CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **87** (3) (2014) 271–285.
http://dx.doi.org/10.5562/cca2402

Original Scientific Article

Integral Sign Change Problem Check in Quantitative Structure-Activity/Property Relationships: A Tutorial

Rudolf Kiralj

Technical College in Bjelovar, Trg Eugena Kvaternika 4, HR-43000 Bjelovar, Croatia
rkiralj@vtsbj.hr, rkiralj@yahoo.com

RECEIVED DECEMBER 6, 2013; REVISED JULY 11, 2014; ACCEPTED NOVEMBER 6, 2014

Abstract. Sign change problem (SCP) in multivariate Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) is the inconsistency in the direction of association between molecular descriptors and the dependent variable. Sign change is observed when the signs of the elements of the reference vector (correlation vector for the data set obtained from variable selection) are compared to the signs of the elements of all correlation and regression vectors related to a model. SCP check in this work, named Integral SCP (ISCP) check, is established to be a general effective anti-SCP procedure, consisting of five check levels. Twelve diverse QSAR/QSPR data sets from literature were tested, and performance of data sets, models and descriptors was assessed by qualitative labeling systems. Most data sets and models did not have satisfactory performance, what is discussed in terms possible data and model remedy.

Keywords: sign change, direction of association between two variables, correlation, molecular descriptor, dependent variable, multivariate model

INTRODUCTION

Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR)[1–5] are usually a multivariate (rarely a univariate) regression equation by which a macroscopic property of interest y in vector form $\mathbf{y}$, usually a measured biological activity (in QSAR) or physico-chemical property (in QSPR) of n chemicals, is modeled from $m \geq 2$ molecular descriptors which form a matrix $\mathbf{X}$. A multivariate regression equation obtained from Multiple Linear Regression (MLR), Partial Least Squares (PLS) regression or other multivariate regression method,[1,2,4,6,7] has a general form in which the vector of the predicted property y ($\hat{\mathbf{y}}$) is calculated from descriptors x_j (vectors $\mathbf{x}_j$), after regression coefficients α and β_j have been determined:

$$\hat{\mathbf{y}} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_m\mathbf{x}_m = \alpha + \sum_j \beta_j\mathbf{x}_j, \quad j = 1, 2, \ldots, m \tag{1}$$

Selected independent variables $\mathbf{x}_j$ *i.e.* molecular descriptor, have important features in relation to $\mathbf{y}$, such as interpretability, easy generation for future applications, and statistically significant correlation to $\mathbf{y}$, which can be expressed via simple linear regression for the j-th descriptor:

$$\hat{\mathbf{y}}_j = \underline{\alpha}_j + \underline{\beta}_j\mathbf{x}_j, \quad j = 1, 2, \ldots, m \tag{2}$$

Regression coefficients are $\underline{\alpha}_j = 0$ and $\underline{\beta}_j = r_j$ when the data are autoscaled. The Pearson correlation coefficient r_j for the j-th descriptor is a statistical index, which measures the degree and direction of the association of variables x_j and y. The final QSPR/QSAR model is obtained after the variable selection procedure and satisfactory performance in various statistical methods known as model validations.[1,4,8–14]

One can notice for a univariate regression (Equation 2) that

$$\text{sign}(\underline{\beta}_j) = \text{sign}(r_j) \tag{3}$$

and, naturally expects for a multivariate model that

$$\text{sign}(\beta_j) = \text{sign}(r_j) \tag{4}$$

i.e., the signs of regression coefficients from the simple (univariate) and multivariate regression equations are

the same for a particular descriptor. In general, it is expected that the signs of all regression and correlation coefficients for the *j*-th descriptor are preserved *i.e.*, they are equal to that of r_j, regardless of data set used (complete, training, external validation or other set). Sign change problem (SCP) or lack of internal inconsistency in a multivariate QSAR/QSPR modeling has been reported and discussed[15,16] as the lack of preservation of these signs. In other words, a descriptor undergoes sign changes when the signs of its correlation and regression coefficients for studied data sets are different from the sign of the respective correlation coefficient r_j for the data set that was obtained from variable selection (reference data set). A multivariate regression model is considered free of SCP when there is no sign change for all of its descriptors. The essence of the SCP check is to ensure that the *j*-th molecular descriptor x_j in a QSPR/QSAR model is physically realistic, *i.e.* it is based on real properties of pure substances. Descriptors must have statistically defined direction of correlation in all linear regressions, *i.e.* the correlation is either positive or negative. Thus, when x_j is increasing, y must either increase or decrease, it is not possible that both trends exist at the same time. This is the natural imperative for chemical problems such as synthesis of new compounds, selection among existing compounds, docking procedures and intermolecular interaction studies.

In the period from 2007 to 2009, the author of this study participated in a line of research on QSAR/QSPR model validation,[4,14,17–19] during which the sign change problem has been identified as a serious obstacle in obtaining chemically realistic regression models. As a natural consequence, a simple SCP check[15] was introduced in 2010 to be a tool for rapid SCP detection and elimination, and then extended to a more advanced SCP check[16] in 2012. SCP check has been already applied by some researchers,[20–24] whilst for other QSAR/QSPR groups it served to make them more aware of the danger of SCP.[25–28] In this work, the previous SCP check[16] is substantially extended to Integral SCP (ISCP) check. The aim of the ISCP check is to detect false or partially deficient regression models in terms of descriptors' sign changes (in further text: the sign changes), and to identify useless or ill-constructed data sets which were used to build such models, and remedy the models and data sets whenever possible. ISCP check consists of five levels, which are introduced to become a general and effective anti-SCP tool, once standard model validations are not efficient to identify, eliminate or minimize the sign changes.[15] For this purpose, ISCP check tutorial is given, twelve examples of diverse QSAR/QSPR models and data sets are tested, and the resulting performance is discussed with possible anti-SCP remedy.

METHODS

ISCP Check Levels: A Tutorial

General Remarks

If a research includes two or more reference regression data sets and models, then the ISCP procedure must be applied for each reference data set and model.

ISCP check should be performed between variable selection and model validations. Based on the performance of data set and model of interest in the ISCP procedure, researcher decides either to proceed to model validation or go back to variable selection and eventual data set modification.

For the ISCP procedure, one should provide the maximum possible number of data sets and models in relation to the model of interest. If the complete data set or split data were not used, they should be employed to build models, regardless that they could not be of the primary interest. More data sets and models based on the same variable selection give better insight into the structure of data and quality of modeling.

Level 1 – Simple SCP Check

ISCP check level 1 was the first SCP check,[15,16] carried out for fifty-two QSAR/QSPR data sets. It consists of calculating correlation vectors of descriptor – y relationships for all data sets, as well as regression vectors of the respective regression models, and comparing all these vectors to the correlation vector for the data set that resulted from variable selection. For example, if the model and the data set obtained from variable selection include all samples (n samples), then these are the reference model and the reference data set, respectively, yielding one correlation and one regression vector. Posterior data split into training set (n_t samples) and external validation set (n_e samples) produces a new regression vector and correlation vector for the training set, whilst the correlation vector for the external validation set is calculated when the set is sufficiently large (for example, having seven or more samples[15]). Then, all correlation and regression vectors are compared to the reference correlation vector element-by-element, including the regression vector of the reference model. If the complete data set was first split, variable selection was carried out and the regression model was built, then the data set and model are the reference ones, and the correlation vector is the reference vector. All other regression and correlation vectors are compared to this one in terms of the signs of their elements, meaning that complete comparison is made for each descriptor. The number or count of sign changes with respect to the reference vector gives the sign change absolute frequency, which has to be zero for a QSAR/QSPR model of acceptable performance.

Table 1. Dependence of the number of multivariate submodels, number of regression coefficients and number of regression coefficients per descriptor on the number of selected descriptors *m*

m	Submodels[(a)]	Regression coefficients	Regr. coeff. per descriptor
2	1	2	1
3	3 + 1 = 4	9	3
4	6 + 4 + 1 = 11	28	7
5	10 + 10 + 5 + 1 = 26	75	15
6	15 + 20 + 15 + 6 + 1 = 57	186	31
7	21 + 35 + 35 + 21 + 7 + 1 = 120	441	63
8	28 + 56 + 70 + 56 + 28 + 8 + 1 = 247	1016	127
9	36 + 84 + 126 + 126 + 84 + 36 + 9 + 1 = 502	2295	255
10	45 + 120 + 210 + 252 + 210 + 120 + 45 + 1 = 101	5110	511

[(a)] Total number of submodels for a given value of *m* is shown as the sum of the numbers of all *l*-variate models, where numbers of bivariate, trivariate, tetravariate, *etc.* models are added sequentially from left to right.

Level 2 – Full SCP Check

ISCP check level 2 has been carried out for three QSPR data sets and described in details previously.[16] The full SCP check consists of calculating regression vectors for all submodels, where a submodel is any model having two or more (at most *m*) selected descriptors. The obtained vectors are compared to the reference vector in the same way as in the simple SCP check. The idea of this check is to confirm the internal consistency of the model that was obtained from variable selection, the reference model: the model with *m* descriptors is always decomposed into submodels with zero sign change frequency. For *m* selected descriptors, the numbers of all combinations giving bivariate, trivariate, *etc.* *l*-variate models up to the final *m*-variate model are binomial coefficients, which can be calculated as *l*-combinations ($l > 2$) for *m* elements, or simply used as elements of the ($m + 1$)-th row of Pascal's triangle, with exception that the first two elements of this row must be discarded. Therefore, the number of all tested multivariate models is $2^m - m - 1$, obviously growing predominantly exponentially with *m*, and meaning that chances for sign changes are greatly augmented for complex regression models. Table 1 shows how the number of multivariate submodels, number of regression coefficients and the number of the coefficients per descriptor, grow with *m*. It is easy to test all submodels in case of MLR, but when a more complex regression method is in question, carrying out the same computational procedure for all submodels can be tedious. For example, PLS requires determination of the optimal number of latent variables for all submodels.

Level 3 – Extended SCP Check

The idea of this new check is to extend the full SCP check to a larger set of descriptors and so, to obtain statistically more reliable report on SCP performance the model of interest and of selected descriptors. This goal can be achieved using a descriptor pool (level 3a) or its well-defined subset (level 3b). For this purpose, two descriptor sets are formed at each level: set of selected descriptors ($\mathbf{X_S}$), and the set containing the pool or its subset with exclusion of selected descriptors ($\mathbf{X_P}$). Then, the extended SCP check is carried out for each selected descriptor to determine its respective sign change frequency. The extended SCP check for a particular descriptor consists of building MLR models: bivariate models for all variables in $\mathbf{X_P}$, then trivariate models for all combinations of two variables from $\mathbf{X_P}$ *etc.* For example, using information from Table 1, one can calculate binomial coefficients so that for $\mathbf{X_P}$ with m_P descriptors there will be m_P bivariate models, $m_P(m_P - 1)/2$ trivariate and $m_P(m_P - 1)(m_P - 2)/6$ tetravariate models for each selected descriptor from $\mathbf{X_S}$. Descriptor sign change frequency is the count of sign changes in regression coefficients for a particular selected descriptor, and not for descriptors from $\mathbf{X_P}$. The complexity of MLR models in the procedure is determined by limiting factors: large number m_P, too long calculation time, computational and memory limits, and the impossibility to build all MLR models for certain descriptor combinations (high multicollinearity, descriptors with mostly constant values *etc.*). The reference model has satisfactory performance in extended ISCP check when its sign change frequency is very small.

Level 4 – Randomization SCP Check

Many published QSAR/QSPR models cannot be well-checked at previous ISCP levels, especially when the number of calculated descriptors is small, there are no available data for all ISCP check levels (external validation set, descriptor pool or its subset), the final regression equation is univariate, and there is only one regression model published, among other difficulties. The idea of this novel SCP check for a model built for *n* samples is to overcome the difficulties, by using descriptors obtained in a large number of random permutations of

the samples' position vector **p** (its transpose is $\mathbf{p}^T$ = [1, 2, 3, …, *n*]). The random descriptors are used in combination with each selected descriptor to build bivariate MLR models. As at the SCP level 3, sign changes are counted for each selected descriptors in all models, and not for random descriptors. In this SCP check a sufficiently large number of random descriptors has to be generated, among which some appear with significant correlation to *y*, so that the sign change can be provoked in deficient QSAR/QSPR models. The maximum number of random descriptors (random vectors) used depends on *m*, *n*, descriptor distributions, descriptor - *y* bivariate distributions and other intrinsic data properties. Therefore, one should scan the reference model, by making usually a set of 10, 25, 50, 100, 200, 500, 1,000, 2,000, 5,000, 10,000, 20,000, 30,000 and much more random descriptors (N_{rv} – number of random vectors). Each set will be used to test all the selected descriptors. For each set *i.e.* for each value of N_{rv}, the correlation coefficient for randomization (ρ_{rd}), defined as the average of the absolute values of the maximum and minimum correlation coefficients for random descriptor - *y* relationships, is calculated. Another parameter is calculated for each value of N_{rv}: the relative sign change frequency for all selected descriptors (f_{SC}). Using a table with values of N_{rv}, f_{SC} and ρ_{rd}, it is possible to identify QSAR/QSPR models with satisfactory performance. If sign change appears at small values of ρ_{rd} or N_{rv}, the model has poor performance, while if sign change occurs only at high values of these parameters, then the model has a satisfactory performance. The level variant 4a is directed to find the smallest value of N_{rv}, whilst the level variant 4b identifies the smallest value of ρ_{rd}, at which sign change starts occurring constantly. Values ρ_{rd} = 0.40 and N_{rv} = 100 are proposed as reasonable empirical thresholds.

