Foundation of Quantum Similarity Measures and Their Relationship to QSPR: Density Function Structure, Approximations, and Application Examples

RAMON CARBÓ-DORCA, XAVIER GIRONÉS

Institute of Computational Chemistry, University of Girona, Girona 17071, Catalonia, Spain

Received 6 September 2002; accepted 26 March 2004
Published online 13 August 2004 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/qua.20191

ABSTRACT: This work presents a schematic description of the theoretical foundations of quantum similarity measures and the varied usefulness of the enveloping mathematical structure. The study starts with the definition of tagged sets, continuing with inward matrix products, matrix signatures, and vector semispaces. From there, the construction and structure of quantum density functions become clear and facilitate entry into the description of quantum object sets, as well as into the construction of atomic shell approximations (ASA). An application of the ASA is presented, consisting of the density surfaces of a protein structure. Based on this previous background, quantum similarity measures are naturally constructed, and similarity matrices, composed of all the quantum similarity measures on a quantum object set, along with the quantum mechanical concept of expectation value of an operator, allow the setup of a fundamental quantitative structure–activity relationship (QSPR) equation based on quantum descriptors. An application example is presented based on the inhibition of photosynthesis produced by some naphthyridinone derivatives, which makes them good herbicide candidates. Int J Quantum Chem 101: 8–20, 2005

Key words: tagged sets; inward matrix product; matrix signature; vector semispaces; σ-shells; convex conditions; generating rules; density functions; quantum objects; quantum object sets; molecular quantum similarity measures; similarity matrices; discrete quantum object sets; QSPR

Correspondence to: R. Carbó-Dorca; e-mail: quantum@iqc.udg.es
Contract grant sponsor: Foundation M. F. de Roviralta.
Contract grant sponsor: Centro de Investigación Cientifica y Tecnológica (CICYT).
Contract grant number: SAF2000-223.

Introduction

During the past 5 years, the theoretical structure of quantum similarity, the study of which started at the beginning of the 1980s [1] has been fully developed. Many possible applications have been provided in the form of several quantitative structure–activity relationships (QSPR), associated with a broad scope of subjects, ranging from biological responses up to molecular reactivity and toxicity, a well as physical properties of molecular, atomic, or nuclear systems (see, e.g., Ref. [2] and the examples therein). The present study gives some new insights into the problem within a brief summary of the mathematical basis, leading, through density function structure and quantum similarity, toward QSPR based on quantum descriptors. To prove further the useful aspects of the theory that have been developed thus far; some applications, concerning density functions and molecular QSPR, are also provided.

Tagged Sets

A given set, the object set, **S**, and another set, made of mathematical elements, hereafter called tags, forming a tag set, **T**, then a tagged set, **Z**, can be constructed [3] using the ordered product:

$$Z = S \times T.$$

Thus:

$$Z = \{\forall \theta \in Z \rightarrow \exists s \in S \wedge \exists t \in T : \theta = (s, t)\}.$$

Such a construction permits a wide variety of applications, including the well-known fuzzy sets [4] as a particular case.

Inward Matrix Products, IMP Inverses, and Matrix Signatures

An inward matrix product (IMP) [5], corresponds to a well-defined matrix operation, which can be easily programmed in any high-level language, such as FORTRAN 90. Considering admitted the matrix addition as can be defined in the usual way, the matrices can then behave almost as a set of scalars. Such an operation as IMP corresponds to a multiplicative internal composition law applicable to matrix or hypermatrix spaces of arbitrary dimensions, producing a new matrix belonging to the same space; that is,

$$\forall \mathbf{A}, \mathbf{B} \in M_{(m\times n)} : \mathbf{P} = \mathbf{A} * \mathbf{B} \rightarrow \mathbf{P} \in M_{(m\times n)}$$

and this is fulfilled whenever the following algorithm applies:

$$\mathbf{P} = \{p_{ij}\}, \quad \mathbf{A} = \{a_{ij}\}, \quad \mathbf{B} = \{b_{ij}\} \rightarrow \forall i, j : p_{ij} = a_{ij}b_{ij}.$$

Defined in such a way, provided that the involved matrix elements are scalars obtained from a field, IMP is associative, distributive with respect to the matrix addition, and commutative.

If they exist, IMP inverses are defined in the following terms:

$$\mathbf{A} = \{a_{ij}\} \in M_{(m\times n)} \rightarrow \exists \mathbf{A}^{[-1]} = \{a_{ij}^{-1}\} \in M_{(m\times n)}$$

such that

$$\mathbf{A}^{[-1]}\mathbf{A} = \mathbf{A}\mathbf{A}^{[-1]} = \mathbf{1} \in M_{(m\times n)} \wedge \mathbf{1} = \{1_{ij} = 1; \forall i, j\}.$$

Matrix **1**, the unity matrix, is constructed by the multiplicative unit of the reference field. The unity matrix is the neutral element of the IMP. That is

$$\forall \mathbf{A} : \mathbf{A} * \mathbf{1} = \mathbf{1} * \mathbf{A} = \mathbf{A}.$$

An IMP regular matrix does not possess zero elements, although pseudo-inverse extensions can be also defined. IMP powers are constructed by the rule:

$$\mathbf{A} * \mathbf{A} * \ldots \mathbf{A} = \mathbf{A}^{[p]} = \{a_{ij}^{p}\}.$$

MATRIX SIGNATURE

The concept of matrix signature appears to be interesting because it is an obvious example of a tagged set. Signatures can be defined by the following prescription over $(m \times n)$ dimensional matrices, constructed in principle over the real field **R**, and it is extended straightforwardly over any matrix or hypermatrix of arbitrary dimensions:

$$\mathbf{S} = Sign(\mathbf{A}) = \{s_{ij} = sign(a_{ij})\} \rightarrow s_{ij} \in \{+1, -1\} \equiv \{1, 0\}.$$

In this manner, as one can write using IMP,

$$\mathbf{A} = Sign(\mathbf{A}) * |\mathbf{A}| \wedge |\mathbf{A}| = \{|a_{ij}|\},$$

where it can be deduced that there are only $2^{m\times n}$ types of matrix signatures of such a dimension. So, taking the set of $(m \times n)$ matrices defined over the positive definite real field, $\mathbf{R}^+$, as objects belonging to an object set made, in turn, by a matrix vector space of the same dimension, a Boolean tagged set can be easily constructed with the aid of the $2^{m\times n}$ matrix signatures, associated with an appropriate tag set. The term "Boolean tagged set" can be applied to these tagged sets with the tag set made by bit strings. Besides this interesting construct, matrix signatures generalize the concept of sign. The usual sign of real numbers taken as scalars, thus as one-dimensional objects, must have just two different elements, while (2×2) matrices, say, necessarily must have $2^4 = 16$ signatures. Signatures may be used alternatively to define matrix classes within a matrix vector space. Moreover, matrix signatures written with their elements as binary digits are just isomorphic to the bit representation of the sequence of natural numbers up to $2^{m\times n}$.

