A Stoichiometric Approach to Quantitative Structure−Property Relationships (QSPR)

Ilie Fishtik* and Ravindra Datta

Fuel Cell Center, Department of Chemical Engineering, Worcester Polytechnic Institute,
Worcester, Massachusetts 01609

Received February 17, 2003

An unusual analogy between the quantitative structure−property relationships (QSPR), stoichiometry, chemical thermodynamics, and kinetics is presented. Namely, the conventional ordinary least-squares (OLS) QSPR analysis is modified so as to explicitly minimize the residuals of the species subject to a set of linear relations among the residuals. The ways the linear relations among the residuals are visualized and defined totally resemble the formalism of chemical stoichiometry and, therefore, were called isostructural reactions. It is further proved that the residuals may be uniquely partitioned into a sum of contributions associated with a set of isostructural reactions that have the same properties as the response reactions (RERs) previously deduced by us from chemical thermodynamics and kinetics. This finding is shown to be a useful tool for a deeper understanding of the QSPR. In particular, the isostructural RERs approach may be effectively used to detect the outliers.

INTRODUCTION

Quantitative structure−property relationships (QSPR) have been demonstrated to be an effective computational tool in understanding the interrelation between the structure of molecules and their properties.[1−3] The QSPR techniques are now being used routinely to predict a large variety of properties of the molecules. Although tremendous progress has been made in discovering new descriptors, employing more effective statistical techniques and using the modern computational capabilities, the construction of a robust and predictive QSPR model is still a challenge: the interrelation between the structure of molecules (descriptors) and properties is too complex to be easily visualized from the massive computer statistical printouts.

Within the conventional regression analysis, the structural and property information of the species is contained in a complicated manner in the QSPR model, e.g., in the regression coefficients, residuals, etc. Clearly, to be able to build better QSPR models, it is necessary to get a deeper insight into the QSPR “machinery” of regression, e.g., to find a way of partitioning the overall statistical outputs into some increments associated explicitly with the structure and properties of the species. In this work, we present a new approach to the QSPR regression analysis along this line. More specifically, we modify the traditional ordinary least squares (OLS) QSPR regression analysis so that the analogy between QSPR, stoichiometry, chemical thermodynamics, and kinetics becomes transparent. Next, we use the concept of response reactions (RERs) deduced by us from chemical thermodynamics[4] and kinetics[5] to derive a hitherto unnoticed identity for the overall statistics as well as a remarkable interpretation of the QSPR models. At this earlier stage we stress that the considerations presented below are strictly applied only to linear (in parameters) QSPR models. Furthermore, they are not meant to enhance the computational power of the existing regression methods. Rather, they represent a tool for an improved comprehension, understanding, and rationalization of the QSPR models.

NOTATION AND DEFINITIONS

Consider a chemical system comprising a training set of n chemical species $B_1, B_2, ..., B_n$. Let $y_i^{exp}(i = 1,2,...,n)$ be an experimentally determined property of the species B_i. Every species in this system is described by a set of q descriptors X_j $(j = 1,2,...,q)$. Let x_{ij} $(j = 1,2,...,q)$ be the value of the jth descriptor X_j of the species B_i. Let further $\mathbf{Y^{exp}}$ be a $(n \times 1)$ vector and $\mathbf{X}$ a $n \times (q + 1)$ matrix

$$\mathbf{Y^{exp}} = (y_1^{exp}, y_2^{exp}, ..., y_n^{exp})^{\mathrm{T}} \tag{1}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} \\ 1 & x_{12} & x_{22} & ... & x_{2q} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nq} \end{bmatrix} \tag{2}$$

It is assumed that $rank\ \mathbf{X} = q + 1$. In what follows, we consider a general linear QSPR model that usually is written in the form

$$\mathbf{Y^{exp}} = \mathbf{Xb} + \mathbf{e} \tag{3}$$

where $\mathbf{b}$ is a $(q + 1) \times 1$ vector of parameters

$$\mathbf{b} = (b_0, b_1, b_2, ..., b_q)^{\mathrm{T}} \tag{4}$$

and $\mathbf{e}$ is a $(n \times 1)$ vector of residuals

$$\mathbf{e} = (e_1, e_2, ..., e_n)^{\mathrm{T}} \tag{5}$$

As well-known,[6] the conventional OLS minimizes the product $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ thus resulting in a vector of parameters

$$\mathbf{b^{calc}} = (\mathbf{X^TX})^{-1}\mathbf{X^TY^{exp}} \tag{6}$$

and a $(n \times 1)$ vector of fitted values

* Corresponding author phone: (508)831-5445; fax: (508)831-5853; e-mail address: ifishtik@wpi.edu.

10.1021/ci0340310 CCC: $25.00
Published on Web 05/20/2003

$$\mathbf{Y}^{\mathbf{calc}} = (y_1^{calc}, y_2^{calc}, ..., y_n^{calc})^{\mathrm{T}} = \mathbf{X}\mathbf{b}^{\mathbf{calc}} \qquad (7)$$

It is to be noted that according to the conventional OLS QSPR regression analysis the vector of residuals may be evaluated only after $\mathbf{Y}^{calc}$ has been calculated

$$\mathbf{e} = \mathbf{Y}^{\mathbf{exp}} - \mathbf{Y}^{\mathbf{calc}} \qquad (8)$$

CHEMICAL STOICHIOMETRY AND QSPR

A common problem in chemical stoichiometry[7] is to generate a set of m linearly independent reactions ρ_j (j = 1,2,...,m) among a given set of species B_i

$$\boldsymbol{\rho} = \boldsymbol{\nu}\mathbf{B} = 0 \qquad (9)$$

where

$$\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_m)^{\mathrm{T}} \qquad (10)$$

$$\mathbf{B} = (\mathrm{B}_1, \mathrm{B}_1, ..., \mathrm{B}_n)^{\mathrm{T}} \qquad (11)$$

and

$$\boldsymbol{\nu} = \begin{bmatrix} \nu_{11} & \nu_{12} & ... & \nu_{in} \\ \nu_{21} & \nu22 & ... & \nu_{2n} \\ ... & ... & ... & ... \\ \nu_{m1} & \nu_{m2} & ... & \nu_{mn} \end{bmatrix}; rank\ \boldsymbol{\nu} = m \qquad (12)$$

The matrix $\boldsymbol{\nu}$ is normally referred to as *stoichiometric* matrix.[7] Usually, the only condition that needs to be satisfied when deriving a set of linearly independent chemical reactions, i.e., the stoichiometric matrix $\boldsymbol{\nu}$, is that of mass-balance. Chemical reactions that are subject to mass-balance constraints only may be, therefore, referred to as the *conventional* chemical reactions. The conventional stoichiometric formalism can be, however, readily generalized so as to define some special classes of chemical reactions. Thus, we may require a chemical reaction to preserve not only the type and number of chemical elements but also other characteristics of the species. For instance, requiring a reaction to preserve the type and number of bonds results in the so-called *isodesmic* reactions.[8] Similarly, requiring a reaction to preserve the type and number of groups results in the so-called *group-additivity* reactions.[9] Generalizing this formalism, we define here a special class of reactions that preserve the stoichiometric coefficients and the type and number of descriptors X_j (j = 1,2,...,q) of the species. Because the stoichiometric coefficients of the species in this new type of stoichiometric relations are functions of the descriptors, that is, functions of the structure of molecules, and because these reactions preserve the number and type of the descriptors, it is appropriate to call them *isostructural reactions*. Of course, the term "isostructural reactions" is somewhat misleading in that conventional chemical reactions always preserve the number and type of atoms, that is, satisfy the mass-balance conditions. By definition, the isostructural reactions are not supposed in general to conserve the type and number of atoms. Only in the special cases when the descriptors represent the number of atoms (conventional reactions), bonds (isodesmic reactions), and groups or fragments (group-additivity reactions), an isostructural reaction will also preserve the type and number of atoms.