Level 5 – t- and F-tests

In a previous study[15] independent variables in QSAR/QSPR were divided into noise or "trash" variables and descriptors, depending whether their absolute values of correlation coefficients with respect to *y* were smaller or greater than the empirical threshold 0.30, respectively. To this criterion a new one is added in this work, *t*- and *F*-tests for determination of statistical significance of descriptor - *y* linear relationships (Equation 2), motivated by the fact that several QSAR/QSPR studies do not incorporate it in usual data set analysis.[15] It consists of two levels: 1) level 5a – *t*-test for $\underline{\alpha}_j$, which is rarely used test in QSAR/QSPR; and 2) level 5b, *t*-test for $\underline{\beta}_j$ and *F*-test, which are two tests not so rarely used in QSAR/QSPR. Although mathematically equivalent, *t*- and *F*-tests do not always give exactly the same results because of differences in propagation of calculation errors. In this work, qualitative labeling system for statistical significance at 95 % confidence interval, based on *p*-values, was adopted from GraphPad statistical software.[29] Variables characterized as having not statistically significant (NSS: $p > 0.10$) or not quite statistically significant (NQSS: $0.05 < p < 0.10$) relationship to *y* are considered noise variables, whilst variables with statistically significant (SS: $0.01 < p < 0.05$), very statistically significant (VSS: $0.001 < p < 0.01$) and extremely statistically significant (ESS: $p < 0.001$) relationship to *y* are considered descriptors. At level 5a, QSAR/QSPR model has poor performance if at least significant fraction of variables fails in this test, whilst the level 5b is a more rigorous criterion. It is important to emphasize that all models *i.e.*, data used for all models inspected, must be checked at level 5. This ISCP check enables identification of noise variables,[15] which can undergo sign changes (*unstable* noise appears due to data splitting or modeling) or can be stable (*hidden* and *real* noise, with significant and not significant contribution to the model, respectively). ISCP check level 5 can also point out descriptors undergoing sign change due to data splitting (*quasi* descriptor), whilst descriptors with sign changes (*anti* descriptors) are idetified at all SCP check levels 1–5. By applying the complete ISCP procedure, identified good (*real)* descriptors are ready to be used in further QSAR/QSPR analysis, and deficient (*quasi* and *anti*) descriptors can be eventually remedied with other data splitting, outlier removal, or descriptor transformation.

Additional Checks

Three additional checks are recommended joint to the ISCP procedure. First, it is the check whether there are descriptors with very small regression coefficients (Equation 1). Such descriptors have no significant contribution to the multivariate model. When using autoscaled data, β_j values are scale-independent.

Other very important test is graphical inspection of bivariate descriptor - *y* plots for all data sets to see whether there are problems with variable distribution, outliers and distinct groups of samples, non-linearity, and artificially high correlations, among others.[15] In general, graphical inspection of relationships between variables is not less important than numerical checks.[30,31]

Third check is for descriptors whose absolute values of correlation coefficients relative to *y* are significantly smaller than the threshold of 0.30, regardless of the results from the ISCP check levels 5a and 5b. It is a practical aid in QSAR/QSPR research to eliminate falsely relevant descriptors.[14,15]

Performance Qualitative Labeling Systems

ISCP check is a complex tool that provides calculation of several statistical indices for descriptors, models and regression coefficients. Usage of these indices is not very simple and therefore, performance qualitative

labeling systems are proposed. Because the labeling systems can be well explained when used for concrete examples, *i.e.* results from ISCP check levels, appropriate places are dedicated to the labeling systems in section Results and discussion. The performance qualitative labeling systems can be used in a QSAR/QSPR research, to compare various models and see which one has the best performance, in the same way as is shown for twelve data sets and models in the present work.

Data Sets and Regression Models

Data Set A

This medium-size QSPR data set was published by Kiralj and Ferreira.[17] It consists of five electronic (E_e, E_{CC}, Δ_{HL}, Q_{C2mul}, Q_{Omul}) and three structural (σ_b, σ_r, D_{CC}) descriptors. Two PLS models were constructed and used to predict δ, the ^{17}O carbonyl chemical shift in substituted benzaldehydes. A PLS model with 50 samples resulted from variable selection, and another model was based on posterior data split, with 40 and 10 samples in the training and external validation set, respectively. Descriptor pool contained 109 variables that were based on chemical knowledge of heteroaromatic compounds. An initial subset of 51 descriptors had absolute values of correlation coefficients for descriptor - *y* relationships ($|r_j|$, Equation 2) greater than 0.60, which was the basis for the ISCP check levels 3a and 3b with 101 and 43 descriptors in the matrix $\mathbf{X_P}$, respectively. The two PLS models were validated previously by various methods,[14] and inspected by the simple SCP check for the reference model, for which no sign change has been observed.[15] In this work, ISCP check levels 2–5 were applied to data set A and the models, and more comprehensive analysis of the descriptors and models in terms of sign changes is reported.

Data Set B

This small QSPR data set for 23 samples (polycyclic aromatic hydrocarbons, PAHs, including benzene) was published by Ferreira,[32] consisting of four descriptors: electronic (EA), steric (SArea) and topological (Log(*W*), X_e). Two PLS models, model 4 from variable selection and its externally validated analogue with 16 training samples,[32] were built with the purpose to predict boiling points T_b of PAHs. Model 4 was later checked by simple and full SCP checks,[16] by which no sign change has been detected. Descriptor pool consisted of 14 variables, generated from chemical knowledge of PAHs, meaning that matrix $\mathbf{X_P}$ contained 10 descriptors. Therefore, the ISCP check levels 3a, 4 and 5 were performed in this work.

Data Set C

This is a QSPR-type (more exactly: LFER, Linear Free Energy Relationship) data set of moderate size (64 samples and five descriptors), which was published by Sprunger *et al.*,[33] and used to build an MLR model (model from Equation 10)[33] to predict $\mathrm{Log}P_{x,\mathrm{CTAB/water}}$, logarithm of micellar phase-water partition coefficient of diverse solutes. Descriptors were rationally generated according to LFER theory and therefore, no additional variables existed in the descriptor pool: electronic (**E**, **S**), steric (**V**) and hydrogen bonding properties (**A**, **B**) of solutes. Only the fitting performance of the model was reported, and no standard model validations were carried out.[33] Data split into 44 training and 20 external validation samples was made, and a new MLR model was constructed by Kiralj and Ferreira and inspected by the simple SCP check,[15] and later the full SCP check.[16] Both sign checks have shown that data set C and the respective MLR models were based on sign changes. In this work, ISCP check levels 4 and 5 were carried out.

Data Set D

This is a larger QSPR data set consisting of five electrotopological descriptors of the MEDV type (Molecular Electronegativity Distance Vector descriptors: x_{15}, x_{25}, x_{26}, x_{27}, x_{36}) for 114 samples, generated by Qin *et al.*[34] and used to predict logBCF, logarithm of bioconcentration factor of diverse nonpolar organic compounds. The reference MLR model (model 2, Equation 7)[34] and its externally validated analogue (Equation 8, with 85 training and 29 external samples)[34] were constructed with very rudimentary validation. The simple and full SCP checks were carried out by Kiralj,[16] revealing the presence of sign changes in the models. Other ISCP check levels were applied in this work, but since there were no data available for the electrotopological descriptor pool (15 descriptors),[34] the ISCP check level 3 could not be performed.

Data Set E

This is a small QSAR-related (more exactly: QSAAR, Quantitative Structure-Activity-Activity Relationship) data set, containing three independent variables: electronic (LUMO) and constitutional (N_O) molecular descriptors, and biological activity (Human liver) for 23 diverse toxic chemicals. It was published by Lessigiarska *et al.*[35] and used to build an MLR model (model 8) to predict logarithm of human toxicity (HAP). The model had only rudimentary validation. Data split into 18 training and 5 external validation samples was made by Kiralj and Ferreira,[15] and a new MLR model was constructed and validated by the simple SCP check, by which sign change was not detected but a hidden noise variable was identified. Descriptor pool[35] had more than 250 variables, and only a part of it was published. In this work, the ISCP check level 3b with only 19 descriptors in $\mathbf{X_P}$, and check levels 4 and 5 were carried out.

Data Set F

This small QSAR data set consists of four electrotopo-

logical descriptors of the MEDV type (x_1, x_7, x_{29}, x_{52}) for 21 samples (cyclooxygenase-2 inhibitors), and was published by Liu *et al.*,[36] and used to build two MLR models: one based on variable selection (Equation 3), and the other one as its externally validated analogue model (Equation 4: 15 and 6 samples in the training and external validation sets, respectively).[36] The models were validated by certain methods, and used to predict 50 % drug activity in logarithmic form, pIC_{50}. Descriptor pool of 91 electrotopological descriptors and its subsets were not published and therefore, in this work only the ISCP check level 3 could not be carried out.

Data Set G

This is a larger QSAR data set for 153 polar narcotics from the phenol class, with 50 % toxic activity against *T. pyriformis* in logarithmic form ($\log 1/IGC_{50}$), as published by Aptula *et al.*[37] for classification purposes. It contains five molecular descriptors: electronic (E_{homo}, E_{lumo}) and constitutional (N_{hdon}) descriptors, lipophilicity ($\log K_{ow}$) and basicity (pK_a). The first MLR model with certain model validation and modest external validation was published by Yao *et al.*[38] More rigorous external validation was carried out by Kiralj and Ferreira,[14] by building two more models, with 75 and 78 samples in the training and external sets and vice versa (roles of the sets were exchanged), and validating with other standard model validations. The simple SCP check for the data set with 75 samples and the respective MLR model[15] has revealed the presence of sign changes, as well as problematic descriptors with no statistically significant relationship to the dependent variable. The original descriptor pool, consisting of seven descriptors generated from chemical knowledge of phenols, was not published.[37] In this work, ISCP check levels 2, 4 and 5 were carried out.

Data Set H

This QSAR data set is small, containing four descriptors (steric: M_{04} and M_{11}; shadow: S_6'; and shadow-structural: P_{5X}) for 21 oral progesterones with progestational activity relative to norethisterone (IC_{50}), as published by Kiralj and Ferreira.[39] It was used to build a PLS model (model Id),[39] which was validated by certain methods. The original publication[39] contained descriptor pool (33 descriptors) and its subset related to model Id (15 descriptors). In this work, all the ISCP check levels were carried out, including the check levels 3a and 3b with 31 and 11 descriptors in the matrix $\mathbf{X_P}$, respectively.

Data Set I

This is a QSPR data set of moderate size, containing eight simple descriptors (S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8), indicator variables with values 0 and 1 for the absence and presence of chloro-substituents, respectively, in 62 polychlorinated naphthalenes, as published by Yin *et al.*[40] It was employed to build an MLR model (model M1/M2, Equations 2, 3)[40] for prediction of a chromatographic retention index (***RI***). The model was validated only by leave-one-out cross-validation. In this work, all the ISCP check levels were performed, with exception of the level 3 due to the lack of the descriptor pool. This type of data set, indicator variables that have only two distinct values, has been discussed previously as not recommendable for QSAR/QSPR studies.[14,15] Data set I is validated in this work because such data set type still appear in current QSAR/QSPR literature.

Data Set J

This QSPR data set was recently published by Ahmadi.[41] It consists of four descriptors of topological (IC5, LP1) and steric (E1v, RDF125m) nature, which were used to predict logarithm of the association constant ($\log K$) of 53 macrocycles with sodium cation, via an MLR model with rudimentary validation. Data set J is the only one in this work as an example in which the reference data set *i.e.*, the data set obtained from variable selection, has been obtained after data splitting of an essential descriptor pool subset. Therefore, the MLR model with 40 training and 13 external validation samples is the reference one, and is tested with the ISCP procedure. The descriptor pool (more than 350 descriptors) was not published.[27] MLR model using the complete data set (53 samples) is built in this work, and was inspected at the ISCP check levels 1, 2, 4 and 5.

Data Set K

This is an example of a larger QSPR data set, with ten TOPS-Mode (Topological Substructural Molecular Design) descriptors (μ_1^{Hyd}, μ_3^{Van}, $\mu_1\mu_2^{Std}$, μ_{10}^{Std}, μ_5^{Ab-R2}, μ_1^{Dip2}, $\mu_1\mu_4^{Dip4}$, $\mu_4^{Ab-\log L16}$, $\mu_4^{Ab-\Sigma\beta 2O}$, μ4Pols) for 232 samples, organic compounds from at least fifteen diverse classes. It was generated by Pérez-Garrido *et al.*[42] and used to build an MLR model to predict logarithm of the stability constants ($\log K$) of β-cyclodextrin with diverse chemicals. The authors presented only an externally validated model (185 training and 47 external samples) that was extensively validated. It is unclear what was the exact size of the descriptor pool.[28] The check level 1 for this data set was reported previously.[15] The model for the complete set *i.e.*, that one resulting from variable selection was not published and therefore. It is constructed in the present work and considered as the reference model. The model is tested by all SCP check levels except for the levels 3a and 3b due to the lack of the descriptor pool and its subsets.

Data Set L

This is even a larger QSAR data set, consisting of four descriptors (constitutional: nX, nCaH; topological: CIC0; and electronic: HOMO) for 460 diverse volatile organic chemicals. It was published by Gramatica *et al.*,[43] and used to build an MLR model to predict logarithm of the rate constant for hydroxyl radical tropo-

spheric degradation of chemicals, –log k(OH). This data set is somewhat similar to data set K: the reference model was not reported by the authors although it was obtained from variable selection and, the published model (model 4)[43] was based on data split (234 training and 236 external samples) and extensively validated. The reference model was tested previously[15] at the ISCP check level 1. In this work, ISCP check levels 2, 4a, 4b, 5a and 5b are performed, with addition of a new model with switched roles of the split subsets (*i.e.* 236 training and 234 external samples), as has been done for data set G. Descriptor pool of 1308 descriptors was not published.[43]

Computational Procedures

Data were autoscaled prior to any calculation, and then used to reproduce selected PLS and MLR models from the literature. The choice of PLS and MLR is justified by the fact that these are the two commonest regression methods employed in QSAR/QSPR.[44] Regression and other statistical analyses as well as diverse calculations were done by using programs Pirouette (version 4.0)[45] and Scilab (version 5.4.0),[46] and statistical significance in *t*- and *F*-tests was checked by online software GraphPad.[29]

For generating random vectors, `grand` function in Scilab was used. The sign change of regression coefficients β_{ij} of the *j*-th descriptor (Equation 1) in ν regressions (i = 1, 2, ..., ν) was counted by introducing the sign matrix with elements $S_{ij} = \mathrm{sign}(\beta_{ij})$, and then using a simple formula:

$$\frac{v - \sum_i S_{ij}}{2} \quad \text{if} \quad r_j > 0 \quad \text{or} \quad \frac{v + \sum_i S_{ij}}{2} \quad \text{if} \quad r_j < 0 \qquad (5)$$

Determination of thresholds in the random SCP check, depending mainly on the number of samples *n*, was carried out as illustrated in Figure 1. It is visible that the correlation coefficient ρ_{rd} has an asymptotic-like behavior for each data set, *i.e.* after a certain region of values of the number of randomized vectors N_{rv} it grows very slowly. For large *n*, it is practically impossible to pass the threshold ρ_{rd} = 0.40 without a large computational time and memory expense. On the other hand, small data sets easily pass the threshold at very small values of N_{rv}. More intuitive plots illustrate the nature of the ISCP check level 4, such as the sign change count (absolute sign change frequency *N*SC) depending on N_{rv} (Figure 2) and on ρ_{rd} (Figure 3), or relative sign change frequency (f_{SC}) depending on N_{rv} (Figure 4). It is visible that models based on problematic data sets exhibit sign changes even for small number of randomized vectors, and this trend is emphasized in linear (Figure 2) and vertical asymptotic form (Figure 3), independently of

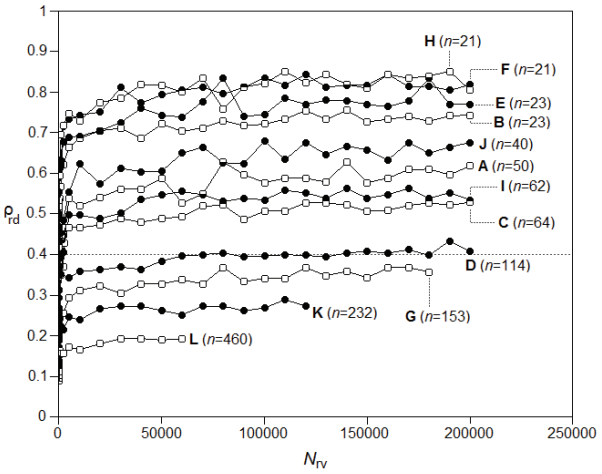

Figure 1. Dependence of the correlation coefficient for randomization (ρ_{rd}) on the number of random vectors (N_{rv}) and number of samples in data sets (*n*). Threshold ρ_{rd} = 0.40 is marked by a dotted line.

the number *n*. In terms of f_{SC}, sign changes stabilize only at high values of N_{rv} (Figure 4). Figures 1 – 4 show no regularity with respect to the number of selected descriptors *m*, which varies from 3 to 8 for data sets A – L. Nine smaller data sets (A – F, H – J) were tested up to N_{rv} = 200,000, whilst for larger data sets (G, K, L) the maximum values of N_{rv} were smaller due to computational time and memory limits.

