Notes on Quantitative Structure-Properties Relationships (QSPR) (1): A Discussion on a QSPR Dimensionality Paradox (QSPR DP) and its Quantum Resolution

RAMON CARBÓ-DORCA, ANA GALLEGOS, ÁNGEL J. SÁNCHEZ
Institut de Química Computacional, Universitat de Girona, Girona 17071, Catalonia, Spain

Received 16 July 2008; Revised 5 September 2008; Accepted 5 September 2008
DOI 10.1002/jcc.21145
Published online 22 October 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Classical quantitative structure-properties relationship (QSPR) statistical techniques unavoidably present an inherent paradoxical computational context. They rely on the definition of a Gram matrix in descriptor spaces, which is used afterwards to reduce the original dimension via several possible kinds of algebraic manipulations. From there, effective models for the computation of unknown properties of known molecular structures are obtained. However, the reduced descriptor dimension causes linear dependence within the set of discrete vector molecular representations, leading to positive semi-definite Gram matrices in molecular spaces. To resolve this QSPR dimensionality paradox (QSPR DP) here is proposed to adopt as starting point the quantum QSPR (QQSPR) computational framework perspective, where density functions act as infinite dimensional descriptors. The fundamental QQSPR equation, deduced from employing quantum expectation value numerical evaluation, can be approximately solved in order to obtain models exempt of the QSPR DP. The substitution of the quantum similarity matrix by an empirical Gram matrix in molecular spaces, build up with the original non manipulated discrete molecular descriptor vectors, permits to obtain classical QSPR models with the same characteristics as in QQSPR, that is: possessing a certain degree of causality and explicitly independent of the descriptor dimension.


Key words: QSPR dimensionality paradox (QSPR DP); fundamental QQSPR equation; quantum similarity matrices; descriptor and molecular spaces; gram matrices

Nomenclature and Notation

This article attempts to study some nuances of a well known set of theoretical and numerical chemical procedures. The nomenclature quantitative structure-properties relationship (QSPR) is preferred and will be used from now on, considering that it conceptually alludes to the well-known QSAR, see for example ref. 1 as a source of the diverse facets of the problem, where A stands for (Biological) Activity, and also includes QSTR, with T for Toxicity. This is due to the fact that, within the general QSPR acrostic, QSAR and QSTR are also referenced, because molecular biological activity is simply a molecular property depending on complex factors. Toxicity is also included as it is a molecular property that can be considered some sort of, sometimes obnoxious, side effect of biological activity.

Some conventional names are previously given in order to propose the nomenclature adopted in the present study. Here, any set of molecules to be studied by means of a QSPR procedure is called a molecular point cloud (MPC). Any MPC is supposed to contain all molecules needed to carry a QSPR study of any kind. The core set (CS) is a subset of the MPC, where every molecule can be also attached to a known numerical value of a chosen property, with a one-to-one correspondence. Also, the term unknown molecular set (UMS) is employed, consisting on a MPC subset, which supposedly contains these MPC elements whose property values are not known. A C-m is an element of the CS whereas a U-m is an element of the UMS respectively.

In the classical way to build up QSPR models, the MPC can be also described as a tagged set, defined with discrete row vector tags, attached to every molecule mI of the MPC in a one-to-one correspondence:

\[ I : m_I \leftrightarrow (d_I) \]

Correspondence to: R. Carbó-Dorca; e-mail: quantumqsar@hotmail.com

Contract/grant sponsor: Spanish Ministry of Education and Science; contract/grant number: CTQ2006-04410/BQU

© 2008 Wiley Periodicals, Inc.
The elements of the row vector $|d_i|$ are the (molecular) descriptors, see ref. 1 for example, of the molecule $m_i$. The descriptor vectors can be ordered in a hipercolumn in order to obtain a matrix:

$$D = \begin{pmatrix} |d_1| \\ \vdots \\ |d_N| \end{pmatrix},$$

which possesses a dimension $(N \times D)$ whenever the chosen cardinality of the MPC is $N$ and the vector space containing the subset of $N$ row descriptors bears dimension $D$. The MPC elements can be considered in this mathematical context as a vector subset of some $D$-dimensional row vector space, and one can refer to it as the descriptor space.

A QSPR Dimensionality Paradox

Introduction

When dealing with usual QSPR problems, nothing opposes in principle to the possibility that the descriptor space dimension $D$ is taken as large as possible. See for extended information ref. 1 which can be taken as a source of a great deal of examples on the subject. Thus, in the current studies if the CS has a cardinality: $N_C$, then it is common that: $N_C \ll D$ and also that: $N \ll D$. This initial descriptor dimension choice loosely ensures the linearly independent description of the molecular elements of the MPC as a whole. Therefore, this option hypothetically guarantees that every different molecule of the MPC has a description tag linearly independent of the rest of these associated to the molecular elements of the MPC.

In summary: in this QSPR context every different molecule bears an intrinsically different descriptor tag. This MPC property is crucial when considering the CS and its potential information content in order to obtain a useful QSPR model.

Molecular Description Linear Independence

At this stage of the discussion, it is interesting to note that, whenever a quantum mechanical framework is chosen for every molecular structure, the problem of the molecular description linear independence has a straightforward solution as the quantum mechanical molecular tags, associated to every element of the MPC, can be chosen as density functions. Furthermore, according to quantum mechanics, quantum mechanical density functions are the containers, from whom all the information about molecular structure properties can be obtained, via the usual statistical expectation values computation. In this way, the finite dimensional discrete molecular descriptor vector arrays of empirical QSPR are substituted by quantum QSPR (QQSPR) infinite dimensional continuous functional elements. Accordingly, the options to manipulate the information inside the descriptor space are, within a QQSPR framework, completely different from the classical QSPR ones, where a finite dimensional descriptor space background is employed.

QSPR Algorithms and Descriptor Dimension Reduction

The QSPR algorithms proceed, generally speaking, see again ref. 1 for example, by defining the matrix of dimension $(D \times D)$:

$$S = D^T D \rightarrow \forall I, J : S_{IJ} = \sum_{k=1}^{N_C} d_{kJ} d_{IK} = |d_I| |d_J|.$$ (2)

Constructed in this way, such a matrix is nothing else than the so-called Gram matrix of the descriptor vectors contained in the columns of the matrix $D$.