The isostructural reactions may be generated as usual, i.e., by solving the conventional system of homogeneous linear equations

$$\boldsymbol{\nu}\mathbf{X} = 0 \qquad (13)$$

or

$$\begin{aligned} &\nu_{j1} + \nu_{j2} + ... + \nu_{jn} = 0 \\ &x_{11}\nu_{j1} + x_{21}\nu_{j2} + ... + x_{n1}\nu_{jn} = 0 \\ &x_{12}\nu_{j1} + x_{22}\nu_{j2} + ... + x_{n2}\nu_{jn} = 0; \\ &... \\ &x_{1q}\nu_{j1} + x_{2q}\nu_{j2} + ... + x_{nq}\nu_{jn} = 0 \\ &j = 1, 2, ..., m \end{aligned} \qquad (14)$$

From the dimension of this system of homogeneous linear equations it is clear that the number of linearly independent isostructural reactions is equal to $m = n - rank\mathbf{X} = n - q - 1$. Let, for instance

$$x = \begin{vmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} \\ 1 & x_{21} & x_{22} & ... & x_{2q} \\ ... & ... & ... & ... & ... \\ 1 & x_{q1} & x_{q2} & ... & x_q \\ 1 & x_{q+1,1} & x_{q+1,2} & ... & x_{q+1,q} \end{vmatrix} \neq 0 \qquad (15)$$

Then, by analogy with the conventional formalism of chemical stoichiometry[4,5,8,9] a set of m linearly independent isostructural reactions may be generated as

$$\nu_{jk} = \begin{vmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} & 0 \\ 1 & x_{21} & x_{22} & ... & x_{2q} & 0 \\ ... & ... & ... & ... & ... & ... \\ 1 & x_{k-1,1} & x_{k-1,2} & ... & x_{k-1,q} & 0 \\ 1 & x_{k1} & x_{k2} & ... & x_{kq} & 1 \\ 1 & x_{k+1,1} & x_{k+1,2} & ... & x_{k+1,q} & 0 \\ ... & ... & ... & ... & ... & ... \\ 1 & x_{q1} & x_{q2} & ... & x_q & 0 \\ 1 & x_{q+1,1} & x_{q+1,2} & ... & x_{q+1,q} & 0 \\ 1 & x_{q+j+1,1} & x_{q+j+1,2} & ... & x_{q+j+1,q} & 0 \end{vmatrix}$$

$$k = 1, 2, ..., q, q + 1$$

$$\nu_{j,q+h+1} = \delta_{j,q+h+1} \begin{vmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} & 0 \\ 1 & x_{21} & x_{22} & ... & x_{2q} & 0 \\ ... & ... & ... & ... & ... & ... \\ 1 & x_{q1} & x_{q2} & ... & x_{qq} & 0 \\ 1 & x_{q+1,1} & x_{q+1,2} & ... & x_{q+1,q} & 0 \\ 1 & x_{q+h+1,1} & x_{q+h+1,2} & ... & x_{q+h+1,q} & 1 \end{vmatrix}$$

$$= \delta_{j,q,+h+1} \begin{vmatrix} 1 & x_{11} & x_{12} & ... & x_{1q} \\ 1 & x_{21} & x_{22} & ... & x_{2q} \\ ... & ... & ... & ... & ... \\ 1 & x_{q1} & x_{q2} & ... & x_q \\ 1 & x_{q+1,1} & x_{q+1,2} & ... & x_{q+1,q} \end{vmatrix} = \delta_{j,q,+h+1}x;$$

$$h = 1, 2, ..., m \qquad (16)$$

where

$$\delta_{j,q+h+1} = \begin{cases} 1 \text{ if } j = h \\ 0 \text{ if } j \neq h \end{cases}$$

Using the properties of the determinants, the set of m linearly independent isostructural reactions obtained above may be presented in a more compact form as

$$\rho_j = \begin{vmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} & B_1 \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} & B_2 \\ \ldots & \ldots & & \ldots & \ldots & \ldots \\ 1 & x_{q+1,1} & x_{q+1,2} & \ldots & x_{q+1,q} & B_{q+1} \\ 1 & x_{q+j+1,1} & x_{q+j+1,2} & \ldots & x_{q+j+1,q} & B_{q+j+1} \end{vmatrix} = 0; \quad j = 1, 2, ..., m \qquad (17)$$

A similar formula is valid for any experimental property change of the isostructural reactions. Thus, the changes in the property $\mathbf{Y}^{\mathbf{exp}}$ of the linearly independent isostructural reactions is characterized by the following vector

$$\boldsymbol{\delta}\mathbf{Y}^{\mathbf{exp}} = (\delta y_1^{exp}, \delta y_2^{exp}, ..., \delta y_m^{exp})^{\mathrm{T}} \qquad (18)$$

where

$$\delta y_j^{exp} = \nu_{j1} y_1^{exp} + \nu_{j2} y_2^{exp} + ... + \nu_{jn} y_n^{exp}; \; j = 1, 2, ..., m \qquad (19)$$

Inserting here the stoichiometric coefficients of the isostructural reactions, eq 16, we obtain

$$\delta y_j^{exp} = \begin{vmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} & y_1^{exp} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} & y_2^{exp} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{q+1,1} & x_{q+1,2} & \ldots & x_{q+1,q} & y_{q+1}^{exp} \\ 1 & x_{q+j+1,1} & x_{q+j+1,2} & \ldots & x_{q+j+1,q} & y_{q+j+1}^{exp} \end{vmatrix}; \quad j = 1, 2, ..., m \qquad (20)$$

The isostructural reactions defined above have a very important property. Namely,

$$\boldsymbol{\nu}\mathbf{Y}^{\mathbf{calc}} = 0 \qquad (21)$$

This result follows from the combination of eqs 7 and 13. Alternatively, combining eqs 8 and 21 gives

$$\boldsymbol{\nu}\mathbf{e} = \boldsymbol{\nu}\mathbf{Y}^{\mathbf{exp}} \qquad (22)$$

or, taking into account eqs 18 and 19

$$\boldsymbol{\nu}\mathbf{e} = \boldsymbol{\delta}\mathbf{Y}^{\mathbf{exp}} \qquad (23)$$

ISOSTRUCTURAL RESPONSE REACTIONS

From general mathematical and stoichiometric considerations, it is clear that the system of homogeneous linear equations defining the isostructural reactions, eqs 13 or 14, has an infinite number of solutions. In other words, the isostructural reactions may de generated arbitrarily and that is a well-known fact in chemical stoichiometry.[7] The arbitrariness of isostructural reactions, as shown below, does not cause any troubles in the numerical QSPR regression analysis. To be able to formulate our main result, however, we define next a set of stoichiometrically unique set of isostructural reactions employing the formalism of the so-called response reactions (RERs).[4,5,8,9] The RERs were shown to have the remarkable property of being stoichiometrically unique. In particular, the RERs are independent of the way they are generated.

