

A Combinatorial Approach to the Variable Selection in Multiple Linear Regression: Analysis of Selwood *et al* Data set -A Case Study.*

Yenamandra S. Prabhakar[♦]
Medicinal Chemistry Division,
Central Drug Research Institute, Lucknow – 226 001, India.

Key words:

Regression analysis; variable selection; combinatorial approach; antimycin A₁ analogues; antifilarial activity

Abstract:

A combinatorial protocol (**CP**) is introduced here to interface it with the multiple linear regression (**MLR**) for variable selection. The efficiency of **CP-MLR** is primarily based on the restriction of entry of correlated variables to the model development stage. It has been used for the analysis of Selwood *et al* data set [16], and the obtained models are compared with those reported from **GFA** [8] and **MUSEUM** [9] approaches. For this data set **CP-MLR** could identify three highly independent models (27, 28 and 31) with Q^2 value in the range of 0.632-0.518. Also, these models are divergent and unique. Even though, the present study does not share any models with **GFA** [8], and **MUSEUM** [9] results, there are several descriptors common to all these studies, including the present one. Also a simulation is carried out on the same data set to explain the model formation in **CP-MLR**. The results demonstrate that the proposed method should be able to offer solutions to data sets with 50 to 60 descriptors in reasonable time frame. By carefully selecting the inter-parameter correlation cutoff values in **CP-MLR** one can identify divergent models and handle data sets larger than the present one without involving excessive computer time.

* CDRI Communication No.6225.

This paper is dedicated to Professor Satya P. Gupta, Chemistry Department, Birla Institute of Technology and Science, Pilani – 333 031, India.

[♦] e-mail: yenpra@yahoo.com Tel.:+91-0522-2212411 Fax.:+91-0522-2223405

1 Introduction:

The chemical structure descriptor data is multidimensional in nature and consists of several properties, which include physicochemical, electronic and topological characteristics of the compound. In case the descriptors are of constituent atoms of the chemical structure, they address the fundamental properties, which are additive and constitutive in nature. Often all these properties of the compounds and their constituent atoms serve as independent descriptors (predictor or explanatory variables) of the chemical structure and contribute to the understanding of various physical, chemical and biological phenomena resulting from them. The derivation of an empirical relationship (model) between a chosen phenomenon of the compounds, a dependent variable, and their structural descriptors provides scope to understand the interrelationship between them. Different models address different sub-structural regions/ attributes in predicting the chosen phenomenon. While the simplest among the models will be the best (principle of parsimony) to explain the chosen phenomenon, the study of a population of various models provide scope to understand the diagnostic aspects of different sub-structural regions as well as in averaging and extrapolating the predictive aspect beyond the individual models. There are several approaches to unravel and address the relationships between dependent and independent descriptors. The principal components in multiple linear regression (**PCR**), partial least squares (**PLS**)[1,2], artificial neural networks (**ANN**)[3-6], cluster significance analysis (**CSA**) in association with stepwise regression [7] and the algorithms namely Genetic Function Approximation (**GFA**) [8], Mutation and Selection Uncover Models (**MUSEUM**) [9,10], combination of Genetic algorithm and Neural Network (**GNN**) [11] and Fast Random Elimination of Descriptors (**FRED**) [12] are some of the useful methods in this regard. The **PCR** and **PLS** methods involve the extraction of orthogonal components of the independent descriptors, which can even outnumber the observations (data points) and as such do not call for any variable selection procedure in the model development. The **ANN** is generally used to model the non-linear relationships between dependent and independent descriptors. In **ANN**, during the training process, the weights of the independent descriptors are adjusted to make the output of the network close to the expected results. In **CSA**, a p-value, the measure of relevance of a descriptor to the activity, provides rating of the independent variables for inclusion in stepwise regression for model development [7]. In **GFA** and **MUSEUM** approaches, sufficiently large 'crossings' or 'mutations' of multiple or single models are attempted in conjugation with predefined selection criteria (fitness criteria) to arrive at a population of significant equations to explain the chosen phenomenon. In **GNN** molecular descriptors are selected by a genetic algorithm to serve the input to a neural network for mapping the activity [11]. **FRED** algorithm is a random descriptor selection (elimination) strategy for model development without direct crossing and mutation of previous generations [12]. Recently some more approaches based on the genetic algorithm, namely Genetic Algorithm guided Selection (**GAS**) [13], Genetic Quantitative Structure-Activity Relationship (**GPQSAR**) and Multiobjective genetic QSAR (**MoQSAR**) [14] have been proposed for variable selection in model development. These methods are reported to be fast and efficient in handling large number of variables in the model development. As genetic and mutation algorithms are probabilistic approaches, they often offer different 'best' solutions (models) in different runs. Also, in both these approaches, the best model evolution criteria has little control on the selection/deletion of

independent variables with inter-correlations which may obscure the diagnostic and also the claimed predictive value of at least the correlated variables of a model. Correlation among the explanatory variables in a multiple regression model is a complex problem [15]. In this background for model development a combinatorial protocol (CP) for variable selection and significance evaluation in multiple linear regression (MLR) is discussed. The layers of significance evaluators, called as filters, incorporated in the procedure make this process efficient and offer unique solutions. It has been applied to Selwood *et al* data set [16] and the results are compared with those obtained from GFA and MUSEUM approaches [8,9].

2 Method:

Combinatorial Protocol: To extract the maximum and diverse structure-property relationship information from the parameter set considered in MLR, a combinatorial strategy with appropriately placed ‘filters’ is adopted to recurrently select the non-repetitive k independent variables, at a time, from a total of p variables for the model development. If we call a group of variables as bundle, then according to the combination rule, a total of pC_k bundles emerge from p variables with k variables in each bundle (original variable bundle, OVB). A variable may contribute to a model in two different ways: (i) by itself alone and / or (ii) by itself and its functionally transformed term together. To find the influence of a selected function of any variable along with its original form in the model development, the k variables of OVB along with their meaningfully transformed functional variables are adopted for the formation of new bundles. We may mention it here that the functionally transformed variable enters the bundle only when its original variable is part of that bundle. The OVB provides k variables to create the functionally transformed variable bundles (TVBs). In this process, the contents and number of variables in the TVBs (s) are varied from **one** to k' to explore the role of functionally transformed variable combinations (kC_s ; $s = 1$ to k' where $k' \leq k$) along with respective OVB. Furthermore, the size of the OVB (k) is also varied from a minimum (begin, b) to maximum (end, e) value with an increment of **one**. The e value will be governed by the number of observations (n) with n/e ratio as large as possible ($1 \leq b \leq e \leq n/e$). This process generates different sizes of OVBs of all variables within the limits of k and joins TVBs of different sizes, to the respective OVBs to form OVB-TVBS. It may be visualized as propagation of roots of a monocotyledon plant and offers scope for the examination and evolution of the best and most meaningful models in multidimensional space. Symbolically the bundles of variables and the search perimeter of the model may be expressed as:

$$\begin{array}{c}
 e \\
 [\\
 k=b \quad i=1
 \end{array}
 \begin{array}{c}
 {}^pC_k \\
 [\\
 i=1
 \end{array}
 (\text{COMB}_k)_i ; \quad
 (\text{COMB}_k)_i \left[\begin{array}{c}
 k' \\
 [\\
 s=1
 \end{array}
 \begin{array}{c}
 {}^kC_s \\
 [\\
 j=1
 \end{array}
 f(\text{COMB}_k)_{i,s,j} ; \right]]]]$$

In this formalism, p is the total number of variables considered in the study, k is the number of variables in a bundle whose value vary from b (beginning) to e (end) and indicate the search perimeter of the models, $(\text{COMB}_k)_i$ refers to the i^{th} OVB with k variables, $f(\text{COMB}_k)_{i,s,j}$ refers to the j^{th} TVB with s functionally transformed variables

of i^{th} **OVB** with **k** variables and joined to it. The pair of outer square brackets (SB) with **k** from **b** to **e** denotes the search domain consisting of varying sizes of **OVB** and corresponding **OVB-TVb** combinations. The immediate next SB, a subset of outer SB, denotes its contents as different **OVBs** of fixed size and respective **OVB-TVb** combinations emerging from them. The semicolon (;) in the formalism demarcates the **OVB** from **OVB-TVb** as distinct variable bundles. The second innermost SB denotes the variation in the size of **TVb** (**s**) from **one** to **k'** ($k' \leq k$) of respective **OVBs**. The innermost SB denotes its contents as different **TVBs** of fixed size emerging from each **OVB**. If no functional transformation is considered in the analysis, all the terms and corresponding SBs after the semicolon will become null and void. If the size of **OVB** (**k**) is restricted to a single value, the outer SB encircles only different **OVBs** of fixed size. **Table 1** shows the bundles formation from three descriptors (for example, **A B C**), $p=3$, with **k** varying from 1(**b**) to 3(**e**) and **s** varying from 1 to **k**. For efficient evaluation of variable bundles and model collection, four layers of 'filters' with statistically sound threshold level are set in terms of inter-parameter correlation cutoff criteria for variables to stay as a bundle (filter-1), *t*-values of regression coefficients of variables associated with a bundle (filter-2), square-root of adjusted multiple correlation coefficient of regression equation (*r*-bar, pronounced as '*r*-bar', equation 1) [15] (filter-3)(**for the clarity of differentiation of square-root of adjusted multiple correlation coefficient and multiple correlation coefficient, *r*-bar is adopted for the former instead of the normal notation of '*r*' with a 'bar' on it**), and Cross-Validated R^2 (Q^2 , equation 3) [9] criteria (filter-4).

$$(\bar{r})^2 = (1.0 - (1.0 - r^2)(n-1)/(n-k-1)) \quad (1)$$

$$r^2 = 1.0 - (\sum(Y_c - Y_o)^2 / \sum(Y_o - Y_m)^2) \quad (2)$$

$$Q^2 = 1.0 - (\sum(Y_p - Y_o)^2 / \sum(Y_o - Y_m)^2) \quad (3)$$

$$\text{SPRESS} = \sqrt{(\sum(Y_p - Y_o)^2 / (n - k - 1))} \quad (4)$$

$$\text{SDEP} = \sqrt{(\sum(Y_p - Y_o)^2 / n)} \quad (5)$$

In above expressions Y_o , Y_m , Y_c and Y_p are observed, mean, calculated and predicted values, respectively, of dependent variable, n is number of observations and k is number of independent variables in regression equation.

The filter-1 controls the entry of **OVBs** with inter-correlated variables in the model development. The default cutoff value for the tolerance of inter-parameter correlation coefficient between pairs of independent variables is set as less than or equal to 0.3 to maintain reasonably good independence among the variables of a bundle. The efficiency of **CP-MLR** is primarily based on this filter. A Pentium-4 personal computer with a 1.5GHz processor can evaluate variable bundles of the order of 10^5 per minute. The second filter (filter-2) evaluates the significance of variables in a bundle in terms of the *t*-values of regression coefficients. A default value of 2.0 is set for this filter; a bundle will pass this filter if the *t*-values of its regression coefficients are more than or equal to the set threshold value. Normally, successive additions of variables to multiple regression equation will increase successive multiple correlation coefficient (*r*) values. In light of this, to compare the internal explanatory power of bundles with different number of variables (variable bundles of different sizes), *r*-bar is adopted in this procedure [15]. Accordingly, filter-3 sets predefined threshold level for *r*-bar. A threshold value for *r*-bar in between 0.7 to 0.8 provides a good starting point for the collection of bundles for further evaluation in the model development process. Only those variable bundles whose

r -bar with the dependent variable is more than or equal to the set threshold level pass will this filter. Also, inflation in r^2 (or r) is associated with the best subset regressions, especially when the total number of predictor variables (p) (corresponding to all bundles) is more than the number of observations (data points) [17]. The number of predictor variables could be very large in many a situations. The inflation in r^2 may be meaningful only when there is increase in the number of relevant descriptors corresponding to the phenomenon under investigation. However, to exclude false or artificial correlations arising out of these situations, it is important to validate the relevance of selected bundles in the model generation. While the first three filters are designed to check the internal consistency of the data, the fourth filter (filter-4) addresses the external consistency in the form of cross-validation of the model with leave-one-out procedure as the default option. Finally, 'goodness of fit' of the model is measured in terms of Q^2 . A Q^2 is considered to be acceptable only when it has a value between zero (no predictive power) and one (perfect predictive power). A value between zero and one for Q^2 will result in the collection of a good number of models with different degree of predictive power. Only those models whose Q^2 value is in the predefined limits are retained for the further study. All computations of this study are performed using self-written programs in GWBASIC. The regression and statistical computation aspects of the self-written programs are compared and verified with well-defined test data sets and commercial software, SYSTAT [18].