RESULTS AND DISCUSSION

Performance of Data Sets and Models in the ISCP Check

General Considerations

A complete example of carrying out the ISCP procedure is given for data set A in Tables S1–S15 in Supplemental Material. ISCP check results for this data set are organized in two tabular forms: summary of data set and model statistics, and summary of descriptor statistics, as shown in Tables 2 and 3, respectively. Statistics summary for other data sets is in Tables S16 – S37 in Supplemental Material. In general, data set and model performance in the ISCP procedure should be used to decide about the tested data set, model of interest and descriptors, as is shown in Tables 1 – 6: to go to the next step (model validation) or go back to variable selection and eventual data modification.

The data set and model statistics summary consists of three types of statistics for each ISCP check level: model statistics, descriptor statistics, and regression coefficient statistics. Each statistics can be expressed in two equivalent forms, as the relative sign change frequency *i.e.*, fraction (sign change count)/(total count) for models, descriptors and regression coefficients, and

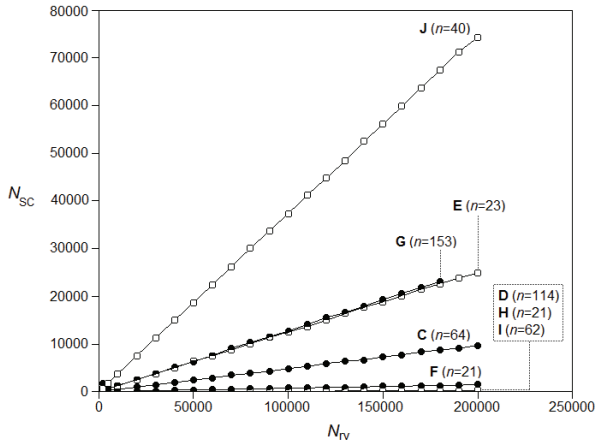

Figure 2. Dependence of the number (absolute frequency) of sign changes (N_{SC}) on the number of random vectors (N_{rv}). The number of sign changes is independent on the number of samples in data sets (n). Four data sets with zero sign changes frequencies (A, B, K, L) are not presented.

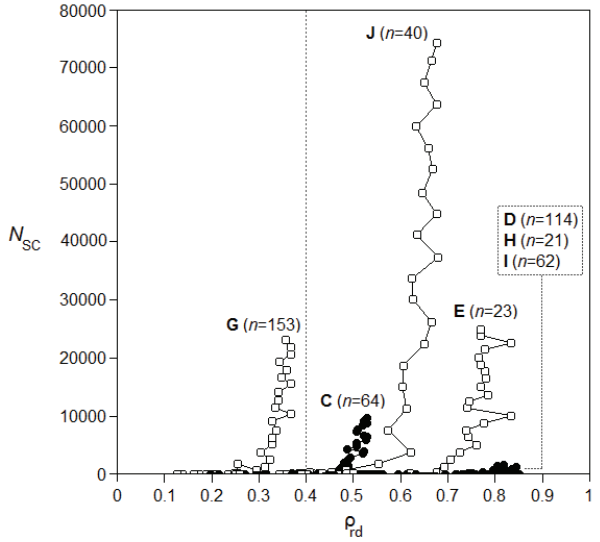

Figure 3. Dependence of the number (absolute frequency) of sign changes (N_{SC}) on the correlation coefficient for randomization (ρ_{rd}). The number of sign changes is independent on the number of samples in data sets (n). Threshold $\rho_{rd} = 0.40$ is marked by a dotted line. Four data sets with zero sign changes frequencies (A, B, K, L) are not presented.

also as the percentage value (given in brackets). Only for the ISCP check level 4b another quantity is reported, the value of ρ_{rd} at which sign change starts occurring regularly. Among the three statistics, descriptor statistics gives an overall sign change appearance, whilst model statistics coincides with regression coefficient statistics for the ISCP check levels 3a, 3b and 4a. Zero sign change count in the ISCP check levels 1, 2 and 4 is a requirement for real models with satisfactory performance. However, it is practically impossible to expect that no sign change will occur at the ISCP levels 3a and

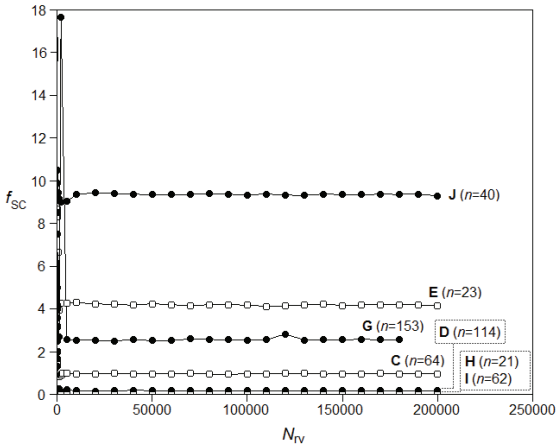

Figure 4. Dependence of the relative frequency of sign changes (f_{SC}, expressed as percentage) on the number of random vectors (N_{rv}). The relative frequency of sign changes is independent on the number of samples in data sets (n). Four data sets with zero sign changes frequencies (A, B, K, L) are not presented.

3b and therefore, up to 10 % relative sign change frequency is reasonable to tolerate for models with satisfactory performance (Table 4).

The most quantitative measure for sign changes in Table 2 is the regression coefficient statistics, which can be expressed in the form of sign change fractions for each test (S_1, S_2, S_{3a}, S_{3b}, S_{4a}, S_{5a}, S_{5b}), whilst S_{4a} is the value of ρ_{rd} from the ISCP level 4a. Values of these indices (Table 2) can be used to characterize data set and model performance, by introducing for example, a five-level performance qualitative labeling system (Table 4): excellent, good, acceptable, poor and extremely poor performance, which correspond to scores 5, 4, 3, 2 and 1, respectively.

Descriptor statistics (Table 3) can aid in deciding about the data set and model at descriptor level, such as excluding, replacing, or transforming descriptors or even excluding samples, carrying out new variable selection, making new data split, excluding outliers, among other actions. Results from each ISCP check level and visual check are taken into account, reporting performance of every descriptor as well as of the descriptor set. For descriptor set performance in visual check,[15] the poorest descriptor performance can be used as a rigorous criterion: problematic (problematic bivariate distribution), acceptable (some changes may improve the distribution), and excellent (no need for change), with scores 1, 2, and 3, respectively. The values of descriptor performance, expressed as percentage sign change frequency for all ISCP check levels (except for the level 4b where two parameters are reported – N_{rv} and ρ_{rd}), can be used together with the visual check performance to finally characterize descriptors by single scores (applying the qualitative labeling system in Table 4).

Table 2. Data set A and its model statistics[(a)] in terms of sign changes

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	0/8 (0 %)	0/2 (0 %)	0/40 (0 %)	$S_1 = 0$	excellent
2 (submodels;8)	7/8 (87.5 %)	34/247 (13.8 %)	43/1016 (4.2 %)	$S_2 = 0.042$	poor
3a (pool;2;101)	2/8 (25.0 %)	6/808 (0.7 %)[(b)]	6/808 (0.7 %)[(b)]	$S_{3a} = 0.007$	good
3b (subset;4;43)	2/8 (25.0 %)	4887/106296 (4.6 %)[(b)]	4887/106296 (4.6 %)[(b)]	$S_{3b} = 0.046$	acceptable
4a (500;0.43)	0/8 (0 %)	0/4000 (0 %)[(b)]	0/4000 (0 %)[(b)]	$S_{4a} = 0$	excellent
4b (200,000;0.63)	>>0.63	>>0.63	>>0.63	$S_{4b} >> 0.63$	excellent
5a(*t*-test;0.05)	2/8 (25.0 %)	1/2 (50.0 %)	2/24 (8.3 %)	$S_{5a} = 0.083$	acceptable
5b(*t*-/*F*-test;0.05)	0/8 (0 %)	0/2 (0 %)	0/24 (0 %)	$S_{5b} = 0$	excellent

[(a)] ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (8), which is the maximum complexity of the multivariate model considered. Level 3a: the maximum complexity (*l* value) of the *l*-variate MLR models considered (2) that could be treated computationally; number of all descriptors excluding the selected descriptors (101). Level 3b: the maximum complexity (*l* value) of the *l*-variate MLR models considered (4) that could be treated computationally; number of descriptors used for testing (43). Level 4a: number of random descriptors (generating 500 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (0.43). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (200,000); the minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (0.63). Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>"only when probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit *i.e.* close to 1).

[(b)] Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases.

Weight for scores for less strict tests (ISCP levels 3a, 3b and 5a) is 0.5, otherwise is 1 (see Table 4). For example, the total score for descriptor E_e from data set A (Table 3) is calculated as the sum of products (score × weight) to which the visual performance score is added: 5×1 (ISCP 1) + 4×1 (ISCP 2)+ 5×0.5 (ISCP 3a) + 4×0.5 (ISCP 3b) + 5×1 (ISCP 4a) + 5×0.5 (ISCP 4b: ρ_{rd} performance) + 5×0.5 (ISCP 4b: N_{rv} performance) + 5×0.5 (ISCP 5a) + 5×1 (ISCP 5b) + 2 (visual check) = 33. When this calculation is carried out for a perfect descriptor (*i.e.* descriptor with the best possible performance), one gets the value of 35.5. In terms of ideal performance, descriptor E_e has relative score of 33/35.5 = 0.93 or 93 %, which can be used to assess the risk of using this descriptor in further modeling. Assuming a new label scheme for the total score (no risk to very low risk: 96 % – 100 %; low risk: 86 % – 95 %; moderate risk: 76 % – 85 %; and high to very high risk: ≤ 75 %), descriptor E_e can be characterized as a low risk descriptor. Calculating the risk scores for other descriptors, one finds that data set A contains one descriptor with none to very low risk, five low risk descriptors, and two moderate risk descriptors (Table 3). In general, a severe action must be made about all high
risk descriptors, and probably a similar action is necessary for moderate risk descriptors, whilst descriptors with lower risk can stay untouched. Data sets with incomplete ISCP validation are somewhat problematic for descriptor risk assessment. In such cases, the total descriptor score can be expressed as a range, with its minimum value (based on results from performed ISCP check levels) and the maximum value (the minimum value is augmented with contributions from missing check levels for a hypothetical, perfect descriptor). For example, data set D could not be checked at the ISCP levels 3a and 3b due to the lack of the descriptor pool and its subsets and therefore, for descriptor x_{15} the minimum score was found 29.5, and the maximum (supposing the perfect descriptor performance in check levels 3a and 3b) was obtained 34.5. Thus, the relative score range is 83 % – 97 %, and its mean (90 % of the maximum possible score) corresponds to a low risk (Table S21 in Supplemental Information).

Summary of all data sets and models statistics is given in Table 5, and summary of all descriptors statistics is in Table 6, as based on data set and model analyses analogue to those in Table 2 and Table 3, respectively (analyses for data sets B – L are shown in Supplemental Material). At first, it is visible that most data sets could not be tested at the ISCP check levels 3a and 3b due to the lack of descriptor pools or pool subsets, and in some cases external validation sets were missing. Table 5 reports data set and model performance with parameters $S_1 - S_{5b}$ in brackets, which agrees well with model characterization from visual check of descriptor - *y* scatterplots (penultimate column). Data set score (last column) is obtained as a sum of products (score × weight) for all ISCP check levels, using the performance qualitative labeling system from Table 4, to which the visual check score is added. For example, the score for data set A (Table 5) is obtained in the following calculation: 5×1 (ISCP 1) + 2×1 (ISCP 2)+ 4×0.5

Table 3. Descriptor statistics[a] in terms of sign changes for data set A and its models.

Level	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	Total
1	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/5 (0 %)	0/40 (0 %)
2	1/127 (0.8 %)	15/127 (11.8 %)	0/127 (0 %)	1/127 (0.8 %)	23/127 (18.1 %)	1/127 (0.8 %)	1/127 (0.8 %)	1/127 (0.8 %)	43/1016 (4.2 %)
3a	0/101 (0 %)	1/101 (0.9 %)	0/101 (0 %)	0/101 (0 %)	5/101 (5.0 %)	0/101 (0 %)	0/101 (0 %)	0/101 (0 %)	6/808 (0.7 %)
3b	33/13287 (0.2 %)	638/13287 (4.8 %)	29/13287 (0.2 %)	186/13287 (1.4 %)	2925/13287 (21.0 %)	0/13287 (0 %)	97/13287 (0.7 %)	979/13287 (7.4 %)	4887/106296 (4.6 %)
4a	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/500 (0 %)	0/4000 (0 %)
4b[b],[c]	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>200000 >>0.63	>>1800000 >>0.63
5a	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	1/3 (33.3 %)	1/3 (33.3 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	2/24 (8.3 %)
5b	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/3 (0 %)	0/24 (0 %)
Visual check	acceptable	acceptable	acceptable	acceptable	acceptable	acceptable	acceptable	acceptable	acceptable
Total score[d]	33 (93 %)	30 (85 %)	34 (96 %)	31 (87 %)	27.5 (77 %)	32.5 (92 %)	32 (90 %)	31 (87 %)	–
Risk[e]	low	moderate	none to very low	low	moderate	low	low	low	–

[a] Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for the level 4b for which ρ_{rd} value is reported (marked with sign ">>" only when probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit *i.e.* close to 1).

