1.000	0.904	0.903	0.825	0.932	0.815	0.860	0.912
(1,1)	0.962	0.865	0.968	1.000	0.903	0.850	0.838	0.961	0.810	0.816	0.926
(2,2)	0.919	0.828	0.930	1.000	0.857	0.839	0.667	0.864	0.724	0.736	0.687
(3,3)	0.920	0.687	0.925	1.000	0.856	0.732	0.634	0.882	0.832	0.674	0.857
(4,4)	0.920	0.687	0.909	1.000	0.855	0.899	0.897	0.882	0.832	0.672	0.857
(5,5)	0.919	0.828	0.930	1.000	0.857	0.839	0.667	0.864	0.724	0.736	0.687
(6,6)	1.000	1.000	1.000	1.000	1.000	0.807	1.000	1.000	1.000	0.831	1.000

IPE	D/L-Thr	D/L-Arg	D/L-Phe	D/L-Cys	D/L-Leu	D/L-Hid	D/L-Tyr	D/L-Ile	D/L-Hie	D/L-Glu	D/L-Met
(0,0)	0.884	0.913	0.949	0.952	0.836	0.918	0.843	0.904	0.888	0.968	0.797
(1,1)	0.889	0.900	0.953	0.960	0.813	0.908	0.770	0.860	0.864	0.901	0.700
(2,2)	0.723	0.807	0.877	0.833	0.805	0.776	0.891	0.766	0.817	0.917	0.843
(3,3)	0.832	0.694	0.799	0.955	0.793	0.801	0.681	0.778	0.805	0.921	0.727
(4,4)	0.832	0.827	0.799	0.948	0.793	0.725	0.647	0.778	0.785	0.921	0.839
(5,5)	0.723	0.807	0.877	0.833	0.805	0.776	0.891	0.766	0.817	0.917	0.843
(6,6)	1.000	0.874	0.967	1.000	1.000	0.899	0.671	1.000	0.851	0.997	1.000

[a] The $S_{\alpha\beta}$ measures correspond to using alignment 1 of Table 6. [b] The $S_{\alpha\beta}$ measures correspond to using alignment 2 of Table 6.

Table 8. Correlation Matrix for the Three Properties Used To Describe Molecular Size for the Amino Acids[a]

	D(N-Tail)	MV	MW
D(N-Tail)	1.00		
MV	0.91	1.00	
MW	0.81	0.94	1.00

[a] See text for the definitions of these properties.

tinguish among close tautomeric isomers such as Hie, Hid, and His. Alignment 1 predicts a composite D/L $S_{\alpha\beta}$ similarity resolution in the following order: Hid ≈ Hie > His, while alignment 2 again outperforms alignment 1 in absolute resolution power and yields the composite RMS resolution order Hie > His ≈ Hid.

Another way to interpret the "superiority" of alignment 2 over alignment 1 to distinguish stereoisomers is to study the dependence of RMS on molecular size. The idea is to examine how well an alignment yields good chiral resolution as a function of molecular size. Accordingly alignment 2 is expected to outperform alignment 1 which does not include any atoms of the amino acid side chains that are responsible for variable molecular size over the amino acids.

Molecular size measurement can be considered in terms of any descriptor that is indicative of the concept of the size of the molecule, such as molecular weight, MW, molecular volume, MV, and/or the distance between the N-amino atom and the last atom of the amino acid side chain, D_{N-Tail}. These three descriptors are highly correlated to each other (see Table 8) for the series of amino acids under study. Thus, any one of these descriptors can be used to characterize molecular size.

Both IPE(0,0) and IPE(1,1) RMS measures are considered in this analysis of chiral resolution as a function of alignment. Figure 6 shows a plot of the differences in RMS and AMS measures, ΔS, for alignments 1 and 2, for all amino acids expressed in terms of their MWs. A positive ΔS means that the RMS measurement distinguishes better between stereoisomers than the AMS measure. The two IPE representations are plotted in Figure 6, and two linear correlation trendlines have been added to facilitate comparisons. Figure 6 confirms that RMS measures are, in general, more appropriate than AMS measures to distinguish between stereoisomers and that RMS chiral resolving power increases over AMS as MW increases. The increasing resolving power of RMS over AMS is indicated by the positive slopes of the linear regression trendlines in Figure 6. Moreover, the fact that the ΔS IPE(1,1) regression line has a more positive slope than the ΔS IPE(0,0) line in Figure 6 means that the RMS IPE(1,1) measures are more sensitive in terms of chiral "resolving power" than the RMS IPE(0,0) measures. Figure 6 indicates that RMS IPE(1,1) measures resolves about 5% more structural chirality than RMS IPE(0,0) measures. This difference increases to about 10% when the MW of the molecule is greater than 150 amu.

B. A Set of Hydrocarbon Chains. A series of hydrocarbon compounds of varying MW were selected for AMS analysis. The purpose of this study was to demonstrate the role of conformation in the expression of molecular similarity measures as well as the ability of the methodology to detect isomerism within constant molecular composition.

1. The Role of Flexibility. The study was carried out for the molecules listed in Table 9a, all linear hydrocarbons with 6−10 carbon atoms, using the IPE(0,0) AMS measure. These molecules are highly flexible and of identical atom-type composition. Overall, these molecules differ from one another only in size (number of $-CH_2-$ units) and in conformational flexibility.

It is common among computational chemists to undertake a three-dimensional computational study utilizing an energy minimized structure of a compound of interest. The extent of energy minimization is often viewed as being proportional to the "quality" of the starting chemical structure. However, it is also common that a high degree of molecular flexibility, and practical time constraints, can prevent the implementation of an exhaustive conformational analysis to identify the global minimum energy conformation. Still, molecular similarity measures should incorporate the conformational information resident to the molecules of a training set or

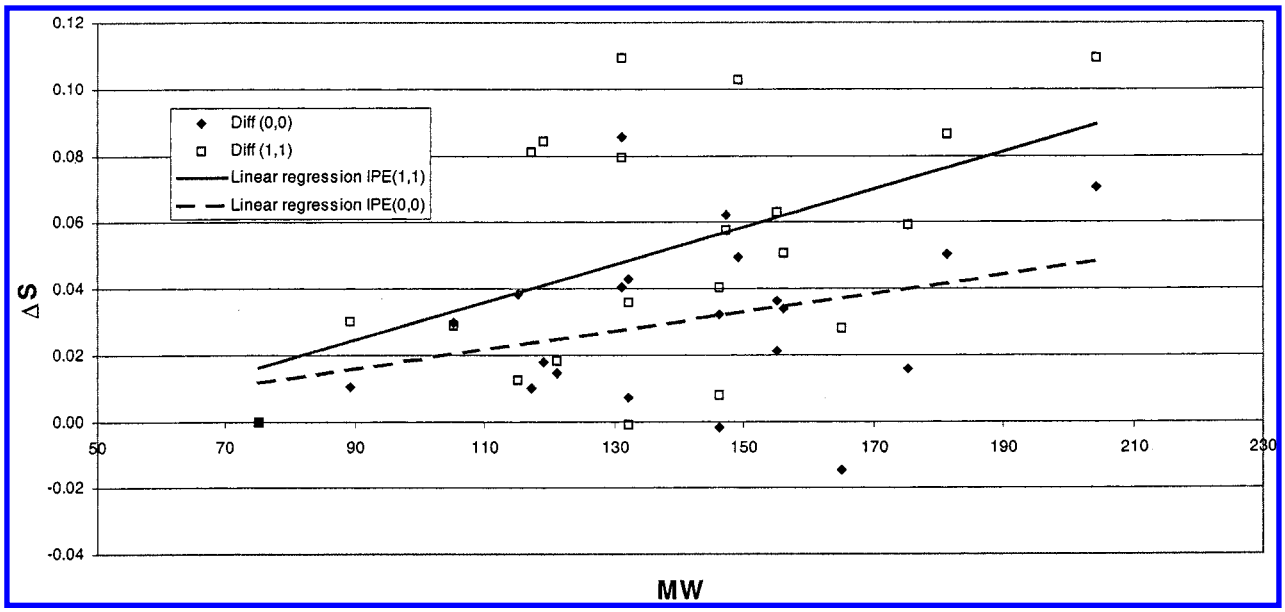


Figure 6. Plot of the differences in the *ALL* RMS $S_{\alpha\beta}$ measures for both alignment 1 and 2, ΔS, for all amino acids listed in Table 3 with respect to their MWs.

Table 9

Part a: A Series of Linear Hydrocarbons, nLC, Studied Using AMS

compound	no. of methylene units	compound	no. of methylene units
n-heptane	5	*n*-decane	8
n-ocatane	6	*n*-undecane	9
n-nonane	7		

Part b: A Diverse Series of Hydrocarbons and Related Compounds Studied using AMS

compound	subgroup classification (see text)
1,2,3-trichlorobenzene	A
1,2,3-trimethylbenzene	A
1,2,4-trichlorobenzene	A
1,2,4-trimethylbenzene	A
1,2-dichlorobenzene	A
1,3,5-trichlorobenzene	A
1,3,5-trimethylbenzene	A
1,3-dichlorobenzene	A
1,4-dichlorobenzene	A
1-butene	B, C
1-pentene	B, C
2,3-"cis"-dimethylpentane	C
2,3-"trans"-dimethylpentane	C
2,4-"cis"-dimethylbutane	C
2,4-"trans"-dimethylbutane	C
2-methylbutane	C
2-methyltoluene	A
3-methyltoluene	A
4-methyltoluene	A
benzene	A, C
butane	C
chlorobenzene	A, C
cis-butene	B
cis-pentene	B
hexane	C
isobutane	C
methane	C
propane	C
toluene	A, C
trans-butene	B, C
trans-pentene	B, C

library. The goal of this particular AMS study is to demonstrate that 4D-QSAR molecular similarity methodology yields molecular similarity measures which are dependent on both the "quality", and the quantity, of the conformational sampling information employed in the study.

For this series of n-linear hydrocarbons, nLC, (linear *n*-alkanes) different MDS were run as a function of the simulation time. Thus, MDS were performed during 0 ps (starting chemical structure), 0.1, 0.5, 1.0, 5.0, and 50 ps for every nLC analogue in Table 9a. Two different initial conformations were defined for each compound: one "correct" and linear, that is, the dihedral angles among four consecutive carbon atoms were set as trans planar. The other conformation is "incorrect" having many bad atom-pair contact regions due to using cis dihedral angles and partial energy minimization.