3 Case Study (Swlwood *et al* data set):

Selwood *et al* have generated a data set of 53 descriptors (physicochemical properties, $p=53$) for 31 antifilarial antimycin A₁ analogues to study their structure-activity relationships [16]. They reduced the number of descriptors with a pattern recognition technique and applied regression analysis on the reduced data set to identify the significant equations. This suggested the importance of melting point (MPNT), octanol-water partition coefficient (LogP) and electrophilic superdelocalizability of atom-10 (ESDL10) in modeling the activity [16]. Apart from Selwood *et al*, various groups adopted different approaches on this data set to find the best QSAR model(s) [4,7-14]. Notable among these are Wikel and Dow's neural networks [4], McFarland and Gans' CSA to identify the most likely determinant variables [7], Rogers and Hopfinger's GFA [8], and Kubinyi's MUSEUM [9] algorithms. The GFA algorithm on Selwood *et al* data resulted in proposing top 20 models of 2 or 3 variables from a population of 300 [8]. From the same data set, with the MUSEUM algorithm, Kubinyi identified 17 models, having 2 to 6 variables [9]. **As Selwood *et al* data set is one of the well studied and worked out data sets, along with result oriented calculations we attempted some time-consuming simulations to minimize the educated guess about search perimeter and to get a clear picture about the approach (CP-MLR). Table 2 lists the names of all 53 descriptors of Selwood *et al* data set along with a serial number that is identical to the one assigned by Kubinyi [9].** An examination of the correlation matrix of descriptors of 20 models identified through GFA algorithm has shown several inter-parameter correlations. In all these 20 models, logP is one common parameter. The other descriptor of these models is provided by either SURFA or MOFIY or PEAXX. All the above four descriptors are highly inter-correlated ($r \geq 0.714$). In fact, even in all the 17 models of Kubinyi the common descriptor is logP. One of the other descriptors was

provided by either VDWVOL, or SURFA or MOFIY or MOFIZ or PEAXX. All these five descriptors are highly correlated with logP ($r \geq 0.714$). Moreover, in some other equations NSDL8, NSDL9 and NSDL10 (inter-parameter correlation range is 0.917 to 0.987) or two of them were selected as part of the descriptor set along with logP and one of the other five descriptors. However, for **MUSEUM** runs Kubinyi used a version of **PLS** analysis, instead of regression analysis, **with number of components equal to number of variables in the equation, which is theoretically and numerically equal to using MLR [9]**. With this in the background, an attempt is made to analyze this data set with **CP-MLR**. For this data the filter-1 was assigned with a value of 0.3, and filter-2 with 2.0. The earlier studies reported a good number of equations with an r equal to or more than 0.8. In terms of \bar{r} a value of 0.74 for filter-3 should select all possible equations within the proximity of 0.8 and above. Also, for this data, previous studies reported models with Q^2 value starting from 0.3 onwards. Accordingly, filter-4 is set to collect all models with Q^2 value between 0.3 and 1.0 ($0.3 \leq Q^2 \leq 1.0$). As a good number of 2 - 6 parameter models was reported earlier, a search is carried out for models up to 6 variables (**OVBs** only) and without any functional transformation terms ($p=53$, $k=2$ to 6, $s=0$) (computation time: ~4:30hr) which resulted in the identification of only two models (equations 6 and 7; the values given in parentheses are 95% confidence levels of regression coefficients).

$$\begin{aligned}
 -\log EC_{50} &= 16.704(\pm 5.572)ATCH1 + 4.454(\pm 2.307)ATCH4 \\
 &\quad - 0.120(\pm 0.109)DIPVX + 0.086(\pm 0.056)DIPVY - 2.742 \\
 r &= 0.811, \quad s = 0.519, \quad Q^2 = 0.538, \quad F = 12.49
 \end{aligned} \tag{6}$$

$$\begin{aligned}
 -\log EC_{50} &= 3.672(\pm 2.296)ATCH4 + 31.334(\pm 10.790)ATCH7 \\
 &\quad - 0.126(\pm 0.107)DIPVX + 0.106(\pm 0.058)DIPVY + 8.088 \\
 r &= 0.803, \quad s = 0.520, \quad Q^2 = 0.515, \quad F = 11.80
 \end{aligned} \tag{7}$$

It prompted us to reconsider the analysis of the data set along with the functionally transformed descriptors. As squared term of a descriptor is the most often and widely used functional transformation of variables in QSAR and QSPR studies, the same is opted for this data set also. The data is reanalyzed with the same threshold levels of the filters, but with added squared terms under **TVBs**. The search perimeter is set as up to 5 original variables (**OVBs**) and up to 3 transformed variables (**TVBs**) ($k=1$ to 5; $s=1$ to 3) (computation time: ~4hr) that resulted in the identification 34 models, which includes the above two models, with t-values of regression coefficients significant at more than 95% level. **The extended searches with 6 and 7 variables in OVBs along with up to 3 variables in TVBs did not yield any more models** (computation time: ~8 and ~30hr, respectively) and indicates that $k=2$ to 5 and $s=1$ to 3 as the optimum perimeter for model extraction for the filter thresholds adopted. **Table 3** lists all the models emerged from the search.

4 Discussion:

In **Table 3**, the descriptors of models are identified with the serial number allotted to them in **Table 2**; a character 's' after descriptor number indicates that both the descriptor and its squared term are involved in the equation. In models 32 to 39, the regression

coefficients of some of the descriptors (identified with an asterisk sign before the descriptor number) and their squared terms have the same mathematical sign (both are positive or both are negative), which is not normally practiced and not expected in QSAR studies, hence they are discounted from considering as good models. In terms of size of **OVB**, the models contain 2 to 5 descriptors with a majority of them emerging from **k=4**. In terms of **OVB** and/ or **OVB-TVb**, the discovered models contain parameters ranging from three (only one model) to seven (only one good model). This study could identify only two models (27 and 28) with Q^2 values above 0.6 as shown.