Vector Semispaces and σ-Shells

The separation of signatures from matrices, creating a new point of view of them as members of a tagged set, also permits to consider the set of objects, formed by matrices with the same dimension, but made of positive definite real elements. These new objects are creating a new kind of sets, which can be called vector semispaces [6]. A matrix object set like the one belonging to the unity matrix signature, can be viewed as a vector space where the additive group has been chosen as a semigroup. A semigroup is an additive group without reciprocal elements, permitting no differences or negative signs. Thus, using a matrix space, one can form an associated matrix semispace by means of the rule:

$$\forall \mathbf{A} \in M_{(m\times n)}(\mathbf{R}) \Rightarrow \exists |\mathbf{A}| \in M_{(m\times n)}(\mathbf{R}^+).$$

NORMS AND SHELLS

In a given matrix semispace, the most suitable norm to be defined for all semispace elements, corresponds to a Minkowski norm, coincident with the sum of all the elements of the chosen matrix:

$$\forall \mathbf{A} \in M_{(m\times n)}(\mathbf{R}^+) : \langle \mathbf{A} \rangle = \sum_i \sum_j a_{ij} \in \mathbf{R}^+.$$

The sum of all matrices that possess the same Minkowski norm σ, constitutes a σ-shell:

$$\forall \mathbf{A} \in S(\sigma) \rightarrow \langle \mathbf{A} \rangle = \sigma.$$

The unit shell or 1-shell is readily defined by means of:

$$\forall \mathbf{A} \in S(1) \rightarrow \langle \mathbf{A} \rangle = 1.$$

This interesting form of partitioning matrix semispaces can be easily extended to other possible semispaces. For example, it can be defined in functional semispaces, made of positive definite and integrable functions. Also, upon multiplying the matrices of any σ-shell by the scalar factor σ^{-1}, this operation transforms any shell element into one belonging to the unit shell, that is:

$$\forall \mathbf{A} \in S(\sigma) \rightarrow \sigma^{-1}\mathbf{A} \in S(1),$$

and conversely:

$$\forall \mathbf{Z} \in S(1) \rightarrow \sigma\mathbf{Z} \in S(\sigma).$$

From here one can infer that the unit shell potentially contains all the elements of the matrix semispace. Alternatively, one can say that from the 1-shell elements, all the semispace elements can be generated. A semispace can thus be viewed as a tagged set, where the objects are the elements belonging to the 1-shell, and the tags could be formed by the elements of $\mathbf{R}^+$.

Norms in Vector Semispaces and Convex Conditions

As was noted earlier, the concept of semispace and σ-shell can be extended into other kinds of objects, like functions, provided that one can describe the following statements, associated with these arbitrarily valued, but integrable, positive definite function sets:

$$\forall \rho(\mathbf{r}) \in V_\infty(\mathbf{R}^+) \rightarrow \langle \rho \rangle = \int_D \rho(\mathbf{r}) d\mathbf{r} \in \mathbf{R}^+.$$

The integral above has the same role of the matrix elements sum or Minkowski norm in matrix semispaces, and in this manner can generate shells as well:

$$iff \quad \langle\theta\rangle = \sigma \rightarrow \theta(\mathbf{r}) \in S(\sigma) \subset V_\infty(\mathbf{R}^+).$$

Such general properties, which exhibit some mathematical objects, have lead to the definition of general properties, which can be summarized into a unique symbol. The discrete convex conditions [7] symbolized by $K_n(\mathbf{w})$, and applied to a known column vector $\mathbf{w}$, belonging to some n-dimensional semispace, have been defined as:

$$K_n(\mathbf{w}) \equiv \left\{ \mathbf{w} \in V_n(\mathbf{R}^+) \wedge \langle\mathbf{w}\rangle = \sum_i w_i = 1 \right\}.$$

Also, continuous convex conditions can be set up using a similar symbol, and applied to functions belonging to some Hilbert semispace:

$$K_\infty(\rho) \equiv \left\{ \rho \in H_\infty(\mathbf{R}^+) \wedge \int_D \rho(\mathbf{r}) d\mathbf{r} = 1 \right\}.$$

As a result of these considerations, convex conditions resume the fact that the object, described within the convex conditions symbol, possess a positive definite structure and belongs to the 1-shell of some known semispace.

Linear Combinations in Vector Semispaces

Linear combinations in vector semispaces have properties of their own, which have important consequences in the construction of density functions. Suppose that one can describe a set of coefficients associated with some convex conditions:

$$K(\{w_i\}) = \left\{ \forall i : w_i \in \mathbf{R}^+ \wedge \sum_i w_i = 1 \right\}.$$

Convex linear combinations of elements belonging to some σ-shell of an arbitrary semispace remain within the σ-shell, as the following reasoning proves:

$$\{\mathbf{A}_i\} \subset S(\sigma) \wedge \mathbf{Z} = \sum_i w_i \mathbf{A}_i \rightarrow \langle\mathbf{Z}\rangle = \sum_i w_i \langle\mathbf{A}_i\rangle$$
$$= \sigma \sum_i w_i \rightarrow \mathbf{Z} \in S(\sigma) \quad iff \quad K(\{w_i\}).$$

Thus, convex linear combinations of 1-shell elements of a given vector semispace can be used to construct all the elements of the semispace. The following property, deducible from this result, has interest in applied quantum mechanics: any convex linear combination of pth order density functions, belonging to some shell, produces again a pth order density function belonging to the same shell.