One can expect, in usual cases, the previous Gram matrix (2) becoming positive semi-definite, that is:

$$\text{Det}|S| = 0 \rightarrow S \geq 0.$$ (3)

It is well known that, when employing the standard QSPR algorithms, in order to avoid over-parameterization of the proposed models, see for example refs. 15–18, the descriptor dimension is reduced, that is choosing:

$$D \rightarrow d \ll N_C < < D.$$ (4)

Any descriptor space dimension reduction can be described by some transformation of the matrix $D$ as defined in eq. (1), which can be written as a matrix equation like:

$$DT = DT,$$

where the transformation matrix $T$, having dimension $(D \times d)$ produces the desired reduced dimension version of the descriptor matrix: $DT$, say, possessing actually a new dimension $(N_C \times d)$. One can also consider this descriptor transformation as a projection of the CS molecular points into some lower dimensional subspace. This is quite evident when dimension reduction procedures like principal component analysis are considered.

The rows of $DT$ still represent the $N_C$ molecular elements of the CS, but now the molecular description is performed within a $d$-dimensional row vector space with: $d \ll N_C$. However, this is equivalent to considering that the CS molecules do not form a linearly independent subset of the MPC discrete description any longer.

A QSPR Dimensionality Paradox

When trying to solve the QSPR problem with the ultimate objective of finding out a way to estimate molecular properties for the UMS elements in the framework of the descriptor space, one usually obtains a linear functional model. Then, at this stage of the QSPR formulation, the statistical procedures to reduce the descriptor dimension create a linear dependence of the MPC as a whole, that is: in both the CS and UMS alike.
Consequently, one is facing the paradox of obtaining a reasonable sized parameter model space, well adapted to some statistical requirements at the expense of a new MPC construct, which possesses the chosen descriptor space reduced dimension \( d \) as a maximal boundary value.

As a result, this disclosed QSRR dimensionality paradox (QSRR DP) may lead to some confuse formal situation involving the usually algebraically correct linearly independent MPC starting point of the problem and the final linearly dependent but statistically correct MPC description.

It can be also stated and stressed that all the obtained QSRR models constructed within finite dimensional descriptor spaces lead to the QSRR DP, as previously discussed. Thus, an algorithm providing a solution of the QSRR DP and bearing a causal add-on feature can be an interesting step forward in the QSRR domain.

### A Simple Illustrative Example of a QSRR DP

A trivial example can be given in order to illustrate the previously described QSRR DP. If the starting point or the descriptor reduction of the QSRR model leads to a linear equation of the type:

\[
\pi = \alpha d + \beta, \tag{3}
\]

where \( \pi \) is the molecular property to be evaluated by the QSRR model, \( \{\alpha, \beta\} \) two parameters to be determined, usually by least squares algorithm, and \( d \) a variable, which can be associated to a scalar, one-dimensional descriptor for each molecule.

Hansch, see for example ref. 20, proposed and applied some time ago such a theoretical background, and numerous successful QSRR examples can be found in the literature bearing this simple linear model structure.\(^{21-23}\) Thus, a discussion on the simplest linear QSRR models can become a useful illustration permitting to grasp the origin of the QSRR DP.

In general, working with these simple linear models becomes the same as accepting that to every molecular structure of the CS: \( C \), which will be employed to compute the parameters in eq. (3), there is a unique descriptor attached:

\[
\forall m_I \in C \rightarrow \exists d_I \land m_I \leftrightarrow d_I. \tag{4}
\]

Thus, in this case, every molecule is described by a unique scalar. As there are \( N_C \) molecules in the CS, this is the same as to consider that every molecular structure is a scalar multiple of some descriptor value: \( d_0 \), acting as an arbitrary scale factor. Therefore, it can be written:

\[
\forall I : d_I = \delta_I d_0. \tag{4}
\]

That is to say that any molecule in the CS, described as above, linearly depends on some background scalar molecule-descriptor value. All molecular structures in the CS are described in scheme (4) as homothetic to a unique ideal molecular structure \( m_0 \), attached to a descriptor: \( d_0 \).

As a result, this molecular characteristic description constitutes a clear example of the QSRR DP: the CS is made of a set with elements constituted by different molecular structures and therefore the whole CS molecular description shall be linearly independent (as it occurs within a quantum mechanical description), but in order to obtain the linear model (3) a homothetic molecular description (4) is used practically, consisting on a set of scalar multiples of a somehow arbitrary background value.

### Ockham’s Razor and Causality as Extra Add-Ons QSRR Features

After the disclosure of the QSRR DP, one must be also aware of the fact that QSRR models, even if in any instance they are found within rational statistical robustness, also result into an attached final equation which is not causal, as it is well-known in the statistical lore.\(^{24}\)

Among other recommendations, causality needs in QSRR have been earlier underlined by Unger and Hansch\(^{21}\) in an explicit manner; causality appears in their work as the need to describe as clear as possible an attached physical mechanism of molecular action, which shall be present in the mind of researchers in the field, before a QSRR model is planned.

Such a causal\(^1\) QSRR intention has been also present in primitive work produced in our laboratory\(^{26}\) and sustained in further publications,\(^{13}\) associated to the initial studies intended to quantum similarity use in QSRR, and kept as a constant idea afterwards (see for example refs. 8 and 9). It seems that the causality idea has been taken into account in classical QSRR in modern times by Kubinyi\(^{23}\) and recently by Estrada,\(^{27}\) showing in this manner the timeliness and ubiquity as well as interest of the physical background which shall be attached QSRR models.

---

\(^1\)Causality is studied exhaustively in Bunge’s classical book.\(^{28}\) To know more see part IV of the mentioned Bunge’s volume, where explicitly the function of the causal principle in science problem is analyzed. Here, following the ideas of ref. 21, when a statistical model is deduced from a known plausible physical formula, then causality is considered that can apply to it. This can be associated to the QSRR model as an added value consisting on a physical relationship between the involved molecular descriptors and a molecular property. In classical QSRR, such a property is present only in very particular cases. In the present work, all computable QQSPR models are attached to a quantum mechanical physical meaning. This is so, because the fundamental QQSPR equation is deducible from basic quantum theory by means of the well-defined quantum mechanical expectation value concept, which in turn can be considered as a tool to compute observable values. Therefore, all the models obtained via QQSPR procedures are causal in this sense. Causality, thus, is a general property of all QQSPR models, in front of the particular causality nature of the models attached to ref. 21.
Unfortunately, causality is rarely taken into account when simple or sophisticated QSRR results are sought; the current literature is full of such noncausal models. As a fuzzy container of this causality problem the reader can peruse the reference book\(^1\) or seeking for a pair of more focused examples, one can observe the articles of Smith and Popelier\(^2\) and Al-Fahemi et al.\(^3\) The causality conundrum and other features can and shall be surely connected to any QSRR problem solving and may be added as well as to the rest of inherent QSRR modeling characteristics, as the above mentioned paradoxical situation produced by the descriptor dimensionality reduction.