The only way a *perfect* molecular similarity match can be assured is to compare a molecule to itself. With this in mind the evolution of the molecular similarity measures between "correct" and "incorrect" structures for each nLC analogue were monitored as a function of MDS time. The results of these MDS time versus molecular similarity measures experiments are summarized in Figure 7. MDS at at 0 ps indicate that the presence of bad atom-pair contact distances can affect the AMS measures depending on the flexibility (conformational freedom) of the molecule. Consequently, if the number of methylene units (size of the nLC analog) increases, then the difference in the AMS $S_{\alpha\beta}$ of a pair of conformations ("correct" and "incorrect") of the same compound also increases. However, as more conformational states are sampled over the course of a MDS, the two initial structures of the same molecule evolve toward a common conformational ensemble distribution and common average conformational state. Correspondingly, the AMS $S_{\alpha\beta}$ measure becomes closer to 1.0. This series of MDS for the set of analogue nLCs shows that after 1.0 ps the AMS $S_{\alpha\beta}$ measure is approximately in the range of 0.90, and that after 30−50 ps the AMS $S_{\alpha\beta}$ measure of all nLCs is near 0.97. A value of 0.97 is probably the practical maximum value considering conformational noise in the MDS. The nLC analogues are relatively small but highly flexible molecules that demonstrate the significant role conformational flexibility can play in a molecular similarity analysis.

2. The Role of Molecular Weight. The effect of increasing MW of an nLC, by changing the number of methylene units, on AMS $S_{\alpha\beta}$ was also studied for the series of nLC analogues given in Table 9a, and the results are illustrated in Figure 8, using *n*-heptane as an arbitrary reference for the AMS $S_{\alpha\beta}$ measures. Figure 8 shows that the *ALL*, IPE-(0,0), $S_{\alpha\beta}$ representation of AMS is a sensitive measure to changes in the number of $-CH_2-$ units. This is another favorable attribute of the 4D-QSAR molecular similarity methodology, because it indicates that the method is sensitive to small variations in MW. When structural variations between a pair of molecules contribute less to their corresponding difference in MWs, in terms of percentage difference, the resultant AMS $S_{\alpha\beta}$ measure shows less difference as well. In this particular example because of the particular chemical composition, only *ALL*, *NP*, *HS* are available in the nLC analogues. The importance of different interaction pharmacophore types of AMS representations will be discussed in the next sections of this paper.

3. AMS for a Broader Series of Common Carbon-Based Compounds. Table 9b lists additional hydrocarbons, and a few related structures, that were used to define an extended hydrocarbon library. This small library consists of 31 compounds which encompass small aliphatic chains, alkenes, aromatic, and haloaromatic compounds. Since the series was generated with the goal of exploring molecular diversity, a few isomers have also been included. Table 9b has been divided into three subgroups: (A) aromatic and chloroaromatic compounds, (B) alkenes, and (C) alkanes. Some of the compounds from subgroups A and B can also be considered part of subgroup C. This subgrouping facilitates analysis and discussion of the AMS calculations.

Subgroup A: Small Aromatic and Related Compounds. This subgroup of compounds includes mono-, di-, and trisubstituted benzene analogues. The substituent group is either methyl or chlorine. Table 10 lists the AMS $S_{\alpha\beta}$ measures for both the IPE(0,0) and IPE(7,7) representations for subgroup A. It can be seen in Table 10 that the *ALL*, IPE(0,0), representation yields very high similarity measures within, and between, members of the trichloro- and dichlorobenzene subsets. The same *ALL* AMS representation gives correspondingly slightly lower S_α values within, and between, members of the trimethyl- and dimethylbenzene analogues.

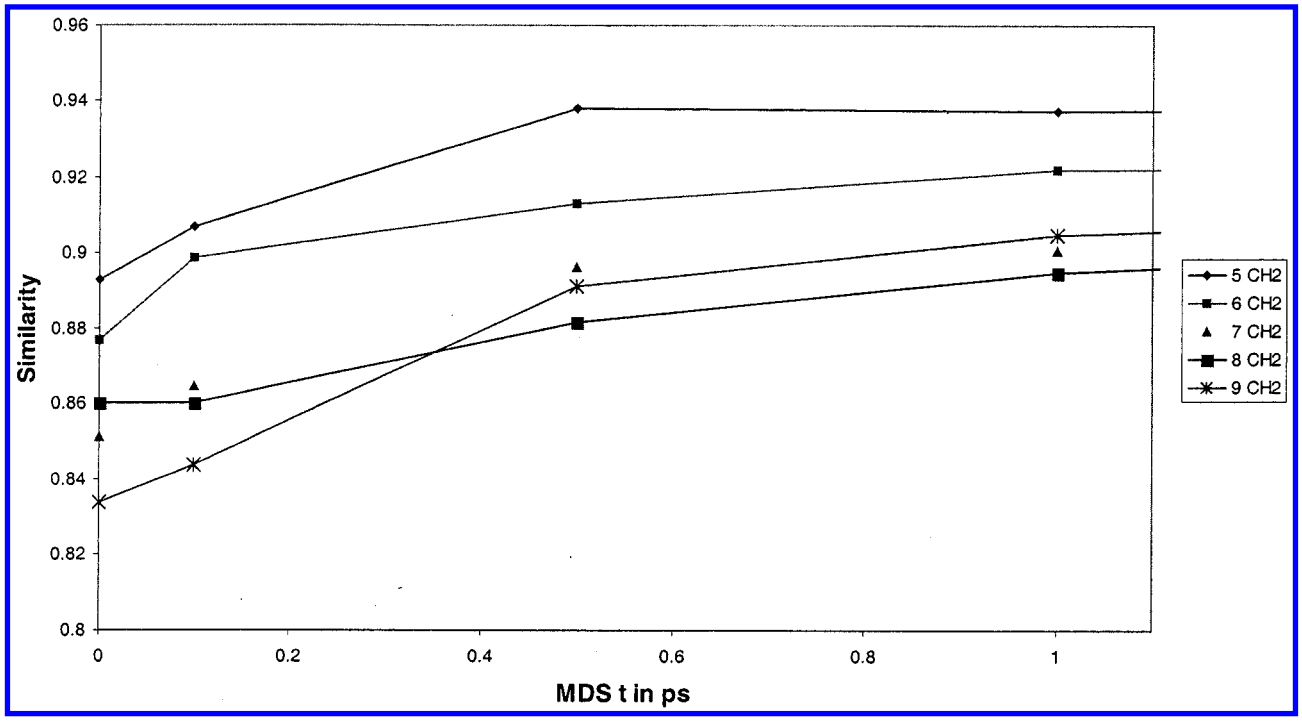


Figure 7. Dependence of *ALL* RMS $S_{\alpha\beta}$ values on molecular dynamics simulation time.

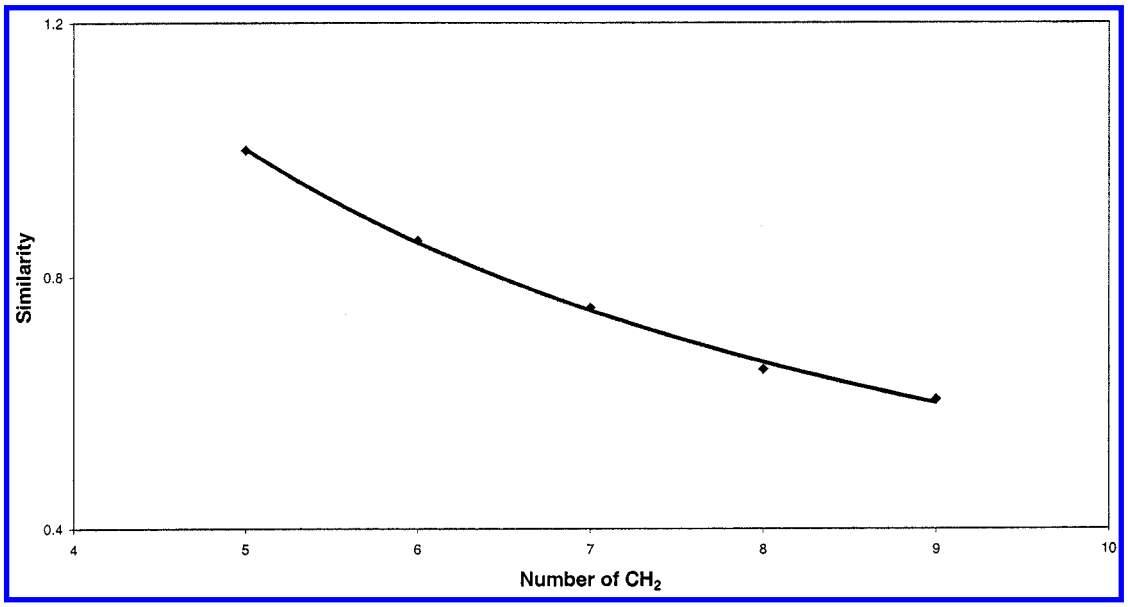


Figure 8. Dependence of *ALL* RMS $S_{\alpha\beta}$ values on the number of methylene units in the *n*-alkane chain molecules.

The slightly lower $S_{\alpha\beta}$ values for the methyl compounds is due to the presence of the extra rotational degree of freedom of each methyl group relative to each chloro group.

The same relative molecular similarity measures ranking within the chloro- and methyl-substituted subsets, respectively, is indicative of the 4D-QSAR based method treating both substituents as similar entities within the subsets. However, the comparison of equivalent molecular similarity measures between chloro- and methyl-analogues reveals that the IPE(0,0) AMS $S_{\alpha\beta}$ are 0.55−0.58 for trisubstitution and 0.59−0.67 for disubstitution. The reason for these low molecular similarity measures between equivalent methyl- and chloro-analogues is that the methodology does not treat chloro- and methyl-substitutions as comparable under the AMS *ALL* IPE representation. In particular, a methyl group has three more atoms than a "chloro-group" for the *ALL* IPE.

There is an IPE representation that leads to higher $S_{\alpha\beta}$ measures between methyl and chlorine substituents, namely the hydrogen suppressed, *HS*, IPE(7,7). The *HS* IPE does not consider molecular hydrogen. Thus, *HS* is a good IPE to use to compare methyl and chlorine substitution patterns. Under the *HS* atom type representation the AMS $S_{\alpha\beta}$ are correspondingly higher than those determined using IPE-(0,0), with $S_{\alpha\beta}$ values greater than 0.93 for corresponding chloro and methyl analogues. Chlorobenze and toluene have a 0.99 $S_{\alpha\beta}$ molecular similarity measure using the *HS* AMS representation.