$$\begin{aligned}
 -\log EC_{50} &= 2.974(\pm 1.771)ATCH4 + 0.183(\pm 0.116)LogP \\
 &+ 3.626(\pm 1.146)NSDL8 - 0.466(\pm 0.155)(NSDL8)^2 \\
 &+ 2.449(\pm 1.500)PEAXY - 0.765(\pm 0.433)(PEAXY)^2 - 7.337 \\
 r &= 0.894, \quad s = 0.413, \quad Q^2 = 0.632 \quad F = 15.99 \quad (27)
 \end{aligned}$$

$$\begin{aligned}
 -\log EC_{50} &= 4.177(\pm 2.182)ATCH4 + 37.093(\pm 10.679)ATCH7 \\
 &- 0.129(\pm 0.097)DIPVX + 0.138(\pm 0.056)DIPVY \\
 &+ 3.156(\pm 2.285)S8IDZ - 0.254(\pm 0.182)(S8IDZ)^2 - 0.092 \\
 r &= 0.859, \quad s = 0.473, \quad Q^2 = 0.616, \quad F = 11.26 \quad (28)
 \end{aligned}$$

$$\begin{aligned}
 -\log EC_{50} &= -0.527(\pm 0.499)ESDL10 + 4.003(\pm 1.581)NSDL8 \\
 &- 0.502(\pm 0.211)(NSDL8)^2 + 4.144(\pm 2.737)S8IDZ \\
 &- 0.308(\pm 0.213)(S8IDZ)^2 - 0.034(\pm 0.023)MPNT \\
 &+ 0.000121(\pm 0.000074)(MPNT)^2 - 17.568 \\
 r &= 0.856, \quad s = 0.488, \quad Q^2 = 0.518, \quad F = 8.98 \quad (31)
 \end{aligned}$$

The above three models share only one common descriptor between two, either ATCH4 or NSDL8; and an examination of the combined correlation matrix of the descriptors of these models indicates that they are highly independent from one another (**Table 4**). All identified models, 26 in number, together share 19 descriptors, and the above three models alone, put together, share 10 descriptors among themselves. From this, the models presented in **Table 3** can broadly categorized into three groups. The first group, from logP (50), and the second group, from MPTN (51), have, respectively, emerged from 2 and 3 size **OVBs**, (**from models 8 and 9, respectively**) **but could not propagate well along the search line**. The descriptors NSDL8 (32) and NSDL9 (33) are part of Models 8 and 9 respectively, and as already stated they are highly correlated ($r=0.987$). Model 28, **a model emerged from OVB of size 5**, belongs to the third group and has its origin in **OVB of size 4 (model 7)**. **Also, model 28 is the last model of the whole search**. Most of the models presented in **Table 3** share many a common features with model 28. Most of the identified models have Q^2 value in the range of 0.4 – 0.5. Majority of the models share ATCH4(4) as one common descriptor. The other common descriptor of the models is either ATCH1(1) or ATCH7(7). The models 12 to 15 are very closely related. Moreover, models 10 and 11 are subsets to models 12 to 15. These models are formed by inter changing of descriptors MOFIY (38) with MOFIZ (39) and NSDL2 (26) with NSDL4 (28). The descriptors MOFIZ and MOFIY are highly correlated ($r=0.997$), and NSDL2 and NSDL4 are also correlated to a considerable extent ($r=0.780$). These models

may be treated as replica of one another. The equations of models 12 and 13 are as shown below.

$$\begin{aligned}
 -\log EC_{50} &= 3.674(\pm 2.132)ATCH4 + 31.164(\pm 10.242)ATCH7 \\
 &\quad - 0.072(\pm 0.062)NSDL4 + 1.475E-4(\pm 8.471E-5)MOFIZ \\
 &\quad - 3.186E-9(\pm 1.616E-10)(MOFIZ)^2 + 6.674 \\
 r &= 0.835, \quad s = 0.498, \quad Q^2 = 0.551, \quad F = 11.55
 \end{aligned}
 \tag{12}$$

$$\begin{aligned}
 -\log EC_{50} &= 3.708(\pm 2.134)ATCH4 + 30.679(\pm 10.292)ATCH7 \\
 &\quad - 0.070(\pm 0.062)NSDL4 + 1.507E-4(\pm 8.528E-5)MOFIY \\
 &\quad - 3.052E-9(\pm 1.539E-9)(MOFIY)^2 + 6.403 \\
 r &= 0.835, \quad s = 0.498, \quad Q^2 = 0.553, \quad F = 11.51
 \end{aligned}
 \tag{13}$$

At a glance, the regression coefficients of MOFIZ and MOFIY in models 12 and 13 may look abnormal. However, the order of the regression coefficient of this descriptor is same in the previous reports as well, except reversal of sign and no associated squared term [8,9]. The smaller magnitude of these terms is logical as their descriptor values are very large ($\sim 10^4$), and for the same reason a squared term of these descriptors may be essential to counter any indefinite linear expansion of their influence. **Table 2** shows the descriptors identified in this study (**Table 3**) along with those of **CSA** [7], **GFA** [8], and **MUSEUM** [9] approaches. Descriptors exclusive to models 32 to 39 are not included in this as these models are discounted. Even though, the present study does not share any models with **CSA** [7], **GFA** [8], and **MUSEUM** [9] results, there are several descriptors common to all these studies, including the present one. LogP is one common descriptor projected by all the previous studies. It is part of the variable bundle of the best model (27) identified in this study without any prior conditioning. The other descriptors common to all the listed approaches are ATCH4 and MOFIY. All the descriptors identified in the **CP-MLR** models offer similar physicochemical inferences like the previous studies on this data set, except where squared terms are involved, this may suggest the necessity of an optimum for the concerned descriptor property.

Variable Bundle formation in CP-MLR: Since inter-parameter correlation cutoff criteria plays a key role in building the efficiency of **CP-MLR**, a study is undertaken to examine trends of bundle formation with its increasing size. The inter-parameter correlation acts with dual character – as associating (at less than or equal to the cutoff limit) and dissociating (above the cutoff limit) force – in the formation of variable bundles. Since limiting the inter-parameter correlation to a cutoff level is the criteria for a variable bundle to be evaluated after filter-1 in **CP-MLR**, we introduce a term ‘population of descriptor’ to analyze the course of bundle formation and its influence on the model development. The population of a descriptor in bundle size **k** is defined as the number of times its occurrence in bundles that satisfied the inter-parameter correlation cutoff criteria. In any data set with **p** descriptors, for a descriptor to enter in to a bundle size **k+1** ($k+1 < p$), at least the concerned descriptor should be within the inter-parameter correlation cutoff limits with **k** other descriptors. Also, for a descriptor to enter in to the next higher size bundle, it should at least have a minimum population of 2 in the current