When such noncausal QSRR results pretend to offer linear relationships simpler than already computed complex ones, one can think about Ockham’s razor choice as a reason to admit the complexity and other QSPR background features. In fact, causal QSRR foundation shall be appropriately balanced with Ockham’s razor choices in such a way that, if a simpler QSRR model is adopted in front of a complex one, but seems to be not causal in appearance, then: one shall find out a physical background attached to or discard it.

The connection between QSRR causality and parsimony principle choices merits a deeper insight than the one given in this subsection. QSRR DP is the main reason of the present study, for such a motive this indubitable quite interesting subject will be developed in a forthcoming article.

**Alternative Gram Matrix in CS Spaces**

There is a possible way to get rid of the previously disclosed QSRR DP also including the mentioned extra added general causal background. The following discussion will provide the basic elements in order to set up the methodological fundamentals leading to a reduction of the QSRR DP.

**Gram Matrix within the MPC**

The Gram matrix as defined in eq. (2) does not seem to be the unique algebraic construction choice to be used in QSRR problems. An alternative matrix can be built from the parameter matrix \(\mathbf{D}\) in the following way:

\[
\mathbf{Z}^{(d)} = \mathbf{D}^T \mathbf{D} \quad \forall \mathbf{I}, \mathbf{J} : z_{\mathbf{IJ}} = \sum_{k=1}^{D} d_k d_{\mathbf{IK}} = \langle \mathbf{d}_I | \mathbf{d}_J \rangle
\]  

resulting into an array of dimension \((N_C \times N_C)\).

In this alternative construct, even a weighting \((D \times D)\) positive definite matrix \(\mathbf{W}\), containing arbitrary elements, can be employed to build up the matrix (5), as follows:

\[
\mathbf{Z}^{(d)}[\mathbf{W}] = \mathbf{D} \mathbf{W} \mathbf{D}^T \quad \forall \mathbf{I}, \mathbf{J} : z_{\mathbf{IJ}} = \sum_{k=1}^{D} \sum_{L=1}^{D} w_{\mathbf{KL}} d_k d_{\mathbf{IL}} = \langle \mathbf{d}_I | \mathbf{W} | \mathbf{d}_J \rangle.
\]  

The matrix \(\mathbf{Z}^{(d)}\) generally defined as in eq. (6), possesses the same characteristics as the parent quantum similarity matrix (QSM) \(\mathbf{Z}\), constructed within the QQSPR theoretical framework\(^{30-32}\) and can be somehow considered as a discrete alternative of the QSM. Indeed, whenever the MPC description is linearly independent in the sense discussed before, the matrix (5) or the weighted matrix (6) is positive definite, becoming the metric or the weighted metric respectively of the \(N\)-dimensional vector subspace, generated by the supposed linearly independent MPC elements. One can obviously refer to this space, generated within the molecular dimensions of the CS, as the (core) molecular space.

**MPC in a QQSPR Framework**

To grasp the connection of the alternative matrix \(\mathbf{Z}^{(d)}\) with the theoretical basis of QQSPR, a brief sketch on quantum similarity applied into QSRR will be given next.

According to the quantum description of molecules, a quantum MPC (QMPC) can be easily constructed whenever a density function tag can be attached to every MPC molecular structure. This tag set construct can be performed just after obtaining the wave function for each molecule, once the Schrödinger equation is solved at some theoretical level, see for example ref. 12, as density functions can be seen as the product of manipulating the squared modules of any wave function. Then, a QMPC is a tagged set with homogeneous, that is: of the same order, density functions acting as tags, a quantum object set construction which was described several years ago.\(^4\)

Then, calling \(\mathcal{R} = \{\mathbf{r}_I | I = 1, N\}\) the set of density functions, attached in a one-to-one correspondence to every molecule in the QMPC, it can be written:

\[
\forall \mathbf{I} : m_I \leftrightarrow \mathbf{r}_I.
\]

As a consequence, the density function tags in the construction of the QMPC have the same role as the descriptor vectors \(\{|\mathbf{d}_I | I = 1, N\}\) in classical MPC.

The difference between classical and quantum descriptions of molecular structures lies in the fact that, while classical descriptors belong to a finite dimensional vector space, the positive semi-definite quantum density functions considered as descriptors belong to an infinite dimensional vector semi-space.\(^4\) This is just the expression of the well-known quantum mechanics theoretical idea, which consists on accepting that the density function contains all the information which one can statistically extract from an associated quantum system.\(^10\)

**The QMPC Gram Matrix and QQSPR**

Once a QMPC is constructed in the way outlined above, the QQSPR framework can be developed as follows (see refs. 33–36 for additional details). First, a quantum CS (QCS), \(\mathcal{C}\), say, is chosen as a subset of a given QMPC. Hence, to
every element of the QCS, a known value of a molecular property set \( \langle p \rangle = \{ p(j) \}_{j=1,N_C} \) is attached, in such a way that:

\[
\forall m_i \in C : p_i \rightarrow p_i.
\]

Then, a QQSPR Hermitian operator, \( \Omega \), is constructed using a first order approach as:

\[
\Omega \approx \sum_j \alpha_j \rho_j,
\]

employing the set of unknown coefficients \( \langle \alpha \rangle = \{ \alpha(j) \}_{j=1,N_C} \), which can be determined via the usual quantum mechanical expectation value computation. That is, connecting the known QCS property values with the expectation values of the QQSPR operator, it can be written:

\[
\forall m_i \in C : p_i = \langle \Omega|W|\rho_i \rangle = \int \Omega W \rho_i dV,
\]

where the additional Hermitian weighting operator \( W \) is a known positive definite operator. Operator \( W \) has been usually chosen in practical developments as a Dirac’s delta function, which is formally the same as choosing \( W \) as a unit operator, or as a Coulomb operator taken often as an alternative choice.

Equation (8) can be transformed with the use of the linear expression (7) of the QQSPR operator:

\[
\forall I : p_i = \sum_j \alpha_j \langle \rho_j|W|\rho_i \rangle.
\]

The set of integrals appearing in eq. (9) can be collected into a \((N_C \times N_C)\) symmetric array, which has been named a quantum similarity matrix (QSM) \( Z \):

\[
Z = \{ z_{ij} = \langle \rho_j|W|\rho_i \rangle = \langle \rho_i|W|\rho_j \rangle = z_{ji} \},
\]

which appears here as the quantum equivalent of the classical weighted Gram matrix, described in eq. (6).

A clarification shall be issued at this point. The description of such a QQSPR operator, \( \Omega \), has to be associated to expectation value computations of complex, in the sense of intricate or entangled, molecular property values, which can be loosely considered as molecular observables.