Subgroup B: A Small Series of Alkenes. Table 11 lists the *ALL* and *HS* AMS $S_{\alpha\beta}$ representations for the six alkenes of subgroup B of Table 9b. Two main features can be observed in the AMS matrices of Table 11. First, the presence of one double bond in a four-carbon molecular system is a

Table 10. AMS $S_{\alpha\beta}$ for Subset A of Table 9b[a]

Subgroup A − *ALL*, IPE(0,0)															
123-tricl-benzene	**1**	1.00													
135-tricl-benzene	**2**	0.99	1.00												
124-tricl-benzene	**3**	0.99	0.98	1.00											
123-trime-benzene	**4**	0.58	0.58	0.58	1.00										
135-trime-benzene	**5**	0.58	0.57	0.58	0.95	1.00									
124-trime-benzene	**6**	0.55	0.55	0.56	0.95	0.92	1.00								
12dicl-benzene	**7**	0.98	0.97	0.98	0.58	0.57	0.55	1.00							
13dicl-benzene	**8**	0.98	0.98	0.98	0.58	0.57	0.55	0.99	1.00						
14dicl-benzene	**9**	0.98	0.97	0.98	0.58	0.57	0.55	0.99	0.99	1.00					
cl-benzene	**10**	0.97	0.96	0.96	0.58	0.57	0.55	0.98	0.98	0.98	1.00				
toluene	**11**	0.77	0.77	0.78	0.72	0.68	0.73	0.77	0.77	0.77	0.76	1.00			
2me-toluene	**12**	0.67	0.66	0.67	0.87	0.84	0.84	0.67	0.66	0.67	0.66	0.81	1.00		
3me-toluene	**13**	0.64	0.64	0.64	0.85	0.82	0.87	0.64	0.64	0.64	0.63	0.81	0.94	1.00	
4me-toluene	**14**	0.59	0.59	0.60	0.83	0.78	0.84	0.59	0.59	0.60	0.59	0.79	0.90	0.93	1.00
benzene	**15**	0.95	0.95	0.95	0.57	0.57	0.54	0.96	0.96	0.96	0.98	0.75	0.65	0.63	0.58	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
HS, IPE(7,7)															
123-tricl-benzene	**1**	1.00													
135-tricl-benzene	**2**	0.98	1.00												
124-tricl-benzene	**3**	0.95	0.94	1.00											
123-trime-benzene	**4**	0.97	0.96	0.94	1.00										
135-trime-benzene	**5**	0.97	0.97	0.93	0.98	1.00									
124-trime-benzene	**6**	0.95	0.93	0.97	0.95	0.94	1.00								
12dicl-benzene	**7**	0.87	0.85	0.86	0.88	0.86	0.87	1.00							
13dicl-benzene	**8**	0.86	0.84	0.87	0.87	0.85	0.89	0.98	1.00						
14dicl-benzene	**9**	0.83	0.81	0.86	0.84	0.82	0.88	0.94	0.95	1.00					
cl-benzene	**10**	0.75	0.73	0.77	0.76	0.74	0.78	0.85	0.85	0.84	1.00				
toluene	**11**	0.74	0.73	0.76	0.76	0.74	0.78	0.85	0.85	0.83	0.99	1.00			
2me-toluene	**12**	0.86	0.85	0.85	0.88	0.85	0.86	0.98	0.96	0.93	0.86	0.86	1.00		
3me-toluene	**13**	0.85	0.83	0.86	0.86	0.85	0.88	0.97	0.98	0.94	0.86	0.86	0.98	1.00	
4me-toluene	**14**	0.82	0.81	0.85	0.84	0.82	0.87	0.94	0.95	0.98	0.86	0.85	0.94	0.95	1.00
benzene	**15**	0.64	0.63	0.61	0.65	0.65	0.63	0.72	0.70	0.68	0.80	0.81	0.73	0.72	0.69	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**

[a] The AMS $S_{\alpha\beta}$ matrices correspond to using IPE(0,0) and IPE(7,7).

Table 11. AMS $S_{\alpha\beta}$ Analysis for Subset B of Table 9b[a]

Subgroup B − *ALL*, IPE(0,0)							
1-butene	**1**	1.00					
cis-2-butene	**2**	0.95	1.00				
trans-2-butene	**3**	0.95	0.92	1.00			
1-pentene	**4**	0.83	0.81	0.83	1.00		
cis-2-pentene	**5**	0.83	0.81	0.83	0.97	1.00	
trans-2-pentene	**6**	0.80	0.79	0.83	0.96	0.95	1.00
		1	**2**	**3**	**4**	**5**	**6**
HS, IPE(7,7)							
1-butene	**1**	1.00					
cis-2-butene	**2**	0.97	1.00				
trans-2-butene	**3**	0.99	0.96	1.00			
1-pentene	**4**	0.78	0.76	0.79	1.00		
cis-2-pentene	**5**	0.80	0.78	0.80	0.96	1.00	
trans-2-pentene	**6**	0.78	0.76	0.79	0.99	0.97	1.00
		1	**2**	**3**	**4**	**5**	**6**

[a] The AMS $S_{\alpha\beta}$ matrices correspond to using IPE(0,0) and IPE(7,7).

sufficient structural change to produce a cis to trans IPE-(0,0) AMS $S_{\alpha\beta}$ value of 0.92 between the two isomers. Differentiation between terminal cis and/or trans double bonds yields an *ALL* AMS $S_{\alpha\beta}$ measure of 0.95 between such isomers. When one double bond is placed in a five-carbon chain the cis/trans isomer *ALL* AMS $S_{\alpha\beta}$ is 0.95. Thus, small structural changes (small changes in the overall MDDM matrix) get incorporated into the molecular similarity measure with the *ALL*, IPE(0,0), representation, and the molecular similarity measure changes little with small changes in MW.

The *HS* IPE $S_{\alpha\beta}$ measures offer no advantage over the *ALL* IPE measures for the alkenes of Table 9b. Contribution of the hydrogens, in particular those of the double bond, to the MDDM matrix cannot be replaced by the carbon "backbone" distance elements for small molecules. This is the reason the pairs of compounds 1-butene/*trans*-2-butene and 1-pentene/*trans*-2-pentene have a high *HS* AMS measure value of 0.99, while for 1-butene/*cis*-2-butene and 1-pentene/*cis*-2-pentene the *HS* AMS value changes to 0.97 and 0.96, respectively. Even when there are small changes in $S_{\alpha\beta}$ measures, they are still significant with respect to the noise introduced by the MDS because of the relatively high rigidity of the subsystems under consideration.

Subgroup C: A Series of Small Hydrocarbons. The corresponding *ALL* and *HS* IPE AMS $S_{\alpha\beta}$ measures for a third, and final, subgroup of Table 9b are listed in Table 12. This library is composed of 19 small molecules containing one to seven carbons. The subset of alkenes (butene and pentene derivatives used to construct Table 11) has been included in building this library to permit molecular comparisons to a group of butane and pentane isomers. Benzene, chlorobenzene, and toluene have also been included to allow comparisons with hexane and heptane isomers.

Once again, the $S_{\alpha\beta}$ of Table 12 demonstrate the ability of AMS to discern between close structural isomers using either the *ALL* or *HS* IPE representations. In Table 12 the AMS $S_{\alpha\beta}$ matrix has been created by positioning the smaller compound, methane, on top and increasingly larger molecules down the column, and left to right on the rows. The first column of Table 12 indicates the relative *ALL* AMS $S_{\alpha\beta}$ with respect to methane, whereas the second column corresponds to the *ALL* AMS $S_{\alpha\beta}$ measures relative to propane. As expected, there is a general decrease in AMS $S_{\alpha\beta}$ measures

Table 12. AMS $S_{\alpha\beta}$ for Subset C of Table 9b[a]

Subgroup C − *ALL*, IPE(0,0)

methane	**1**	1.00																	
propane	**2**	0.40	1.00																
butane	**3**	0.28	0.76	1.00															
isobutene	**4**	0.30	0.75	0.88	1.00														
1-butene	**5**	0.34	0.88	0.86	0.81	1.00													
cis-butene	**6**	0.33	0.86	0.83	0.81	0.95	1.00												
trans-butene	**7**	0.32	0.84	0.86	0.77	0.95	0.92	1.00											
2-me-butane	**8**	0.23	0.64	0.79	0.78	0.69	0.69	0.67	1.00										
1-pentene	**9**	0.24	0.69	0.91	0.81	0.79	0.77	0.79	0.85	1.00									
cis-2-pentene	**10**	0.25	0.70	0.89	0.83	0.79	0.78	0.79	0.84	0.97	1.00								
trans-2-pentene	**11**	0.23	0.68	0.88	0.79	0.76	0.75	0.79	0.83	0.96	0.95	1.00							
hexane	**12**	0.15	0.51	0.66	0.60	0.58	0.57	0.61	0.76	0.73	0.73	0.75	1.00						
benzene	**13**	0.33	0.83	0.79	0.83	0.89	0.90	0.86	0.71	0.73	0.72	0.72	0.53	1.00					
toluene	**14**	0.24	0.69	0.87	0.85	0.77	0.78	0.76	0.85	0.93	0.95	0.93	0.72	0.75	1.00				
Cl-benzene	**15**	0.32	0.82	0.80	0.83	0.89	0.89	0.87	0.71	0.74	0.73	0.72	0.54	0.98	0.76	1.00			
2,3cis-dime-pentane	**16**	0.16	0.48	0.61	0.57	0.54	0.51	0.52	0.74	0.66	0.67	0.66	0.85	0.51	0.66	0.52	1.00		
2,3trans-dime-pentane	**17**	0.16	0.48	0.61	0.58	0.54	0.52	0.52	0.74	0.66	0.67	0.65	0.85	0.52	0.67	0.52	0.98	1.00	
2,3cis-dime-butane	**18**	0.20	0.52	0.66	0.66	0.58	0.57	0.57	0.82	0.71	0.72	0.70	0.86	0.58	0.73	0.58	0.86	0.86	1.00
2,3trans-dime-butane	**19**	0.18	0.53	0.66	0.68	0.58	0.58	0.56	0.84	0.72	0.72	0.71	0.87	0.60	0.75	0.61	0.85	0.86	0.95	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**