bundle. It means that a descriptor with a population less than 2 in bundle size k will not show its presence beyond this size. In other words, in normally or approximately normally distributed descriptors, for a given inter-parameter correlation cutoff limit, the study of populations of descriptors with respect to bundle size will give some idea about their population in higher size bundles. The population of descriptors in bundle size 2 will be the simplest among all. This indicates the degree of relatedness of each and every descriptor with all others. At this stage a descriptor can attain a maximum population of $p-1$ (totally unrelated character) and a minimum of **zero** (totally related character). The descriptor with higher population will go a long way in the bundle formation and may survive higher bundle sizes. As the attention is on the bundles and not on its components, the descriptors with lower populations act like dissociating forces for the next generation and set limits to the maximum bundle size that can be achieved by them as well as by the other descriptors. In the beginning stages of bundle formation the associating force of inter-parameter correlation takes lead and contribute to the increase in the population. With the increments in the bundle size, the dissociating force of inter-parameter correlation develops and soon reaches a matching level with its counterpart. The population reaches an optimum at this matching level of the dual character of inter-parameter correlation. After this matching level of dual character, the dissociating force, unlike the associating one, develops rapidly with additional descriptors in the bundle. At this stage addition of more descriptors to increase the bundle size aborts the whole bundle and soon the population becomes insignificant or zero. This collective behavior of descriptors in the formation of bundles has clear implications in the model development. If a high (H) population descriptor is significant to the model, it will try to show its presence from the very beginning as it has the maximum mixing opportunity to do so. If a low (L) population descriptor is relevant to the model, then it will be having better opportunity to show its presence up to reaching the optimum level. After the optimum population level, the L population descriptor opportunities to enter a bundle diminish rapidly. If the L population descriptors are not significant to the model, they will not be making any significant presence in the bundles beyond the optimum population. One can safely omit such descriptors without loss of information and enrich the data set with relevant and contributing descriptors to repeat the study. To demonstrate the above-discussed behavior of descriptors and bundles, a simulation has been carried out on Selwood *et al* data set in a stepwise manner to discover the populations of descriptors with respect to bundle size. For each k , all those bundle which satisfy the filter-1 cutoff criteria are counted for their contents to form the respective descriptor populations. If the population of a descriptor is less than 2 in bundle size k , then its population in bundle size $k+1$ is made zero and deleted from the list of descriptors from $k+1$ onwards. We made only one approximation in $k=9$ which saved us from lot of computation time without altering the results. Six descriptors have registered populations below 14 at $k=8$ (identified with * in **Table 5**), at $k=9$ we equated them to zero (which they may have become on their own, had they been allowed to mix with others). By doing so we reduced the computation time to 4 days from 14 days. The total simulation (from $k=2$ to 10) took about eleven and half days; the major breakups are as follows: $k=6$, 4hr, $k=7$, 1day, $k=8$, 6 days, $k=9$, 4 days and $k=10$, 6hr. **Table 5** projects the so obtained populations of Selwood *et al* data set descriptors for bundle sizes 2 to 10 with inter-parameter correlations less than or equal to 0.3. The columns of this table indicate the population of

each descriptor for a fixed bundle size, and the rows indicate the variation in the population of each descriptor with respect to varying bundle sizes. The mean of the population of all the descriptors at bundle size $k=2$ (mean=32.566) is taken to categorize them as low (L) and high (H) population ones. The descriptors with population more than or equal to 33 are designated as H group and the remaining are designated as L group. The 26 good models (**Table 3**) identified in **CP-MLR** share 19 descriptors among them. Of these 19 descriptors 10 come under H group and 9 fall in L group. From **Table 5**, it is clear that optimum descriptor population is at bundle sizes 5 (beginning) and 6 (end). Our search beyond bundle size $k=5$ coupled with **TVBs** did not yield any more models. In fact all the identified descriptors, excepting four, have reached their optimum population at bundle size $k=5$. Three H group descriptors (~2-22% increase in population) and one L group descriptor (~12% increase in population) touched optimum at bundle size $k=6$. **Table 6** shows a translation of the models listed in **Table 3** in the form of 'H' 'L' character strings corresponding to the group the respective descriptor belongs to. As we conjectured, all the identified descriptors of L group showed a significant presence much before bundle size $k=5$ where they reached the optimum population level. Also, the contributing H group descriptors exhibited their presence significantly all through the bundles till reaching optimum population level. In fact the last good model of the search (model 28) is rich with four H group descriptors and one L group descriptor. Beyond $k=6$ (the optimum population level) the fall in population is vary rapid as evidenced from bundle sizes 7, 8 and 9. We probed the descriptors, with non-zero population, at $k=9$ for the possibility of getting a model with 9 variables in **OVBs** and up to 3 variables in **TVBs** without any success (computation time: ~3:30hr). Among all the descriptors, the exception is the availability of an L group descriptor (ATCH10) at $k=10$ with a population of 5. ATCH10 population from bundle size 2 to 10 has a gradual change without steep increments or decrements. The reason for its availability at $k=10$ is that it maintained unrelated character with all H group descriptors – otherwise it would have disappeared much earlier. At $k=10$ only 5 bundles are left behind to be formed from 14 descriptors. Of these 14 descriptors, the population of 5 descriptors is below the critical number 2 making them ineligible to go to the next round ($k=11$). It means that only 9 descriptors are left behind to go to the next stage, $k=11$, which they can not go due to the shortage of company of two more descriptors to make the bundle. So, they disappeared prematurely at $k=10$ itself. Even though it is not examined, the probability of finding a ten descriptor model out of five bundles left out at $k=10$ has a far remote possibility. Hence the simulation logically explained the behavior of descriptors and bundles in **CP-MLR** in model development.