Finally, it shall be commented that the same formal results could be obtained when, instead of density functions, shape functions are used in eqs. (11)–(12). As it is well-known, shape functions are just density functions scaled by the inverse of the number of electrons, which is the inverse of the density function Minkowski norm value. Therefore, shape functions have a unit Minkowski norm. Use of shape functions in the QQSPR set up will scale the elements of the similarity matrix by the inverse product of the two numbers of electrons, belonging to the involved density functions. Such a scaling will obviously provide different QQSPR operator coefficients, which will adapt to the new quantum similarity integral values. The rest of what follows remains formally invariant.

**Matrix Form of the Fundamental QQSPR Equation**

It is now trivial from the expression (9) to deduce the matrix form of the QQSPR fundamental equation:

\[
\langle p \rangle = \langle \omega \rangle Z,
\]

and owing to the symmetry of the QSM the equivalent equation can be easily written in column space:

\[
[p] = Z(\omega).
\]

Because the density set \( R \), being attached to a set of different molecular structures, is linearly independent, any subset has the same property and thus, the matrix \( Z \) is a weighted positive definite metric matrix. In practice, to arrive to this QSM characteristic, some Aufbau procedure has to be followed.

Consequently, the QSM is a nonsingular matrix, possessing a well-defined inverse: \( Z^{-1} \). In this way, eq. (11) has a unique solution:

\[
\langle \omega \rangle = Z^{-1}[p].
\]

Equations (7) and (12) permit to obtain an approximate first order linear form of the QQSPR operator \( \Omega \). High order constructs have been also studied but here they will not be taken into account.

**The Connection between Empirical QSPR and QQSPR**

To associate the construction of the Gram matrices of both classical and quantum QSPR, the discrete molecular description of the first could be considered as a discrete approximation of the molecular density functions. That is, writing:

\[
\forall I : p_i \approx \langle d_i \rangle \rightarrow Z \approx Z^{(d)},
\]

one formally considers the empirical matrices \( Z^{(d)} \) as possible approximations of the QSM \( Z \).

From this point of view, the fundamental QSPR equation can be constructed by substitution of the QSM in eq. (11) by the approximate empirical Gram matrix (6) and solved afterwards in the same way. As a result, all the discussion which follows can be applied to obtain predictive classical QSPR models within the molecular space as well.

Therefore, from now on, only the QSM symbol will be used, but everything obtained in this context is still valid; when using empirical QSPR, only the following substitution is needed:

\[
Z \leftarrow Z^{(d)}.
\]
Notes on Quantitative Structure-Properties Relationships

A Possible Reduction of the QSPR DP

QQSPR and Classical QSPR

In both classical and quantum QSPR procedures, the use of the fundamental QQSPR equation written in the form (11), permits to reproduce exactly the CS properties, once the linear coefficients are obtained employing the eq. (12). Such a result is valid in both theoretical levels, classical and quantum, but becomes useless from the point of view of employing the whole of CS elements to guess approximate values of the property for UMS. As far as we know, it has not been until recent times that several discussions of the possible solutions of the QQSPR problem appeared in the literature.33–36

Such a computational step was necessary in the applied quantum similarity area, as QSM have been employed within a classical statistical computational framework as a universal and unbiased source of molecular descriptors, see for example ref. 37. In order to use in full the practical potentiality of the QQSPR as a causal source\(^1\) of computing approximate UMS, some preliminary considerations are to be made.

At this discussion stage, it must be stressed again the important fact that the same procedure can be applied within classical molecular parameter metric matrices, as defined in eq. (6) or QSM as constructed in eq. (10). Starting from the fundamental QQSPR eq. (11), only the quantum mechanical models can be associated to some sort of causal relationship\(^8\) as eq. (11) is deduced from the construction of the QQSPR operator (7) and the usual quantum expectation value definition (8).

The classical Gram matrices: \( Z^{(d)} \), when used replacing the QSM and solving the resultant fundamental QQSPR equation, may provide approximate results, where the causal nature of the models perhaps may be considered to persist, see for example footnotes † and §. Furthermore, this causal character may be assumed despite the empirical choice of the parameters leading to the approximate Gram matrices, on the contrary to classical QSPR models obtained via statistical linear regression procedures. This can be so because empirical QQSPR models rely on the structure of the fundamental QQSPR equation, which can be considered causal, according the definition of footnote\(^1\).

Moreover, the approximate classical Gram matrices \( Z^{(d)} \) do not explicitly depend on the descriptor space dimension D as obviously happens in the classical QSPR definition (2) for the Gram matrix \( S \) in descriptor space.

Thus, in principle, one can use any kind and number of descriptors to build up the empirical Gram matrix in molecular space, submitted to the unique restriction consisting on that the approximate matrix \( Z^{(d)} \), constructed in this way, becomes positive definite. Such a property is quite interesting for the practical application of classical QSPR methodologies, precisely those which employ as descriptors grid values of a chosen function, computed in the three dimensional neighborhood of a molecular structure. Such computational framework occurs in comparative molecular field analysis (CoMFA) (see, for example, ref. 41) or in the so-called 3D QSAR in general (see for a recent account ref. 42 or the sophisticated procedure developed by So and Karplus in refs. 43 and 44).

Approximate Solution of the Fundamental QQSPR Equation

An exhaustive description of the QQSPR problem solution has been described in previous publications\(^{34–36}\) and in an encyclopedia chapter,\(^{32}\) therefore, minute details will be omitted. A broad summary can be described as follows.

Suppose that a QCS is known and the attached similarity matrix is computed as it has been previously discussed. Then, suppose a known U-m belonging to the QUMS of the QMPC (where the QCS also belongs to), therefore the similarity integrals between the U-m and the QCS elements can be computed easily as:

\[
\forall m \in C : z_{ql} = \langle \rho_{U} | W | \rho_{l} \rangle, \tag{13}
\]

where \( \rho_{U} \) is the U-m density function. The similarity integrals can be collected into a row vector:

\[
\langle z \rangle = \{z_{ql}\}
\]

and the U-m quantum self-similarity integral, see for example ref. 45, can also be computed as:

\[
\theta = \langle \rho_{U} | W | \rho_{U} \rangle.
\]

With this previous information, an extended fundamental QQSPR equation can be constructed:

\[
\begin{pmatrix}
Z & |z\rangle \\
\langle z| & \theta
\end{pmatrix}
\begin{pmatrix}
|\omega\rangle \\
\langle \omega|
\end{pmatrix} =
\begin{pmatrix}
|p\rangle \\
\langle p|
\end{pmatrix}, \tag{14}
\]

where an extra row and column have been added to the original eq. (11). Also, a new QQSPR operator coefficient, \( \omega \), has to be added and the U-m unknown property value, \( \pi \), is considered as an extra parameter. Therefore to the \( N_C + 1 \) unkown coefficients one shall add an extra one: the property value to be evaluated.