HS, IPE(7,7)

methane	**1**	1.00																	
propane	**2**	0.00	1.00																
butane	**3**	0.00	0.72	1.00															
isobutene	**4**	0.00	0.70	0.88	1.00														
1-butene	**5**	0.00	0.74	0.97	0.90	1.00													
cis-butene	**6**	0.00	0.76	0.94	0.93	0.97	1.00												
trans-butene	**7**	0.00	0.74	0.98	0.89	0.99	0.96	1.00											
2-me-butane	**8**	0.00	0.58	0.78	0.76	0.78	0.79	0.77	1.00										
1-pentene	**9**	0.00	0.56	0.79	0.69	0.78	0.76	0.79	0.93	1.00									
cis-pentene	**10**	0.00	0.59	0.80	0.73	0.80	0.78	0.80	0.96	0.96	1.00								
trans-pentene	**11**	0.00	0.57	0.79	0.70	0.78	0.76	0.79	0.93	0.99	0.97	1.00							
hexane	**12**	0.00	0.43	0.61	0.54	0.61	0.60	0.61	0.74	0.79	0.78	0.79	1.00						
benzene	**13**	0.00	0.49	0.63	0.68	0.63	0.65	0.63	0.80	0.75	0.78	0.75	0.78	1.00					
toluene	**14**	0.00	0.42	0.57	0.57	0.57	0.58	0.57	0.73	0.71	0.72	0.71	0.77	0.81	1.00				
Cl-benzene	**15**	0.00	0.41	0.56	0.57	0.57	0.58	0.56	0.73	0.70	0.72	0.71	0.77	0.80	0.99	1.00			
2,3cis-dime-pentane	**16**	0.00	0.38	0.55	0.50	0.54	0.53	0.54	0.70	0.69	0.69	0.69	0.81	0.72	0.89	0.90	1.00		
2,3trans-dime-pentane	**17**	0.00	0.38	0.54	0.51	0.54	0.54	0.54	0.71	0.69	0.69	0.69	0.81	0.73	0.90	0.91	0.99	1.00	
2,3cis-dime-butane	**18**	0.00	0.46	0.64	0.62	0.63	0.63	0.63	0.82	0.80	0.81	0.80	0.85	0.87	0.85	0.85	0.83	0.84	1.00
2,3trans-dime-butane	**19**	0.00	0.46	0.64	0.63	0.63	0.65	0.63	0.84	0.80	0.81	0.80	0.86	0.88	0.87	0.87	0.83	0.84	0.97	1.00
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**

[a] The AMS $S_{\alpha\beta}$ matrices correspond to using IPE(0,0) and IPE(7,7).

with increasing molecular weight. To better interpret the resulting AMS $S_{\alpha\beta}$ matrices and to understand the type of information provided by *ALL* and *HS* IPE representations, a few compound families have been individually analyzed.

Butane. The *ALL* AMS IPE representation yields the following order of $S_{\alpha\beta}$ measures relative to butane: *isobutane* > *1-butene* ≈ *trans*-butene > *cis*-butene > 2-methylbutane. The *HS* AMS $S_{\alpha\beta}$ order differs from the *ALL* IPE representation since it gives higher $S_{\alpha\beta}$ measures to the alkenes as compared to the alkane structures: *trans*-butene > 1-butene > *cis*-butene > isobutene > 2-methylbutane. This finding corroborates the view that the *HS* AMS representation tends to correlate with both molecular shape and 2D molecular (chemical) structure.

2-Methylbutane. Compared to 2-methylbutane the following *ALL* AMS $S_{\alpha\beta}$ order is obtained: *1-pentene* ≈ *cis*-2-pentene ≈ *trans-pentene* ≈ *2,3-dimethylbutane*. All $S_{\alpha\beta}$ values are around 0.85. To the contrary, when the *HS* IPE is utilized the AMS $S_{\alpha\beta}$ measures are higher than 0.93 when compared to the pentene subset: *cis*-2-pentene > 1-pentene ≈ *trans-pentene* > *2,3-dimethylbutane*.

Hexane. This compound, when used as a molecular comparison reference, results, on average, in greater differences between the $S_{\alpha\beta}$ measures of the *ALL* and *HS* AMS representations than the other two shape reference molecules. The *ALL* AMS $S_{\alpha\beta}$ values are in the order *2,3-dimethylbutane* ≈ *2,3-dimethylpentane* > *toluene* ≫ *benzene* ≈ *chlorobenzene*. The *ALL* AMS $S_{\alpha\beta}$ for the pair hexane/benzene is 0.53. As discussed above, *HS* IPE representations tend to correlate with 2D-graph (chemical structure) features of molecules. Thus, it is expected that the *HS* IPE AMS $S_{\alpha\beta}$ value for the pair of compounds hexane/benzene should be greater than that of the *ALL* IPE. In fact, the *HS* $S_{\alpha\beta}$ is 0.78, and the relative ranking of the $S_{\alpha\beta}$ measures for the compound subset is *2,3-dimethylbutane* > *2,3-dimethylpentane* > *benzene* ≈ *toluene* ≈ *chlorobenzene*. 2,3-"Trans" and/or "cis"-dimethyl alkanes correspond to the initial structures used to perform the AMS analysis. Since both "trans" and "cis" starting conformers lead to about the same $S_{\alpha\beta}$ values, this is another corroboration of the findings from nLC calculations.

C. Bioisosterism and Conformational Flexibility. An AMS analysis of a library of approximately 3300 randomly diverse compounds identified a molecule with limited *chemical structure* similarity to two compounds, which are analogues to one another, to be highly similar in both *ALL* and *HS* IPE similarity measures to each of the two analogues.

Table 13. *ALL* AMS $S_{\alpha\beta}$ and *ARO* AMS $S_{\alpha\beta}$ Measures for a Subset of 28 Organic Compounds from a Diverse Library of of Approximately 3300 Compounds

Part a[a]: *ALL* AMS $S_{\alpha\beta}$

1	**1**	1.000																												
1121	**2**	0.008	1.000																											
1122	**3**	0.001	0.710	1.000																										
1123	**4**	0.045	0.208	0.090	1.000																									
1124	**5**	***0.005***	***0.631***	***0.723***	***0.141***	***1.000***																								
1126	**6**	0.036	0.430	0.248	0.824	***0.371***	***1.000***																							
1127	**7**	***0.005***	***0.635***	***0.730***	0.144	**0.996**	***0.375***	***1.000***																						
1430	**8**	0.212	0.000	0.000	0.002	***0.000***	0.001	***0.000***	1.000																					
1431	**9**	0.162	0.000	0.000	0.001	***0.000***	0.000	***0.000***	**0.962**	1.000																				
150	**10**	0.000	0.194	0.461	0.014	***0.159***	0.043	***0.168***	0.000	0.000	1.000																			
175	**11**	***0.004***	***0.629***	***0.783***	0.137	***0.970***	***0.366***	**0.970**	***0.000***	***0.000***	***0.235***	***1.000***																		
2000	**12**	0.618	0.007	0.001	0.105	***0.004***	0.059	***0.004***	0.308	0.241	0.000	***0.003***	1.000																	
2050	**13**	0.117	0.432	0.166	0.334	***0.267***	0.475	***0.271***	0.002	0.001	0.029	***0.273***	0.112	1.000																
2090	**14**	0.011	0.678	0.565	0.124	***0.709***	0.311	***0.711***	0.000	0.000	0.164	***0.747***	0.008	0.509	1.000															
2100	**15**	0.006	0.479	0.474	0.041	***0.694***	0.139	***0.685***	0.000	0.000	0.090	***0.622***	0.003	0.249	0.730	1.000														
2102	**16**	0.020	0.658	0.498	0.146	***0.811***	0.356	***0.800***	0.000	0.000	0.068	***0.690***	0.011	0.407	0.705	0.786	1.000													
2103	**17**	0.020	0.677	0.502	0.125	***0.770***	0.316	***0.770***	0.000	0.000	0.083	***0.696***	0.011	0.470	*0.806*	*0.841*	**0.953**	1.000												
2104	**18**	0.021	0.659	0.482	0.124	***0.762***	0.312	***0.757***	0.000	0.000	0.076	***0.675***	0.011	0.464	0.788	*0.843*	**0.964**	**0.993**	1.000											
2105	**19**	0.171	0.093	0.031	0.644	***0.090***	0.542	***0.090***	0.014	0.008	0.002	***0.080***	0.263	0.272	0.073	0.030	0.119	0.100	0.100	1.000										
2106	**20**	0.023	0.000	0.000	0.009	***0.000***	0.002	***0.000***	0.068	0.057	0.000	***0.000***	0.070	0.001	0.000	0.000	0.000	0.000	0.000	0.018	1.000									
2108	**21**	0.015	0.553	0.437	0.117	***0.783***	0.307	***0.753***	0.000	0.000	0.048	***0.652***	0.009	0.300	0.623	0.725	*0.894*	*0.806*	*0.820*	0.108	0.000	1.000								
2903	**22**	0.003	0.734	*0.850*	0.194	***0.788***	0.437	***0.807***	0.000	0.000	0.360	***0.883***	0.003	0.314	0.701	0.458	0.570	0.597	0.570	0.083	0.000	0.467	1.000							
2905	**23**	0.002	0.517	0.621	0.043	***0.552***	0.137	***0.554***	0.000	0.000	0.309	***0.627***	0.001	0.247	*0.809*	0.665	0.470	0.559	0.546	0.018	0.000	0.401	0.601	1.000						
2906	**24**	0.257	0.241	0.079	0.320	***0.217***	0.424	***0.217***	0.008	0.005	0.006	***0.185***	0.232	0.746	0.323	0.199	0.382	0.390	0.397	0.435	0.002	0.324	0.168	0.116	1.000					
3000	**25**	0.013	0.418	0.364	0.070	***0.540***	0.194	***0.536***	0.000	0.000	0.105	***0.592***	0.007	0.452	*0.857*	0.686	0.583	0.707	0.697	0.052	0.000	0.474	0.504	0.717	0.290	1.000				
3019	**26**	0.002	0.205	0.149	0.413	***0.092***	0.395	***0.099***	0.000	0.000	0.107	***0.120***	0.007	0.155	0.103	0.023	0.061	0.063	0.058	0.096	0.000	0.039	0.257	0.065	0.060	0.051	1.000			
400	**27**	0.199	0.081	0.016	0.346	***0.040***	0.271	***0.039***	0.014	0.008	0.001	***0.030***	0.217	0.176	0.035	0.018	0.080	0.065	0.066	0.541	0.038	0.074	0.034	0.007	0.288	0.019	0.039	1.000		
500	**28**	0.000	0.246	0.398	0.006	***0.201***	0.024	***0.205***	0.000	0.000	0.437	***0.254***	0.000	0.058	0.342	0.310	0.148	0.202	0.190	0.001	0.000	0.114	0.280	0.655	0.016	0.284	0.022	0.001	1.000	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	