[b] The minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude).

[c] The minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρrd reached if sign change has not been observed (reported as a lower limit >>0.63, meaning that the value of ρ_{rd} must be far from the lower limit *i.e.* close to 1).

[d] Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5.

[e] Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modeling. Total risk means the risk of taking into account all selected descriptors.

(ISCP 3a) + 3×0.5 (ISCP 3b) + 5×1 (ISCP 4a) + 5×0.5 (ISCP 4b) + 3×0.5 (ISCP 5a) + 5×1 (ISCP 5b) + 2 (visual check for descriptor set) = 27. When this calculation is carried out for a hypothetical, perfect data set (*i.e.* data set with the best possible performance), one obtains 35.5. Therefore, data set A has relative score of 27/35.5 = 0.76 or 76 % of the maximum possible score. The score for data set D is given as an interval, for which the lowest value is obtained from in calculation for the performed ISCP check levels (11.5), and the highest value is the sum of the lowest value and the sum of maximum values for the ISCP levels 3a and 3b (5.0), yielding the relative score 32 % – 46 % with the mean of 39 %. As a simple rule, data sets with relative score greater than 75 % can be considered as worth to be considered for further modeling.

Data sets can be characterized also by the average descriptor score and overall risk of using descriptors, as is shown in Table 6, where descriptors are classified according to their risk levels. The average descriptor score is obtained simply from descriptor scores (for data set A these values are shown in Table 3), and the same five-level risk system can be applied to the average descriptor score, with certain corrections when necessary. These corrections take into account the presence of one or more descriptors with the poorest possible performance, what was not adequately included in the performance qualitative labeling system (Table 4). The poorest possible performance includes: sign change frequency of 75 % – 100 % (ISCP check levels 1, 2, 3a, 3b and 4a), extremely small values of N_{rv} (N_{rv} < 25, ISCP check level 4b), and extremely poor *t*- and *F*-test performance (no statistical significance in 100 % cases, ISCP check levels 5a and 5b). For example, the value of the average descriptor score for data set F is 0.86, what would correspond to low overall risk of using

Table 4. Performance qualitative labeling system for all ISCP check levels.[(a),(b)]

Qualitative label	ISCP 1, ISCP 2	ISCP 3a, ISCP 3b	ISCP 4a	ISCP 4b	ISCP 5a	ISCP 5b	Score
excellent	0 [0 %]	0 [0 %]	0 [0 %]	> 0.6 [ρ_{rd} > 0.6, N_{rv} > 10000]	0 [0 %]	0 [0 %]	5 [5]
good	< 0.001 [≤ 0.5 %]	< 0.01 [≤ 1 %]	< 0.0001 [≤ 0.1 %]	0.5 – 0.6 [ρ_{rd}: 0.5 – 0.6, N_{rv} > 1000 – 10000]	< 0.01 [≤ 1 %]	< 0.001 [≤ 0.5 %]	4 [4]
acceptable	0.001 – 0.01 [0.5 % – 5 %]	0.01 – 0.1 [1 % – 10 %]	0.0001 – 0.001 [0.1 % – 1 %]	0.4 – 0.5 [ρ_{rd}: 0.4 – 0.5, N_{rv} > 100 – 1000]	0.01 – 0.1 [1 % – 10 %]	0.001 – 0.01 [0.5 % – 5 %]	3 [3]
poor	0.01 – 0.1 [5 % – 25 %]	0.1 – 0.3 [10 % – 50 %]	0.001 – 0.01 [1 % – 5 %]	0.3 – 0.4 [ρ_{rd}: 0.3 – 0.4, N_{rv} > 25 – 100]	0.1 – 0.5 [10 % – 50 %]	0.01 – 0.1 [5 % – 25 %]	2 [2]
extremely poor	> 0.1 [> 25 %]	> 0.3 [> 50 %]	> 0.01 [> 5 %]	< 0.3 [ρ_{rd} < 0.3, N_{rv} < 25]	> 0.5 [> 50 %]	> 0.1 [> 25 %]	1 [1]
Weight	1 [1]	0.5 [0.5]	1 [1]	1 [1 or 0.5 + 0.5]	0.5 [0.5]	1 [1]	–

[(a)] For the ISCP check levels 1, 2, 3a, 3b, 3c, 4a, 4b, 5a and 5b regression coefficient performance indices are S_1, S_2, S_{3a}, S_{3b}, S_{4a}, S_{4b}, S_{5a}, S_{5b}, respectively, where S_{4b} is the value of ρ_{rd} at which sign change starts appearing continuously with the increase of the number of random vectors, and other indices are fractions of regression coefficients of selected descriptors with sign changes. The values of these indices are given as out of brackets.

[(b)] For the ISCP check levels 1, 2, 3a, 3b, 3c, 4a, 5a and 5b percentage sign change frequencies for descriptors are given in brackets. For the ISCP 4b level the values of ρ_{rd} and N_{rv} at which sign changes start occurring continuously are given in brackets. For both values weight is 0.5, whilst at all other ISCP levels the weight is always 1.

descriptors. However, there are three descriptors (underlined in Table 6) with the poorest possible performance at the ISCP check level 5b (Table S25 in Supplemental Information), and therefore, the estimated overall risk must be shifted by one level down, *i.e.* to moderate risk. The proposed performance qualitative labeling systems and their usage should be understood as an aid to find the best regression model in a QSAR/QSPR study, and they might be refined in future studies.

Performance of Data Sets and Models in the ISCP Check

The final scores for data sets and models (Table 5) and descriptors with the overall risk (Table 6) show that the best performance data sets and models are A, B and H. Among them, data set B has the best performance, with no necessity for any change in descriptors. Data set H has one moderate risk descriptor (M_{11}) which has, together with another one (S_6'), the poorest possible performance at the ISCP check level 5a. This deficiency could be repaired, by which the published model would be refined. Data set A has satisfactory performance at all ISCP check levels except at the level 2, and this is probably due to many descriptors used (eight) and bimodal normal distribution of y.[17] The reported reference model could be refined by reducing the number of descriptors and perhaps including a new descriptor, once the model was validated with several standard model validations and additional checks known at the time of its publication.[14,15,17]

All other data sets, *i.e.* C – G, I – L are problematic, having at least one moderate or high risk descriptor (Table 6), unsatisfactory visual performance and relatively low score (Table 5). In such cases, it is wise to keep only descriptors that have no risk, very small or small risk for regression modeling, add new descriptors via variable selection, and then test a new model carrying out the complete ISCP check.

It can be noted from statistics summaries in Tables 5 and 6 that the sign change frequency does not depend on data set size, *i.e.* on the number of selected descriptors (m) and the number of samples for the reference model (n). Summary of data set and model characterization (Table 5) and summary of descriptor characterization (Table 6) are very useful diagnostics, because ISCP acts as a model validation which indicates what should be the next step after a model has been constructed *i.e.*, either model validation or a new variable selection. High and even moderate risk descriptors in multivariate modeling should be removed or replaced, or eventually modified. One should also inspect particular ISCP check level performances, such as those for data set A (Tables 2 and 3), before the final decision about the data set, model and descriptors.

Finally, variable selection should end when all selected descriptors show low, very low or no risk to multivariate modeling, and the models have satisfactory performance: excellent performance in all ISCP check levels, with exception of ISCP checks 3a and 3b, in which the performance should be at least good or acceptable (Table 4).

Table 5. Summary of all data set and model statistics in terms of sign changes[(a)]

Data set	ISCP 1	ISCP 2	ISCP 3a	ISCP 3b	ISCP 4a	ISCP 4b	ISCP 5a	ISCP 5b	Visual[(b)]	Score[(c)]
A	excellent [0]	poor [0.042]	good [0.007]	acceptable [0.046]	excellent [0]	excellent [>>0.63]	acceptable [0.083]	excellent [0]	acceptable	27 (76 %)
B	excellent [0]	excellent [0]	poor [0.120]	NA	excellent [0]	excellent [>>0.75]	poor [0.250]	excellent [0]	acceptable	29 – 31.5 (82 % – 89 %)
C	extr. poor [0.150]	extr. poor [0.200]	NA	NA	poor [0.009]	extr. poor [0.21]	excellent [0]	extr. poor [0.400]	problematic	11.5 – 16.5 (32 % – 46 %)
D	extr. poor [0.160]	poor [0.042]	NA	NA	poor [0.002]	extr. poor [0.25]	excellent [0]	extr. poor [0.200]	problematic	11.5 – 16.5 (32 % – 46 %)
E	excellent [0]	excellent [0]	NA	good [0.003]	extr. poor [0.083]	acceptable [0.43]	extr. poor [1.00]	extr. poor [0.333]	problematic	19.5 – 22 (55 % – 62 %)
F	excellent [0]	excellent [0]	NA	NA	excellent [0]	excellent [0.61]	excellent [0]	extr. poor [0.750]	problematic	24.5 – 29.5 (69 % – 83 %)
G	extr. poor [0.267]	extr. poor [0.293]	NA	NA	extr. poor [0.025]	extr. poor [0.13]	poor [0.400]	extr. poor [0.600]	problematic	9 – 14 (25 % – 39 %)
H	excellent [0]	excellent [0]	acceptable [0.034]	acceptable [0.068]	excellent [0]	excellent [0.76]	poor [0.500]	excellent [0]	problematic	31.5 (89 %)
I	excellent [0]	acceptable [0.002]	NA	NA	excellent [0]	acceptable [0.50]	excellent [0]	excellent [0.375]	problematic	21.5 – 26.5 (61 % – 75 %)
J	poor [0.10]	excellent [0]	NA	NA	extr. poor [0.085]	extr. poor [0.25]	acceptable [0.083]	extr. poor [0.333]	problematic	13.5 – 18.5 (38 % – 52 %)
K	extr. poor [0.120]	extr. poor [0.486]	NA	NA	excellent [0]	excellent [>>0.29]	excellent [0]	extr. poor [0.167]	problematic	16.5 – 21.5 (46 % – 61 %)
L	extr. poor [0.125]	poor [0.091]	NA	NA	excellent [0]	excellent [>>0.19]	excellent [0]	poor [0.083]	problematic	18.5 – 23.5 (52 % – 66 %)

[(a)] Sign change statistics expressed as data set and model performance at all ISCP check levels: a) in qualitative manner (excellent, negligible for slight negative performance, tolerable, bad and very bad) and quantitative manner (indices for each ISCP check level). For the ISCP check levels 1, 2, 3a, 3b, 3c, 4a, 4b, 5a and 5b indices are S_1, S_2, S_{3a}, S_{3b}, S_{4a}, S_{4b}, S_{5a}, S_{5b}, respectively, where S_{4b} is the value of ρ_{rd} at which sign change starts appearing continuously with the increase of the number of random vectors, and other indices are fractions of regression coefficients of selected descriptors with sign changes. NA – information not available. Performance qualitative labeling system consists of five levels: excellent, good, acceptable, poor and extremely poor (extr. poor) performance of data and models.

[(b)] Visual check of selected descriptors in all data sets via descriptor - *y* scatterplots with general description: excellent, acceptable (some or all scatterplots have modest distribution problems), and problematic (some scatterplots are not acceptable for regression).

[(c)] Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Influence of the number of samples *n* (in statistics: sample size) on statistical significance of descriptor - *y* relationships deserves a special attention in QSAR/QSPR. As Capraro[47] says, "given a large enough sample, one would always achieve statistical significance", because the value of a statistical test and its corresponding *p*-value depend not only on effect size and the level of α selected, but also on the number *n*.[48,49] Effect size, defined by Kenny,[50] is "the measure of the strength of effect as opposed to its *p*-value". In terms of descriptor - *y* relationships, it is the strength of association between a descriptor and *y*, which is measured by the Pearson's product-moment correlation coefficient.[48,49,51–54] Standard test statistic for one variable or for relationship between two variables, such as *t*- and *F*-test statistic, can be expressed as a product of sample size and effect size.[48,49] The Pearson correlation coefficient is data set size-independent and therefore, should be reported together with the *p*-value from statistical significance testing of the relationship between two variables. Effect of data set size on test statistic and probability is well noticeable for large data sets. In this work, both data sets with the largest *n*, K ($n = 232$) and L ($n = 460$), possess one descriptor characterized with statistically significant relationship to *y* at the ISCP level 5b (*t*- and *F*-tests), but with low absolute values of the respective correlation coefficients: μ_1^{Dip2} (data set K) with extreme statistical significance and $r = 0.26$, and nCaH (data set L) with statistical significance and $r = 0.11$. Such descriptors should not be used in further modeling, because they aid in producing misleading, falsely good models. Although the mentioned values of *r* are

Table 6. Summary of descriptor statistics[(a)] in terms of risk of sign changes in multivariate modelling.

Data set	No risk to very low risk	Low risk	Moderate risk	High to very high risk	Score[(b)]	Overall risk[(c)]
A	1 [Q_{Omul}]	5 [E_e, Δ_{HL}, σ_r, D_{CC}, Q_{C2mul}]	2 [E_{cc}, σ_b]	0	0.88	low
B	0	4 [EA, X_e, SArea, Log(W)]	0	0	0.91	low
C	0	2 [**E**, **V**]	0	3 [**A**, **B**, **S**]	0.68	high to very high
D	0	2 [x_{15}, x_{25}]	0	3 [x_{26}, $\underline{x}_{27}$, x_{36}]	0.75	high to very high
E	0	2 [Human liver, LUMO]	0	1 [NO]	0.77	[moderate to high]
F	0	3 [x_1, $\underline{x}_7$, $\underline{x}_{29}$]	1 [$\underline{x}_{52}$]	0	0.86	[moderate to high]
G	0	1 [log K_{ow}]	1 [E_{LUMO}]	3 [pKa, E_{HOMO}, N_{Hdon}]	0.60	high to very high
H	2 [P_{5X}, M_{04}]	1 [$\underline{M}_{11}$]	1 [$\underline{S}_6$']	0	0.92	[moderate]
I	0	5 [S_2, S_3, S_4, S_5, S_8]	2 [$\underline{S}_7$, $\underline{S}_6$]	1 [$\underline{S}_1$]	0.82	[moderate to high]
J	0	2 [E1v, LP1]	1 [IC5]	1 [RDF125m]	0.74	high to very high
K	0	2 [μ_1^{Hyd}, μ_3^{Van}]	4 [$\underline{\mu_1\mu_4}^{Dip4}$, $\mu_4^{Ab-logL16}$, $\mu_4^{Ab-\Sigma\beta 2O}$, $\underline{\mu_4}^{Pols}$]	4 [$\mu_1\mu_2^{Std}$, μ_{10}^{Std}, μ_5^{Ab-R2}, μ_1^{Dip2}]	0.74	high to very high
L	0	3 [Homo, nX, CIC0]	0	1 [nCaH]	0.80	moderate

[(a)] Descriptors are named in the same way as in original publications. They are considered as bearing no risk to very low risk, low risk, moderate risk, and high to very high risk of the sign changes for the use in multivariate modeling.
[(b)] Average score is expressed in terms of the sum of score contributions of descriptors divided by the number of descriptors.
[(c)] Overall risk means the risk of taking into account all selected descriptors, based on average score. Overall risk is given in square brackets when it must be higher than predicted from the average score, due to the presence of descriptors with the poorest possible performance in the ISCP tests (underlined descriptors).

statistically significant *i.e.*, have 95 % probability of being different from zero, they measure weak associations of the two descriptors to y. Threshold $r = 0.3$ for moderately strong associations in QSAR studies[14,15] is also recommended in statistical literature.[48,51,52]

ISCP and Other Approaches to Treat Sign Changes

PLS and MLR are still the two commonest regression methods in QSAR/QSPR.44 QSAR/QSPR models made by using PLS, MLR and other simpler regression methods can be rather effectively checked for sign changes using the ISCP procedure. Hence, there is no need to abandon the mentioned regression methods and use more complicated ones which treat sign changes up to a certain extent.