By analogy with the conventional, isodesmic, and group-additivity RERs we define an *isostructural RER* as an isostructural reaction that involves no more than $rank\mathbf{X} + 1 = q + 2$ species. Let $B_{i_1}, B_{i_2}, ..., B_{i_{q+1}}, B_{i_{q+2}}$ be the $q + 2$ species involved in an isostructural RER where $i_1, i_2, ..., i_{q+1}, i_{q+2}$ is an $(q + 2)$-tuple set of integers satisfying the condition $1 \leq i_1 < i_2 < ... < i_{q+1} < i_{q+2} \leq n$. Such an isostructural RER is denoted by $\theta(B_{i_1}, B_{i_2}, ..., B_{i_{q+1}}, B_{i_{q+2}})$ and its general equation is

$$\theta(B_{i_1}, B_{i_2}, ..., B_{i_{q+1}}, B_{i_{q+2}}) = \begin{vmatrix} 1 & x_{i_1 1} & x_{i_1 2} & \ldots & x_{i_{1q}} & B_{i_1} \\ 1 & x_{i_2 1} & x_{i_2 2} & \ldots & x_{i_2 q} & B_{i_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{i_{q+1} 1} & x_{i_{q+1} 2} & \ldots & x_{i_{q+1} q} & B_{i_{q+1}} \\ 1 & x_{i_{q+2} 1} & x_{i_{q+2} 2} & \ldots & x_{i_{q+2} q} & B_{1_{q+2}} \end{vmatrix} = 0 \qquad (24)$$

In a more conventional way an isostructural RER may be presented as

$$\theta(B_{i_1}, B_{i_2}, ..., B_{i_{q+1}}, B_{i_{q+2}}) = \sum_{k=1}^{q+2} \nu_{i_k}(\theta) B_{i_k} = 0$$

where

$$\nu_{i_k}(\theta) = \begin{vmatrix} 1 & x_{i_1 1} & x_{i_1 2} & \ldots & x_{i_{1q}} & 0 \\ 1 & x_{i_2 1} & x_{i_2 2} & \ldots & x_{i_2 q} & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{i_{k-1} 1} & x_{i_{k-1} 2} & \ldots & x_{i_{k-1} q} & 0 \\ 1 & x_{i_k 1} & x_{i_k 2} & \ldots & x_{i_k q} & 1 \\ 1 & x_{i_{k+1} 1} & x_{i_{k+1} 2} & \ldots & x_{i_{k+1} q} & 0 \\ \ldots & \ldots & .. & \ldots & \ldots & \ldots \\ 1 & x_{i_{q+1} 1} & x_{i_{q+1} 2} & \ldots & x_{i_{q+1} q} & 0 \\ 1 & x_{i_{q+2} 1} & x_{i_{q+2} 2} & \ldots & x_{i_{q+2} q} & 0 \end{vmatrix} \qquad (25)$$

Observe that the set of linearly independent isostructural reactions generated above are also isostructural RERs. A complete enumeration of isostructural RERs may be achieved by considering all of the possible choices of $q + 2$ species from a total of n. Hence, the total number N of isostructural RERs does not exceed

$$N = \frac{n!}{(q + 2)!(n - q - 2)!}$$

The changes in the experimental property $\mathbf{Y}^{\mathbf{exp}}$ of the isostructural RERs and denoted by $\delta y^{exp}(\theta) = \delta y^{exp}(B_{i_1}, B_{i_2}, ..., B_{i_s}, B_{i_{q+1}}, B_{i_{q+2}})$ are interrelated with the properties of the species y_i^{exp} involved in an isostructural RER via

$$\delta y^{exp}(\theta) = \nu_{i_1}(\theta) y_{i_1}^{exp} + \nu_{i_2}(\theta) y_{i_2}^{exp} + ... + \nu_{i_{q+1}}(\theta) y_{i_{q+1}}^{exp} + \nu_{i_{q+2}}(\theta) y_{i_{q+2}}^{exp}$$

or, taking into account eq 25

$$\delta y^{exp}(\theta) = \begin{vmatrix} 1 & x_{i_1 1} & x_{i_1 2} & \ldots & x_{i_1 q} & y_{i_1}^{exp} \\ 1 & x_{i_2 1} & x_{i_2 2} & \ldots & x_{i_2 q} & y_{i_2}^{exp} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{i_{q+1} 1} & x_{i_{q+1} 2} & \ldots & x_{i_{q+1} q} & y_{i_{q+1}}^{exp} \\ 1 & x_{i_{q+2} 1} & x_{i_{q+2} 2} & \ldots & x_{i_{q+2} q} & y_{i_{q+2}}^{exp} \end{vmatrix} \qquad (26)$$

Remind that by definition, eqs 14, the isostructural reactions preserve the stoichiometric coefficients and the type and number of the descriptors X_j ($j = 1,2,...,q$) of the species. Clearly, the isostructural RERs have the same properties, that is,

$$\sum_{k=1}^{q+2} \nu_{i_k}(\theta) = 0 \qquad (27)$$

$$\sum_{k=1}^{q+2} \nu_{i_k}(\theta) x_{i_k j} = 0 \text{ for any } j = 1, 2, ..., q \qquad (28)$$

From eq 21 we further deduce another important property of the isostructural RERs

$$\sum_{k=1}^{q+2} \nu_{i_k}(\theta) y_{i_k}^{calc} = 0 \qquad (29)$$

AN ALTERNATIVE APPROACH TO THE QSPR REGRESSION ANALYSIS

Next, we reformulate the QSPR regression analysis in terms of isostructural reactions. The idea is to calculate the vector of residuals **e** without any preliminary evaluations of the vector of parameters **b** and, consequently, without the regression equation. Thus, we can evaluate directly the vector of residuals **e** by minimizing $\mathbf{e}^{T}\mathbf{e}$ subject to the linear constraints given by eq 23. For this purpose we employ the method of Lagrange's undetermined multipliers and minimize the Lagrangean function

$$F = \mathbf{e}^{T}\mathbf{e} + \boldsymbol{\lambda}^{T}(\boldsymbol{\nu}\mathbf{e} - \boldsymbol{\delta}\mathbf{Y}^{\mathbf{exp}}) \qquad (30)$$

with respect to **e** and $\boldsymbol{\lambda}$. The procedure results in a system of linear equations

$$2\mathbf{e}^{T} + \boldsymbol{\lambda}^{T}\boldsymbol{\nu} = 0 \qquad (31)$$

$$\boldsymbol{\nu}\mathbf{e} = \boldsymbol{\delta}\mathbf{Y}^{\mathbf{exp}} \qquad (32)$$

It is seen that within this approach the residuals **e** may be directly evaluated without any need to evaluate the vector of parameters! It should be noticed that although the stoichiometric matrix $\vec{\nu}$ is generated arbitrarily, the solution for **e** is unique. That is, **e** is independent of the choice of $\vec{\nu}$. A proof of this statement is presented below. We also stress that there are no mathematical advantages in the above modification of the QSPR regression analysis. Moreover, as shown later on, the vector of residuals **e** obtained from eqs 31 and 32 is *equivalent* with the vector of residuals obtained by applying the conventional OLS QSPR regression analysis, i.e., given by eq 8. The importance of this approach lies in its ability to provide a remarkable interpretation of the QSPR regression analysis that is discussed below.

Since the QSPR analysis may be performed without generating the regression equation a natural question that arises in this respect is how to estimate the properties of the species from the test set, say, species B_{n+1}? The property of this species may be evaluated from any conceivable isostructural reaction involving the species B_{n+1}, including, of course, an isostructural RER. Let an arbitrary isostructural reaction involving the species B_{n+1} be

$$\rho = \sum_{i=1}^{n} \nu_i B_i + \nu_{n+1} B_{n+1} = 0 \qquad (33)$$

Because for any isostructural reaction, according to eq 21, we have

$$\sum_{i=1}^{n} \nu_i y_i^{calc} + \nu_{n+1} y_{n+1}^{calc} = 0 \qquad (34)$$

then

$$y_{n+1}^{calc} = -\frac{1}{\nu_{n+1}} \sum_{i=1}^{n} \nu_i y_i^{calc} \qquad (35)$$

Of course, this procedure is valid provided eq 35 is independent of the choice of the isostructural reaction involving the species B_{n+1}. A proof of the independence of the y_{n+1}^{calc} on the choice of the isostructural reaction is presented in Appendix A.