Part b[b]: *ARO* AMS $S_{\alpha\beta}$

1	**1**	1.000																												
1121	**2**	0.358	1.000																											
1122	**3**	0.024	0.371	1.000																										
1123	**4**	0.013	0.202	0.787	1.000																									
1124	**5**	***0.011***	***0.172***	***0.721***	***0.982***	***1.000***																								
1126	**6**	0.013	0.199	0.782	1.000	***0.984***	1.000																							
1127	**7**	***0.013***	***0.200***	***0.785***	***1.000***	***0.982***	***1.000***	***1.000***																						
1430	**8**	0.020	0.323	**0.973**	0.881	***0.813***	0.877	***0.880***	1.000																					
1431	**9**	0.016	0.265	**0.935**	**0.922**	***0.875***	**0.920**	***0.923***	**0.974**	1.000																				
150	**10**	0.331	0.678	0.266	0.244	***0.219***	0.242	***0.240***	0.277	0.233	1.000																			
175	**11**	***0.447***	***0.781***	***0.229***	***0.172***	***0.148***	***0.170***	***0.169***	***0.224***	***0.178***	***0.929***	***1.000***																		
2000	**12**	0.301	**0.913**	0.424	0.299	***0.258***	0.295	***0.296***	0.407	0.342	0.864	***0.873***	1.000																	
2050	**13**	0.016	0.245	0.875	**0.978**	***0.947***	**0.976**	***0.978***	**0.936**	**0.972**	0.248	***0.186***	0.332	1.000																
2090	**14**	0.018	0.284	**0.936**	**0.939**	***0.891***	**0.936**	***0.939***	**0.974**	**0.989**	0.257	***0.201***	0.366	**0.987**	1.000															
2100	**15**	0.476	**0.935**	0.226	0.108	***0.088***	0.106	***0.107***	0.192	0.149	0.608	***0.769***	0.801	0.135	0.162	1.000														
2102	**16**	0.465	**0.950**	0.246	0.119	***0.098***	0.117	***0.117***	0.207	0.165	0.601	***0.748***	0.808	0.149	0.178	**0.993**	1.000													
2103	**17**	0.718	0.245	0.010	0.003	***0.003***	0.003	***0.003***	0.007	0.005	0.152	***0.262***	0.165	0.004	0.006	0.390	0.362	1.000												
2104	**18**	0.730	0.244	0.010	0.003	***0.003***	0.003	***0.003***	0.007	0.005	0.157	***0.268***	0.167	0.004	0.006	0.388	0.360	**0.995**	1.000											
2105	**19**	0.328	0.631	0.269	0.274	***0.255***	0.271	***0.271***	0.279	0.248	0.922	***0.856***	0.790	0.279	0.281	0.540	0.546	0.128	0.132	1.000										
2106	**20**	0.189	0.011	0.000	0.000	***0.000***	0.000	***0.000***	0.000	0.000	0.033	***0.040***	0.013	0.000	0.000	0.018	0.017	0.116	0.125	0.030	1.000									
2108	**21**	0.002	0.071	0.609	0.783	***0.811***	0.785	***0.785***	0.664	0.747	0.071	***0.046***	0.104	0.780	0.740	0.032	0.036	0.000	0.000	0.085	0.000	1.000								
2903	**22**	0.391	0.779	0.282	0.234	***0.207***	0.231	***0.231***	0.281	0.235	**0.966**	***0.965***	**0.910**	0.248	0.263	0.716	0.711	0.192	0.197	**0.936**	0.030	0.070	1.000							
2905	**23**	0.000	0.001	0.047	0.091	***0.099***	0.092	***0.092***	0.059	0.080	0.001	***0.000***	0.001	0.083	0.073	0.000	0.000	0.000	0.000	0.001	0.000	0.278	0.001	1.000						
2906	**24**	0.010	0.170	0.753	**0.984**	***0.970***	**0.985**	***0.986***	0.853	**0.918**	0.203	***0.139***	0.256	**0.959**	**0.917**	0.089	0.098	0.002	0.002	0.230	0.000	0.799	0.193	0.111	1.000					
3000	**25**	0.425	0.852	0.275	0.198	***0.171***	0.196	***0.195***	0.266	0.215	**0.928**	***0.976***	**0.937**	0.217	0.238	0.822	0.810	0.246	0.249	0.850	0.030	0.057	**0.972**	0.000	0.164	1.000				
3019	**26**	0.284	0.590	0.252	0.258	***0.235***	0.257	***0.255***	0.271	0.231	**0.980**	***0.872***	0.799	0.255	0.258	0.514	0.507	0.117	0.121	**0.936**	0.031	0.076	**0.937**	0.001	0.216	0.867	1.000			
400	**27**	0.216	0.790	0.504	0.407	***0.369***	0.404	***0.404***	0.499	0.452	0.809	***0.721***	**0.929**	0.440	0.468	0.625	0.652	0.093	0.094	0.772	0.009	0.163	0.821	0.003	0.367	0.811	0.768	1.000		
500	**28**	0.009	0.191	0.882	**0.911**	***0.873***	**0.909**	***0.911***	**0.929**	**0.949**	0.174	***0.129***	0.255	**0.945**	**0.950**	0.101	0.111	0.003	0.003	0.185	0.000	0.856	0.175	0.119	0.890	0.155	0.177	0.339	1.000	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	

[a] *ALL* AMS $S_{\alpha\beta}$ values larger than 0.90 are in boldface while those values larger than 0.80 are in italics. The rows and columns corresponding to the three compounds (1124, 1127, and 175) discussed in the text are in both boldface and italics. [b] *ARO* AMS $S_{\alpha\beta}$ values larger than 0.90 are in boldface while those values larger than 0.80 are in italics. The rows and columns corresponding to the three compounds (1124, 1127, and 175) discussed in the text are in both boldface and italics.

That is, the composite molecular shape profiles of all three molecules are predicted to be highly similar to one another. The high molecular similarity, however, does not extend over all IPE types, and low molecular similarity measures are also found for some IPEs. Table 13 contains the AMS $S_{\alpha\beta}$ measures for the *ALL*, IPE(0,0), and *ARO*, IPE(6,6), atom types for a common region of the molecular similarity matrices of the 3300 compound library. The rows and columns which are both italicized and boldface in Table 13a allow identification of three compounds, 175, 1124, and 1127, that have high $S_{\alpha\beta}$ values for the *ALL* IPE. These three compounds are *isosteric* to one another. That is, 175, 1124, and 1127 have nearly identical conformationally averaged spatial molecular shapes to one another when atom typing is not included in the analysis. There are other pairs of compounds spread out over this AMS matrix having *ALL* $S_{\alpha\beta}$ higher than 0.75. However, only pairs having $S_{\alpha\beta} > 0.90$ are flagged in this particular study.

As can be seen from Table 13a and Table 9, the high *ALL* AMS $S_{\alpha\beta}$ for compound pair 1124 and 1127 derives from their near common chemical structures. In contrast, compound 175 has a different bonding topology and chemical composition than 1124 and 1127. Yet, it has high *ALL* AMS $S_{\alpha\beta}$ values, with both 1124 and 1127. However, there is a higher $S_{\alpha\beta}$ set of measures for the *ALL* IPE than for the *ARO* IPE, see Table 13b. Only analogues 1124 and 1127 have an *ARO* AMS $S_{\alpha\beta}$ value greater than 0.90. These differences in $S_{\alpha\beta}$ for *ALL* and *ARO* IPEs can be easily corroborated from inspection of the chemical structures and comparing the number of aromatic groups that are present in each of these three compounds.

This type of finding for the *ALL* and *ARO* $S_{\alpha\beta}$ of 175 with 1124 or 1125 can be viewed as a *"partial biosteric match"* for drug discovery and is a direct consequence of the deconvolution of molecular similarity information into pharmacophoric group similarity measures. If the biological activity of a compound depends on a few distinct features that can be represented as IPEs, then AMS analysis should be able to identify these features in terms of partial, that is IPE, molecular similarity representations. It is also clear that a pair of compounds α and β must present high AMS $S_{\alpha\beta}$ measures for all IPE representations in order to be considered *"completely bioisosteric"* compounds. That is why the term "*partial bioisosteric match"* is used above since only the *ALL* IPE is confirmed to yield high $S_{\alpha\beta}$ values. As defined in eq 8, the underlying reason for the high AMS $S_{\alpha\beta}$ values for *ALL* IPE resides in the corresponding eigenvalues of these three compounds. These eigenvalues are listed and plotted in Table 14 and Figure 9, respectively.

In Figure 9 the *ALL* eigenvalues for compounds 175, 1124, and 1127 are compared to each other and to a randomly selected "compound 1" from the 3300 compound library. This type of graphical representation is useful for verifying that a high AMS $S_{\alpha\beta}$ value is derived from a close numerical match of the eigenvalues, as is the case for 175, 1124, and 1127 but not the case, for example, for 175 and 1. In general, and because of the definition given in eq 8, an overall good numerical match of corresponding eigenvalues of α and β is needed to achieve a high AMS $S_{\alpha\beta}$ value.

To investigate the extent of *partial bioisosterism* between compounds 1124 (1125 gives same results as 1124) and 175, all possible AMS IPE $S_{\alpha\beta}$ measures were computed. Table 15 contains these $S_{\alpha\beta}$ values for the *ALL* and self, IPE(x,x), representations for compounds 175 and 1124. The highest cross-term, IPE(x,y), AMS $S_{\alpha\beta}$ measures (higher than 0.80) are also included in Table 15. Both compounds are "similar" from the point of view of the *ALL* (0.97), *NP* (0.83), and *HS* (0.93) $S_{\alpha\beta}$ IPE representations. Conversely, 1124 and 175 are not similar, or bioisosteric, with respect to the $S_{\alpha\beta}$ of the other self-IPEs. Pair combinations involving *ALL*, *ARO*, and *HS* IPEs also show up in the highest value cross-term $S_{\alpha\beta}$ measures. In general, compounds 1124 and 175 are quite similar in overall average conformational shape but of limited, or low, similarity with respect to other IPEs.