Regression methods that employ orthogonalized variables generated from original descriptors[55,56] such as PLS, successfully deal with descriptor multicollinearity, but are not effective against sign changes and besides, yield models that are difficult to interpret. It has been shown previously[15] that sign changes in PLS models are originated from multicollinearity and increased model complexity. In this work, examples of PLS models (A, B and H) also incorporate sign changes. In other words, the use of orthogonalized descriptors does not prevent the appearance of sign changes, which are visible when the model is expressed in terms of original descriptors.

Modern shrinkage regression methods,[57,58] such as ridge regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression and its variants indirectly deal with sign changes, but do not completely solve the sign change problem. In all these methods descriptors with small coefficients are discarded, what enables partial solution of SCP, because eliminated descriptors can easily undergo sign changes during the regression modeling. In LASSO and its variants, descriptors whose regression coefficients are unstable (*i.e.* have large variations in size and even change signs) are also discarded. In fact, shrinkage regressions tend to minimize descriptors multicollinearity. Shrinkage methods are rather automated procedures in terms of variable selection. LASSO and its variants are based on preserving regression vector signs with respect to the MLR regression vector as the reference vector (Equation 1, with $\alpha = 0$) that serves as the first estimate of regression coefficients,[57,58] and not with respect to the correlation vector from univariate regressions (Equation 2). Besides, the complete procedure is carried out for each data set. Therefore, shrinkage regressions do not preserve regression vector signs with respect to one reference vector, which would characterize the reference

data set. Consequently, the sign change may occur in shrinkage regressions.

ISCP introduced in this work can be considered as a general anti-SCP tool. A model based on a regression method that partially treats SCP should be also subjected to the ISCP check.

Spectral-SAR[59–61] is a novel, challenging approach in terms of its QSAR/QSPR philosophy and theoretical background, calculation of new goodness-of-fit indices, and model interpretation. Spectral-SAR regression models are Hansch-type equations with hydrophobic, electronic and steric descriptors, which are orthogonalized via the Gram-Schmidt algorithm. The spectral-SAR methodology should be further tested for eventual sign changes, which probably should be interpreted somewhat differently than in standard QSAR/QSPR. Similar can be said about alert-QSAR,[62] in which residual analysis is employed with the purpose to minimize residuals correlation with descriptors.

ISCP is based on the assumption that relationships between descriptors and *y* are linear, statistically significant, and sufficiently strong. However, when descriptor - *y* relations are not linear, and the non-linearity is included in a multivariate model but not in the univariate regressions, sign change may appear. Catastrophe-QSAR[63] uses Thom's polynomials to model non-linearity in multivariate regression (Equation 2), showing that the sign change in linear terms of a descriptor is caused by inadequate univariate regressions (Equation 1). Catastrophe-QSAR is an example of a new challenge for sign change treatment and interpretation in non-linear QSAR/QSPR.

CONCLUSION

The five-level integral sign change problem check established in this work can be considered as an effective anti-sign change problem methodology, acting as a new model validation that detects sign changes in QSAR/QSPR data sets and models. A detailed tutorial for the complete procedure and accompanying checks was presented. The procedure was applied to twelve QSAR/QSPR data sets and models, resulting data and model performance was reported and discussed in terms of data and model remedy. Performance qualitative labeling systems are proposed as an aid to characterize data set and model performance, and simplify human decision in this stage of modeling *i.e.*, the choice between model validation and new variable selection with eventual data modification. Future research will be directed to further development and refinement of the proposed integral sign change problem check. The descriptor sign change problem is an issue to which no sufficient attention is paid in QSAR/QSPR research, which certainly contributes to generation of statistically false, deficient and low predictable regression models.

Supplementary Materials. – Supporting informations to the paper are enclosed to the electronic version of the article. These data can be found on the website of *Croatica Chemica Acta* (http://public.carnet.hr/ccacaa).

Acknowledgements. This work was supported by Grant "Developing methods for modeling properties of bioactive molecules and proteins" (No. 098-1770495-2919) awarded by the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

1. M. M. C. Ferreira, *J. Braz. Chem. Soc.* **13** (2002) 742–753.
2. D. Livingstone, Data Analysis for Chemists: Applications to QSAR and Chemical Product Design, Oxford University Press, Oxford, 2002.
3. R. D. Cramer, *J. Comput. Aided Mol. Des.* **26** (2012) 35–38.
4. M. M. C. Ferreira and R. Kiralj, Métodos Quimiométricos em Relações Quantitativas Estrutura-Atividade (QSAR), in: C. A. Montanari (Ed.), Química Medicinal: Métodos e fundamentos em planejamento de fármacos, EdUSP, São Paulo, 2011, pp. 387–453.
5. T. Puzyn, J. Leszczynski, and M. T. D. Cronin (Eds.), *Recent Advances in QSAR Studies*, Challenges and Advances in Computational Chemistry and Physics Vol. 8, Springer, Dordrecht, 2010.
6. K. R. Beebe, R. Pell, and M. B. Seasholtz, *Chemometrics: a practical guide*, Wiley, New York, 1998.
7. H. Martens and T. Naes, *Multivariate Calibration*, 2nd ed., Wiley, New York, 1989.
8. *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models*, OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69, OECD, Paris, 2007. [Last access on October 23, 2013 at http://www.oecd.org/fr/securitechimique/risques/38130292.pdf.
9. P. Gramatica, *QSAR Comb. Sci.* **26** (2007) 694–701.
10. J. C. Dearden, M. T. D. Cronin, and K. L. E. Kaiser, *SAR QSAR Environ. Res.* **20** (2009) 241–266.
11. R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese, and R. K. Agrawal, *Int. J. Drug Des. Disc.* **2** (2011) 511–519.
12. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, and P. Gramatica, *Environ. Health Perspect.* **111** (2003) 1361–1375.
13. T. Scior, J. L. Medina-Franco, Q.-T. Do, K. Martínez-Mayorga, J. A. Yunes Rojas, and P. Bernard, *Curr. Med. Chem.* **16** (2009) 4297–4313.
14. R. Kiralj and M. M. C. Ferreira, *J. Braz. Chem. Soc.* **20** (2009) 770–787.
15. R. Kiralj and M. M. C. Ferreira, *J. Chemom.* **24** (2010) 681–693.
16. R. Kiralj, Radovi zav. znan. umj. rad Bjelovar **6** (2012) 179–208.
17. R. Kiralj and M. M. C. Ferreira, *J. Phys. Chem. A* **112** (2008) 6134–6149.
18. R. F. Teófilo, R. Kiralj, H. J. Ceragioli, A. C. Peterlevitz, V. Baranauskas, L. T. Kubota, and M. M. C. Ferreira, *J. Electrochem. Soc.* **155** (2008) D640–D650.
19. N. Hernández, R. Kiralj, M. M. C. Ferreira, and I. Talavera, *Chemom. Intell. Lab. Syst.* **98** (2009) 65–77.
20. E. B. de Melo and M. M. C. Ferreira, *J. Chem. Inf. Model.* **52** (2012) 1722–1732.
21. E. G. Barbosa, K. F. M. Pasqualoto, and M. M. C. Ferreira, *J. Comput. Aided Mol. Des.* **26** (2012) 1055–1065.
22. L.-T. Qin, S.-S. Liu, F. Chen, Q.-F. Xiao, and Q.-S. Wu, *Chemosphere* **90** (2013) 300–305

23. M. A. C. Fresqui, M. M. C. Ferreira, and M. Trsic, *Anal. Chim. Acta* **759** (2013) 43–52.
24. L.-T. Qin, S.-S. Liu, F. Chen, and Q.-S. Wu, *J. Sep. Sci.* **36** (2013) 1553–1560.
25. N. Omidikia and M. Kompany-Zareh, *Chemom. Intell. Lab. Syst.* **128** (2013) 56–65.
26. S. Bagheri, N. Omidikia, and M. Kompany-Zareh, *Chemom. Intell. Lab. Syst.* **128** (2013) 135–143.
27. X. Wang, Y. Sun, L. Wu, S. Gu, R. Liu, L. Liu, X. Liu, and J. Xu, *Chemom. Intell. Lab. Syst.* **134** (2014) 1–9.
28. C. F. Matta, *J. Comput. Chem.* **35** (2014) 1165–1198.
29. *GraphPad QuickCalcs: Statistical distributions and interpreting P values*, GraphPad Software, Inc., La Jolla, CA, 2013. [Last access on June 23, 2014 at http://www.graphpad.com/quickcalcs/DistMenu.cfm]
30. F. J. Anscombe, *Am. Stat.* **27** (1973) 17–21.
31. A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences*, Wiley, Chichester, 2004.
32. M. M. C. Ferreira, *Chemosphere* **44** (2001) 125–146.
33. L. M. Sprunger, J. Gibbs, W. E. Cree Jr., and M. H. Abraham, *QSAR Comb. Sci.* **28** (2009) 72–88.
34. L.-T. Qin, S.-S. Liu, and H.-L. Liu, *Mol. Divers.* **14** (2010) 67–80.
35. I. Lessigiarska, A. P. Worth, T. I. Netzeva, J. C. Dearden, and M. T. D. Cronin, *Chemosphere* **65** (2006) 1878–1887.
36. S.-S. Liu, S.-H. Cui, Y.-Y. Shi, and L.-S. Wang, *J. Mol. Des.* **1** (2002) 610–619.
37. A. O. Aptula, T. I. Netzeva, I. V. Valkova, M. T. D. Cronin, T. W. Schultz, R. Kühne, and G. Schüürmann, *Quant. Struct.-Act. Relat.* **21** (2002), 12–22.
38. X. J. Yao, A. Panaye, J. P. Doucet, R. S. Zhang, H. F. Chen, M. C. Liu, Z. D. Hu, and B. T. Fan, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1257–1266.
39. R. Kiralj and M. M. C. Ferreira, *J. Braz. Chem. Soc.* **14** (2003), 20–26.
40. C. Yin, S. Liu, and X. Wang, *J. Chin. Chem. Soc.* **50** (2003) 875–879.
41. S. Ahmadi, *Macroheterocycles* **5** (2012) 23–31.
42. A. Pérez-Garrido, A. M. Helguera, M. N. D. S. Cordeiro, and M. G. Escudero, *J. Pharm. Sci.* **98** (2009) 4557–4576.
43. P. Gramatica, P. Pilutti, and E. Papa, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1794–1802.
44. L. C. Yee and Y. C. Wei, *Current Modeling Methods Used in QSAR/QSPR*, in: M. Dehmer, K. Varmuza, and D. Bonchev (Eds.), *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, Wiley-VCH, Weinheim, 2012, pp. 1–31.
45. *Pirouette 4.0 rev. 2*, Infometrix, Inc., Woodinville, WA, 2009.
46. *Scilab 5.4.0*, Scilab Enterprises, S.A.S., Versailles, 2013.
47. R. M. Capraro, *Significance Level*, in: N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, vols. 1 & 2, Sage, Thousand Oaks, CA, 2007, pp. 889–891.
48. W. Rodriguez, *Effect Size*, in: N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, vols. 1 & 2, Sage, Thousand Oaks, CA, 2007, pp. 300–304.
49. C. J. Ferguson, *Prof. Psycho. Res. Pract.* **40** (2009) 532–538.
50. D. A. Kenny, *Statistics for the Social and Behavioral Sciences*, Little, Brown, Boston, MA, 1987, p. 394.
51. A. J. Onwuegbuzie, L. Daniel, and N. L. Leech, *Pearson Product-Moment Correlation Coefficient*, in: N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, vols. 1 & 2, Sage, Thousand Oaks, CA, 2007, pp. 750–755.
52. J. M. Maher, J. C. Markey, and D. Ebert-May, *CBE Life Sci. Educ.* **12** (2013) 345–351.
53. J. H. Steiger, *Psychol. Methods* **9** (2004) 164–182.
54. M. Grahramanlou-Holloway, *Meta Analysis*, in: N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*, vols. 1 & 2, Sage, Thousand Oaks, CA, 2007, pp. 595–598.
55. S. C. Peterangelo and P. G. Seybold, *Int. J. Quant. Chem.* **96** (2004) 1–9.
56. P. R. Duchowicz, F. M. Fernández, and E. A. Castro, *Orthogonalization methods in QSPR – QSAR Studies*, in: E. A. Castro (Ed*.), QSPR-QSAR Studies on Desired Properties for Drug Design*, Research Signpost, Trivandrum, India, 2010, pp. 189–203.
57. R. Tibshirani, *J. R. Statist. Soc. B* **58** (1996) 267–288.
58. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistic, Springer, New York, 2009, pp. 61–79.
59. M. V. Putz and A.-M. Lacrămă, *Int. J. Mol. Sci.* **8** (2007) 363–391.
60. M. V. Putz, A.-M. Putz, M. Lazea, L. Ienciu, and A. Chiriac, *Int. J. Mol. Sci.* **10** (2008) 1193–1214.
61. M. V. Putz, A.-M. Putz, M. Lazea, L. Ienciu, and A. Chiriac, *J. Theor. Comput. Chem.* **6** (2009) 1235–1251.
62. M. V. Putz, C. Ionaşcu, A.-M. Putz, and V. Ostafe, *Int. J. Mol. Sci.* **12** (2011) 5098–5134.
63. M. V. Putz, M. Lazea, A.-M. Putz, and C. Duda-Seiman, *Int. J. Mol. Sci.* **12** (2011) 9533–9569.