THE MAIN RESULT

The formal solution of eqs 31 and 32 is (Appendix B)

$$\mathbf{e} = \boldsymbol{\nu}^{T}\mathbf{g}^{-1}\boldsymbol{\delta}\mathbf{Y}^{\mathbf{exp}} \qquad (36)$$

where **g** is a square matrix of order m defined as

$$\mathbf{g} = ||g_{rs}||;\ g_{rs} = g_{sr} = \sum_{i=1}^{n} \nu_{ri}\nu_{si};\ r,s = 1, 2, ..., m \qquad (37)$$

Notice, the determinant of the matrix **g** is necessarily a positive value[4] and, therefore, **g** is nonsingular.

The mathematical form of eq 36 is precisely the same as the fundamental equations of chemical kinetics and thermodynamics[4,5] (see Appendix C for details). Based on this analogy, we are now in a position to formulate our main finding, namely

$$e_i = \frac{1}{\Delta} \sum_{\theta} \nu_i(\theta)\delta y^{exp}(\theta);\ i = 1, 2, ..., n \qquad (38)$$

where

$$\Delta = \frac{1}{m} \sum_{\theta} g(\theta) \qquad (39)$$

$$g(\theta) = \sum_{i=1}^{n} \nu_i^{2}(\theta) \qquad (40)$$

INTERPRETATION

As can be seen from eq 38, the residuals of the species may be uniquely partitioned into a sum of contributions associated with isostructural RERs. It may be remembered that an isostructural RER involves no more than $q + 2$ species where q is the number of descriptors. Consequently, the partition of the residuals into contributions associated with isostructural RERs is in fact a partition into contributions coming from all possible subsets of $q + 2$ species from a total of n. Each of these contributions has a very simple form. Thus, the contribution of every isostructural RER is given by a product of two terms. One of them is the stoichiometric coefficient $\nu_i(\theta)$ of the species B_i while the other, $\delta y^{exp}(\theta)$, is the change in the property $\mathbf{Y^{exp}}$ in a particular isostructural RER θ. Notice that the stoichiometric coefficients $\nu_i(\theta)$ of the isostructural RERs are solely functions of the descriptors (structure) while the changes in the properties of RERs $\delta y^{exp}(\theta)$ are both functions of the descriptors (structure) and property $\mathbf{Y^{exp}}$. Since the isostructural RERs are stoichiometrically unique, the independence of the solution of eqs 31 and 32 of the choice of the stoichiometric matrix ν becomes obvious.

Equation 38 is a powerful tool for the rationalization, comprehension, and interpretation of the QSPR models. A complete list of isostructural RERs, i.e., their stoichiometric coefficients $\nu_i(\theta)$, along with their changes in the property $\delta y^{exp}(\theta$, provides detailed information about the model behavior. Thus having a complete list of isostructural RERs, one can easily determine the isostructural RERs that have the smallest or highest contributions to the residuals, or, equivalently, a subset of species whose structure is highly or poorly correlated with the property.

Although the stoichiometric coefficients $\nu_i(\theta)$ and changes in the property of the isostructural RERs $\delta y^{exp}(\theta)$ incorporate detailed information about the QSPR model, for the purpose of evaluating the dominant isostructural RERs, it is useful to define a quantity that characterizes the individual contributions of the isostructural RERs. We, therefore, define the *error* $s(\theta)$ of an isostructural RER in a QSPR model as

$$s(\theta) = \sqrt{\frac{1}{n-q-1}\frac{1}{\Delta^2}\sum_{i=1}^{n}(\nu_i(\theta)\delta y^{exp}(\theta))^2} \qquad (41)$$

The reasoning behind this formula is that the quantity $\nu_i(\theta)\delta y^{exp}(\theta)/\Delta$ represents the contribution to the residual of the species B_i coming from the isostructural RER θ, while $n - q - 1$ is the number of degrees of freedom. The sum in eq 41 runs over the species involved in an isostructural RER. Bearing in mind that $m = n - q - 1$ and taking into account eqs 39 and 40 we obtain

$$s(\theta) = \frac{\sqrt{\sum_i \nu_i^2(\theta)\delta y^{exp}(\theta)^2}}{\sum_\theta\sum_i \nu_i^2(\theta)} \qquad (42)$$

The evaluation of the errors of isostructural RERs provides a useful rank of isostructural RERs. Thus, the isostructural RERs with the highest errors involve the species whose structure is poorly correlated with the property, while the isostructural RERs with the smallest errors involve the species whose structure is highly correlated with the property. In particular, the outliers, if any, are among the species involved in the isostructural RERs with the highest errors.

EQUIVALENCE OF THE TWO APPROACHES

Equation 38 is mathematically equivalent with the conventional OLS analysis. First, we observe that

$$\sum_{i=1}^{n} e_i y_i^{calc} = 0 \qquad (43)$$

whatever is the linear model. To prove it, eq 38 is substituted into eq 43

$$\frac{1}{\Delta}\sum_{i=1}^{n} y_i^{calc}\sum_\theta \nu_i(\theta)\delta y^{exp}(\theta) = \frac{1}{\Delta}\sum_\theta \delta y^{exp}(\theta)(\sum_{i=1}^{n}\nu_i(\theta)y_i^{calc})$$

The expression in the parenthesis is equal to zero by virtue of eq 29. Second, if there is a b_0 term in the model, it is also true that

$$\sum_{i=1}^{n} e_i = 0 \qquad (44)$$

Indeed

$$\sum_{i=1}^{n} e_i = \frac{1}{\Delta}\sum_{i=1}^{n}\sum_\theta \nu_i(\theta)\delta y^{exp}(\theta) = \frac{1}{\Delta}\sum_\theta \delta y^{exp}(\theta)(\sum_{i=1}^{n}\nu_i(\theta))$$

This expression is equal to zero by virtue of eq 27.

Finally we prove that the conventional QSPR and the isostructural RERs analyses are mathematically equivalent. Consider first eq 38 using a vector notation

$$\mathbf{e} = \mathbf{Y^{exp}} - \mathbf{Y^{calc}} \qquad (45)$$

where

$$\mathbf{e} = \left(\frac{1}{\Delta}\sum_\theta \nu_1(\theta)\delta y^{exp}(\theta), \frac{1}{\Delta}\sum_\theta \nu_2(\theta)\delta y^{exp}(\theta), ..., \frac{1}{\Delta}\sum_\theta \nu_n(\theta)\delta y^{exp}(\theta)\right)^{T}$$

Multiplying both parts of eq 45 by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and taking into account eq 6 we obtain

$$\mathbf{b^{calc}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y^{calc}} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \qquad (46)$$

By virtue of eqs 27 and 28 we further have

$$\mathbf{X}^T\mathbf{e} = 0 \qquad (47)$$

Thus, eq 46 takes the form

$$\mathbf{b^{calc}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y^{calc}} \qquad (48)$$

Exactly the same result is obtained if both parts of eq 7 are multiplied by $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

AN EXAMPLE

The described above formalism is next illustrated with the help of an example. Consider the normal boiling points (NBPs) of a set of six organic fluorides investigated by Balaban et al.[10] These authors developed a linear QSPR model employing two descriptors, namely, the valence connectivity indices $^0\chi^1$ [11,12] and the average distance-sum

Table 1. NBPs (°C), Valence Connectivity $^1\chi^v$, and Balaban's J_s Indices of Several Fluorides[a]

	$^1\chi^v$	J_x	NBP_{exp}	res
1. Me-F (B_1)	0.3206	0.6054	−78.0	3.04
2. Et-F (B_2)	0.6801	0.9143	−38.0	−5.74
3. Pr-F (B_3)	0.9058	1.0550	3.0	0.17
4. Bu-F (B_4)	1.0899	1.1346	33.0	−1.41
5. sBu-F (B_5)	1.0685	1.2334	25.0	3.34
6. I-C_5-F (B_6)	1.2453	1.1860	63.0	0.60

[a] Data from ref 10.