Equation (14) can be easily written as two separate parts in the following way:

\[
Z(\omega) + \omega(z) = |p\rangle
\]

\[
\langle z| \omega + \omega \theta = \pi
\]

thus, the first equality can be used to evaluate the coefficient vector \( |\omega\rangle \):

\[
|\omega\rangle = Z^{-1}(|p\rangle - \omega|z\rangle). \tag{16}
\]

Substitution of this result into the eq. (15) second equality, produces the expression which can be used to obtain a value of the U-m unknown property:
\[ \pi = \alpha \theta + \langle z \langle Z^{-1} | p \rangle - \omega \langle z \rangle \rangle = \alpha \langle \theta - \langle z \langle Z^{-1} | z \rangle \rangle + \langle z \langle Z^{-1} | p \rangle \rangle. \]  
(17)

Equation (17) proves the obvious linear dependence of the U-m unknown property from the unknown QQSPR operator linear coefficient \( \omega \).

It is also obvious that the only way to obtain an estimate of the property via eq. (17), is to find an approximate value of the unknown coefficient \( \omega \).

The first equation of the system (15) provides the generic source of the approximate values to be associated to the unknown coefficient \( \omega \). Although there are several ways to obtain an approximate guess of the coefficient value, a plausible procedure may be based on accepting an extremum restriction, which can be imposed to the vector difference appearing in eq. (16).

This can be performed by defining the norm of the difference vector:

\[ |d| = |p| - \omega|z| \rightarrow \langle d | d \rangle = (p | p) - 2\omega(z | p) + \omega^2(z | z) \]

and making it extremal:

\[ \frac{\partial \langle d | d \rangle}{\partial \omega} = 2(\omega(z | z) - (z | p)) = 0 \quad \rightarrow \quad \omega^{(\text{opt})} = \frac{(z | p)}{(z | z)} \]

producing an approximate optimal value of the U-m unknown property, once used in eq. (17):

\[ \pi^{(\text{opt})} = \left( \frac{z}{z} \right) \left( \theta - \langle z \langle Z^{-1} | z \rangle \rangle + \langle z \langle Z^{-1} | p \rangle \rangle \right). \]  
(18)

It is easy to see, when terms in eq. (18) are rearranged, that it can be rewritten as:

\[ \pi^{(\text{opt})} = \left( \frac{\theta - \langle z \langle Z^{-1} | z \rangle \rangle}{z \langle z \rangle} \right) (z | 1 + Z^{-1} | p). \]

Therefore, defining the row vector:

\[ \langle t \rangle = \left( \frac{\theta - \langle z \langle Z^{-1} | z \rangle \rangle}{z \langle z \rangle} \right) (z | 1 + Z^{-1} | p) = \langle t \rangle \]

one can write the optimal property value of the U-m as the linear functional:

\[ \pi^{(\text{opt})} = \langle t | p \rangle = \sum_{l} t_{l} p_{l}. \]

This last result shows that the optimal value of the property is a linear combination of the CS property values, which is a common outcome, shared with the usual statistical QSAR procedures, computed under the molecular description paradox.

A Simple QQSPR Application Algorithm

There are no limitations, but a large flexibility choice among the number of C-m elements used in building up the QCS and the number of U-m elements belonging to the QUMS, which can be considered when a QQSPR model is to be computed. A general algorithm has been already described\(^{34}\); hence, it will not be repeated here. The same can be said for the alternative ways of using the first part of eq. (15) in order to obtain the optimal estimate of the U-m properties.\(^{34}-36\) However, it is interesting to give the simplest case one can face as a working example, where only one element of both sets is taken into account, constituting in this manner the simplest procedure among all the possible QQSPR algorithms that can be set up.

A QQSPR 1:1 Algorithm

As it has been discussed in the previous section 4 above, the fundamental QQSPR equation can be solved approximately for the practical evaluation of U-m unknown properties in particular, using the elements contained within the QCS as a whole. This situation can be named as a \( N_{C}:1 \) algorithm. However, in previous published work even a possible \( N_{C}:N_{U} \) algorithm was described,\(^{34}\) involving the whole QCS and \( N_{U} \) elements of the QUMS.

Practice with QSM has shown that the elementary 1:1 algorithm, involving just one C-m and one U-m, was sufficient to obtain predictive models,\(^{36}\) which turned to be in some extent quantum equivalents of the well-known Hansch procedures.

In such a 1:1 procedure, the augmented fundamental QQSPR eq. (14) acquires a convenient form, like:

\[ \left( \begin{array}{c}
z_{CC} \\
z_{CU}
\end{array} \right) \left( \begin{array}{c}
\omega_{C} \\
\omega_{U}
\end{array} \right) = \left( \begin{array}{c}
p_{C} \\
p_{U}
\end{array} \right) \]  
(19)

where the subscripts \( C \) and \( U \) refer to some pair of specifically chosen C-m and U-m, respectively. The 1:1 algorithm associated to eq. (19) is quite suitable, because the extra complexity of the Aufbau problem\(^{46}-49\) involved in the construction of the quantum similarity matrix can not be present between pairs of QSM elements. However, the quantum similarity measure \( z_{CU} \) has to be optimized by means of any of the described superposition algorithms,\(^{46}-49\) as the unique extra preparatory numerical task to be done. In the same manner as in the \( N_{C}:1 \) algorithm case, the final form of the approximate U-m property value can be obtained as in eq. (18), but taking into account the reduced dimensions of the problem, transforming into:

\[ p^{(\text{opt})}_{U} = \left( \frac{p_{C}}{z_{CU}} \right) \left( z_{CU} - \frac{z_{CU}}{z_{CC}} \right) + \frac{z_{CU}p_{C}}{z_{CC}} \]

which in turn can be straightforwardly reordered into the expression:

\[ p^{(\text{opt})}_{U} = \left( \frac{z_{CU}}{z_{CC}} \right) p_{C}. \]  
(20)

Then, from this simple result, which can be repeatedly used over every one of the QCS C-m elements, one can obtain an average of
the estimates (20) of any chosen U-m in terms of the whole $N_C$ elements of the QCS:

$$
\left\langle p_{U}^{(\text{opt})}\right\rangle = \frac{z_{U}}{N_C} \left( \sum_{i=1}^{N_C} p_i \right).
$$

(21)

**A Hansch-Like Model**

Having deduced the eqs. (20) and (21) it is also easy to obtain a Hansch-like model.20–22 (see also ref. 23 for a modern account), to estimate the U-m properties. There is the possibility to use the following technique which will resemble to a leave-one-out procedure:

1. Use as U-m elements every element of the QCS in turn, employing the remaining $N_C-1$ to obtain the averaged type (21), this will provide a set of $N_C$ property estimates: \{ $p_{U}^{(\text{opt})}$ \}. These estimates can be expressed as:

$$
\forall m_X \in C : \left\langle p_{X}^{(\text{opt})}\right\rangle = \frac{z_{X}}{N_C-1} \left( \sum_{i \neq X}^{N_C} p_i \right).
$$

(22)

2. Use the known experimental values \{ $p_1$ \} of the QCS and the averaged optimal properties \{ $p_{X}^{(\text{opt})}$ \} obtained by using eq. (22), acting as estimates of the experimental property values, to obtain by a least squares procedure the parameters $\{ \alpha, \beta \}$ associated to the linear relationship:

$$
p_{X}^{(\text{opt})} = \alpha + \beta \left\langle p_{X}^{(\text{opt})}\right\rangle.
$$

(23)

The sets \{ $p_1$, $p_{X}^{(\text{opt})}$ \}, and \{ $p_{X}^{(\text{opt})}$ \} can be employed to calibrate the model (23) by the usual statistical means and a regression coefficient can be computed in this way. It must be noted here that the set \{ $p_{X}^{(\text{opt})}$ \} has not necessarily to be considered as a molecular one-dimensional descriptor, but a first QSPR estimate known property values associated to the C-m elements of the QCS. Thus, no QSPR DP is present in such a step of the QQSRR procedure, when just experimental properties are compared with estimated properties of the MPC elements.

3. The computation of an approximate value for any U-m property $p_{U}^{(\text{opt})}$ can be obtained by means of the linear eq. (23) following the using steps:

a. Employ eqs. (20) and (21) to obtain \{ $p_{X}^{(\text{opt})}$ \}. 
b. Utilize eq. (23) to estimate the U-m property:

$$
p_{U}^{(\text{opt})} = \alpha + \beta \left\langle p_{U}^{(\text{opt})}\right\rangle.
$$

**Final Remarks**

The algorithm proposed above can be seen as the outcome of a sequence of computational or numerical experiments, performed using the QCS elements in order to estimate a value of some property attached to a given U-m. The average associated to eq. (21) is the primary outcome of such $N_C$ experiments, which can be refined employing the Hansch-like linear eq. (23). The arithmetic average result can be analyzed via the usual statistical procedures,24 as it can be associated to a variance:

$$
\text{var}(p_{U}^{(\text{opt})}) = \frac{1}{N_C} \sum_{i=1}^{N_C} \left( \frac{z_{U} p_i}{z_{U}} - \left\langle p_{U}^{(\text{opt})}\right\rangle \right)^2 \quad = \left\langle p_{U}^{(\text{opt})} \right\rangle^2 - \left\langle p_{U}^{(\text{opt})}\right\rangle^2
$$

where:

$$
\left\langle p_{U}^{(\text{opt})} \right\rangle^2 = \frac{1}{N_C} \left( \sum_{i=1}^{N_C} \left| \frac{p_i}{z_{U}} \right|^2 \right).
$$

Thus, the confidence range and other tests can be easily designed for the evaluated properties.

**Practical Application Algorithm for Empirical QSPR Problems**

The previous algorithm has been applied in a recent work, where a QQSRR problem with a QSM $Z$ was studied as an example.56 Here, the computational scheme, which has been followed to test the present molecular space QSPR design within a discrete computation of the empirical matrix $Z^{(d)}$ (instead of the quantum QSM $Z$), will be commented. Of course, the authors do not pretend to present an exhaustive account of all the procedures which can be devised within this kind of discrete QQSRR procedures, that is the definitive set up, but a preliminary sketch, which can be employed to further refine and amplify the application of a causal QSPR DP free procedure, as it has been previously described in this article.

**A Discrete QQSRR Program**

The implementation of the theoretical QQSRR development presented here has to be taken as a possible preliminary step of the multiple possibilities, which are contained within the mathematical elements so far developed. Practice will tell which refinements can be added to the present implementation.

A general program has been constructed and applied to four examples, on a variety of available different discrete molecular descriptors from diverse sources. In the program, any of the $N_C - \ell : \ell = 1, N_C - 1$ possible algorithms have been implemented in a simple and elegant programming way using a nested do loop technique.50–52

In the program tests, the elements of the CS, taken one by one in turn, have been considered as U-m. As the algorithm parameter $\ell$ increases, the number of possible U-m property evaluations increases in a combinatorial explosive manner. The present implementation does not take into account the increasing number of trials and obviously only the feasible simple com-
puter processor abilities have been explored in full. This makes each U-m statistical evaluation as one acquired from the whole sampling universe. A future random sampling has to be clearly envisaged, if other than the few first values of the algorithm parameter are to be explored. The structure of nested do loops as already implemented permit easily the program code migration to parallel computing machinery.

Even if the descriptor space dimension $D$ seems to be greater than the dimension of the molecular space, the corresponding Gram matrix has been constructed within every algorithmic level in a fashion reminiscent of the Montecarlo procedures employed in statistical mechanics simulations, see for example ref. 53. The algorithm below has been followed in order to choose the optimal molecular descriptors entering a given QQSPR computation level.

**Algorithm: Montecarlo Descriptor Choice to be Used in Empirical QQSPR Procedures**

a. Choose the QQSPR algorithm level $\ell$. Randomly choose a minimal set of $N_C$ (or greater) number of descriptors making linearly independent the used MPC. Assign to each chosen descriptor a null importance real tag. Let $R_{\ell}^2 = 0$.
b. Construct the Gram matrix with the chosen descriptor set.
c. Solve the QQSPR equation via the algorithm $N_C - \ell : 1$ for every molecule in the CS.
d. Compute $R^2$, the regression coefficient between the experimental and the estimated optimal properties. Compute: $\Delta = |R^2 - R_{\ell}^2|$. Assign importance to the descriptors: importance + $\Delta$ for the accepted descriptor and importance $- \Delta$ for the discarded one.
e. If: $\Delta > \varepsilon$ then go to step f
else, stop: the optimal set of descriptors is chosen.
f. If: $R^2 > R_{\ell}^2$ go to step g; else,
1. choose a random number $x \in (0,1]$.
2. if: $x > \Delta$ discard the choice and go to step h; else,
g. Accept the descriptor choice. Let: $R_{\ell+1}^2 = R^2$.
h. Order the used and not yet used descriptors from minimal to maximal importance, taking into account that they are considered as two independent sets for ordering purposes.
i. Choose at random a descriptor not yet used. Discard at random a used descriptor. Substitute the second by the first. Go to step b.