Quantum Density Functions and Quantum Objects Sets

The origin of quantum mechanical density functions is now well known. The first description of the possible sense of the squared module of a wave function appears to be attributed to Born [8]; however, in an early study, Schrödinger [9] also referred to a density equation as a first step, from which one can variationally deduce the equation that bears his name, as this variational final form the source of state energies and wave functions pairs. A generating rule [7] can thus be set, once a given system state wave function, belonging to some Hilbert space, is known:

$$R(\Psi \rightarrow \rho) = \{\forall \Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow$$
$$\exists \rho(\mathbf{r}) = |\Psi|^2 \in H_\infty(\mathbf{R}^+)\}.$$

Thus, quantum mechanical density functions are elements of a Hilbert semispace, which in any case one shall consider definite positive and normalized in the usual sense, that is, submitted to the convexity conditions: $K_\infty(\rho)$. However, suppose a homogeneous set of such quantum mechanical density functions is known, $P = \{\rho_I(\mathbf{r})\} \subseteq K_\infty(\mathbf{R}^+)$, using a convex set of coefficients: $K(\{w_I\})$. A new density function can be obtained of the same characteristics as those contained in the set P:

$$\rho(\mathbf{r}) = \sum_I w_I \rho_I(\mathbf{r}) \rightarrow K_\infty(\rho)$$

simply because of the definite positive nature of both coefficients and density function basis set, as well as

$$\int_D \rho(\mathbf{r})d\mathbf{r} = \sum_I w_I \int_D \rho_I(\mathbf{r})d\mathbf{r} = \sum_I w_I = 1.$$

A finite dimensional generating rule can be invoked to construct sets of convex coefficients. It is only necessary to construct the symbol:

$$R_n(\mathbf{x} \rightarrow \mathbf{w}) = \{\forall \mathbf{x} \in V_n(\mathbf{C}) \rightarrow \exists \mathbf{w} = \mathbf{x}^* * \mathbf{x} \in V_n(\mathbf{R}^+)\},$$

where the IMP has been employed to construct the elements of the finite dimensional semispace. This means that any convex set of coefficients $K_n(\{w_I\})$ can be generated by the simple rule:

$$\forall I : w_I = |x_I|^2 \wedge \sum_I w_I = 1 \rightarrow \sum_I |x_I|^2 = 1,$$

which, when collected into vectors, obeys the norm properties:

$$\langle \mathbf{w} \rangle = 1 \rightarrow \langle \mathbf{x}^* * \mathbf{x} \rangle = \mathbf{x}^+\mathbf{x} = \langle \mathbf{x}|\mathbf{x} \rangle = 1.$$

This can be associated with the fact that, while Minkowski norms are adequate to semispaces, Euclidean norms have to be applied to the generating complex space. Thus, a set of elements of the 1-shell of a given n-dimensional vector semispace corresponds to a set of discrete probability functions. This can be generated, in turn, by a set of n-dimensional complex vectors, in the same manner as the set of quantum mechanical density functions that represent continuous probability distributions, when constructed by the squared modules of Hilbert space wave functions. In fact, and as a consequence of the previous properties and definitions, any vector space of arbitrary dimension can be employed as the generating source of vector semispace elements. Moreover, semispace subset pairs, made up of finite and infinite dimensional elements, can be used together to construct new elements, with the same characteristics as those associated with the building blocks themselves.

QUANTUM OBJECT SETS

Once a quantum mechanical probability density distribution, $\rho \in P$, is known for an arbitrary state of a given submicroscopic system, $s \in S$, the pair, $\theta = \{s; \rho\}$, constitutes an element of a tagged set, where the objects are the submicroscopic systems and the tags are made of quantum density functions. These tagged set elements can be called quantum objects and the corresponding tagged sets quantum object sets. It can accordingly be written: $\Theta = S \times P$.

In this context, quantum object sets can be made of submicroscopic systems and some built-in density function set, which does not necessarily have to be exact, but can even be approximately constructed by means of an appropriate basis set of density functions and a convex set of coefficients. This is the case for atoms, say, where an atomic shell approximation (ASA) [10] can be envisaged, owing to the atomic spherical symmetry. For instance, suppose that we know the density function for a given atom, $\rho_A(\mathbf{r})$. Suppose also that we know a set of spherical functions belonging to some functional semispace, $\Sigma = \{\sigma_I(\mathbf{r})\} \subseteq F_\infty(\mathbf{R}^+)$, and a set of convex coefficients: $K(\{w_I\})$. It is straightforward to see that the following construction holds:

$$\rho_A(\mathbf{r}) \approx \rho_A^a(\mathbf{r}) = \sum_I w_I \sigma_I(\mathbf{r}) \in F_\infty(\mathbf{R}^+).$$

Both coefficients and nonlinear parameters, possible embedded in the basis density functions, can be optimized to fit the original atomic density. A description of a correct fitting procedure can be found in several papers [10]. In the same way, one can imagine interating the above fitting procedure to obtain approximate molecular density functions for, once a set of approximate atomic functions is known, i.e., $P = \{\rho_A^a\}$, one can also choose a set of convex coefficients: $K(\{\omega_A\})$ to represent a molecular density function: ρ_M. In this case,

$$\rho_M(\mathbf{r}) \approx \rho_M^a(\mathbf{r}) = \sum_{A \in M} \omega_A \rho_A^a(\mathbf{r} - \mathbf{R}_A),$$

where the sum runs over the atoms of the molecule. Such approximations have been employed in crystallography, where an approximate molecular density function is constructed in this way. However, the coefficients of the atomic densities are chosen as the atomic charges, producing the so-called "promolecular density."

ASA DENSITY SURFACES

Promolecular densities represent an outstanding tool to generate approximate DF. Using this approach, several calculations can benefit from