In a further study, to validate the **CP-MLR's** filter-directed search in identifying diverse models arising out of relaxed filter-1, a higher cutoff value of 0.79 is assigned to it. The filter-2, filter-3 and filter-4 are assigned the values of 2.0, 0.8 and 0.36 ($0.36 \leq Q^2 \leq 1.0$), respectively. With these, a search is carried out on Selwood *et al* data set for models in terms of 3 to 6 variables in **OVBs** ($k=3$ to 6, $s=0$), which resulted in the identification of 11 three-parameter models, 64 four-parameter models, 180 five-parameter models and 318 six-parameter models. For each k value, a few representative models out of this search are listed in **Table 7** along with their status. Many of Rogers and Hopfinger [8] and Kubinyi [9] models are part of the identified models of this search. A five-parameter

model (model 54; ATCH4, DIPVX, MOFIZ, LogP, SUMF or 4,11,39,50,52) identified in this search has the highest Q^2 value (0.699) among this class of models so far reported. Also, it is interesting to note that among the six-parameter models the best one (model 56) is identical with that of Kubinyi's best six-parameter model (4,11,38,48,50,52) with Q^2 value 0.754. Within the search perimeter there is no model better than this – for the opted filter values, it is an unequivocal assertion of the search procedure. Also, to explore the information content of 'identified' vs. 'left behind' descriptors (with respect to activity, the dependent variable), the 53 explanatory descriptors of Selwood *et al* data set are partitioned into two mutually exclusive subsets – one containing the 19 descriptors (data set A) of 26 good models listed in **Table 3** and the other containing the remaining 34 descriptors (data set B). A **CP-MLR** search (filter-1 as 0.79; filter-2 as 2.0; filter-3 as 0.74; filter-4 as $0.3 \leq Q^2 \leq 1.0$) on data set B for 2 to 6 variable models ($k=2$ to 6, $s=0$) in terms of **OVBs** did not result in the identification of even a single model. This clearly indicates that the descriptors of data set B have little or insufficient information (corresponding to the activity) among themselves – independently as well as collectively – to come up the level of model formation. On the other hand, one may efficiently reuse data set A in **CP-MLR** with relaxed filters for high Q^2 value models or in **PLS** for further development etc. **Table 8** lists a few such models of the lot emerged from the **CP-MLR** search (filter-1 as 0.79; filter-2 as 2.0; filter-3 as 0.80; **filter-4 as $0.63 \leq Q^2 \leq 1.0$** ; 53 models identified) of data set A in terms of 3 to 6 variables in **OVBs** and up to 2 variables in **TVBs** ($k=3$ to 6 and $s=1$ to 2). Model 71 (4,11,39,50s,51s,52; $Q^2=0.782$)(**Table 8**), a closely related one to Kubinyi's best models (models 56 to 59 in **Table 7**), has emerged from this search. These studies further explain the scope of **CP-MLR** in identifying models under varying conditions.

5 Conclusions:

In summary, the **CP-MLR** is a filter directed search algorithm for discovering divergent models. From the models point of view, in brief, the advantage is that it can identify models with highly independent explanatory descriptors. This provides good diagnostic and predictive aspects of the identified sub-structural regions independent of others influences. Also, the models identified are more divergent compared to those identified through **GFA** and **MUSEUM** approaches. In fact, being probability based approaches, **GFA** and **MUSEUM** is normally expected to provide divergent models with multiple convergences. Instead all the results of **GFA** and **MUSEUM** ended with single convergence in a narrow area. Comparisons of the distribution of descriptors in the models of respective procedures, including the present one, make this amply clear. Probably, **GFA** and **MUSEUM** approaches may suit better for data sets much larger than, say 2 to 3 times, that of Selwood *et al* data set. Inflation in r^2 in best subset regression is a phenomenon with all the approaches and it is more pronounced in **MUSEUM** results compared to that of **GFA** and **CP-MLR**. Coming to the predictive aspect, **MUSEUM** approach produced a model with Q^2 value of 0.745 and **CP-MLR** identified a model with Q^2 value of 0.632. Considering the fact that the procedure adopted in identifying the **CP-MLR** model is rigorous, the associated Q^2 may probably be the more realistic estimate. The provision of modulation of filter cutoff/ threshold values in **CP-MLR** gives it the flexibility to identify divergent models including those involving intercorrelated descriptors and high Q^2 values. We discounted all those models

from being considering as good ones where the regression coefficients of one or more descriptors and their squared terms have the same mathematical sign (both are positive or both are negative). These situations are normally avoided in view of the associated complexities in attributing physical meaning to such behavior of the descriptors in the diagnostic aspect of the model. However, in **GPQSAR** and **MoQSAR** approaches, in some cases transformed variable(s) alone has become part of the explanatory descriptors of the model(s) [14]. Unlike genetic and mutation approaches, the models obtained through **CP-MLR** are unique for the given filter values.

Apart from practical and economic considerations, one purpose of models is to forecast the events or phenomena of interest with far less number of observations (from the experimental domain) compared to the actual scope it holds. The collection of descriptors for an observation can be very large or infinite, but the number of observations sets the upper limits for the number of descriptors that can enter in to a model. For QSAR / QSPR studies, Unger and Hansch suggest five to six observations per descriptor to avoid chance correlations [19]. This provides a healthy estimate of search perimeter of descriptors that one can attempt to discover a model. In this background **CP-MLR** can comfortably and practically handle 50 to 60 descriptors within reasonable time frame. And it may produce results a little above the upper limit without involving excessive computer time. The cross-validation runs of this study are controlled through monitoring the accumulating predictive residual sum of squares in successive resamplings and the process is subjected to premature termination in cases of accumulated predictive residual sum of squares exceeding the sum of squared deviations of Y_o . The computation of externally standardized residuals with the help of leverage of data [20] may be helpful in the calculation of PRESS without successive resamplings. Incorporation of this shortcut in the computation may further reduce the time taken by **CP-MLR** for analysis. Also, the computation time may reduce marginally if one adopts some pre-processed data instead of processing the data while computing. The simulation carried out on Selwood *et al* data set demonstrates the behavior of descriptors/ bundles in **CP-MLR** in model development. The combination rule suggests the members of the bundle family, but whether the member is to be admitted or not will be decided by filter-1. By adjusting the inter-parameter correlation cutoff limits one can shift the optimum population of descriptors and change the course of bundle formation and associated results. A smaller cutoff value may be helpful in handling larger data, as this shifts the optimum towards left. By carefully selecting the filter-1 value one can handle data sets larger than the present one within reasonable time frame. Integration of this knowledge, as a routine, in the procedure makes it even more efficient, especially when one wants to perform a search beyond the descriptors' optimum population levels. We have practically examined and found that an integration of this kind will not significantly increase the basic computational time; rather it may become a knowledge base of descriptor behavior to alter the course of computation and even considerably reduce the time.

6 Acknowledgements:

The author expresses his sincere thanks to the referees for the helpful suggestions in the revision of the manuscript.