Table 14. AMS Eigenvalues for Compounds 1124, 1127, and 175 for the *ALL*- [IPE(0,0)] Representation[a]

1124	1127	175	1124	1127	175
0.347	0.347	0.341	0.008	0.008	0.008
0.146	0.147	0.145	0.008	0.008	0.008
0.075	0.075	0.080	0.007	0.007	0.008
0.044	0.045	0.043	0.007	0.007	0.007
0.043	0.043	0.041	0.007	0.007	0.007
0.032	0.032	0.035	0.007	0.007	0.007
0.028	0.029	0.032	0.006	0.006	0.006
0.025	0.025	0.025	0.006	0.006	0.006
0.020	0.020	0.020	0.005	0.005	0.006
0.016	0.016	0.016	0.005	0.005	0.005
0.015	0.015	0.014	0.005	0.005	0.005
0.013	0.013	0.014	0.005	0.005	0.005
0.012	0.012	0.012	0.005	0.005	0.005
0.011	0.011	0.011	0.004	0.004	0.004
0.011	0.011	0.011	0.004	0.004	0.004
0.010	0.010	0.010	0.004	0.004	0.004
0.009	0.009	0.010	0.004	0.004	0.004
0.009	0.009	0.009	0.004	0.004	0.004
0.008	0.008	0.008	0.004	0.004	0.004
0.008	0.008	0.008	0.004	0.004	0.004

[a] The eigenvalues are ordered from largest to smallest and listed side by side for comparison.

A good relationship between biological activity and *partial bioisosterism* should be realized only if the IPEs of $S_{\alpha\beta}$ having high values are responsible for the activity. For example, if in this case the biological activity depends on hydrogen bonding donor groups, IPE = *HBD* properties, compounds 175 and 1124 and/or 1127 cannot be considered as bioisosteric since the corresponding $S_{\alpha\beta} = 0$.

Partial bioisosterism (in particular with regard to IPE-(7,7) which is mainly related to "molecular shape") can be interpreted through the comparison of ensemble average spatial similarities, which is represented in AMS as an average common set of spatial distances between atom-pairs of given IPE types for a pair of molecules. If RMS is employed, then the identification of similar spatial occupancies of common three-dimensional pharmacophoric spatial regions, grid cells, serves as the representation of ensemble average similarity. If an ensemble average spatial comparison is done, a question that automatically arises is "Can these compounds be significantly overlaid?". The response to this question, based on the CEP sampling component of this methodology, should be "not necessarily". Moreover, since AMS has been used to compute the $S_{\alpha\beta}$ of Table 15, no alignments can be considered to establish the overlay of any single conformer of 175 and of 1124. However, the possibility of having a common set of spatial features exists, and, due to the fact that overlapping chemical structures is an established practice among chemists, the hybrid idea of

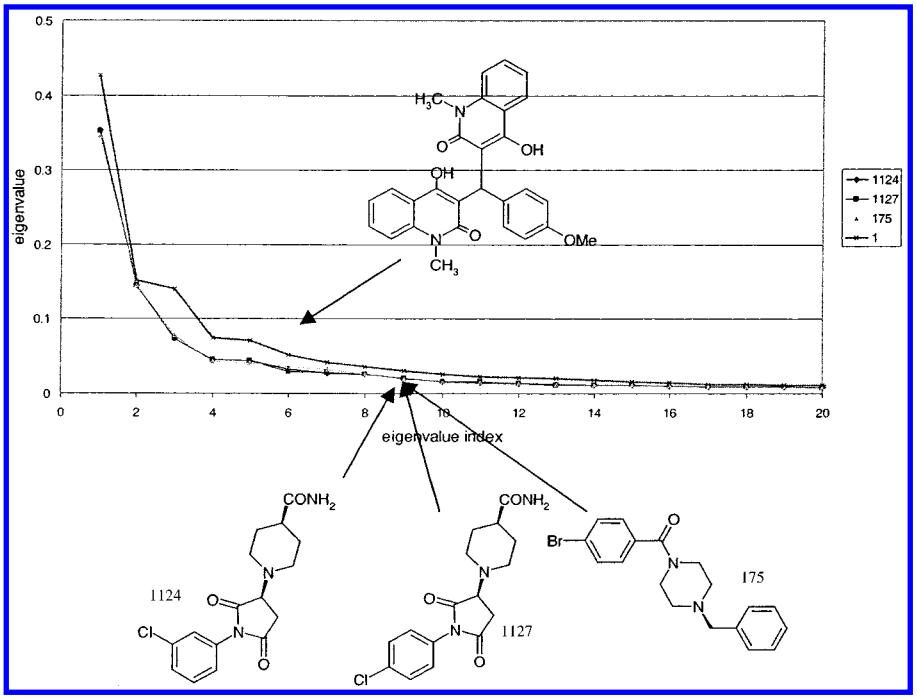


Figure 9. Plot of the AMS eigenvalues versus the corresponding eigenvalue index for largest to smallest eigenvalue ranking.

Table 15. AMS $S_{\alpha\beta}$ of Different IPE Representations for the Pair of Compound 175 and 1124[a]

symbol	numerical code	$S_{\alpha\beta}$
ALL	(0,0)	**0.970**
NP	(1,1)	0.834
P+	(2,2)	*0.000*
P−	(3,3)	*0.454*
HBA	(4,4)	*0.375*
HBD	(5,5)	*0.000*
ARO	(6,6)	*0.216*
HS	(7,7)	**0.926**
ALL/NP	(0,1)	0.843
ALL/HS	(0,7)	**0.942**
NP/HS	(1,7)	0.872

[a] See text for details. The AMS $S_{\alpha\beta}$ measures greater than 0.9 are in bold and those $S_{\alpha\beta}$ less than 0.5 are in italics.

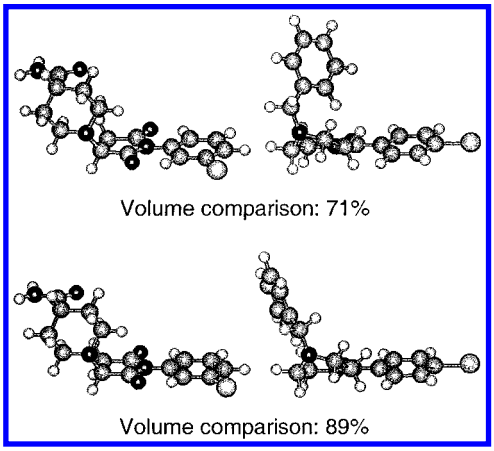


Figure 10. Molecular comparisons of the two "isosteric" compounds 175 (right) and 1124 (left), using common overlap volume as a measure of molecular similarity.

mixing AMS (alignment-independent) measures with a RMS (alignment-dependent) graphical analysis was also explored in this study.

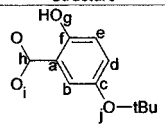
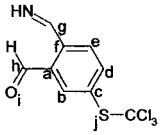
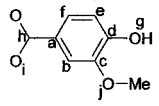
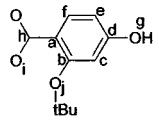
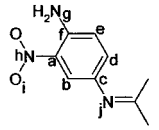
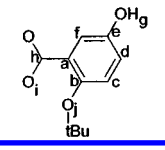
To overlay conformers of 175 and 1124, a defined alignment must be selected. Since there are multiple ways to explore and define alignments for a chemical training set[13a] and to also simplify the graphical spatial comparison, a three-point alignment was chosen based on the common halo-aromatic ring of the two molecules, which is illustrated in Figure 10. This simple alignment seems to work quite well. AM1 optimized structures of 1124 and 175 were overlayed, using the alignment rule, and their common overlap molecular volume was utilized as a measure of common spatial similarity. Seventy-nine percent of the molecular volume of 175 matches that of 1124. One of the torsional angles of compound 175 was changed to increase the common overlap molecular volume (Figure 10b). The energy of the conformer corresponding to the torsion angle alteration is about 0.2 kcal/mol less stable than the global minimum energy conformer (see Figure 10a). However, this higher energy conformer of 175 has a common overlap volume with 1124 of 89%. The corresponding *HS* AMS $S_{\alpha\beta}$ value is 0.93.

This molecular overlay exercise illustrates the inherent ambiguity involved in selecting a single conformation for each of a pair of molecules when making a 3D molecular similarity comparison. That is, how does one select the amount of conformational destabilization which is acceptable to achieve a given level of molecular similarity? This ambiguity is overcome, or at least canceled out, by performing the conformational ensemble averaging of 4D-QSAR molecular similarity analysis.

These molecular overlays demonstrate that common overlap volume can be correlated to a high *HS* AMS $S_{\alpha\beta}$ measure. This finding, in turn, suggests that AMS $S_{\alpha\beta}$ measures can be a highly correlated subset to the RMS (alignment-dependent) information. The resulting concept of using AMS $S_{\alpha\beta}$ measures as a guide to alignment selection will be developed in a future paper, where receptor-independent (RI) 4D-QSAR models, based on both AMS and RMS analyses are built and compared to one another.

D. A "Standard Library" for Investigating Methods Measuring Molecular Similarity. A library composed of

Table 16. "Standard Library" for Evaluating Molecular Similarity Methods and the Definitions of the Three Ordered Atom Alignments Used in the RMS Analysis

System	Structure	Alignment definitions
1a		1- a b c 2- a c e 3- f c j 4- h a d 5- i a b 6- c f g
1b		1- a b c 2- a c e 3- f c j 4- h a d 5- i a b 6- c f g
1c		1- a b c 2- a c e 3- f c j 4- h a d 5- i a b 6- a d g
1d		1- a b c 2- a c e 3- e b j 4- h a d 5- i a b 6- a d g
1e		1- a b c 2- a c e 3- f c j 4- h a d 5- i a b 6- f c g
1f		1- a b c 2- a c e 3- e b j 4- h a d 5- i a b 6- b e g

two small sets of compounds have been suggested as a standard library[23] to evaluate and compare different methods of computing and measuring molecular similarity. Both AMS and RMS analyses have been applied to this data set of compounds. The results of these analyses suggest that investigators have biased their ranking of molecular similarity over these compounds by *implicitly assuming* alignments.

The first series of compounds in the standard library is listed in Table 16. These six compounds were used to compare both the AMD and RMS methodologies, while the second set of 13 compounds, given in Table 17, was used to evaluate the effective dependence of *AMS* $S_{\alpha\beta}$ measures with respect to the different embedded pharmacophore groups composed of two hydroxyls at different substituent sites over the common tryciclic aromatic ring core structure.

Small organic molecules, all containing a common aromatic ring and different sets of substituents, compose the set of compounds in Table 16. Six different three ordered-atom alignments were defined for this series. The alignments were selected to span pharmacophoric space and are also given in Table 16. Alignments 1 and 2 are based only on atoms of the aromatic ring, whereas the other four alignments include two atoms on the common aromatic ring and one substituent atom. Both AMS and RMS analyses were performed on this set of compounds. *ALL* RMS $S_{\alpha\beta}$ values for the six compounds, for each of the six alignments, were computed and correlated to one another and to the *ALL* and *HS* AMS $S_{\alpha\beta}$ measures.