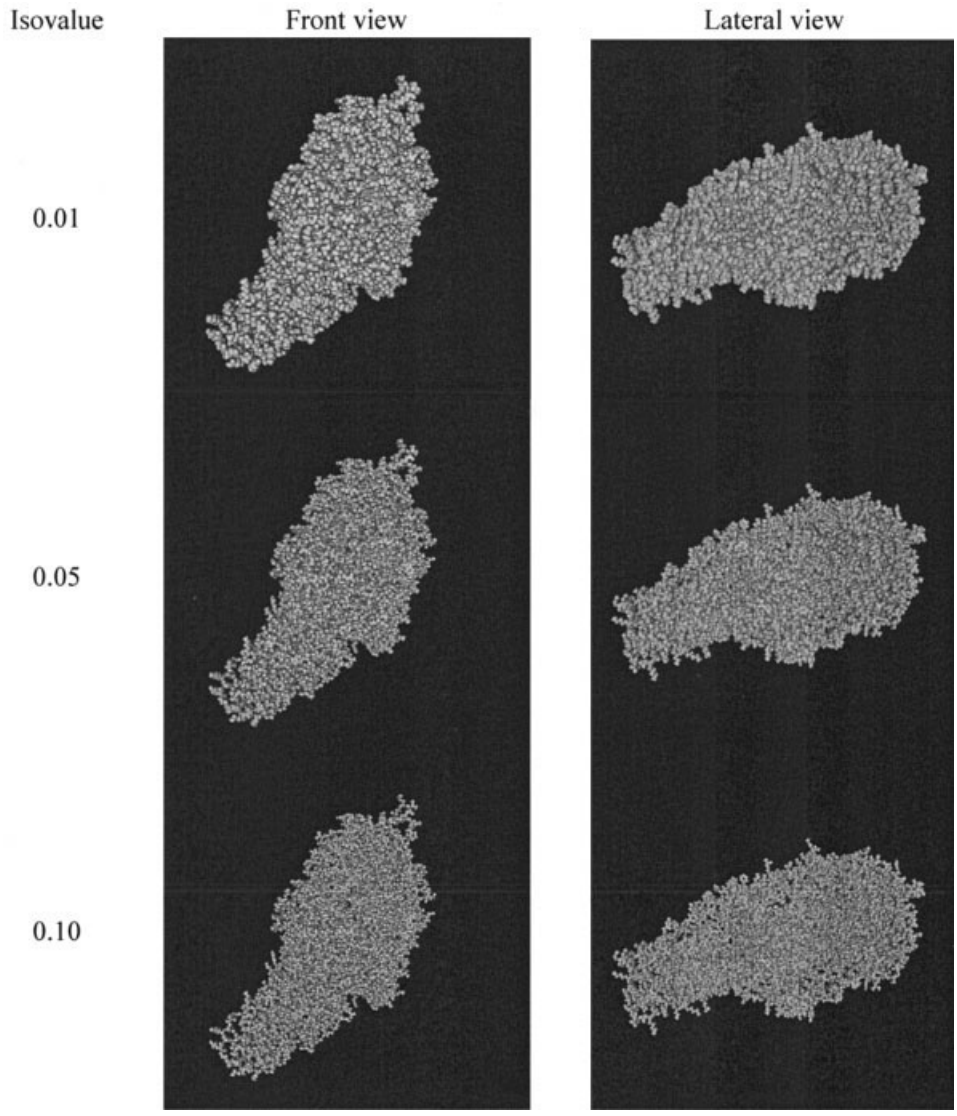

FIGURE 1. Plots of 1PRC isodensity contours at different levels.

considerably increased speed, obtaining results very close to those from regular electronic structure computations, as in Hartree–Fock procedures (see, for example, Ref. [11]). As an illustrative example of the computational performance of promolecular densities, isodensity plots of the photosynthetic reaction center of *Rhodopseudomonas viridis* are presented in Figure 1. This protein is composed of 20352 atoms, after including the H atoms, which were missing in the experimental X-ray structure [12]. The molecular coordinates were retrieved from the Protein Data Bank (PDB Id: 1PRC) (for more details, see http://www.rcsb.org/pdb/). The corresponding isodensity contours were generated using fitting parameters to the 6-311G basis set [10].

As can be seen in Figure 1, at low isodensity levels, the outer electronic density distribution can easily be plotted. Decreasing the isovalue points, the bonding patterns become present along the structure. Finally, at high density isocontours, the surface collapses the density around atomic locations. Although present-day computer speed does not permit comparison of the contours presented above with those generated from ab initio densities, some comparisons

between smaller molecules were previously published [13], providing very similar results. However, it must be said that those surfaces calculated from promolecular densities required only a small fraction of computation time.

Quantum Similarity Measures, Similarity Matrices, and Discrete Quantum Object Sets

Owing to the previous definitions and properties of vector semispaces, it should not be difficult to find a natural way to compare elements of quantum object sets to obtain information about the degree of similarity or dissimilarity among two or more of them. Such a procedure may be used in chemical structures in countless ways and applications. In addition, if the computational algorithm appears to be unconstrained by the nature of the submicroscopic objects studied, the comparison technique can be applied to nuclei, atoms, or molecules, without changes other than those related to the nature of the respective density functions.

A quantum similarity measure (QSM) [14] can be associated with a function, which transforms a pair or more of quantum object tags into a positive real number. This is the same as starting with a quantum object set, $\Theta = S \times P$, and finding a function such as

$$Z : P \times P \times \ldots P \rightarrow \mathbf{R}^{+}.$$

A simple way to express this possible transformation, involving two density functions, in the most usual form, is expressed as an integral:

$$Z_{AB}(\Omega) = \int\int \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1, \mathbf{r}_2)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2,$$

where $\{\rho_A; \rho_B\}$ are the density function tags of quantum objects A and B, while $\Omega(\mathbf{r}_1; \mathbf{r}_2)$ is a positive definite weight operator. Overlap similarity measures are obtained when the operator is chosen as the Dirac delta function: $\delta(\mathbf{r}_1 - \mathbf{r}_2)$, while Coulomb similarity measures appear when choosing: $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$. These two operators seem the most suitable and popular ones for similarity comparisons between molecules. However, alternative operator choices are also possible. For example, the use of a third density function tag attached to another quantum object, as a possible operator, produces a triple density QSM [15]:

$$Z_{AB;C} = \int \rho_A(\mathbf{r})\rho_C(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r}.$$

Finally, it must be said that the: $Z_{AA}(\Omega)$ type integrals can be called quantum self-similarity measures.

MOLECULAR SUPERPOSITION

In the molecular context, the approximate density functions, discussed in the previous paragraph, acquire a fundamental role. Indeed, molecular similarity integrals between two or more molecular structures are not invariant upon the relative spatial positions of the molecular structures and have to be calculated within an optimal repositioning of the quantum objects involved. As the integrals attached to the QSM produce, in any case, real positive definite results, the problem can be expressed as a way to determine the molecular relative positions, which give a maximal QSM. For an overlap QSM, this situation can be expressed by means of the equation:

$$\max_{\mathbf{T};\Phi} Z_{AB}(\mathbf{T}; \Phi) = \max_{\mathbf{T};\Phi} \int \rho_A(\mathbf{r})\rho_B(\mathbf{r}|\mathbf{T}; \Phi)d\mathbf{r},$$

where it is implicitly supposed that the density tag of molecule B is translated and rotated by the six possible ways: $(\mathbf{T}; \Phi)$, which are shown as explicit parameters in the integral [15]. Molecular superposition has also been studied in a simpler geometrical way, resulting in a purely geometrical procedure: the topogeometric algorithm (TGSA) [16], which avoids the cumbersome calculation of the similarity integrals for molecular superposition, providing the user with an alternative superposition to QSM-directed alignments, as TGSA is based on an alignment that uses the common molecular backbone of both studied structures.