SUPPLEMENTAL MATERIAL

Integral Sign Change Problem Check in Quantitative Structure-Activity/Property Relationships: A Tutorial

Rudolf Kiralj

Technical College in Bjelovar, Trg Eugena Kvaternika 4, 43000 Bjelovar, Croatia

Example for ISCP: Data set A (Tables S1 – S15)

Table S1. Basic information about data set A.

Item	Values
Data sets*	-Complete data set (n_c = 50) -Subsets after splitting: training (n_t = 40) and external validation (n_e = 10) sets
Selected descriptors	8 (E_e, E_{cc}, Q_{Omul}, Δ_{HL}, σ_b, σ_r, D_{CC}, Q_{C2mul})
QSPR models	-PLS model for the complete data set (n_c = 50) -Proposed PLS model for the training set (n_t = 40) with external validation (n_e = 10)
Reference data set	Complete data set (obtained from variable selection)
Reference model	PLS model for the complete data set (obtained from variable selection)

*Number of samples in the complete, training and external validation set is marked by n_c, n_t and n_e, respectively.

Table S2. ISCP level 1: Comparing correlation and regression vectors to the reference vector for data set A.

Vector*	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	$R^{2\#}$	$Q^{2\#}$	LVs(%)$^{\#}$
Correlation	**-0.8561**	**-0.8920**	**0.9282**	**-0.8267**	**0.8619**	**-0.8905**	**0.9069**	**0.8915**	-	-	-
(n_c=50), r_c											
Regression	-0.1036	-0.0636	0.1222	-0.2276	0.0372	-0.2876	0.0882	0.1124	0.9154	0.8951	2(92.33%)
(n_c=50), β_c											
Correlation	-0.8445	-0.8842	0.9176	-0.8435	0.8580	-0.8989	0.8976	0.8863	-	-	-
(n_t=40), r_t											
Regression	-0.1095	-0.0559	0.1071	-0.2156	0.0374	-0.3178	0.0764	0.1153	0.9105	0.8857	2(92.61%)
(n_t=40), β_t											
Correlation	-0.9166	-0.9589	0.9753	-0.7463	0.8984	-0.8526	0.9769	0.9172	-	-	-
(n_e=10), r_e											

*Correlation vectors: r_c - for complete data set, r_t - for training set, r_e - for external validation set. Regression vectors: β_c - for complete data set, β_t - for training set. The reference vector is the correlation vector r_c, typed in bold

$^{\#}R^2$ – correlation coefficient of multiple determination; Q^2 – cross-validated correlation coefficient; LVs - number of latent variables with corresponding percentage (%) of the original information.

Observation based on Table S1: No sign change was noticed in this SCP check, i.e. the sign change frequency at the ISCP level 1 is equal to zero for all descriptors.

Table S3. ISCP level 2: Counting sign change absolute frequencies for *l*-variate submodels and descriptors of data set A.

Regressions*	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	SCP#
Bivariate (28)	0	2	0	0	2	0	1	0	5
Trivariate (56)	0	5	0	0	8	0	0	0	12
Tetravariate (70)	1	5	0	1	9	0	0	1	12
Pentavariate (56)	0	3	0	0	4	0	0	0	4
Hexavariate (28)	0	0	0	0	0	1	0	0	1
Heptavariate (8)	0	0	0	0	0	0	0	0	0
Octavariate (1)	0	0	0	0	0	0	0	0	0
Total multivariate (247)	1	15	0	1	23	1	1	1	34

*Bivariate and higher *l*-variate PLS regression models and their numbers (given in brackets).
#Total SCP count (absolute sign change frequency) in *l*-variate PLS regression models

Table S4. ISCP level 2: Final sign change (SC) statistics for data set A.

Count	Total	With SC	%With SC
No. Multivariate models	247	34	13.78%
No. Regression coefficients	1016	43	4.23%
No. Selected descriptors	8	7	87.50%

Observation based on Tables S3 and S4: Sign change was noticed in this SCP check, i.e. the sign change frequency at the ISCP level 2 is equal to 4.23% in terms of regression coefficients, what is not a substantially problematic sign change frequency.

Table S5. ISCP level 3a: Checking the reference model using the descriptor pool for data set A.

Item	Values
No. Selected descriptors	8
No. Descriptors in descriptor pool	109
No. Descriptors in the matrix $\mathbf{X_p}$	109 - 8 = 101
No. Bivariate MLR regressions*	8×101 = 808
No. Regression coefficients of interest	8×101 = 808

*Calculations could not be performed for higher MLR regressions: several trivariate MLR models could not be constructed because of the matrix singularity problem.

Table S6. ISCP level 3a: Final sign change statistics for data set A (bivariate MLR models).

Regressions	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	SCP*
Bivariate (808)	0	1	0	0	5	0	0	0	6

*Total sign change count: 6 in 808 regression coefficients or 0.74%.

Observation based on Tables S5 and S6: Sign change was noticed in this SCP check, i.e. the sign change frequency at the ISCP level 3a is equal to 0.74% in terms of regression coefficients, what is not a negligible sign change frequency.

Table S7. ISCP level 3b: Checking the reference model using the descriptor pool subset for data set A.

Item	Values
No. Selected descriptors	8
No. Descriptors in descriptor pool subset*	51
No. Descriptors in matrix $\mathbf{X_p}$	51 - 8 = 43
No. Bivariate MLR regressions	8×43 = 344
No. Regression coefficients (bivariate regressions)	8×43 = 344
No. Trivariate MLR regressions	8×43×42/2 = 7224
No. Regression coefficients (trivariate regressions)	8×43×42/2 = 7224
No. Tetravariate MLR regressions[#]	8×43×42×41/6 = 98728
No. Regression coefficients (tetravariate regressions)	8×43×42×41/6 = 98728

*Descriptor pool subset contains descriptors whose absolute values of correlation coefficients related to the dependent variable are $|r|>0.60$.
[#]Calculations could not be performed for higher MLR regressions because of time- and memory-consuming problems.

Table S8. ISCP level 3b: Final sign change statistics for data set A (bivariate to tetravariate MLR models).

Regressions*	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	SCP[#]
Bivariate (344)	0	1	0	0	5	0	0	0	6
Trivariate (7224)	0	35	0	4	162	0	5	29	235
Tetravariate (98728)	33	602	29	182	2758	0	92	950	4646
Total multivariate (106296)	33	638	29	186	2925	0	97	979	4887

*Bivariate and higher *l*-variate MLR regression models and their numbers (given in brackets).
#Total sign change count (absolute sign change frequency) in *l*-variate MLR regression models: 4887 in 106296 models or 4.60%.

Observation based on Tables S7 and S8: Sign change was noticed in this SCP check, i.e. the sign change frequency at the ISCP level 3b is equal to 4.60% in terms of regression coefficients, what is a very small sign change frequency.

Table S9. ISCP levels 4a and 4b: sign change (SC) and ρ_{rd} statistics*,# for data set A.

N_{rv}	r_{min}	r_{max}	ρ_{rd}	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	SC	%SC
10	-0.2825	0.2119	0.2472	0	0	0	0	0	0	0	0	0	0%
25	-0.2386	0.2182	0.2284	0	0	0	0	0	0	0	0	0	0%
50	-0.4026	0.3352	0.3689	0	0	0	0	0	0	0	0	0	0%
100	-0.4026	0.3352	0.3689	0	0	0	0	0	0	0	0	0	0%
250	-0.4026	0.3678	0.3852	0	0	0	0	0	0	0	0	0	0%
500	**-0.4026**	**0.4520**	**0.4273**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0%**
1000	-0.4201	0.4698	0.4450	0	0	0	0	0	0	0	0	0	0%
2000	-0.4275	0.4698	0.4486	0	0	0	0	0	0	0	0	0	0%
5000	-0.5951	0.4806	0.5379	0	0	0	0	0	0	0	0	0	0%
10000	-0.5275	0.5131	0.5203	0	0	0	0	0	0	0	0	0	0%
20000	-0.5557	0.5265	0.5411	0	0	0	0	0	0	0	0	0	0%
30000	-0.5951	0.5265	0.5608	0	0	0	0	0	0	0	0	0	0%
40000	-0.5951	0.5271	0.5611	0	0	0	0	0	0	0	0	0	0%
50000	-0.5967	0.5787	0.5877	0	0	0	0	0	0	0	0	0	0%
60000	-0.5498	0.5214	0.5272	0	0	0	0	0	0	0	0	0	0%
70000	-0.5673	0.5306	0.5490	0	0	0	0	0	0	0	0	0	0%
80000	-0.6248	0.6318	0.6283	0	0	0	0	0	0	0	0	0	0%
90000	-0.5971	0.5957	0.5964	0	0	0	0	0	0	0	0	0	0%
100000	-0.5809	0.5739	0.5774	0	0	0	0	0	0	0	0	0	0%
110000	-0.5967	0.5787	0.5877	0	0	0	0	0	0	0	0	0	0%
120000	-0.5967	0.5787	0.5877	0	0	0	0	0	0	0	0	0	0%
130000	-0.5760	0.5816	0.5788	0	0	0	0	0	0	0	0	0	0%
140000	-0.6248	0.6318	0.6283	0	0	0	0	0	0	0	0	0	0%
150000	-0.5809	0.5739	0.5774	0	0	0	0	0	0	0	0	0	0%

160000	-0.5638	0.6121	0.5880	0	0	0	0	0	0	0	0	0	0	0%
170000	-0.5759	0.6426	0.6093	0	0	0	0	0	0	0	0	0	0	0%
180000	-0.5958	0.6246	0.6102	0	0	0	0	0	0	0	0	0	0	0%
190000	-0.5820	0.6111	0.5966	0	0	0	0	0	0	0	0	0	0	0%
200000	**-0.6243**	**0.6120**	**0.6182**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0%**

*N_{rv} – No. random vectors; r_{min}, r_{max} – the minimum and maximum correlation coefficients for correlations between the random vectors and the dependent variable; ρ_{rd} – the correlation coefficient for randomization, given as the average, $\rho_{rd} = (|r_{min}| + |r_{max}|)/2$; SC - total sign change (SC) count (absolute SC frequency); %SC – relative SC frequency, defined as the ration of the absolute SC frequency and the number of random vectors.

#Statistics for the ISCP level 4a is typed in bold black: for $\rho_{rd} = 0.43$ the SC frequency is equal to zero. Statistics for the ISCP level 4b is typed in bold red: this is not a definitively statistics but statistics indicating that non-zero SC frequency probably occurs at very high values of ρ_{rd}, certainly at $\rho_{rd} > 0.62$. Calculations stopped at N_{rv} = 200,000 because of time- and memory-consuming problems.

Observation based on Table S9: Sign change at the ISCP level 4a is equal to zero in terms of regression coefficients, and probably is the same for the SC frequency at the ISCP level 4b up to very high values of the random coefficient ρ_{rd} (to achieve this limit N_{rv} should be of the order of magnitude of millions).

Table S10. ISCP level 5a: *t*-Test values for intercepts of univariate descriptor – *y* relationships for data set A.

Data set	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}
Complete (n_c=50)	19.7691	14.5924	40.773	2.6224	-3.8729	123.8859	-13.4885	85.0262
Training (n_t=40)	17.6656	11.8573	34.1011	3.1878	-2.9619	115.9003	-11.0634	79.1452
External (n_e=10)	8.9877	8.6020	27.7945	0.0708	-2.4471	41.6592	-7.5292	29.9819

Table S11. ISCP level 5a: Statistical significance of intercepts of univariate descriptor – *y* relationships for data set A (*t*-test*).

Data set	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}
Complete (n_c=50)	ESS	ESS	ESS	SS	ESS	ESS	ESS	ESS
Training (n_t=40)	ESS	ESS	ESS	VSS	**NQSS**	ESS	ESS	ESS
External (n_e=10)	ESS	ESS	ESS	**NSS**	SS	ESS	ESS	ESS

*Not acceptable statistical significance is typed in bold.

Observation based on Tables S10 and S11: Two (Δ_{HL} and σ_b) out of eight descriptors are not characterized by statistically significant intercepts, what does not still mean significantly bad performance of the data set A at this ISCP level (sign change frequency is only 2/24 = 8.33%).

Table S12. ISCP level 5b: *t*-Test (up) and *F*-test (down) for slopes of univariate descriptor – *y* relationships for data set A (*t*- and *F*-tests).

Data set	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}
Complete (n_c=50)	-11.4769	-13.6732	17.2870	-10.1788	11.7785	-13.5616	14.9116	13.6347
	131.7189	186.9570	298.8413	103.6071	138.7331	183.9171	222.3561	185.9040
Training (n_t=40)	-10.0002	-11.0813	14.0186	-9.9058	9.7805	-12.0855	12.2760	12.1371
	100.0031	122.7953	196.5214	98.1249	95.6575	146.0602	150.7014	147.3089
External (n_e=10)	-5.8866	-8.1286	13.2662	-3.5165	6.0128	-5.4895	8.2213	5.8832
	34.6519	66.0734	175.9908	12.3657	36.1536	30.1342	67.5895	34.6115

Table S13. ISCP level 5b: Statistical significance of slopes of univariate descriptor – *y* relationships for data set A (*t*- and *F*-tests*).

Data set	E_e	E_{cc}	Q_{Omul}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}
Complete (n_c=50)	ESS	ESS	ESS	ESS	ESS	ESS	ESS	ESS
Training (n_t=40)	ESS	ESS	ESS	ESS	ESS	ESS	ESS	ESS
External (n_e=10)	ESS	ESS	ESS	VSS	ESS	ESS	ESS	ESS

*Results for the *t*- and *F*-tests for data set A are exactly the same for all descriptors.

Observation based on Tables S12 and S13: All descriptors are characterized by statistically significant slopes at this ISCP level.

Table S14. Legend for statistical significance for ISCP levels 5a and 5b.*

Abbreviation	Wording	*p*-value range (for 95% confidence limit)
NSS	Not statistically significant	$p > 0.10$
NQSS	Not quite statistically significant	$0.05 < p < 0.10$
SS	Statistically significant	$0.01 < p < 0.05$
VSS	Very statistically significant	$0.001 < p < 0.01$
ESS	Extremely statistically significant	$p < 0.001$

*Statistical significance levels NSS and NQSS are considered as not acceptable in the ISCP analyses in this work. The qualitative labelling of statistical significance levels is from the GraphPad QuickCalcs software (GraphPad Software, Inc., La Jolla, CA, 2013. Last access on June 23, 2014 at http://www.graphpad.com/quickcalcs/DistMenu.cfm].