Table 18 is the correlation matrix of the $S_{\alpha\beta}$ measures for the six compounds of Table 16. From an inspection of this table a high correlation is noted for *ALL* AMS $S_{\alpha\beta}$ measures

Table 17. A Small Set of Positional Isomers Incorporating Relative Differences in the Intramolecular Distances between Hydroxy Substituents While Presenting a Constant Aromatic Core

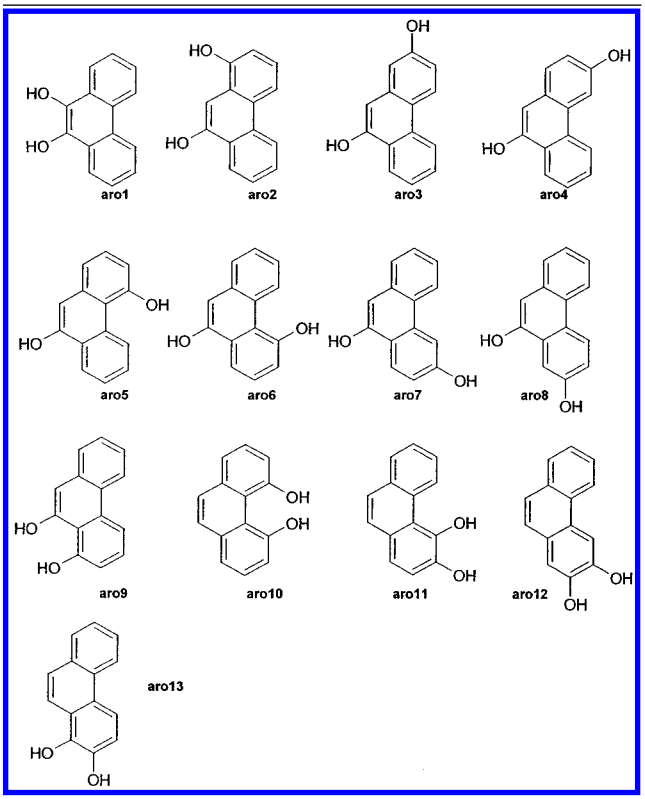


Table 18. Correlation Tables for *ALL* and *NP* RMS $S_{\alpha\beta}$ Values and *ALL* and *HS* AMS $S_{\alpha\beta}$ Measures[a]

	Al 1	*Al 2*	*Al 3*	*Al 4*	*Al 5*	*Al 6*	*AMS (0,0)*	*AMS (7,7)*
				RMS IPE(0,0) or (*ALL*, *ALL*)				
Al 1	1							
Al 2	**0.96**	1						
Al 3	*0.87*	*0.88*	1					
Al 4	−0.31	−0.36	−0.45	1				
Al 5	−0.15	−0.24	−0.31	0.83	1			
Al 6	0.19	0.22	−0.14	0.60	0.27	1		
AMS *ALL*	**0.96**	**0.98**	**0.92**	−0.39	−0.33	0.19	1	
AMS *HS*	0.72	0.66	0.32	0.02	0.25	0.45	0.56	1
				RMS IPE(1,1) or (*NP*, *NP*)				
Al 1	1							
Al 2	**0.96**	1						
Al 3	−0.11	−0.19	1					
Al 4	0.71	0.86	−0.26	1				
Al 5	−0.20	−0.15	0.62	0.05	1			
Al 6	0.19	0.20	0.22	0.36	−0.10	1		
AMS *ALL*	0.80	0.69	0.10	0.45	−0.34	0.63	1	
AMS *HS*	0.89	0.87	−0.13	0.66	0.10	−0.13	0.52	1

[a] The correlation coefficients are determined by correlating the *ALL and NP* RMS $S_{\alpha\beta}$ measures of alignment I, Al I, to the RMS $S_{\alpha\beta}$ measures of alignment J, Al J, or to the AMS $S_{\alpha\beta}$ measures. >0.90 correlation coefficients are in bold. >0.80, but <0.90, correlation coefficients are in italics.

and *ALL* RMS $S_{\alpha\beta}$ values for alignments 1, 2, and 3 but not alignments 4−6. A more modest correlation is found between the *HS* AMS $S_{\alpha\beta}$ values and the *HS* RMS $S_{\alpha\beta}$ for alignments 1 and 2, and poor correlations are found for alignments 3−6. If the alignment correlation analysis employs the nonpolar IPE, then only alignments 1 and 2 correlate well using the *HS* and *ALL* AMS $S_{\alpha\beta}$ values. The fact that only alignments

Table 19. AMS Analysis using Selected IPEs, *Atom Types*, for a Subset of Three Compounds Listed in Table 14[a]

atom	type	(0,0) *ALL*		atom	type	(6,6) *ARO*	
1a	1.000			**1a**	1.000		
1d	0.934	1.000		**1d**	1.000	1.000	
1f	*0.973*	**0.986**	1.000	**1f**	1.000	1.000	1.000
atom	type	(1,1) *NP*		atom	type	(7,7) *HS*	
1a	1.000			**1a**	1.000		
1d	*0.942*	1.000		**1d**	*0.945*	1.000	
1f	0.873	**0.976**	1.000	**1f**	0.919	**0.980**	1.000
atom	type	(3,3) *P*−		atom	type	(1,4) *NP,HBA*	
1a	1.000			**1a**	1.000		
1d	0.941	1.000		**1d**	0.713	1.000	
1f	*0.962*	**0.998**	1.000	**1f**	**0.935**	*0.896*	1.000
atom	type	(4,4) *HBA*		atom	type	(1,5) *NP,HBD*	
1a	1.000			**1a**	1.000		
1d	0.941	1.000		**1d**	0.869	1.000	
1f	*0.962*	**0.998**	1.000	**1f**	*0.957*	**0.974**	1.000
atom	type	(0,5) *ALL, HBD*		atom	type	(0,4) *ALL, HBA*	
1a	1.000			**1a**	1.000		
1d	0.890	1.000		**1d**	0.749	1.000	
1f	*0.959*	**0.982**	1.000	**1f**	**0.925**	*0.913*	1.000

[a] The highest AMS $S_{\alpha\beta}$ value of each (3 × 3) molecular similarity matrix is in bold while the second highest AMS $S_{\alpha\beta}$ value is in italics.

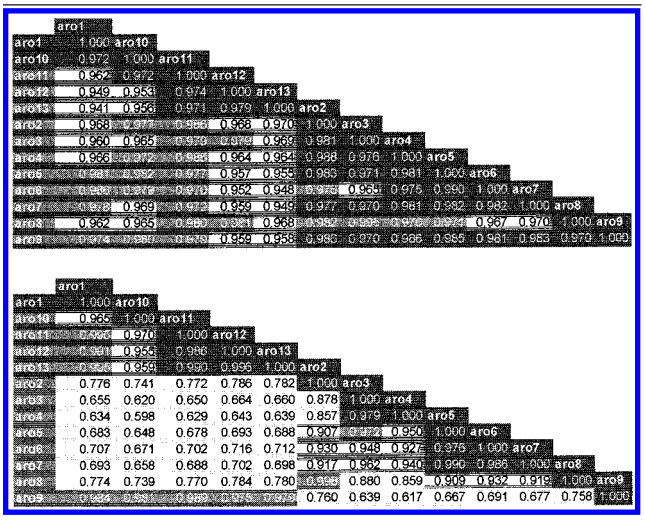
1 and 2 perform well for the *ALL* and *HS* AMS $S_{\alpha\beta}$ measures is not too surprising since both alignments are mainly based on atoms of the aromatic ring.

A general conclusion from this particular molecular similarity analysis is that the relative ordering of molecular similarity predicted by the *unbiased* AMS measures can be reproduced by the *alignment-dependent* RMS $S_{\alpha\beta}$ values but is quite dependent upon the alignment selected. That is, RMS $S_{\alpha\beta}$ measures can be highly dependent upon alignment but embed the information inherent to the AMS $S_{\alpha\beta}$ measures.

Compounds **1a**, **1d**, and **1f** (see Table 16) can be used to exemplify some *alignment-based* preconceptions most chemists have about chemical similarity. These three compounds differ only in the relative positions of their substituents (carboxylate, hydroxy, and *tert*-butyloxy groups). Chemists most often group these in pairs such as (**1a**/**1f**) which share the same groups at the *para*-position and (**1d**/**1f**) which share the same *ortho*-substitution. These two pairs of grouped compounds are selected based on the adoption of *implicit* alignments. Chemists *assume* alignments when comparing molecules. Hence, it is interesting and informative to utilize AMS (alignment independent) to determine which of the possible pairs of compounds in Table 16 have the highest molecular similarity for different choices of IPEs.

Table 19 contains *ALL*, *self*, and *cross-term* representations for selected IPEs using AMS analysis for pairs of the three compounds **1a**, **1d**, and **1f** of Table 16. Each of these three compounds has the same substituents, but the (OH, O-tBu) pair is located at different combinations of positions on the same ring of each compound. The compound pair (**1d**/**1f**) has the largest AMS $S_{\alpha\beta}$ values for IPEs (0,0), (1,1), (3,3), (4,4), (7,7), (0,5), and (1,5) but not for IPEs (0,4) and (1,4). The AMS $S_{\alpha\beta}$ ranked order most often found is (**1d**/**1f**) > (**1a**/**1f**) > (**1a**/**1d**) for IPEs (0,0), (3,3), (4,4), (0,5), and (1,5) followed by (**1a**/**1f**) > (**1d**/**1f**) > (**1a**/**1d**) for IPEs (0,4) and (1,4) and (**1d**/**1f**) > (**1a**/**1d**) > (**1a**/**1f**) for IPEs (1,1) and (7,7). Surprisingly, at least from the point of view of "chemical intuition", nonpolar[1] and hydrogen-suppressed[7] IPEs predict that (**1a**/**1d**) are more similar to one another than (**1a**/**1f**) by 4−7%.

Table 20. AMS Analysis for the Set of Compounds Given in Table 17[a]



[a] The upper molecular similarity matrix presents the *ALL* AMS $S_{\alpha\beta}$ values, while the lower matrix contains the *P*-AMS $S_{\alpha\beta}$ measures which correspond to the oxygens of the OH groups. Both matrices are gray scaled based on the AMS $S_{\alpha\beta}$ values.