6 References:

1. Stahle, L, and Wold. S., Multivariate data analysis and experimental design in biomedical research In Ellis, G.P., and West, G.B., (Eds). "*Progress in Medicinal Chemistry, Vol.25*". Elsevier Science Publishers, 1988, pp. 292-338.
2. Rawlings, J.O., "*Applied regression Analysis: A research tool*". Wadsworth & Brooks, 1988.
3. Gasteiger, J., and Zupan, J., Neural Networks in Chemistry. *Angew. Chem., Intl. Ed. Engl.*, 32, 503-527 (1993).
4. Wikel, J.H., and Dow, E.R., The use of neural networks for variable selection in QSAR, *Bioorg. Med. Chem. Lett.*, 3, 645-651 (1993).
5. Manallack, D.T., Ellis, D.D., and Livingstone, D.J., Analysis of Linear and Non-Linear QSAR Data using Neural Networks. *J. Med. Chem.*, 37, 3758-3767 (1994).
6. Kovesdi, I., Dominguez-Rodriguez, M.F., Orfi, L., Naray-Szabo, G., Varro, A., Gy.Papp, J., and Matyus, P., Application of Neural Networks in Structure-Activity Relationships. *Med. Res. Rev.*, 19, 249-269 (1999).
7. McFarland, J.W., and Gans, D.J., On identifying likely determinants of biological activity in high dimensional QSAR problems, *Quant. Struct.-Act. Relat.*, 13, 11-17 (1994).
8. Rogers, D., and Hopfinger, A.J., Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships, *J. Chem. Inf. Comput. Sci.*, 34, 854-866 (1994).
9. Kubinyi, H., Variable selection in QSAR studies. 1. An evolutionary algorithm, *Quant. Struct.-Act. Relat.*, 13, 285-294 (1994).
10. Kubinyi, H., Variable selection in QSAR studies. II. A high efficient combination of sydtematic search and evolution, *Quant. Struct.-Act. Relat.*, 13, 393-401 (1994).
11. So, S.-S., and Karplus, M., Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks, *J. Med. Chem.*, 39, 1521-1530 (1996).
12. Waller, C.L., and Bradley, M.P., Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies, *J. Chem. Inf. Comput. Sci.*, 39, 345-355 (1999).
13. Cho, S.J., and Hermsmeier, M.A., Genetic algorithm guided selection: Variable selection and subset selection, *J. Chem. Inf. Comput. Sci.*, 42, 927-936 (2002).

14. Nicolotti, O., Gillet, V.J., Fleming, P.J., and Green, D.V.S., Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs, *J. Med. Chem.*, *45*, 5069-5080 (2002).
15. Armitage, P., and Berry, G., “*Statistical Methods in Medical Research 2nd Edition*”, Blackwell Scientific Publications, Oxford, 1990, pp.296-357.
16. Selwood, D.L., Livingstone, D.J., Comley, J.C.W., O’Dowd, A.B., Hudson, A.T., Jackson, P., Jandu, K.S., Rose, V.S., and Stables, J.N., Structure-activity relationships of antifilarial antimycin analogues: A multivariate pattern recognition study, *J. Med. Chem.*, *33*, 136-142 (1990) (data set downloaded from: <http://www.disat.unimib.it/chm/Datasets.htm#Selwood>).
17. Rencher, A.C., and Pun, F.C., Inflation of R² in best subset regressions, *Technometrics*, *22*, 49-53 (1980).
18. SYSTAT, Version 7.0: SPSS Inc., 444 North Michigan Avenue, Chicago, IL 60611.
19. Unger, S.H., and Hansch, C., On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of β -halo- β -arylalkylamines, *J. Med. Chem.*, *16*, 745-749 (1973).
20. Velleman, P.F., and Welsch, R.E., Efficient computing of regression diagnostics. *The American Statistician*, *35*, 234-242 (1981).

Table 1: The **OVB** and **OVB-TVB** descriptor bundle for three descriptors (A B C) i.e., $p=3$, with k varying from 1(**b**) to 3(**e**) in the combinatorial protocol; $f()$ stands for functional transformation of a descriptor from the corresponding **OVB**.

OVB	OVB-TVB	
k=1	k=1	s=1
A	A	$f(A)$
B	B	$f(B)$
C	C	$f(C)$
k=2	k=2	s=1
AB	AB	$f(A)$
	AB	$f(B)$
		s=2
	AB	$f(A)f(B)$
		s=1
AC	AC	$f(A)$
	AC	$f(C)$
		s=2
	AC	$f(A)f(C)$
		s=1
BC	BC	$f(B)$
	BC	$f(C)$
		s=2
	BC	$f(B)f(C)$
k=3	k=3	s=1
ABC	ABC	$f(A)$
	ABC	$f(B)$
	ABC	$f(C)$
		s=2
	ABC	$f(A)f(B)$
	ABC	$f(A)f(C)$
	ABC	$f(B)f(C)$
		s=3
	ABC	$f(A)f(B)f(C)$

Table 2: Descriptors of Selwood *et al* data set and descriptors identified in models from different approaches: **CSA** – C [6], **GFA** – G [7], **MUSEUM** – M [8] and **CP-MLR** – P.

S. No.	Descriptor Name ^a	Descriptors Identified in Models
1	ATCH1	G P
2	ATCH2	
3	ATCH3	G
4	ATCH4	C G M P
5	ATCH5	C G M
6	ATCH6	G
7	ATCH7	P
8	ATCH8	
9	ATCH9	
10	ATCH10	
11	DIPVX	C M P
12	DIPVY	P
13	DIPVZ	
14	DIPMOM	
15	ESDL1	
16	ESDL2	
17	ESDL3	G M
18	ESDL4	
19	ESDL5	
20	ESDL6	
21	ESDL7	
22	ESDL8	
23	ESDL9	
24	ESDL10	P
25	NSDL1	
26	NSDL2	P
27	NSDL3	
28	NSDL4	P
29	NSDL5	
30	NSDL6	
31	NSDL7	
32	NSDL8	M P
33	NSDL9	M P
34	NSDL10	M
35	VDWVOL	M
36	SURFA	G M
37	MOFIX	M
38	MOFIY	C G M P
39	MOFIZ	M P
40	PEAXX	G M P
41	PEAXY	P
42	PEAXZ	M
43	MOLWT	
44	S81DX	
45	S81DY	P
46	S81DZ	P
47	S81CX	M
48	S81CY	M
49	S81CZ	
50	LogP	C G M P
51	MPNT	G P
52	SUMF	G M P
53	SUMR	

^a,ATCH1 to ATCH10 are partial atomic charges for atoms 1 to 10; DIPVX, DIPVY and DIPVZ are dipole moments of vectors X, Y and Z respectively; DIPMOM is dipole moment; ESDL1 to ESDL10 are electrophilic superdelocalizability for atoms 1 to 10; NSDL1 to NSDL10 are nucleophilic superdelocalizability for atoms 1 to 10; VDWVOL is van der Waals volume; SURFA is surface area; MOFIX, MOFIY and MOFIZ are principal moments of inertia along X, Y, and Z axis respectively; PEAXX, PEAXY and PEAXZ are principal ellipsoid axes; MOLWT is molecular weight; S81DX, S81DY and S81DZ are substituent dimensions; S81CX, S81CY and S81CZ are substituent centers; LogP is partition coefficient; MPNT is melting point.