connectivity, adapted for heteroatoms based on their electronegativities (Balaban's index, J_x[13,14])

$$NBP^{calc} = -(96.8 \pm 15) + (208 \pm 23)\ {}^0\chi^1 - (84 \pm 32)\ J_x$$

The list of species, their NBPs, and the values of the descriptors as well as the results of the conventional QSPR analyses are presented in Table 1. Our starting point is the matrix **X** that in this case is

$$\mathbf{X} = \begin{array}{cc} & \begin{array}{ccc} & {}^0\chi^1 & J_x \end{array} \\ & \left(\begin{array}{ccc} 1 & 0.3206 & 0.6054 \\ 1 & 0.6801 & 0.9143 \\ 1 & 0.9058 & 1.0550 \\ 1 & 1.0899 & 1.1346 \\ 1 & 1.0685 & 1.2334 \\ 1 & 1.2453 & 1.1860 \end{array}\right) \begin{array}{c} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \end{array} \end{array}$$

Because $rank\mathbf{X} = 3$, the number of linearly independent isostructural reactions is equal to $m = n - rank\mathbf{X} = 6-3 = 3$. Consider, for instance, the determinant formed by the first three rows of $\boldsymbol{X}$

$$x = \begin{vmatrix} 1 & 0.3206 & 0.6054 \\ 1 & 0.6801 & 0.9143 \\ 1 & 0.9058 & 1.0550 \end{vmatrix} = -0.02$$

Since $x \neq 0$, according to eq 16, an appropriate set of linearly independent isostructural reactions is

$$\rho_1 = \begin{vmatrix} 1 & 0.3206 & 0.6054 & B_1 \\ 1 & 0.6801 & 0.9143 & B_2 \\ 1 & 0.9058 & 1.0550 & B_3 \\ 1 & 1.0899 & 1.1346 & B_4 \end{vmatrix} = 0$$

$$\rho_2 = \begin{vmatrix} 1 & 0.3206 & 0.6054 & B_1 \\ 1 & 0.6801 & 0.9143 & B_2 \\ 1 & 0.9058 & 1.0550 & B_3 \\ 1 & 1.0685 & 1.2334 & B_5 \end{vmatrix} = 0$$

$$\rho_3 = \begin{vmatrix} 1 & 0.3206 & 0.6054 & B_1 \\ 1 & 0.6801 & 0.9143 & B_2 \\ 1 & 0.9058 & 1.0550 & B_3 \\ 1 & 1.2453 & 1.1860 & B_6 \end{vmatrix} = 0$$

Evaluating the determinants we have

$$\rho_1 = 0.41\ B_1 - 1.89\ B_2 + 2.48\ B_3 - 1.00\ B_4 = 0$$

$$\rho_2 = -0.91\ B_1 + 1.63\ B_2 + 0.27\ B_3 - 1.00\ B_5 = 0$$

$$\rho_3 = 0.95\ B_1 - 3.97\ B_2 + 4.02\ B_3 - 1.00\ B_6 = 0$$

In what follows, we divide *all* of the isostructural reactions (as well as their characteristics) by x in order to obtain stoichiometric coefficients close to 1. Let T_i^{exp} (i = 1,2,...,6) be the experimental NBPs of the species B_i (i = 1,2,...,6). Then, the changes in the experimental NBPs of the above isostructural reactions according to eq 20 are given by

$$\delta T_1^{exp} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & T_1^{exp} \\ 1 & 0.6801 & 0.9143 & T_2^{exp} \\ 1 & 0.9058 & 1.0550 & T_3^{exp} \\ 1 & 1.0899 & 1.1346 & T_4^{exp} \end{vmatrix} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & -78.0 \\ 1 & 0.6801 & 0.9143 & -38.0 \\ 1 & 0.9058 & 1.0550 & 3.0 \\ 1 & 1.0899 & 1.1346 & 33.0 \end{vmatrix} = 13.94$$

$$\delta T_2^{exp} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & T_1^{exp} \\ 1 & 0.6801 & 0.9143 & T_2^{exp} \\ 1 & 0.9058 & 1.0550 & T_3^{exp} \\ 1 & 1.0685 & 1.2334 & T_5^{exp} \end{vmatrix} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & -78.0 \\ 1 & 0.6801 & 0.9143 & -38.0 \\ 1 & 0.9058 & 1.0550 & 3.0 \\ 1 & 1.0685 & 1.2334 & 25.0 \end{vmatrix} = -15.42$$

$$\delta T_3^{exp} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & T_1^{exp} \\ 1 & 0.6801 & 0.9143 & T_2^{exp} \\ 1 & 0.9058 & 1.0550 & T_3^{exp} \\ 1 & 1.2453 & 1.1860 & T_6^{exp} \end{vmatrix} = \begin{vmatrix} 1 & 0.3206 & 0.6054 & -78.0 \\ 1 & 0.6801 & 0.9143 & -38.0 \\ 1 & 0.9058 & 1.0550 & 3.0 \\ 1 & 1.2453 & 1.1860 & 63.0 \end{vmatrix} = 25.74$$

Now, if $e_i = T_i^{exp} - T_i^{calc} (i = 1,2,...,6)$, then we can minimize $\sum_{i=1}^{6} e_i^2$ explicitly subject to

$$0.41\ e_1 - 1.89\ e_2 + 2.48\ e_3 - 1.00\ e_4 = \delta T_1^{exp}$$

$$-0.91\ e_1 + 1.63\ e_2 + 0.27\ e_3 - 1.00\ e_5 = \delta T_2^{exp}$$

$$0.95\ e_1 - 3.97\ e_2 + 4.02\ e_3 - 1.00\ e_6 = \delta T_3^{exp}$$

According to eqs 31 and 32 this gives

Table 2. Complete List of Isostructural RERs, Their Changes in the Experimental NBPs, and Errors

Isostructural RERs	$\delta T^{exp}(\theta)$	$s(\theta)$
1. $\theta(B_1, B_2, B_3, B_4) = 0.41\ B_1 - 1.89\ B_2 + 2.48\ B_3 - 1.00\ B_4 = 0$	13.94	0.17
2. $\theta(B_1, B_2, B_3, B_5) = -0.91\ B_1 + 1.63\ B_2 + 0.27\ B_3 - 1.00\ B_5 = 0$	−15.42	0.12
3. $\theta(B_1, B_2, B_3, B_6) = 0.95\ B_1 - 3.97\ B_2 + 4.02\ B_3 - 1.00\ B_6 = 0$	25.74	0.54
4. $\theta(B_1, B_2, B_4, B_5) = -2.36\ B_1 + 4.56B_2 + 0.27\ B_4 - 2.48\ B_5 = 0$	−42.00	0.87
5. $\theta(B_1, B_2, B_4, B_6) = 0.69\ B_1 - 2.23\ B_2 + 4.02\ B_4 - 2.48\ B_6 = 0$	7.72	0.15
6. $\theta(B_1, B_2, B_5, B_6) = 3.91\ B_1 - 7.65\ B_2 + 4.02\ B_5 - 0.27\ B_6 = 0$	69.04	2.38
7. $\theta(B_1, B_3, B_4, B_5) = -1.04\ B_1 + 4.56\ B_3 - 1.63\ B_4 - 1.89\ B_5 = 0$	−6.39	0.12
8. $\theta(B_1, B_3, B_4, B_6) = 0.15\ B_1 - 2.23\ B_3 + 3.97\ B_4 - 1.89\ B_6 = 0$	−6.66	0.12
9. $\theta(B_1, B_3, B_5, B_6) = 2.05\ B_1 - 7.65\ B_3 + 3.97\ B_5 + 1.63\ B_6 = 0$	19.18	0.63
10. $\theta(B_1, B_4, B_5, B_6) = 0.86\ B_1 - 7.65\ B_4 + 2.23\ B_5 + 4.56\ B_6 = 0$	23.62	0.79
11. $\theta(B_2, B_3, B_4, B_5) = -1.04\ B_2 + 2.36\ B_3 - 0.91\ B_4 - 0.41\ B_5 = 0$	6.26	0.06
12. $\theta(B_2, B_3, B_4, B_6) = 0.15\ B_2 - 0.69\ B_3 + 0.95\ B_4 - 0.41\ B_6 = 0$	−2.58	0.01
13. $\theta(B_2, B_3, B_5, B_6) = 2.05\ B_2 - 3.91\ B_3 + 0.95\ B_5 + 0.91\ B_6 = 0$	−8.70	0.15
14. $\theta(B_2, B_4, B_5, B_6) = 0.86\ B_2 - 3.91\ B_4 + 0.69\ B_5 + 2.36\ B_6 = 0$	4.31	0.07
15. $\theta(B_3, B_4, B_5, B_6) = 0.86\ B_3 - 2.05\ B_4 + 0.15\ B_5 + 1.04\ B_6 = 0$	4.18	0.04