**End of Montecarlo Descriptor Choice**

The preceding algorithm at each QQSPR algorithmic level produces a model like the one of eq. (23) associated to a maximal regression coefficient. There is no compulsive choice for the search of the optimal descriptor set employed at a given algorithmic level. For sure, more efficient choices could be made, but we were interested in testing a new optimum search adapted to the problem, so no other computational possibilities were explored yet.

Another remark should be made concerning the number of descriptors chosen. In fact, there is no compulsive restriction on this number, except that the Gram matrix $Z^{\ell0}$ must remain positive definite. In the present calculations the number of chosen descriptors at any optimization step has been usually kept equal to $N_C$, the number of CS molecules employed in any MPC case of study. Although in the examples below, some tests with larger descriptor sets have been tested.

**Computational Examples and Discussion of the Results**

In this final section some examples have been chosen to illustrate the applicability of the previous mathematical background. This section does not pretend nor being exhaustive for obvious reasons, previously discussed in the preceding section, nor definitive, as the present results are to be taken just as raw tests to describe in a practical way both the new features of QQSPR algorithms and employed to show the acceptable results produced in this way.

Moreover, the examples given in this section have been chosen not to present unpolluted results, all of them with nice relationships and perfect regression lines and coefficients. They have been chosen in order to show that some of the problems of empirical QQSPR model search can turn out to be similar to the classical QSAR procedures.

**Description of the Present CS Examples Chosen**

**Introduction**

Four CS have been studied in order to assess the potential of the QQSPR algorithms. They have been chosen mainly because of the availability of a set of attached discrete descriptors and other reasons, as it will be explained below. All the present figures have the associated regression line in numerical form printed within the figure. On the $R^2$ values of each graphic, statistical significance tests, using $t$ and $F$ distributions, have been performed, resulting in significative regression at the 99.9% and 99% levels of both parameters, respectively.

**Justification of the Chosen MPC**

The first example, which studies the Cramer steroid set with two activities, has been chosen because it was worth to try to assess the effect of different activities associated to the same CS on the QQSPR model search. The extended experimental activity set, as a result of the present study, does not seem to possess an ideal structure, even if it has been used as a common benchmark dataset in various QSAR model searching. The second example has been employed because of the wide and perhaps ill-conditioned range of experimental activities, in order to make apparent what can be expected in a QQSPR model search for these cases. No experimental data manipulation has been performed to picture the effect of the original property distribution. Finally, the third example has been used in order to test empirical QQSPR modeling in a case where a well-defined UMS is known. This third case, therefore, corresponds to a full deployment of the proposed QQSPR varied algorithms. The specific results, obtained in this complete way, demonstrate the potentialities of the present approach.
Results

Cramer Steroids. Two steroid biological activity datasets, for which descriptors were calculated by using EPI Suite [EpiSuite v3.20 (2000–2007) US Environmental Protection Agency, Syracuse Research Corporation], have been studied. EpiSuite includes a database of experimental environmental and physicochemical properties, and several modules which estimate a variety of endpoints such as atmospheric oxidation, bioconcentration, biodegradation, aquatic toxicity, Henry’s law constant, aqueous hydrolysis, octanol-air and octanol-water partition, melting point, boiling point, and vapor pressure, soil sorption, and water solubility. In the present study, only the estimated physicochemical properties were used, due to missing experimental values for all database compounds.

Two computations have been performed on Cramer steroids, according to the extent and quality of the experimental activity values available. Details are commented in the two following subsections:

a. The Original Set of Cramer steroids: This example was chosen because the original 21 steroid CS activities reported have been already studied within a density function descriptor at the 1:1 algorithm level. The optimal 20:1 resulting relationship, which follows the trend as has been commented earlier, is presented in Figure 1.

b. The Extended Set of Cramer Steroids: The CS made by Cramer 31 steroid set has also been chosen as a complementary source of information to assess the QQSPR algorithm in a case which has been studied in many ways. Although it could be interesting to compare the results for both CS, the results for the present extended experimental activities are too irregular to merit more than a comment and a graph. The presence of quite a number of equal activities for some C-m, provides a result like the ones displayed in Figures 2 and 3 for the algorithms 2:1 and 29:1, respectively.

Even if both $R^2$ values are statistically acceptable, the point dispersion shown in both cases induces to dismiss the results as a valid way to obtain a possible reliable guess of activities. These results are shown to prove how sensitive is the proposed QQSPR algorithmic family to weird experimental data.

A Set of Endocrine Disruptors. This set was chosen because it has already been studied with classical QSPR and a large set of descriptors was available. All the descriptors have been computed by using the TSAR for Windows software (TSAR v3.2; Oxford Molecular, Oxford, UK). Among them one can find calculated physicochemical properties, topological indices, and some indicator variables. Some of the physicochemical

![Figure 1. Optimal 20:1 relationship for the original set of Cramer steroids.](image1)

![Figure 2. Optimal 2:1 relationship for the extended set of Cramer steroids.](image2)

![Figure 3. Optimal 29:1 relationship for the extended set of Cramer steroids.](image3)
properties encode information regarding steric, electronic, and hydrophobic features. Indicator variables account for the presence or absence of explicit structural features, such as atom counts.

The quality of the activity of the molecules involved does not seem to be very good as many structures of the CS had practically no activity, and the rest bear quite small values, while a few present large values. One can consider this experimentally ill-conditioned set as a difficult one, even in a classical QSRR computational framework. A graph for one of the best tests obtained so far for this CS is presented in Figure 4 at the level 58:1, where the experimental activity trends can be quite well appreciated.

A Set of TIBO Compounds. The TIBO antimalarial set, described in ref. 66, has been chosen here because the parameter set is obtained from topological quantum similarity indices (TQSI), see for example refs. 37 and 67. It has also been studied classically in the original article and employing the mentioned TQSI.67 The TIBO set has a well-defined CS and collateral set, whose activities are known, which can easily bear the role of the UMS. The detailed computational collection, issued from the QQSPR algorithms and performed within the TIBO set, will be described as follows. In the family of TIBO derivatives, 46 compounds have been optimized, constructing the Gram matrix with 46 descriptors and employing an algorithm 46:1, producing the following results, shown in Figure 5:

After algorithm optimization, the 46 optimal descriptors have been used together with the associated experimental activities to estimate the properties of a UMS made with 24 TIBO derivatives, which have not entered the previous optimization process. Such estimation process has been performed with the algorithms: \( N : 1(N = 1, nC) \). In the following graphical description, one can easily see that the correlation between experimental and computed activities augments as \( N \) and the corresponding number of activity estimations \( NC \) follow the trend (see Fig. 6):

As an example of the performed calculations, the results for the algorithm level 6:1 are presented below in Figure 7:

**General Behavior of the Discrete Descriptor

QQSPR Algorithms**

In all cases of the four presented examples, one can observe a similar behavior pattern, so the resulting general features will be discussed now.