The IPE types that allow the best resolution (the largest composite differences among the $S_{\alpha\beta}$ values) for the three pairs of compounds are

(0,5), or (1,5), for (**1d**/**1f**) > (**1a**/**1f**) [0.96] > (**1a**/**1d**)

(1,1) for (**1a**/**1f**) > (**1d**/**1f**) > (**1a**/**1d**) and

(0,4) or (1,4) for (**1d**/**1f**) > (**1a**/**1d**) > (**1a**/**1f**)

In contrast, least similarity resolutions are realized for IPE types (0,0), (3,3), (4,4), and (7,7).

In summary, AMS $S_{\alpha\beta}$ measures predict the molecular similarity order (**1d**/**1f**) > (**1a**/**1f**) > (**1a**/**1d**) as most frequent over the IPEs. This relative ranking of molecular similarity for several IPEs is generally consistent with chemical intuition, even though it would not be easy to reach consensus among chemists to select (**1d**/**1f**) over (**1a**/**1f**) in terms of molecular similarity.

The second part of this particular study involves the compounds listed in Table 17. These compounds have a common aromatic core and a constant pair of OH substituents whose locations on the ring system as well as their intramolecular distance varies over the compounds. The goal of this particular analysis is to demonstrate the robustness of AMS $S_{\alpha\beta}$ measures to differentiate among pharmacophoric groups. The two hydroxyl groups are assumed to be members of the pharmacophore.

The *ALL* and *P*- AMS $S_{\alpha\beta}$ measures for the compounds in Table 17 are presented in Table 20. It can be seen that the *ALL* and *P*- AMS $S_{\alpha\beta}$ matrices in Table 20 are very different from one another. The IPE(0,0), *ALL*, predicts AMS

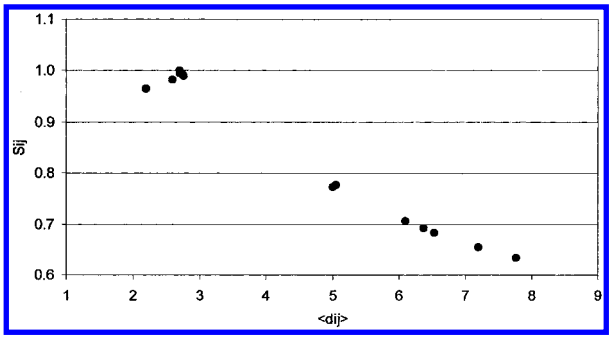


Figure 11. Dependence of the IPE(3,3) AMS $S_{\alpha\beta}$ measures versus the intramolecular distance of the IPE(3,3) "pharmacophore" groups, the two hydroxyl groups.

$S_{\alpha\beta}$ values greater than 0.95 for all the compounds. In contrast, the polar negative (*P*-) representation, IPE(3,3), only includes the hydroxyl groups and yields corresponding AMS $S_{\alpha\beta}$ measures in the range of 0.60−1. The near equal *ALL* AMS $S_{\alpha\beta}$ values is due, to a great extent, to the presence of the same aromatic ring system across the set of compounds. Still, the *ALL* AMS $S_{\alpha\beta}$ measures can differentiate between compounds such as **aro1** and **aro10** (AMS $S_{\alpha\beta} = 0.97$) and compounds **aro11**, **aro12**, and **aro13**. In fact, these three latter analogues are better distinguished by the *ALL* AMS $S_{\alpha\beta}$ measure than for the IPE(3,3) representation due to the constant distance between the two OH groups across these three analogues.

The larger differentiation of molecular similarity from *P*-AMS $S_{\alpha\beta}$ measures, as compared to the *ALL* IPE representation, is due to the different substituent locations of the OH groups on the tricyclic ring system. Taking compound **aro1** as a reference, it is seen that the IPE(3,3) AMS $S_{\alpha\beta}$ values vary with the distance $<d_{ij}>$ (defined as the Boltzmann average intramolecular distance between OH groups, see Figure 11), as is expected from the definition of a molecular similarity matrix element given by eq 1.

This study indicates that the search for patterns in AMS $S_{\alpha\beta}$ values across molecular similarity matrices of different IPE pairs is equivalent to searching for target compounds having common pharmacophores as specified by the IPEs. The molecular similarity matrix representation provides a visual tool to rapidly inspect and compare large amounts of molecular structure data.

DISCUSSION

The paradigm for estimating molecular similarity based on the foundations of 4D-QSAR analysis introduces four intuitive, but by and large neglected, components into a molecular similarity measure. First is the concept of relative (or alignment-dependent) and absolute molecular similarities. Chemists use a reference (alignment) when comparing (similarity analysis) molecules. This reference, or alignment, can include features (physical properties and molecular descriptors) used in the large majority of similarity methods currently employed. The similarity measure depends on the choice of alignment. *The 4D-QSAR method of molecular similarity analysis permits a "pure" absolute molecular similarity measure, independent of any alignment bias, to be made*.

Second, different degrees of molecular similarity are inherent to a pair of molecules depending on the atom types selected to establish the similarity estimate. Chemists nearly always base their estimate of molecular similarity on the structure of the whole molecule. 4D-QSAR molecular similarity permits estimation of molecular similarity measures on the basis of atom-types composing the molecules. Thus, *the 4D-QSAR method of molecular similarity intrinsically allows multiple measures of molecular similarity to be made based on atom type and pharmacophore.*

Third, 4D-QSAR molecular similarity analysis not only considers a 3D molecular structure but also the complete conformational ensemble of states of a molecule in estimating molecular similarity. Indeed, consideration of only a single conformational state of a molecule in making a molecular similarity measure can cause more harm than good. One need only consider the all trans conformation of an *n*-alkane as compared to the same *n*-alkane with a single gauche conformation replacing a trans state in the "middle" of the molecule. Very different spatial shapes result for these two conformations of the same molecule. Conformational sampling is required to overcome the complications inherent to the 3D comparisons of flexible molecules.

Finally, the eigenvalues of the AMS matrix of a molecule are *fundamental 3D properties* of the molecule which embed conformational sampling and can be employed in any type of QSAR/QSPR analysis. This extended use of the fundamental properties (eigenvalues) of 4D-QSAR molecular similarity analysis will be demonstrated in a forthcoming paper.

The complete set of $S_{\alpha\beta}$ across all pairs of IPEs, for either AMS or RMS, can be used to define a molecular diversity space of a library. The AMS $S_{\alpha\beta}$ across all pairs of IPEs corresponds to the most general molecular diversity space. The RMS $S_{\alpha\beta}$ for all pairs of IPEs defines a molecular diversity space consistent with, and constrained to, the selected common alignment.

ACKNOWLEDGMENT

This work was supported, in part, from resources of the Laboratory of Molecular Modeling and Design at the University of Illinois at Chicago. We appreciate the help from, and useful discussions with, Greg Makara and Edward Wintner, of NeoGenesis Drug Discovery and Dan Pernich, James Ruiz and Debby Camper of Dow AgroSciences, over the course of this study. We also acknowledge financial support from Dow AgroSciences, NeoGenesis Drug Discovery, and The Chem21 Group, Inc.

REFERENCES AND NOTES

(1) Nalewajski, R. F.; Parr, R. G. Information theory, atoms in molecules and molecular similaritypp. *Proc. Natl. Acad. Sci. U.S.A. − Paper Edition* **2000**, *97(16)*, 8879−9225.

(2) Pollack, N.; Cunningham, A. R.; Klopman, G.; Rosenkranz, H. S. Chemical Diversity Approach for Evaluating Mechanistic Relatedness among Toxicological Phenomena. *SAR QSAR Environ. Res.* **1999**, *10-(6)*, 533−544.

(3) Xu, J. Chemical Diversity Exploration and Combinatorial Chemistry in Drug Discovery. *Prog. Chem. −Beijing* **1999**, *11(3)*, 286−299.

(4) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D. Quasi-orthogonal basis sets of molecular graph descriptors as a chemical diversity measure. *J. Chem. Inf. Computer Sci.* **2000**, *40(1)*, 126−134.

(5) Popelier, L. A. Quantum Molecular Similarity. 1. BCP Space. *J. Phys. Chem. A* **1999**, *103(15)*, 2883−2890.

(6) Willett, P. Chemoinformatics − similarity and diversity in chemical libraries. *Curr. Opin. Biotech.* **2000**, *11(1)*, 85−88.

(7) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38(6)*, 983−996 (see page 986).
(8) Szarek, S. J.; Voiculescu, D. Volumes of restricted minkowski sums and the free analogue of the entropy power inequality. *Comm. Math. Phys.* **1996**, *178(3)*, 563−570.
(9) Gabarro-Arpa, J.; Revilla, R. Clustering of a molecular dynamics trajectory with a hamming distance. *Computers Chem*. **2000**, *24(6)*, 693−698.
(10) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New perspectives in lead generation II: evaluating molecular diversity. *Drug Discovery Today* **1996**, *1*, 71−78.
(11) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using MDL keys as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(12) Miller M. M.; Sheridan, R. F.; Kearsley, S. K. SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505−1514.
(13) (a) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223. (b) Mason, J.D.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. L. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(14) (a) Silverman, B. D.; Platt, D. E. Comparative molecular field moment analysis (CoMMA). *J. Med. Chem.* **1996**, *39*, 2129−2140. (b) Silverman, B. D. Three-dimensional moments of molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1470−1476.
(15) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caufield, T. J.; Kowala, M. C.; Valentine, M. T.; Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem*. **1999**, *42*, 3919−3933.
(16) (a) Hopfinger, A. J.; Duca, J. S. Extraction of pharmacophore information from high-throughput screens. *Curr. Opin. Biotech.* **2000**, *11(1)*, 97−103. (b) *Pharmacophore Perception, Development and Use in Drug Design* Guner O. F., Ed.; International University Line: La Jolla, CA, 2000.
(17) Doherty, D. C. MOLSIM User's Guide; The Chem21 Group, Inc.: 1780 Wilson Dr., Lake forest, IL, 1997.
(18) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(19) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; deAlencastro, R. B. Four-dimensional quantitative structure−activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.
(20) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding free energy by 4D-QSAR analysis: application to a set of glucose inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150.
(21) Hopfinger, A. J.; Reaka, A. Venkatarangan, P.; Duca, J. S.; Wang S. Construction of a Virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase *b. J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.
(22) Press, W. H.; Teukolsky, S. A.; Vettering, W. T.; Flannery, B. P. *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: London, 1997; pp 463.
(23) (a) Phelan, J. C. Libraries to leads: which molecules should we make? *Cheminformatics Symp*. **1998**, June 15−16, Boston, MA. (b) Mount, J.; Ruppert, J.; Welsh, W.; Jain, A. N.; Icepick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60−66.

CI0100090