$$2e_1 + 0.41\,\lambda_1 - 0.91\,\lambda_2 + 0.95\,\lambda_3 = 0$$

$$2e_2 - 1.89\,\lambda_1 + 1.63\,\lambda_2 - 3.97\,\lambda_3 = 0$$

$$2e_3 + 2.48\,\lambda_1 + 0.27\,\lambda_2 + 4.02\,\lambda_3 = 0$$

$$2e_4 - 1.00\,\lambda_1 = 0$$

$$2e_5 - 1.00\,\lambda_2 = 0$$

$$2e_6 - 1.00\,\lambda_3 = 0$$

$$0.41\,e_1 - 1.89\,e_2 + 2.48\,e_3 - 1.00\,e_4 = 13.94$$

$$-0.91\,e_1 + 1.63\,e_2 + 0.27\,e_3 - 1.00\,e_5 = -15.42$$

$$0.95\,e_1 - 3.97\,e_2 + 4.02\,e_3 - 1.00\,e_6 = 25.74$$

The solution of this linear system of equations is

$$e_1 = 3.04\; e_4 = \text{-}1.41$$

$$e_2 = \text{-}5.74\; e_5 = 3.34$$

$$e_3 = 0.17\; e_6 = 0.60$$

As expected, these residuals are identical with those obtained from the Balaban's et al QSPR model.[10]

Let us now rationalize this simple model in terms of isostructural RERs. First, we generate a complete set of isostructural RERs. According to the formalism described above, an isostructural RER in this system involves no more than *rank***X** + 1 = 3 + 1 = 4 species. Thus, any four species from a total of six species define an isostructural RER. Consequently, the total number of isostructural RERs is equal to 6!/4!/2! = 15. For instance, the species B_1, B_2, B_4, and B_6 define the following isostructural RER

$$\theta(B_1, B_2, B_4, B_6) = \begin{vmatrix} 1 & 0.3206 & 0.6054 & B_1 \\ 1 & 0.6801 & 0.9143 & B_2 \\ 1 & 1.0899 & 1.1346 & B_4 \\ 1 & 1.2453 & 1.1860 & B_6 \end{vmatrix} = 0$$

or, after evaluating the determinant

$$\theta(B_1, B_2, B_4, B_6) = 0.69\ B_1 - 2.23\ B_2 + 4.02\ B_4 - 2.48\ B_6 = 0$$

Similarly, the change in the experimental NBPs of this isostructural RER is

$$\delta T^{exp}(\theta) = \delta T^{exp}(B_1, B_2, B_4, B_6) = \begin{vmatrix} 1 & 0.3206 & 0.6054 & -78.0 \\ 1 & 0.6801 & 0.9143 & -38.0 \\ 1 & 1.0899 & 1.1346 & 33.0 \\ 1 & 1.2453 & 1.1860 & 63.0 \end{vmatrix} = 7.72$$

A complete list of isostructural RERs along with their changes in the experimental NBPs and errors is presented in Table 2. With this information in hand we can now uniquely partition the residuals of the species into contributions coming from isostructural RERs. The results of such analyses are presented in Table 3. From Table 3, we immediately deduce that the most significant contributions to the residuals of the species come from five isostructural RERs, namely, $\theta(B_1, B_2, B_3, B_6)$, $\theta(B_1, B_2, B_4, B_5)$, $\theta(B_1, B_2, B_5, B_6)$, $\theta(B_1, B_3, B_5, B_6)$, and $\theta(B_1, B_4, B_5, B_6)$. It is seen that all of these dominant isostructural RERs involve the species B_1. This observation suggests that removing B_1 will substantially improve the regression results. Indeed, as can be seen from Table 4, removing B_1 results in a model that with a st. error equal to 1.64. This value should be compared with the st. error of the full model equal to 4.31. Notice, that although the residual of the species B_2 in the original model is the highest ($e_2 = -5.74$), elimination of B_2 from the model does not result in the lowest st. error! In other words, a high value of the residual of a species does not necessarily mean that that this species is an outlier.

DISCUSSION AND CONCLUDING REMARKS

From the above discussion, it follows that the conventional OLS QSPR analysis may be modified so as to explicitly minimize the residuals subject to a set of linear relations among them. This approach is similar and, to some extent, is suggested by chemical reaction thermodynamics. As well-known,[7] the equilibrium condition in a multiple chemical reaction system may be formulated in two different, yet equivalent, ways. According to the nonstoichiometric approach, the Gibbs free energy of a chemical reaction system is minimized subject to mass balance. Alternatively, the Gibbs free energy may be minimized directly if the mass-balance constrains are eliminated via a set of stoichiometrically independent reactions. While the reactions may be

Table 3. Species Contributions to the Residuals Associated with Isostructural RERs[a]

	$\delta T^{exp}(\theta)$	$s(\theta)$	e_1	e_2	e_3	e_4	e_5	e_6
1. $\theta(B_1, B_2, B_3, B_4)$	13.94	0.17	0.04	−0.17	0.22	−0.09	0.00	0.00
2. $\theta(B_1, B_2, B_3, B_5)$	−15.42	0.12	0.09	−0.16	−0.03	0.00	0.10	0.00
3. $\theta(B_1, B_2, B_3, B_6)$	**25.74**	**0.54**	**0.15**	**−0.64**	**0.65**	**0.00**	**0.00**	**−0.16**
4. $\theta(B_1, B_2, B_4, B_5)$	**−42.00**	**0.87**	**0.62**	**−1.20**	**0.00**	**−0.07**	**0.66**	**0.00**
5. $\theta(B_1, B_2, B_4, B_6)$	7.72	0.15	0.03	−0.11	0.00	0.20	0.00	−0.12
6. $\theta(B_1, B_2, B_5, B_6)$	**69.04**	**2.38**	**1.70**	**−3.32**	**0.00**	**0.00**	**1.75**	**−0.12**
7. $\theta(B_1, B_3, B_4, B_5)$	−6.39	0.12	0.04	0.00	−0.18	0.07	0.08	0.00
8. $\theta(B_1, B_3, B_4, B_6)$	−6.66	0.12	−0.01	0.00	0.09	−0.17	0.00	0.08
9. $\theta(B_1, B_3, B_5, B_6)$	**19.18**	**0.63**	**0.25**	**0.00**	**−0.92**	**0.00**	**0.48**	**0.20**
10. $\theta(B_1, B_4, B_5, B_6)$	**23.62**	**0.79**	**0.13**	**0.00**	**0.00**	**−1.14**	**0.33**	**0.68**
11. $\theta(B_2, B_3, B_4, B_5)$	6.26	0.06	0.00	−0.04	0.09	−0.04	−0.02	0.00
12. $\theta(B_2, B_3, B_4, B_6)$	−2.58	0.01	0.00	0.00	0.01	−0.02	0.00	0.01
13. $\theta(B_2, B_3, B_5, B_6)$	−8.70	0.15	0.00	−0.11	0.21	0.00	−0.05	−0.05
14. $\theta(B_2, B_4, B_5, B_6)$	4.31	0.07	0.00	0.02	0.00	−0.11	0.02	0.06
15. $\theta(B_3, B_4, B_5, B_6)$	4.18	0.04	0.00	0.00	0.02	−0.05	0.00	0.03
		sum:	3.04	−5.74	0.17	−.41	3.34	0.60
		st. error:	4.31					

[a] The dominant isostructural RERs are in bold.