The optimal regression coefficient increases as the algorithms increase the number of C-m employed to estimate the rest. The algorithm: \( N_{C-1}:1 \) provides an optimally chosen descriptor
regression, which virtually produces a unit regression coefficient. This result indicates that with a reasonable number of CS elements, the U-m property values can be obtained with reasonable high accuracy. However, in this case there are no means to compute uncertainty intervals for the U-m property estimations. In the rest of algorithms and better on the ones with high values of the parameter $\ell$, the corresponding statistical estimates of each arithmetic mean property value can be clearly attached to an uncertainty interval.

In any case, the optimal descriptor choice permits to compute the necessary U-m matrix elements, in order to solve the approximate QQSPR problem and estimate the unknown property value.

**Conclusions**

A paradox, named here as the QSPR dimensionality paradox, affecting the usual QSPR procedures, has been described and its reduction by means of the use of the algorithmic family attached to QQSPR ideas has been explained.

As a result of the preliminary programmed tests, in order to assess the immediate application of some of the possible the QQSPR algorithms, an alternative to classical QSPR procedures has been described in terms of a simple choice of such algorithms derived from quantum mechanics and quantum similarity.

As the procedures chosen are based on the well-known quantum perception of expectation values, the advantage of the QQSPR algorithmic family consists on obtaining causal QSPR models also devoid of the QSPR dimensionality paradox.

Another add-on benefit of QQSPR algorithms consists on their general application either at the infinite dimensional quantum density function or at empirical discrete descriptor levels.

The statistical scheme of the QQSPR algorithms becomes very simple, once the random mechanisms of optimal choice of the descriptors are put forward.

Results are comparable to classical QSPR, but QQSPR algorithms possess other characteristic features. Some QQSPR characteristics seem difficult to be found in classical QSPR models and possibly bear a number of nuances, still waiting to be developed.

A universal causal procedure, devoid of the dimensionality paradox, has been defined for QSAR model search purposes.

**Acknowledgments**

A. Gallegos stage in our laboratory is associated to the research contract 2006 BP-B1 00171 by DURSI. Advice from Professor E. Besalú, who has critically read the manuscript and his altruistic provision of data on the TIBO molecular descriptors as well are warmly acknowledged. The enlightening comments of the referees, whose suggestions contributed to increase the quality of the work, are warmly acknowledged.

**Glossary**

Here, a succinct list of the terms appearing in this work follows. The italicized words appearing in the definitions refer to other terms already defined within the glossary.

C-m: Can be also referred as C-molecule or core set molecule. It is a molecular structure belonging to a MPC, whose experimental properties are known and thus belonging to a CS.

CS: Core set, any tagged set of well-defined molecular structures, possessing known experimental values of some properties. The tags of its object set elements are used as a cornerstone to build up the fundamental QQSPR equation.

Density Function: A non-negative definite function which can be derived from quantum mechanical theoretical procedures. Usually the term is used for the first-order density function, obtained from the squared module of the molecular wave function after integration over all electron coordinates but one. The Minkowski norm of the (first order) density function is the number of electrons.

Descriptor Spaces: Vector spaces with elements made by column or row matrices, with elements constructed in turn by a parameter (molecular descriptor), whose values are attached to some set of molecular structures.

Gram Matrix: A symmetric matrix whose elements are the scalar products of a known set of vectors or functions.

Metric Matrix: A Gram matrix made with a set of linearly independent vectors or functions.

Minkowski Norm: A norm obtained from the complete sum of the (absolute values) of the elements of a vector or a matrix. When the vector is a function, this norm is just the integral (of the absolute value) of the function. The absolute value does not apply when the vector elements or the function are non-negative definite, as occurs in density or shape functions.

Molecular Descriptors: A set of theoretical, empirical or experimental parameters, which is associated to some well-defined molecular structure. In quantum mechanics the essential molecular descriptor is the density function, which is assumed to contain all the information which can be obtained for the associated molecule. Shape functions can be also employed for such a purpose.

Molecular Spaces: Vector spaces constructed by column or row matrices, whose elements are molecular descriptors. In a quantum mechanical framework the vectors are the molecular density functions, attached to precise molecular structures.

MPC: A molecular point cloud is a tagged set of molecular structures associated to some tag vector made of molecular descriptors. It can be considered as a subset of the molecular space.

Object Set: One of the two parts of a tagged set, whose elements are well defined; for example: molecular structures. Its elements are called objects.

QCS: A Quantum core set.

QMP: A Quantum molecular point cloud is a MPC whose elements are described by density or shape function tags.

QQSPR: Quantum QSPR, the set of algorithms described in this article. They are essentially based on the quantum mechanical expectation value concept and the quantum mechanical description of submicroscopic systems by means of density or shape functions.

QQSPR Fundamental Equation: The equation deduced from the use of the quantum mechanical expectation value, when a QQSPR operator is set with the tag elements of a QCS.

QQSPR Operator: The Hermitian operator constructed by linear (or higher order) combinations of the density or shape function tags belonging to some QCS.
QSPR: Quantitative structure-properties relationships, the term used in this article referring to the procedures employed to construct any predictive computational functional model (usually linear) between molecular structure and molecular descriptors.

QSPR DP: QSPR dimensionality paradox, the paradox studied in the present article. It appears in classical QSPR procedures and consists of the fact that, starting with a well-defined linearly independent object set, the necessary reduction of molecular descriptor space dimension, due to statistical manipulations, produces a linearly dependent MPC.

Quantum Similarity Matrix: The metric matrix, constructed with the density or shape function tags of a CS, entering the fundamental QQSPR equation.

QUMS: A Quantum UMS.

Shape Function: A first order density function, scaled with the inverse of the number of electrons. Its Minkowski norm is the unity.

Tag Set: One of the two parts of a tagged set containing information about the elements of the object set. The information form can be gathered as a vector made of bits, by numerical values of any kind or just consisting of functions. For example, in a molecular tagged set, the object set is made of molecular structures and the tag set is constructed using the existing one-to-one correspondence of every object with the attached quantum mechanical molecular density functions.

Tagged Set: A set with elements made of the Cartesian product of any object set and a tag set. The tag set contains information on the elements of the object set. A MPC is just a tagged set, whose elements are molecular descriptors.

Tag Vector: An element of the tag set part of a tagged set.

U:m: Can be referred as U-molecule or unknown molecule. They are the elements of a UMS.

UMS: Unknown molecular set, it is a subset of a MPC, whose experimental properties are unknown and that will be estimated by means of a QQSPR procedure.

References