SIMILARITY MATRICES

Once known a quantum object set, Θ, of cardinality, $\#(\Theta) = N$, a matrix can be constructed with the values of all the QSM between pairs of quantum objects, choosing the simpler way to proceed. This is so because choosing triple density QSM will pro-

duce a matrix of three indices, and so on. Such matrices are called similarity matrices. In the most usual case, dealing with QSM between quantum object pairs, a similarity matrix becomes a square ($N \times N$) symmetric matrix, $\mathbf{Z}$. As QSM are always definite positive real numbers, any quantum similarity matrix could be considered as an element belonging to the unity signature class. One can conclude that similarity matrices constitute elements of some ($N \times N$)-dimensional matrix semispace. One can then easily write

$$\mathbf{Z} = \{Z_{AB}\} \rightarrow \forall A, B : Z_{AB} \in \mathbf{R}^+ \wedge \mathbf{Z} \in V_{(N\times N)}(\mathbf{R}^+).$$

A similarity matrix could be also viewed as a metric of the density tags, but only in the case in which no maximization of the QSM has taken place. Computed in this way, a metric similarity matrix will be positive definite but, when the QSM are optimized, the matrix no longer possesses this property, and some eigenvalues could be negative. In any case, equivalent column or row partitions of a similarity matrix can be easily performed. A column partition may be written as

$$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_N),$$

where $\{\mathbf{z}_I\}$ collect the elements of the matrix associated with the QSM between the Ith quantum object and the rest, including itself. A new tagged set can be constructed in this way, substituting in the original quantum object set the density tags by the columns of the similarity matrix, that is,

$$\Theta = S \times P \Rightarrow D = S \times \mathbf{Z} \rightarrow d_I = (s_I; \mathbf{z}_I) \in D,$$

such a tagged set can be called a discrete quantum object set. Such tagged sets can be employed to describe the quantum objects as N-dimensional semispace entities, and used accordingly, for instance, to classify or find out some ordering among the quantum object set elements, using any of the well-known algorithms, described for this purpose [17]. Because of this possibility, QSM become a set of unbiased and universal descriptors attached to quantum objects. They are unbiased because they are obtained from density function tags, so to construct them, the user needs to do little more than choose the weight operator. Even operators, as they are necessarily positive definite, can be employed in an optimal way, by using convex mixtures of them and optimizing the mixing coefficients. The QSM can be also considered universal, because there are no exceptions or limitations to the computation of QSM among the elements of any quantum object set, other than the knowledge of the density function tags.

Similarity Matrix Transformations: Similarity Indices and Stochastic Matrices

From the early QSM definition, the literature has described transformations of the resultant measures into the so-called similarity indices [1]. Discussions of the meaning of these indices have also been present in the literature (e.g., see Ref. [18]). Recently, other possible alternatives have been put forward [19], and some applications given [20]. The obvious transformation of the similarity matrix elements, associated with a pair of quantum objects A and B, corresponds to the computation of the cosine of the angle subtended by the involved density functions tags, within the Hilbert semispace, where they belong:

$$r_{AB} = \frac{z_{AB}}{\sqrt{z_{AA} z_{BB}}},$$

where z_{AB}, is the QSM between both quantum objects, and z_{AA}, z_{BB} the corresponding self-similarity measures. This similarity index has been referred to in the literature as the Carbó similarity index. The interesting feature of such an index, when compared with the QSM itself, is the fact that one can easily see that $r_{AB} \in (0, 1]$. The upper value corresponds to a similarity between two exact objects, while the index will approach nullity as the objects become dissimilar. Other possibilities are also well described; see the discussion in Ref. [18] and the extension of the similarity index areas in Ref. [21].

Another possible index corresponds to a dissimilarity index, which can be expressed as an Euclidean distance; employing the same symbols as before, an index can thus be defined varying inversely as the Carbó index previously defined:

$$d_{AB} = \sqrt{z_{AA} + z_{BB} - 2z_{AB}}.$$

Similarity indices can be employed as quantum object descriptors in the same way as the initial similarity measures have been described.

Besides this reciprocal forms of the two similarity indices and all their variants [18, 21], simple transformations have been put forward that produce a nonsymmetric matrix, when applied over any similarity matrix. Suppose a diagonal matrix, whose elements are formed by the Minkowski norm of the rows (or columns) of a known similarity matrix:

$$\mathbf{D} = Diag(\langle \mathbf{z}_1 \rangle, \langle \mathbf{z}_2 \rangle, \ldots \langle \mathbf{z}_N \rangle).$$

A stochastic column matrix **S** can be obtained from the similarity matrix **Z**, simply by postmultiplication by the inverse of **D**:

$$\mathbf{S} = \mathbf{Z}\mathbf{D}^{-1};$$

the transposition of the previous stochastic matrix is simply a stochastic matrix by rows. The interesting feature now is the fact that the stochastic column matrix corresponds to a new similarity matrix whose columns can be considered discrete probability distributions and, as such, the correspondence between the quantum object tags, made up of continuous quantum density distributions and columns of such a stochastic matrix, appears to be more adequately adapted to the problem than direct correspondence with similarity matrix columns. However, the unique distinction between raw similarity measures and stochastic elements is a normalization factor. Of course, the partition of the stochastic column matrix in its columns:

$$\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_N)$$

indicates that the column set $\{\mathbf{s}_I\} \subset S(1)$, belongs to the 1-shell of the corresponding vector semispace. Discrete quantum object sets can be made in this way, using as tags the columns of stochastic matrices.

Expectation Values and Quantitative Structure–Property Relationships: Fundamental Equation

One of the most interesting applications of quantum similarity consists of the possibility of constructing a matrix equation, which produces a way to obtain a relationship between the involved quantum objects structures and their properties. The value of such a finding is well known since, in chemistry, apparently equivalent relationships have been sought for a century and a half. However, quantum mechanics and similarity show that this situation is only apparent, and that the resemblance between classical and the quantum similarity deduced equations ends when it is realized that within quantum similarity theory appears the possibility to set up a causal relationship between the structure and properties of submicroscopic objects. Such a performance cannot be claimed at all by the usual QSPR models.