Table 4. Effect of Eliminating the Species B_1 from the QSPR Model

	$\delta T^{exp}(\theta)$	$s(\theta)$	e_2	e_3	e_4	e_5	e_6
11. $\theta(B_2, B_3, B_4, B_5)$	6.26	0.42	−0.22	0.50	−0.19	−0.09	0.00
12. $\theta(B_2, B_3, B_4, B_6)$	−2.58	0.08	−0.01	0.06	−0.08	0.00	0.03
13. $\theta(B_2, B_3, B_5, B_6)$	−8.70	0.97	−0.61	1.16	0.00	−0.28	−0.27
14. $\theta(B_2, B_4, B_5, B_6)$	4.31	0.49	0.12	0.00	−0.58	0.10	0.35
15. $\theta(B_3, B_4, B_5, B_6)$	4.18	0.25	0.00	0.12	−0.29	0.02	0.15
		sum:	−0.72	1.85	−1.14	−0.25	0.26
		st. error:	1.64				

selected arbitrarily, the position of the equilibrium is (and must be!) independent of the choice of the reactions. This independence becomes obvious when the thermodynamic functions are uniquely partitioned into a sum of contributions associated with RERs. Based on this analogy and using the RERs thermodynamic formalism it appears that the QSPR regression analysis may be performed and interpreted from a quite unusual point of view. Namely, it is possible to partition the residuals into a sum of contributions associated with a unique set of specially defined isostructural RERs. As a consequence, one can build and test a QSPR model without actually deriving a regression equation.

The isostructural RERs approach is an interpretative concept. That is, the partitioning of the residuals into contributions coming from isostructural RERs provides detailed "insider" information about the QSPR model that is unavailable within the conventional OLS regression analysis. Clearly, this information may be used to get a deeper understanding of the mathematically complex and, often, musky interrelations between the structure and properties of the species and, ultimately, to construct better QSPR models. The main result of the present approach is a subset of dominant isostructural RERs. Thus, one may determine a small subset of species that are mutually poorly correlated and, hence, are responsible for a low performance of the QSPR model. These species are not always outliers, and, hence, their deletion may not improve the quality of the QSPR model. Rather, one should concentrate on these species in order to determine the particularities of their structure that caused the poor correlation and/or make an effort to find other descriptors that may improve their mutual correlation. It is mainly in this spirit that the isostructural RERs approach developed may be more effectively used.

APPENDIX A: INDEPENDENCE OF THE CALCULATED PROPERTIES OF THE SPECIES FROM THE TEST SET ON THE CHOICE OF ISOSTRUCTURAL REACTIONS

Consider two different isostructural reactions involving the *n* species $B_1,B_2,...,B_n$ from the training test and the species B_{n+1} from the test set

$$\rho' = \sum_{i=1}^{n} \nu'_i B_i + \nu'_{n+1} B_{n+1} = 0$$

$$\rho'' = \sum_{i=1}^{n} \nu''_i B_i + \nu''_{n+1} B_{n+1} = 0$$

For simplicity, let us assume that these reactions are isostructural RERs. According to eqs 27 and 29 for these reactions we have

$$\sum_{i=1}^{n} \nu'_i + \nu''_{n+1} = 0 \tag{A1}$$

$$\sum_{i=1}^{n} \nu''_i + \nu''_{n+1} = 0 \tag{A2}$$

$$\sum_{i=1}^{n} \nu'_i y_i^{calc} + \nu'_{n+1} y_{n+1}^{calc} = 0$$

$$\sum_{i=1}^{n} \nu''_i y_i^{calc} + \nu''_{n+1} y_{n+1}^{calc} = 0$$

The independence of y_{n+1}^{calc} on the choice of the isostructural reaction implies that

$$\frac{1}{\nu'_{n+1}}\sum_{i=1}^{n}\nu'_i y_i^{calc} = \frac{1}{\nu''_{n+1}}\sum_{i=1}^{n}\nu''_i y_i^{calc}$$

or

$$\sum_{i=1}^{n}\begin{vmatrix}\nu'_i & \nu'_{n+1}\\ \nu''_i & \nu''_{n+1}\end{vmatrix} y_i^{calc} = 0$$

The LHS of this equation is equal to zero by virtue of eq A1 and A2.

APPENDIX B: SOLUTION OF EQS 31 AND 32

Let

$$\mathbf{e} = [\mathbf{e_1}, \mathbf{e_2}]$$

$$\boldsymbol{\nu} = [\boldsymbol{\nu_1}, \boldsymbol{\nu_2}]$$

such that

$$\mathbf{e_1} = (e_1, e_2, ..., e_m)^{\mathrm{T}} \tag{B1}$$

$$\mathbf{e_2} = (e_{m+1}, e_{m+2}, ..., e_n)^{\mathrm{T}} \tag{B2}$$

$$\boldsymbol{\nu_1} = \begin{bmatrix}\nu_{11} & \nu_{12} & ... & \nu_{1m}\\ \nu_{21} & \nu_{22} & ... & \nu_{2m}\\ ... & ... & ... & ...\\ \nu_{m1} & \nu_{m2} & ... & \nu_{mm}\end{bmatrix} \tag{B3}$$

$$\boldsymbol{\nu_2} = \begin{bmatrix}\nu_{1,m+1} & \nu_{1,m+2} & ... & \nu_{1n}\\ \nu_{2,m+1} & \nu_{2,m+2} & ... & \nu_{2n}\\ ... & ... & ... & ...\\ \nu_{m,m+1} & \nu_{m,m+2} & ... & \nu_{mn}\end{bmatrix} \tag{B4}$$

It is assumed that the submatrix $\boldsymbol{\nu_1}$ is nonsingular. Thus, eqs 31 and 32 may be written as

$$2\mathbf{e_1}^{\mathrm{T}} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\nu_1} = 0 \tag{B5}$$

$$2\mathbf{e_2}^{\mathrm{T}} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\nu_2} = 0 \tag{B6}$$

$$\boldsymbol{\nu_1}\mathbf{e_1} + \boldsymbol{\nu_2}\mathbf{e_2} = \boldsymbol{\delta}\mathbf{Y} \tag{B7}$$

Solving eq B5 for λ

$$\boldsymbol{\lambda} = -2(\boldsymbol{\nu_1})^{-1}\mathbf{e_1} \tag{B9}$$

and substituting the result into eq B6 gives

$$\mathbf{e_2} = \boldsymbol{\nu_2}^{\mathbf{T}}(\boldsymbol{\nu_1}^{\mathbf{T}})^{-1}\mathbf{e_1} \tag{B9}$$

Substituting further eq B9 into eq B7 we have

$$[\boldsymbol{\nu_1} + \boldsymbol{\nu_2}\boldsymbol{\nu_2}^{\mathbf{T}}(\boldsymbol{\nu_1}^{\mathbf{T}})^{-1}]\mathbf{e_1} = \boldsymbol{\delta}\mathbf{Y} \tag{B10}$$

Define further the following square matrix of order m

$$\mathbf{g} = ||g_{rs}||;\ g_{rs} = g_{sr} = \sum_{i=1}^{n}\nu_{ri}\nu_{si};\ r,s = 1, 2, ..., m \tag{B11}$$