Table S15. Additional checks for the data set A.

Check	No. checks	Result
Check for very small regression coefficients (< 0.001 for autoscaled data) for all descriptors in all models	8×3 = 12	No problematic descriptors were found
Visual inspection of bivariate distribution of descriptor – *y* scatterplots for all data sets and subsets	8×3 = 24	No serious distribution problems were found: acceptable scatterplots

In Continuation:

Data set & model and descriptor statistics for data sets B – L (Tables S16 – S37)

Table S16. Data set B and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	0/4(0%)	0/2(0%)	0/16(0%)	$S_1 = 0$	excellent
2 (submodels;4)	0/4(0%)	0/11(0%)	0/44(0%)	$S_2 = 0$	excellent
3a (pool;11;10)	4/4(100.0%)	493/4092(12.0%)*	493/4092(12.0%)*	$S_{3a} = 0.120$	poor
4a (25;0.41)	0/4(0%)	0/100(0%)*	0/100(0%)*	$S_{4a} = 0$	excellent
4b (200000;0.75)	>>0.75	>>0.75	>>0.75	$S_{4b} >> 0.75$	excellent
5a(*t*-test;0.05)	1/4(25.0%)	2/2(100.0%)	2/8(25.0%)	$S_{5a} = 0.250$	poor
5b(*t*-/*F*-test;0.05)	0/4(0%)	0/2(0%)	0/8(0%)	$S_{5b} = 0$	excellent

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (4), which is the maximum complexity of the multivariate model considered. Level 3a: the maximum complexity (*l* value) of the *l*-variate MLR models considered (11) that could be treated computationally; number of all descriptors excluding the selected descriptors (10). Level 4a: number of random descriptors (generating 25 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.41). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (in this case, it is 200,000); the minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (in this case, 0.73). Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>"only when probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).

Table S17. Descriptor statistics* in terms of sign changes for data set B and its models.

Level	EA	X_e	SArea	Log(*W*)	Total
1	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/32(0%)
2	0/7(0%)	0/7(0%)	0/7(0%)	0/7(0%)	0/28 (0%)
3a	82/1023 (8.0%)	210/1023 (20.5%)	194/1023 (19.0%)	7/1023 (0.7%)	493/4092(12.0%)
4a	0/25(0%)	0/25(0%)	0/25(0%)	0/25(0%)	0/100(0%)
4b[#,&]	>>200000	>>200000	>>200000	>>200000	>>800000
	>>0.75	>>0.75	>>0.75	>>0.75	>>0.75
5a	0/2(0%)	0/2(0%)	2/2(100.0%)	0/2(0%)	2/8(25.0%)
5b	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/8(0%)
Visual check	acceptable	excellent	excellent	excellent	acceptable
Total score[##]	30.5 – 33 (86 – 93%)	31.5 – 34 (89 – 96%)	29.5 – 32 (83 – 90%)	32 – 34.5 (90 – 97%)	-
Risk**	low	low	low	low	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (in this case, reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (in this case, reported as a lower limit >>0.63, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors. [##]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S18. Data set C and its model statistics# in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	2/5(40.0%)	2/2(100.0%)	3/20(0.2%)	$S_1 = 0.150$	poor
2 (submodels;5)	2/5(40.0%)	14/26(53.8%)	15/75(20.0%)	$S_2 = 0.200$	poor
4a (2500;0.43)	2/5(40.0%)	116/12500(0.9%)*	116/12500(0.9%)*	$S_{4a} = 0.009$	poor
4b (10;0.21)	0.21	0.21	0.21	$S_{4b} = 0.21$	extremely poor
5a(*t*-test;0.05)	0/5(0%)	0/2(0%)	0/15(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	2/5(40.0%)	2/2(100.0%)	6/15(40.0%)	$S_{5b} = 0.400$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. #ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (5), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 2500 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.43). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 2500), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.43), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S19. Descriptor statistics* in terms of sign changes for data set C and its models.

Level	**E**	**S**	**A**	**B**	**V**	Total
1	0/4(0%)	2/4(50.0%)	1/4(25.0%)	0/4(0%)	0/4(0%)	3/20(15.0%)
2	0/15(0%)	10/15(66.7%)	5/15(33.3%)	0/15(0%)	0/15(0%)	15/75(20.0%)
4a	0/2500(0%)	0/2500(0%)	93/2500(3.7%)	23/2500(0.9%)	0/2500(0%)	116/12500(0.9%)
4b[#,&]	>>200000	>>200000	25	100	>>200000	10
	>>0.53	>>0.53	0.21	0.25	>>0.53	0.21
5a	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/15(0%)
5b	0/3(0%)	0/3(0%)	3/3(100.0%)	3/3(100.0%)	0/3(0%)	6/15(40.0%)
Visual check	acceptable	acceptable	problematic	acceptable	excellent	problematic
Total score[##]	29 – 34 (82 – 96%)	21 – 26 (59 – 73%)	10 – 15 (28 – 42%)	20.5 – 25.5 (58 – 72%)	28.5 – 33.5 (80 – 94%)	-
Risk**	low	high to very high	high to very high	high to very high	low	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). #The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 10), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). &The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.21), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.53, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).
**Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors.
##Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S20. Data set D and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	2/5(40.0%)	2/2(100.0%)	4/25(16.0%)	$S_1 = 0.160$	poor
2 (submodels;5)	7/8(87.5%)	34/247(13.8%)	43/1016(4.2%)	$S_2 = 0.042$	poor
4a (80000;0.40)	1/8(12.5%)	92/400000(0.02%)*	92/400000(0.02%)*	$S_{4a} = 0.002$	poor
4b (500;0.25)	0.25	0.25	0.25	$S_{4b} = 0.250$	extremely poor
5a(*t*-test;0.05)	0/5(0%)	0/2(0%)	0/15(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	1/5(20.0%)	2/2(100.0%)	3/15(20.0%)	$S_{5b} = 0.200$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (5), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 8000 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.40). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 500), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.25), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S21. Descriptor statistics* in terms of sign changes for data set D and its models.

Level	x_{15}	x_{25}	x_{26}	x_{27}	x_{36}	Total
1	0/5(0%)	0/5(0%)	2/5(40.0%)	2/5(40.0%)	0/5(0%)	4/25(16.0%)
2	0/15(0%)	0/15(0%)	4/15(26.7%)	7/15(46.7%)	1/15(6.7%)	12/75(16.0%)
4a	0/80000(0%)	0/80000(0%)	0/80000(0%)	92/80000(0.12%)	0/80000(0%)	92/400000(0.02%)
4b[#,&]	>>200000	>>200000	>>200000	500	>>200000	500
	>>0.43	>>0.43	>>0.43	0.25	>>0.43	0.25
5a	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/15(0%)
5b	0/3(0%)	0/3(0%)	0/3(0%)	3/3(100.0%)	0/3(0%)	3/15(20.0%)
Visual check	acceptable	acceptable	problematic	problematic	acceptable	problematic
Total score[##]	29.5 – 34.5 (83 – 97%)	29.5 – 34.5 (83 – 97%)	20.5 – 25.5 (58 – 72%)	12.5 – 17.5 (35 – 49%)	28.5 – 33.5 (80 – 94%)	-
Risk**	low	low	high to very high	high to very high	high to very high	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).

[#]The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 500), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.25), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.43, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).

**Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors.

[##]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S22. Data set E and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	0/3(0%)	0/2(0%)	0/12(0%)	$S_1 = 0$	excellent
2 (submodels;3)	0/3(0%)	0/4(0%)	0/9(0%)	$S_2 = 0$	excellent
3b (subset;6;19)	2/3(66.7%)	14/49791(0.03%)*	14/49791(0.03%)*	$S_{3b} = 0.003$	good
4a (20;0.43)	1/3(33.3%)	5/60(8.3%)*	5/60(8.3%)*	$S_{4a} = 0.083$	poor
4b (20;0.43)	0.43	0.43	0.43	$S_{4b} = 0.430$	acceptable
5a(*t*-test;0.05)	3/3(100.0%)	2/2(100.0%)	6/6(100.0%)	$S_{5a} = 1.000$	extremely poor
5b(*t*-/*F*-test;0.05)	1/3(33.3%)	2/2(100.0%)	2/6(33.3%)	$S_{5b} = 0.333$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (3), which is the maximum complexity of the multivariate model considered. Level 3b: the maximum complexity (*l* value) of the *l*-variate MLR models considered (6) that could be treated computationally; number of descriptors used for testing (19). Level 4a: number of random descriptors (generating 20 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.43). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 20), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.43), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S23. Descriptor statistics* in terms of sign changes for data set E and its models.

Level	Human liver	LUMO	N_O	Total
1	0/4(0%)	0/4(0%)	0/4(0%)	0/12(0%)
2	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)
3b	0/16597(0%)	6/16597(0.3%)	8/16597(0.5%)	14/49791(0.03%)
4a	0/20(0%)	0/20(0%)	5/20(25.0%)	5/60(8.3%)
4b[#,&]	>>200000	>>200000	20	20
	>>0.83	>>0.83	0.43	0.43
5a	2/2(100.0%)	2/2(100.0%)	2/2(100.0%)	6/6(100.0%)
5b	0/2(0%)	0/2(0%)	2/2(100.0%)	2/6(33.3%)
Visual check	excellent	acceptable	problematic	problematic
Total score[##]	31 – 33.5 (87 – 94%)	29.5 – 32 (83 – 90%)	17.5 – 20 (49 – 56%)	-
Risk**	low	low	high to very high	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 20), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.43), or the maximum ρ_{rd} reached if sign change has not been observed(reported as a lower limit >>0.83, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).
**Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors.

[##]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S24. Data set F and its model statistics# in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	0/4(0%)	0/2(0%)	0/16(0%)	$S_1 = 0$	excellent
2 (submodels;4)	0/4(0%)	0/2(0%)	0/28(0%)	$S_2 = 0$	excellent
4a (25;0.40)	0/4(0%)	0/100(0%)*	0/100(0%)*	$S_{4a} = 0$	excellent
4b (500;0.61)	0.61	0.61	0.61	$S_{4b} = 0.610$	excellent
5a(*t*-test;0.05)	0/4(0%)	0/2(0%)	0/8(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	3/4(75.0%)	2/2(100.0%)	6/8(75.0%)	$S_{5b} = 0.750$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. #ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (4), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 25 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.40). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 500), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.61), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S25. Descriptor statistics* in terms of sign changes for data set F and its models.

Level	$\boldsymbol{x}_1$	$\boldsymbol{x}_7$	$\boldsymbol{x}_{29}$	$\boldsymbol{x}_{52}$	Total
1	0/4(0%)	0/4(0%)	0/4(0%)	0/4(0%)	0/16(0%)
2	0/7(0%)	0/7(0%)	0/7(0%)	0/7(0%)	0/28(0%)
4a	0/25(0%)	0/25(0%)	0/25(0%)	0/25(0%)	0/100(0%)
4b#,&	>>200000	150000	5000	500	500
	>>0.84	0.82	0.73	0.61	0.61
5a	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/8(0%)
5b	0/2(0%)	2/2(100.0%)	2/2(100.0%)	2/2(100.0%)	6/8(75.0%)
Visual check	problematic	problematic	problematic	problematic	problematic
Total score##	28.5 – 34 (80 – 96%)	28.5 – 34 (80 – 96%)	28 – 33 (79 – 93%)	27.5 – 32.5 (77 – 92%)	-
Risk**	low	low	low	moderate	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). #The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 500, 5000 and 150000), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). &The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.61, 0.73 and 0.81), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.84, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors. ##Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S26. Data set G and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;3)	3/5(60.0%)	3/3(100.0%)	8/30(26.7%)	$S_1 = 0.267$	extremely poor
2 (submodels;5)	3/5(60.0%)	17/26(65.4%)	22/75(29.3%)	$S_2 = 0.293$	extremely poor
4a (150000;0.37)	2/5(40.0%)	19297/750000(2.6%)*	19297/750000(2.6%)*	$S_{4a} = 0.025$	extremely poor
4b (10;0.13)	0.13	0.13	0.13	$S_{4b} = 0.13$	extremely poor
5a(*t*-test;0.05)	2/5(40.0%)	3/3(100.0%)	6/15(40.0%)	$S_{5a} = 0.400$	poor
5b(*t*-/*F*-test;0.05)	3/5(60.0%)	3/3(100.0%)	9/15(60.0%)	$S_{5b} = 0.600$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (3). Level 2: No. descriptors (5), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 150000 random descriptors is still not sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.37). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 10), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.13), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S27. Descriptor statistics* in terms of sign changes for data set G and its models.

Level	logK_{ow}	pK_a	E_{LUMO}	E_{HOMO}	N_{Hdon}	Total
1	0/6(0%)	1/6(16.7%)	0/6(0%)	4/6(66.7%)	3/6(50.0%)	8/30(26.7%)
2	0/15(0%)	6/15(40.0%)	0/15(0%)	8/15(53.3%)	8/15(53.3%)	22/75(29.3%)
4a	0(0%)[##]	(≈0.4%)[##]	0(0%)[##]	(12-13%)[##]	0(0%)[##]	(≈2.6%)[##]
4b[#,&]	>>160000	50	>>160000	10	>>160000	10
	>>0.37	0.15	>>0.37	0.13	>>0.37	0.13
5a	0/3(0%)	3/3(100.0%)	0/3(0%)	3/3(100.0%)	0/3(0%)	6/15(40.0%)
5b	0/3(0%)	3/3(100.0%)	3/3(100.0%)	3/3(100.0%)	0/3(0%)	9/15(60.0%)
Visual check	excellent	problematic	problematic	problematic	acceptable	problematic
Total score[###]	30.5 – 35.5 (86 – 100%)	11.5 – 16.5 (32 – 46%)	24.5 – 29.5 (69 – 83%)	6 – 11 (17 – 31%)	21.5 – 26.5 (61 – 75%)	-
Risk**	low	high to very high	moderate	high to very high	high to very high	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 10 and 50), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>160,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.13 and 0.15), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.37, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors. [##]This performance was predicted from plots (Figures 1-4 in the article) based on the fact that $\rho_{rd} = 0.40$ was not reached up to 150,000 random vectors applied. [###]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S28. Data set H and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;1)	0/4(0%)	0/1(0%)	0/8(0%)	$S_1 = 0$	excellent
2 (submodels;4)	0/4(0%)	0/1(0%)	0/28(0%)	$S_2 = 0$	excellent
3a (pool;2;29)	1/4(25.0%)	4/116(3.4%)*	4/116(3.4%)*	$S_{3a} = 0.034$	acceptable
3b (subset;2;11)	1/4(25.0%)	3/44(6.8%)*	3/44(6.8%)*	$S_{3b} = 0.068$	good
4a (25;0.42)	0/4(0%)	0/100(0%)*	0/100(0%)*	$S_{4a} = 0$	excellent
4b (70000;0.76)	0.76	0.76	0.76	$S_{4b} = 0.76$	excellent
5a(*t*-test;0.05)	2/4(50.0%)	1/1(100.0%)	2/4(50.0%)	$S_{5a} = 0.500$	poor
5b(*t*-/*F*-test;0.05)	0/4(0%)	0/2(0%)	0/4(0%)	$S_{5b} = 0$	excellent

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (1). Level 2: No. descriptors (4), which is the maximum complexity of the multivariate model considered. Level 3a: the maximum complexity (*l* value) of the *l*-variate MLR models considered (2) that could be treated computationally; number of all descriptors excluding the selected descriptors (29). Level 3b: the maximum complexity (*l* value) of the *l*-variate MLR models considered (2) that could be treated computationally; number of descriptors used for testing (11), from a descriptor pool subset (set Id from the original publication). Level 4a: number of random descriptors (generating 25 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.42). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 70,000), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.76), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S29. Descriptor statistics* in terms of sign changes for data set H and its models.