EXPECTATION VALUES

To prove this, one must recall the usual statistical way of computing expectation values of operators, which can be expressed by the integral both in quantum mechanics and theoretical statistics. Moreover, it can be associated with a scalar product [7]:

$$\langle \omega \rangle = \int \Omega(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} = \langle \Omega | \rho \rangle.$$

Suppose, now, that a quantum object set is known. The operator can then be expressed approximately in terms of the quantum object density function tags, in such a way as:

$$\Omega(\mathbf{r}) \approx \sum_I w_I \rho_I(\mathbf{r})$$

an expression that can be substituted in the expectation value of some property π_A of a quantum object A:

$$\pi_A = \langle \Omega | \rho_A \rangle \approx \sum_I w_I \langle \rho_I | \rho_A \rangle = \sum_I w_I Z_{IA},$$

where use has been made of the definition of the QSM between the elements of the quantum object set and the chosen quantum object.

QSPR FUNDAMENTAL EQUATION

Owing to the symmetry of the similarity matrix elements, as well as the validity of the expression above, for all the quantum objects involved, it can be deduced that the following equation must hold:

$$\mathbf{Z}\mathbf{w} = |\pi\rangle,$$

where **Z** is the similarity matrix of the quantum objects; the vector, $\mathbf{w} = \{w_I\}$, collects the coefficients approximating the unknown operator, related to the studied observable property, expressed in the basis of the density function tags; and $|\pi\rangle = \{\pi_I\}$ is a vector collecting the properties associated with every quantum object. This linear system constitutes the QSPR fundamental equation based on quantum descriptors.

Discussion of the characteristics of the above equation is well documented [7, 14], even with the possibility to add nonlinear terms and the nontrivial solutions leading to an approximate semispace vector **w** [22].

An interesting consequence of the previously established considerations can be put forward as a final result. The similarity matrix columns belong to some *N*-dimensional semispace. Thus, the property vector shall be also seen as a linear combination of the similarity matrix columns. In other words, the fundamental QSPR equation can also be written:

$$\sum_I w_I \mathbf{z}_I = |\pi\rangle.$$

So, in the case in which the properties vector can be considered to belong to a vector semispace, the coefficient vector **w** must necessarily belong to some shell in the semispace of the adequate dimension, because

$$\alpha = \langle|\pi\rangle\rangle = \sum_I w_I \langle \mathbf{z}_I\rangle.$$

Thus, no generality is lost, considering that the similarity matrix columns are normalized such that they belong to the 1-shell, considering their columns normalized as in a stochastic transformation:

$$\alpha = \langle|\pi\rangle\rangle = \sum_I w_I = \langle \mathbf{w}\rangle.$$

Consequently, in this case, both the property and the QSPR fundamental equation solution vectors must belong to the same semispace α-shell. An adequate way to obtain such α, necessarily constrained, solution will be by means of a modified least squares technique, quite similar to the optimization procedures to be applied in the approximate density-fitting algorithms.

STOCHASTIC TRANSFORMATION OF THE QSPR FUNDAMENTAL EQUATION

The use of stochastic column matrices, instead of the similarity matrix counterparts, corresponds to a simple matrix transformation, as when one writes the QSPR fundamental equation within the stochastic transform as:

$$\mathbf{S}\mathbf{v} = |\pi\rangle.$$

Then, one can easily write

$$\mathbf{Z}\mathbf{D}^{-1}\mathbf{v} = |\pi\rangle.$$

Thus, for the solution vectors of both systems, the following straightforward relationship is obtained:

$$\mathbf{w} = \mathbf{D}^{-1}\mathbf{v}.$$

Then, suppose that the properties vector are normalized to be contained in the unit shell, using:

$$\mathbf{p} = \langle|\pi\rangle\rangle^{-1}|\pi\rangle \in S(1).$$

Thus, the conclusion is reached that

$$\mathbf{S}\mathbf{v} = \mathbf{p},$$

has a solution, if existing, lying in the unit shell too. Then, the existing solutions to the stochastic QSPR equation will be formed by a set of complex scalars, or using the usual notations:

$$\mathbf{S} = \{\mathbf{s}_I\} \wedge K(\mathbf{s}_I) \wedge K(\mathbf{p}) \rightarrow \mathbf{S}\mathbf{v} = \mathbf{p} \rightarrow K(\mathbf{v}).$$

This result appears to indicate that the solutions of the stochastic QSPR fundamental equation exist if the properties vector can be expressed as a convex combination of the columns of initial stochastic matrix, which represented the involved quantum objects.

This can be put in terms of still more precise definitions. A convex cone, $C(\mathbf{S})$, described by the column set of the stochastic matrix **S** can be defined as the set of all the possible convex combinations of these columns:

$$C(\mathbf{S}) = \left\{ \mathbf{u} \in C \middle| \mathbf{u} = \sum_i \omega_i \mathbf{s}_i \wedge K(\{\omega_i\}) \right\},$$

and it is easily seen that a convex cone, so defined, can always be considered a subset of the unit shell:

$$C(\mathbf{S}) \subseteq S(1).$$

Thus, solutions of the stochastic QSPR fundamental equation, if they exist, will belong to the convex cone generated by the columns of the stochastic matrix associated to the quantum objects. Thus, it appears to be important to know whether there is a convex solution.

A test designed to determine whether a convex solution can be expected could be put forward in the following way. Suppose a distance between pairs of elements is defined in the unit shell:

$$\forall \mathbf{a}, \mathbf{b} \in S(1) \rightarrow \exists d(\mathbf{a}, \mathbf{b}) \in \mathbf{R}^{+},$$

a convex solution then exists, provided that the following properties hold:

$$\forall i : d(\mathbf{p}, \mathbf{s}_i) \leq \max_{i,j}\{d(\mathbf{s}_i, \mathbf{s}_j)\} \rightarrow \exists \mathbf{v} \in K(1) : \mathbf{S}\mathbf{v} = \mathbf{p}.$$

The existence of the solution permits us to consider that the normalized properties vector can be observed as a point inside the polyhedron formed by the stochastic matrix columns.

Application Example: Modeling the Inhibition of Photosystem II by 1,8-Naphthyridin-4-ones

On the other hand, if simpler approaches are applied, it is easier to use well-known statistical algorithms, as in classical QSPR search, to deal with the QSPR fundamental equation. In this way, we have dealt with a large number of application examples, as can be seen within recent [20] and earlier [23] references.