The determinant of the matrix $\mathbf{g}$ is denoted by Δ

$$\Delta = \mathrm{Det}\mathbf{g} = \begin{vmatrix}g_{11} & g_{12} & ... & g_{1m}\\ g_{21} & g_{22} & ... & g_{2m}\\ ... & ... & ... & ...\\ g_{m1} & g_{m2} & ... & g_{mm}\end{vmatrix} \tag{B12}$$

This determinant is necessarily a positive value, and, therefore, the inverse of $\mathbf{g}$ exists. We observe further that the matrices $\boldsymbol{\nu}$ and $\mathbf{g}$ are interrelated via

$$\mathbf{g} = \boldsymbol{\nu}\ \boldsymbol{\nu}^{\mathbf{T}} \tag{B13}$$

or, taking into account equations B3 and B4

$$\mathbf{g} = \boldsymbol{\nu_1}\boldsymbol{\nu_1}^{\mathbf{T}} + \boldsymbol{\nu_2}\boldsymbol{\nu_2}^{\mathbf{T}} \tag{B14}$$

and, consequently

$$\boldsymbol{\nu_2}\boldsymbol{\nu_2}^{\mathbf{T}} = \mathbf{g} - \boldsymbol{\nu_1}\boldsymbol{\nu_1}^{\mathbf{T}} \tag{B15}$$

Substituting eq B15 into eq B10 and solving for $\mathbf{e_1}$ we have

$$\mathbf{e_1} = \boldsymbol{\nu_1}^{\mathbf{T}}\mathbf{g}^{-\mathbf{1}}\boldsymbol{\delta}\mathbf{Y} \tag{B16}$$

Combination of eqs B9 and B16 gives

$$\mathbf{e_2} = \boldsymbol{\nu_2}^{\mathbf{T}}\mathbf{g}^{-\mathbf{1}}\boldsymbol{\delta}\mathbf{Y} \tag{B17}$$

Thus

$$\mathbf{e} = [\mathbf{e_1}, \mathbf{e_2}] = [\boldsymbol{\nu_1}^{\mathbf{T}}\mathbf{g}^{-\mathbf{1}}\boldsymbol{\delta}\mathbf{Y}, \boldsymbol{\nu_2}^{\mathbf{T}}\mathbf{g}^{-\mathbf{1}}\boldsymbol{\delta}\mathbf{Y}] = \boldsymbol{\nu}^{\mathbf{T}}\mathbf{g}^{-\mathbf{1}}\boldsymbol{\delta}\mathbf{Y} \tag{B18}$$

APPENDIX C: ANALOGY WITH CHEMICAL THERMODYNAMICS AND KINETICS

Consider a set of m linearly independent conventional chemical reactions, eq 9. Each of these reactions is characterized by its extent ξ_j (j = 1,2,...,m), affinity A_j (j = 1,2,...,m), and rate r_j (j = 1,2,...,m). The latter is defined as $r_j = d\xi/dt$. As well-known from chemical thermodynamics and kinetics,[15] at constant temperature and pressure, these quantities are interrelated as

$$\frac{dG}{dt} = -\mathbf{A}^{\mathbf{T}}\mathbf{r} \tag{C1}$$

$$\frac{d\mathbf{c}}{dt} = \nu^{\mathbf{T}}\mathbf{r} \tag{C2}$$

where G is the Gibbs free energy and $\mathbf{A}$, $\mathbf{r}$, and $\mathbf{c}$ are the reaction affinities, reaction rates, and species concentrations vectors

$$\mathbf{A} = (A_1, A_2, ..., A_m)^{\mathrm{T}}$$

$$\mathbf{r} = (r_1, r_2, ..., r_m)^{\mathrm{T}}$$

$$\mathbf{c} = (c_1, c_2, ..., c_n)^{\mathrm{T}}$$

In turn, the affinities and the rates are interrelated via[15]

$$-\frac{1}{RT}\frac{d\mathbf{A}}{dt} = \mathbf{Gr} \tag{C3}$$

where $\mathbf{G}$ is the Hessian matrix of the Gibbs free energy

$$\mathbf{G} = ||G_{rs}||;\ G_{rs} = G_{sr} = \frac{1}{RT}\frac{\partial^2 G}{\partial \xi_r \partial \xi_s};\ r,s = 1, 2\ ,...,\ m \tag{C4}$$

For ideal systems chemical thermodynamics provides the following result[15]

$$G_{rs} = G_{sr} = \sum_{i=1}^{n} \frac{\nu_{ri}\nu_{si}}{c_i};\ r,s = 1, 2, ..., m \tag{C5}$$

Because the determinant of the Hessian matrix **G** is necessarily a positive value eq C3 may be solved for **r** and the result substituted in eqs C1 and C2

$$\frac{dG}{dt} = \frac{1}{RT}\mathbf{A}^{\mathbf{T}}\mathbf{G}^{-\mathbf{1}}\frac{d\mathbf{A}}{dt} \tag{C6}$$

$$\frac{d\mathbf{c}}{dt} = -\frac{1}{RT}\boldsymbol{\nu}^{\mathbf{T}}\mathbf{G}^{-\mathbf{1}}\frac{d\mathbf{A}}{dt} \tag{C7}$$

It is this form of the fundamental equations of chemical thermodynamics and kinetics that allows their partition into a sum of contributions associated with the conventional RERs.[4,5] Taking into account that for $c_i = 1$ ($i = 1,2,...,n$) eq C5 becomes

$$G_{rs} = G_{sr} = \sum_{i=1}^{n} \frac{\nu_{ri}\nu_{si}}{c_i} = \sum_{i=1}^{n} \nu_{ri}\nu_{si} = g_{rs} = g_{sr}$$

and, hence

$$\mathbf{G} = \mathbf{g}$$

the mathematical analogy between eqs C6 and C7 and eq 36 is self-evident.

Note Added after ASAP Posting. This article was released ASAP on 5/20/2003 with an incorrect column head in Table 2. The correct version was posted on 5/23/2003.

REFERENCES AND NOTES

(1) Karcher, W. Basic concepts and aims of QSAR studies. In *Quantitative Structure/Activity Relationships (QSAR) in Toxicology*; Coccini, T., Giannoni, L., Karcher, W., Manzo, L., Roi, R., Eds.; Commission of the European Communities: Luxembourg, 1992; pp 5−25.

(2) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(3) Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 1: Methodology. *Drug Discov. Today* **1997**, *2*, 457−467. Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 2: Applications and Problems. *Drug Discov. Today* **1997**, *2*, 538−546.

(4) Fishtik, I.; Gutman, I.; Nagypal, I. Response reactions in chemical thermodynamics. *J. Chem. Soc. Faraday Trans.* **1996**, *92*, 3625−3532.

(5) Fishtik, I.; Datta, R. A thermodynamic approach to the systematic elucidation of unique reaction routes in Catalytic Reactions. *Chem. Eng. Sci*. **2000**, *55*, 4029−4043.

(6) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons, New York, 1998.

(7) Smith, W. R.; Missen, R. W. *Chemical Reaction Equilibrium Analysis: Theory and Algorithms*; John Wiley & Sons: New York, 1982.

(8) Fishtik, I.; Datta, R.; Liebman, J. F. Response reactions − a mathematical well-defined way to obtain accurate thermochemistry from ab initio calculations, *J. Phys. Chem. A* **2003**, *107*, 695−705.

(9) Fishtik, I.; Datta, R.; Liebman, J. F. Group Additivity Methods in Terms of Response Reactions, *J. Phys. Chem. A* **2003**, in press.

(10) Balaban, A. T.; Basak, C. S.; Mills, D. Normal boiling points of 1, *ω*-alkanedinitriles: the highest increment in a homologous series. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 769−774.

(11) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(12) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.

(13) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *80*, 399−404.

(14) Balaban, A. T. Chemical graphs. 48. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.

(15) De Donder, T.; Van Rysselberghe, P. *Thermodynamic Theory of Affinity;* Stanford University Press: Stanford, 1936.

CI0340310