Level	S_6'	P_{5X}	M_{04}	M_{11}	Total
1	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/8(0%)
2	1/7(0%)	1/7(0%)	1/7(0%)	1/7(0%)	0/28(0%)
3a	4/29(13.8%)	0/29(0%)	0/29(0%)	0/29(0%)	4/116(3.4%)
3b	3/11(27.3%)	0/11(0%)	0/11(0%)	0/11(0%)	3/44(6.8%)
4a	0/25(0%)	0/25(0%)	0/25(0%)	0/25(0%)	0/100 (0%)
4b[#,&]	>>200000; >>0.85	>>200000; >>0.85	160000; 0.84	160000; 0.84	70000; 0.76
5a	1/1(100.0%)	0/1(0%)	0/1(0%)	1/1(100.0%)	2/4(50.0%)
5b	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/4(0%)
Visual check	problematic	excellent	acceptable	acceptable	problematic
Total score[##]	28 (79%)	35.5 (100%)	34.5 (97%)	32.5 (92%)	-
Risk**	moderate	no risk to very low	no risk to very low	low	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 70,000 and 16,000), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude).[&]The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.76 the model and 0.84), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.85, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the

risk of taking into account all selected descriptors. [##]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S30. Data set I and its model statistics# in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;1)	0/8(0%)	0/2(0%)	0/40(0%)	$S_1 = 0$	excellent
2 (submodels;8)	1/8(12.5%)	2/247(0.8%)	2/1016(0.2%)	$S_2 = 0.002$	acceptable
4a (2000;0.41)	0/8(0%)	0/16000(0%)*	0/16000(0%)*	$S_{4a} = 0$	excellent
4b (30000;0.50)	0.50	0.50	0.50	$S_{4b} = 0.50$	acceptable
5a(*t*-test;0.05)	0/8(0%)	0/1(0%)	0/8(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	1/8(12.5%)	1/1(100.0%)	3/8(37.5%)	$S_{5b} = 0.375$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. #ISCP parameters. Level 1: No. models considered (1). Level 2: No. descriptors (8), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 2000 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.41). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 30,000), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.50), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ was determined).

Table S31. Descriptor statistics* in terms of sign changes for data set I and its models.

Level	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	Total
1	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/2(0%)	0/16(0%)
2	2/127(1.6%)	0/127(0%)	0/127(0%)	0/127(0%)	0/127(0%)	0/127(0%)	0/127(0%)	0/127(0%)	2/1016(0.2%)
4a	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/2000(0%)	0/16000(0%)
4b[#,&]	>>200000	>>200000	>>200000	>>200000	>>200000	>>200000	30000	>>200000	30000
	>>0.56	>>0.56	>>0.56	>>0.56	>>0.56	>>0.56	0.50	>>0.56	0.50
5a	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/8(0%)
5b	1/1(100.0%)	0/1(0%)	0/1(0%)	0/1(0%)	0/1(0%)	1/1(100.0%)	1/1(100.0%)	0/1(0%)	3/8(37.5%)
Visual check	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic
Total score[##]	22.5 – 27.5 [63 – 77%]	28.5 – 33.5 [80 – 94%]	28.5 – 33.5 [80 – 94%]	28.5 – 33.5 [80 – 94%]	28.5 – 33.5 [80 – 94%]	24.5 – 29.5 [69 – 83%]	24.5 – 29.5 [69 – 83%]	28.5 – 33.5 [80 – 94%]	-
Risk**	high to very high	low	low	low	low	moderate	moderate	low	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). #The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 30,000), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). &The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.50), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.68, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors. ##Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S32. Data set J and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	1/4(25.0%)	2/2(100.0%)	2/20(10.0%)	$S_1 = 0.10$	poor
2 (submodels;4)	0/4(0%)	0/2(0%)	0/28(0%)	$S_2 = 0$	excellent
4a (50;0.40)	1/4(25.0%)	17/200(8.5%)*	17/200(8.5%)*	$S_{4a} = 0.085$	extremely poor
4b (10;0.25)	0.25	0.25	0.25	$S_{4b} = 0.25$	extremely poor
5a(*t*-test;0.05)	1/4(25.0%)	1/2(50.0%)	1/12(8.3%)	$S_{5a} = 0.083$	acceptable
5b(*t*-/*F*-test;0.05)	2/4(50.0%)	2/2(100.0%)	4/12(33.3%)	$S_{5b} = 0.333$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (2). Level 2: No. descriptors (4), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 50 random descriptors is sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.40). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously (in this case, 10), or the maximum number of bivariate models tested if sign change has not been observed; the minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.25), or the maximum ρ_{rd} reached if sign change has not been observed. Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">" if only probable region for ρ_{rd} was determined).

Table S33. Descriptor statistics* in terms of sign changes for data set J and its models.

Level	E1v	IC5	LP1	RDF125m	Total
1	0/5(0%)	0/5(0%)	0/5(0%)	2/5(40.0%)	2/20(10.0%)
2	0/7(0%)	0/7(0%)	0/7(0%)	0/7(0%)	0/28(0%)
4a	0/50(0%)	0/50(0%)	0/50(0%)	17/50(34.0%)	17/200(8.5%)
4b[#,&]	>>200000	>>200000	>>200000	10	10
	>>0.68	>>0.68	>>0.68	0.25	0.25
5a	0/3(0%)	1/3(33.3%)	0/3(0%)	0/3(0%)	1/12(8.3%)
5b	0/3(0%)	1/3(33.3%)	0/3(0%)	3/3(100.0%)	4/12(33.3%)
Visual check	acceptable	acceptable	problematic	problematic	problematic
Total score[##]	29.5 – 34.5 (83 – 97%)	25 – 30 (70 – 85%)	28.5 – 33.5 (80 – 94%)	12.5 – 17.5 (35 – 49%)	-
Risk**	low	moderate	low	high to very high	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously (in this case, 10), or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>200,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously (in this case, 0.25), or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.68, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).
**Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors.
[##]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S34. Data set K and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;2)	3/10(30.0%)	2/2(100.0%)	6/50(12.0%)	$S_1 = 0.120$	extremely poor
2 (submodels;10)	8/10(80.0%)	989/1013(97.6%)	2481/5110(48.6%)	$S_2 = 0.486$	extremely poor
4a (120000;0.29)	0/10(0%)	0/1200000(0%)*	0/1200000(0%)*	$S_{4a} = 0$	excellent
4b (120000;0.29)	>>0.29	>>0.29	>>0.29	$S_{4b} >> 0.29$	excellent
5a(*t*-test;0.05)	0/8(0%)	0/1(0%)	0/30(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	3/10(30.0%)	2/2(100.0%)	5/30(16.7%)	$S_{5b} = 0.167$	extremely poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (3). Level 2: No. descriptors (10), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 120000 random descriptors is still not sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.29). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (in this case, 120,000); the minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (in this case, 0.29). Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>"only when probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **This performance was predicted from plots (Figures 1-4 in the article) based on the fact that no sign change has been observed up to 120,000 random vectors applied (memory limit reached).

Table S35. Descriptor statistics* in terms of sign changes for data set K and its models.

Level	$\mu_1\mu_2^{Std}$	μ_{10}^{Std}	$\mu_5^{Ab\text{-}R2}$	μ_1^{Hyd}	μ_1^{Dip2}	μ_3^{Van}	$\mu_1\mu_4^{Dip4}$	$\mu_4^{Ab\text{-}logL16}$	$\mu_4^{Ab\text{-}\Sigma\beta2O}$	μ_4^{Pols}	Total
1	2/5(40.0%)	0/5(0%)	2/5(40.0%)	0/5(0%)	2/5(40.0%)	0/5(0%)	0/5(0%)	0/5(0%)	0/5(0%)	0/5(0%)	6/50(12.0%)
2	455/511 (89.0%)	60/511 (11.7%)	314/511 (61.4%)	0/511(0%)	466/511 (91.2%)	0/511(0%)	387/511 (75.7%)	187/511 (36.6%)	241/511 (47.2%)	371/511 (72.6%)	2481/5110 (48.6%)
4a	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]
4b[#,&]	>>120000	>>120000	>>120000	>>120000	>>120000	>>120000	>>120000	>>120000	>>120000	>>120000	>>1200000
	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29	>>0.29
5a	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/30(0%)
5b	0/3(0%)	1/3(33.3%)	0/3(0%)	0/3(0%)	1/3(33.3%)	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	3/3(100.0%)	5/30(16.7%)
Vis. ch.	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic	problematic
Total score[###]	20.5 – 25.5 [58 – 72%]	21.5 – 26.5 [61 – 75%]	20.5 – 25.5 [58 – 72%]	29.5 – 34.5 [83 – 97%]	20.5 – 25.5 [58 – 72%]	28.5 – 33.5 [80 – 94%]	24.5 – 29.5 [69 – 83%]	24.5 – 29.5 [69 – 83%]	24.5 – 29.5 [69 – 83%]	24.5 – 29.5 [69 – 83%]	-
Risk**	high to very high	high to very high	high to very high	low	high to very high	low	moderate	moderate	moderate	moderate	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). [#]The minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>120,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude). [&]The minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.29, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).

**Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors.
##This performance was predicted from plots (Figures 1-4 in the article) based on the fact that no sign change has been observed up to 120,000 random vectors applied (memory limit reached).
###Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).

Table S36. Data set L and its model statistics[#] in terms of sign changes.

ISCP level	Descriptors	Models	Regr. coefficients	Calculated index	Performance
1 (models;3)	0/4(0%)	0/2(0%)	3/24(12.5%)	$S_1 = 0.125$	extremely poor
2 (submodels;4)	1/4(25.0%)	4/11(36.4%)	4/44(9.1%)	$S_2 = 0.091$	poor
4a (60000;0.19)	0/4(0%)	0/240000(0%)*	0/24000(0%)*	$S_{4a} = 0$	excellent**
4b (60000;0.19)	>>0.19	>>0.19	>>0.19	$S_{4b} >> 0.19$	excellent**
5a(*t*-test;0.05)	0/4(0%)	0/3(0%)	0/12(0%)	$S_{5a} = 0$	excellent
5b(*t*-/*F*-test;0.05)	1/4(25.0%)	1/3(33.3%)	1/12(8.3%)	$S_{5b} = 0.083$	poor

*Only parameters (regression coefficients) for the selected descriptors are taken into account; parameters for other (real and random) variables are not of interest; therefore, the models and descriptors have the same statistics in such cases. [#]ISCP parameters. Level 1: No. models considered (3). Level 2: No. descriptors (4), which is the maximum complexity of the multivariate model considered. Level 4a: number of random descriptors (generating 60,000 random descriptors is still not sufficient to obtain correlation coefficient with respect to *y* around 0.40); correlation coefficient for randomization, ρ_{rd} (in this case, 0.19). Level 4b: the minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (in this case, 60,000); the minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (in this case, 0.19). Level 5a and 5b: confidence level α, the probability threshold (0.05). Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>"only when probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **This performance was predicted from plots (Figures 1-4 in the article) based on the fact that no sign change has been observed up to 60,000 random vectors applied (memory limit reached).

Table S37. Descriptor statistics* in terms of sign changes for data set L and its models.

Level	HOMO	nX	CIC0	nCAH	Total
1	0/6(0%)	0/6(0%)	0/6(0%)	3/6(50.0%)	3/24(12.5%)
2	0/11(0%)	0/11(0%)	0/11(0%)	4/11(36.4%)	4/44(9.1%)
4a	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]	0(0%)[##]
4b[#,&]	>>60000	>>60000	>>60000	>>60000	>>240000
	>>0.19	>>0.19	>>0.19	>>0.19	>>0.19
5a	0/3(0%)	0/3(0%)	0/3(0%)	0/3(0%)	0/12(0%)
5b	0/3(0%)	0/3(0%)	0/3(0%)	1/3(33.3%)	1/12(8.3%)
Visual check	acceptable	problematic	acceptable	problematic	problematic
Total score[###]	29.5 – 34.5 (83 – 97%)	28.5 – 33.5 (80 – 94%)	29.5 – 34.5 (83 – 97%)	16.5 – 21.5 (46 – 61%)	-
Risk**	low	low	low	high to very high	-

*Reported parameters for all ISCP levels are relative sign change frequencies expressed as ratios and percentages (in brackets), except for level 4b for which the value of ρ_{rd} is reported (marked with sign ">>" if only probable region for ρ_{rd} was determined, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1).

[#]The minimum number of bivariate models at which sign change starts occurring continuously, or the maximum number of bivariate models tested if sign change has not been observed (reported as a lower limit >>60,000, meaning that the true number of random vectors must be greater than the limit by one or more orders of magnitude).

[&]The minimum ρ_{rd} at which sign change starts occurring continuously, or the maximum ρ_{rd} reached if sign change has not been observed (reported as a lower limit >>0.19, meaning that the value of ρ_{rd} must be far from the lower limit i.e. close to 1). **Descriptors are characterized as bearing low, moderate and high risk of the sign change problem for the use in multivariate modelling. Total risk means the risk of taking into account all selected descriptors. [##]This performance was predicted from plots (Figures 1-4 in the article) based on the fact that no sign change has been observed up to 60,000 random vectors applied (memory limit reached).

[###]Total score is expressed in terms of the sum of score contributions along the ISCP check levels (rules given in Table 4), and as percentage of the maximum value of 35.5 (given in brackets).