Owing to this possibility, this final section presents a QSPR example using quantum similarity-based molecular descriptors. The present example study consists of a set made by 20 1,8-naphthyridin-4-ones, which inhibit photosystem II, thereby blocking the electron transport in a chloroplast [24]. Given that a good portion of present-day herbicides formerly act as inhibitors of photosynthesis, the naphthyridinone derivatives discussed represent a novel class of potential herbicides.

The QSPR protocol can be summarized as follows:

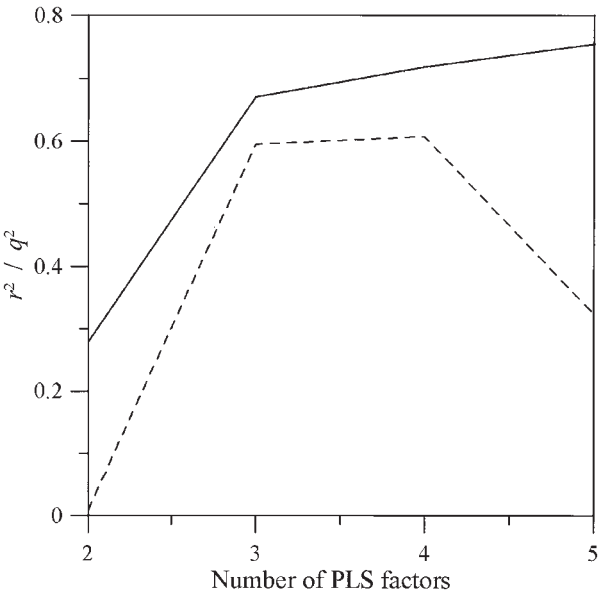

FIGURE 2. Evolution of r^2 (continuous line) and q^2 (dashed line) vs. number of PLS factors.

1. *Modeling molecular structures and geometry optimization:* In this case, molecules have been constructed and optimized at the AM1 level, using AMPAC software (6.55, Semichem, Shawnee, KS).
2. *Molecular alignment and quantum similarity measures:* The optimized geometries have been aligned by pairs using the TGSA [16] procedure. Once superposed, a Coulomb coupling QSM has been carried out over molecular pairs, thus obtaining a symmetric (20 × 20) similarity matrix.
3. *Quantum similarity descriptors and QSAR:* The similarity matrix, with the elements transformed into Carbó similarity indices, has been used as a source of molecular descriptors in a partial least-squares (PLS) [25] technique, to relate the molecular descriptors to the inhibition rate, expressed as pI_{50}.

Several models have been constructed using a variable number of PLS factors, and the evolution of the correlation, expressed in terms of r^2, and prediction capacity, in q^2 terms, is presented in Figure 2. As can be observed in Figure 2, a sharp increase in both magnitudes occurs up to the third factor added, a minor improvement appears up to the fourth, and whereas the correlation increases, the prediction capacity decreases due to overfitting.

So, as an optimal balance between the results and factors used, also taking into account the risk of chance correlations, a three-factor model is chosen, yielding the following QSAR equation:

$$pI_{50} = 1.571 \cdot \mathbf{f}_1 + 3.441 \cdot \mathbf{f}_2 + 4.699 \cdot \mathbf{f}_3.$$

$$r^2 = 0.670 \qquad q^2 = 0.595 \qquad s = 0.376$$

A cross-validated versus experimental values plot is presented in Figure 3. As can be deduced from Figure 3, most of the compounds are predicted correctly within a narrow margin.

Finally, to state the absence of chance correlation in the proposed QSAR model, a random test has been carried out, randomly permuting the inhibition activity vector 1,000 times. As can be seen in Figure 4, a clear separation between the original unperturbed model and the randomly permuted ones is present. Most of the random permutations yield a negative prediction capacity, pointing out the total absence of chance correlations.

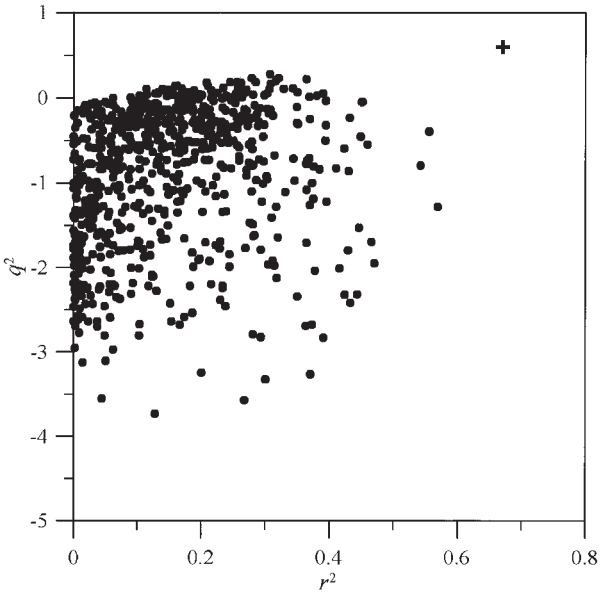

FIGURE 4. Random test results. +, original model; ●, permuted models.

Conclusions

Based on simple mathematical ideas and emerging from the quantum mechanical description of submicroscopic systems, which can now be described as quantum objects, naturally appears in this work to be the theoretical formalism of quantum similarity, taking the form of a quantum mechanical geometric extension. After some reasoning, however, the theory is shown to contain more than such a primary point of view. Properly handled, quantum similarity provides valuable molecular descriptors, which by means of the statistical expectation value concept, permits us to demonstrate that, in some cases, a sound relationship can be found between the quantum object structures and their properties.

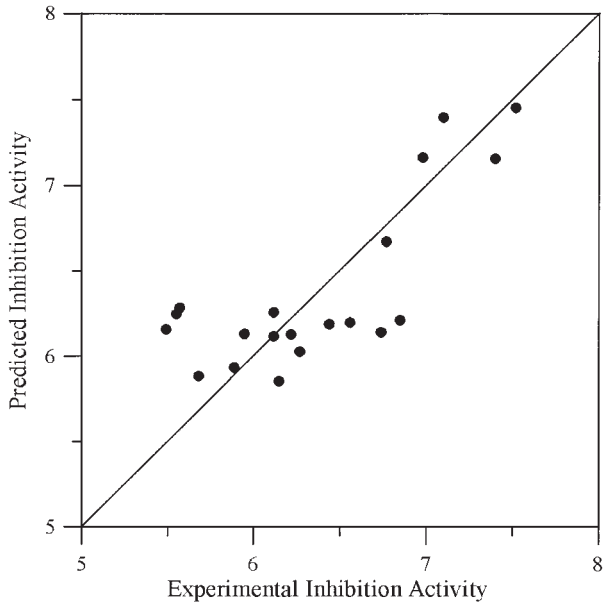

FIGURE 3. Cross-validated activities vs experimental inhibition activity.

ACKNOWLEDGMENTS

X. Gironés acknowledges the University of Girona for a predoctoral fellowship.

References

1. Carbó, R.; Leyda, L.; Arnau, M. Int J Quantum Chem 1980, 17, 1185–1189.
2. Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. In Lecture Notes in Chemistry; Springer-Verlag: Berlin, 2000; Vol. 73, 123 pp.
3. Carbó-Dorca, R. J Math Chem 1998, 23, 353–364.
4. Zadeh, L. A. Inform Control 1965, 8, 338–353.
5. (a) Carbó-Dorca, R. In Proceedings of ECCOMAS; CIMNE: Barcelona, 2000; (b) Carbó-Dorca, R. J Mol Struct (Theochem)

2001, 537, 41–54; (c) Sen, K.; Carbó-Dorca, R. J Mol Struct (Theochem) 2000, 501, 173–176.

6. Carbó-Dorca, R. Advances in Molecular Similarity; JAI: London, 1998; Vol. 2, Chapter 2, 43–72.
7. Carbó-Dorca, R.; Amat, L.; Besalu, E.; Gironés, X.; Robert, D. Fundamentals of Molecular Similarity; Kluwer/Academic/Plenum: New York, 2001, Chapter 12, 187–320.
8. Born, M. Atomic Physics; Blackie & Son: London, 1945.
9. Schrödinger, E. Éditions Jacques Gabay: Paris, 1988.
10. (a) Amat, L.; Carbó-Dorca, R. J Chem Inf Comput Sci 2000, 40, 1188–1198; (b) Amat, L.; Carbó-Dorca, R. J Comput Chem 1999, 20, 911–920; (c) Amat, L.; Carbó-Dorca, R. J Comp Chem 1997, 18, 2023–2039.
11. Amat, L.; Carbó-Dorca, R. Int J Quantum Chem 2002, 87, 59–67.
12. Deisenhofer, J.; Epp, O.; Sinning, I.; Michel, H. J Mol Biol 1995, 246, 429–457.
13. (a) Gironés, X.; Amat, L.; Carbó-Dorca, R. J Mol Graph Mod 1998, 16, 190–196; (b) Gironés, X.; Carbó-Dorca, R.; Mezey, P. G. J Mol Graph Mod 2001, 19, 343–348; (c) Gironés, X.; Amat, L.; Carbó-Dorca, R. J Chem Inf Comput Sci 2002, 42, 847–852.
14. (a) Carbó-Dorca, R.; Besalú, E. J Mol Struct (Theochem) 1998, 451, 11–23; (b) Carbó-Dorca, R.; Besalu, E. Contrib Sci 2000, 1, 399–422.
15. Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Mol Eng 1992, 2, 43–64.
16. Gironés, X.; Robert, D.; Carbó-Dorca, R. J Comput Chem 2001, 22, 255–263.
17. (a) Carbó, R.; Calabuig, B. Int J Quantum Chem 1992, 42, 1681–1683; (b) Carbó, R.; Calabuig, B. Int J Quantum Chem 1992, 42, 1695–1709; (c) Carbó, R.; Calabuig, B. J Mol Struct (Theochem) 1992, 254, 517–531.
18. Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. J Math Chem 1996, 19, 47–56.
19. Carbó-Dorca, R. Int J Quantum Chem 2000, 79, 163–177.
20. (a) Besalú, E.; Gironés, X.; Amat, L.; Carbó-Dorca, R. Acc Chem Res 2002, 35, 289–295; (b) Gironés, X.; Carbó-Dorca, R. J Chem Inf Comput Sci 2002, 42, 317–325; (c) Amat, L.; Besalú, E.; Carbó-Dorca, R. J Chem Inf Comput Sci 2001, 41, 978–991; (d) Gironés, X.; Gallegos, A.; Carbó-Dorca, R. J Comput-Aided Mol Des 2001, 15, 1053–1063.
21. (a) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Advances in Molecular Similarity; JAI: London, 1998; Vol. 2, Chapter 1, 1–42; (b) Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Advances in Molecular Similarity; JAI: London, 1996; Vol. 1, Chapter 1, 1–42; (c) Carbó, R.; Besalú, E.; Calabuig, B.; Vera, L. Adv Quantum Chem 1994, 25, 255–313.
22. Carbó-Dorca, R.; Besalu, E. Int J Quantum Chem 2002, 88, 167–182.
23. (a) Gallegos, A.; Robert, D.; Gironés, X.; Carbó-Dorca, R. J Comput-Aided Mol Des 2001, 15, 67–80; (b) Robert, D.; Carbó-Dorca, R. Int J Quantum Chem 2000, 77, 685–692; (c) Robert, D.; Gironés, X.; Carbó-Dorca, R. J Chem Inf Comput Sci 2000, 40, 839–846; (d) Carbó-Dorca, R.; Amat, L.; Besalu, E.; Gironés, X.; Robert, D. J Mol Struct (Theochem) 2000, 504, 181–228; (e) Robert, D.; Amat, L.; Carbó-Dorca, R. Int J Quantum Chem 2000, 80, 265–282; (f) Gironés, X.; Gallegos, A.; Carbó-Dorca, R. J Chem Inf Comp Sci 2000, 40, 1400–1407; (g) Robert, D.; Amat, L.; Carbó-Dorca, R. J Chem Inf Comp Sci 1999, 39, 333–344; (h) Robert, D.; Carbó-Dorca, R. SAR QSAR Environ Res 1999, 10, 401–422; (i) Mezey, P. G.; Ponec, R.; Amat, L.; Carbó-Dorca, R. Enantiomers 1999, 4, 371–378; (j) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. J Chem Inf Comp Sci 1998, 38, 624–631.
24. Šoškić, M. J Chem Inf Comput Sci 2001, 41, 1316–1321.
25. Wold, S.; Sjöström, M.; Eriksson, L. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Screiner, P. R., Eds.; John Wiley & Sons: Chichester, UK, 1994; Vol. 4, p. 2006–2021.