Table 3: Models identified in CP-MLR along with statistical parameters.^a

Model No.	Model ^b	Normal stat. ^c				Cross-validation stat.		
		<i>r</i>	<i>r</i> -bar	<i>s</i>	F	Q^2	SPRESS	SDEP
6	1, 4, 11, 12	0.811	0.778	0.519	12.49	0.538	0.603	0.532
7	4, 7, 11, 12	0.803	0.768	0.529	11.80	0.515	0.618	0.566
8	32s, 50	0.786	0.758	0.538	14.54	0.416	0.666	0.621
9	33s, 50, 51	0.814	0.781	0.516	12.74	0.379	0.699	0.641
10	4, 7, 38s	0.796	0.760	0.537	11.23	0.494	0.632	0.578
11	4, 7, 39s	0.794	0.757	0.540	11.08	0.488	0.635	0.582
12	4, 7, 28, 39s	0.835	0.798	0.498	11.55	0.551	0.607	0.545
13	4, 7, 28, 38s	0.835	0.798	0.498	11.51	0.553	0.605	0.543
14	4, 7, 26, 39s	0.834	0.797	0.499	11.45	0.552	0.606	0.544
15	4, 7, 26, 38s	0.834	0.796	0.500	11.39	0.556	0.603	0.542
16	1s, 4, 38s	0.819	0.777	0.520	10.17	0.461	0.665	0.597
17	1s, 4, 39s	0.816	0.775	0.522	10.00	0.454	0.669	0.601
18	4, 11, 38s, 52	0.815	0.773	0.524	9.91	0.471	0.658	0.591
19	4, 7, 26, 45s	0.814	0.771	0.526	9.80	0.509	0.634	0.569
20	4, 11, 39s, 52	0.814	0.771	0.526	9.79	0.464	0.663	0.595
21	4, 7, 12, 46s	0.810	0.766	0.531	9.52	0.517	0.629	0.565
22	4, 7, 26, 40s	0.808	0.764	0.533	9.42	0.358	0.725	0.651
23	1s, 4, 45s	0.806	0.762	0.536	9.28	0.473	0.657	0.590
24	4, 7, 11, 45s	0.793	0.745	0.551	8.50	0.441	0.677	0.608
25	4, 7, 12, 41s	0.793	0.744	0.552	8.46	0.484	0.651	0.584
26	1, 4, 11, 45s	0.791	0.743	0.553	8.38	0.413	0.693	0.623
27	4, 32s, 41s, 50	0.894	0.866	0.413	15.99	0.632	0.561	0.493
28	4, 7, 11, 12, 46s	0.859	0.820	0.473	11.26	0.616	0.572	0.504
29	1s, 4, 11, 45s	0.843	0.799	0.501	9.85	0.494	0.657	0.578
30	4, 11s, 45s, 52	0.824	0.774	0.524	8.46	0.488	0.661	0.582
31	24, 32s, 46s, 51s	0.856	0.806	0.488	8.98	0.518	0.655	0.564
32	4, *7s, 11	0.812	0.780	0.517	12.63	0.508	0.623	0.570
33	4, *7s, 11, 45s	0.856	0.816	0.480	10.98	0.546	0.623	0.548
34	4, 12, *21s, 25, 53	0.808	0.752	0.545	7.50	0.442	0.690	0.607
35	4, 12, *15s, 25, 53	0.804	0.745	0.550	7.30	0.411	0.709	0.624
36	4, 12, *21s, 31, 53	0.801	0.744	0.553	7.18	0.420	0.703	0.619
37	4, *19s, 32s, 46s	0.862	0.815	0.479	9.49	0.570	0.619	0.533
38	1, 4, 12, *22s, 41s	0.856	0.807	0.488	9.02	0.394	0.735	0.633
39	4, 14, *16s, 26s, 52	0.827	0.766	0.531	7.09	0.420	0.719	0.619

^a, filter-1 as 0.3; filter-2 as 2.0; filter-3 as 0.74; filter-4 as $0.3 \leq Q^2 \leq 1.0$.

^b, the number corresponds to the descriptor serial number given in **Table 2**. A character 's' after the descriptor number indicates that both the normal and squared terms are involved in the model. An '*' before the descriptor number indicates that the regression coefficients of the descriptor and its squared term have the same mathematical sign.

^c, in all the models the number of observations is 31; *s* is standard error of the estimate; F is F-ratio between the variance of calculated and observed activities; remaining statistics are calculated according to equations 1 to 5.

Table 4: Correlation matrix of the descriptors of models 27, 28 and 31.

		1	2	3	4	5	6	7	8	9	10
1	ATCH4	1.000									
2	ATCH7	-0.098	1.000								
3	DIVVX	0.278	0.012	1.000							
4	DIPVY	-0.172	-0.209	0.063	1.000						
5	ESDL10	-0.039	-0.119	-0.398	-0.266	1.000					
6	NSDL8	-0.018	0.545	-0.027	-0.197	0.018	1.000				
7	PEAXY	-0.043	0.077	-0.443	-0.040	0.198	0.111	1.000			
8	S81DX	0.240	0.198	-0.081	-0.068	0.254	0.010	0.480	1.000		
9	LogP	0.108	0.484	-0.253	-0.317	0.225	0.169	0.262	0.562	1.000	
10	MPNT	-0.190	0.171	-0.210	0.547	-0.105	0.288	0.168	-0.290	-0.235	1.000

Table 5: Populations of Descriptors of Selwood *et al* data set for bundle size (**k**) 2 to 10 for inter-parameter correlation cutoff 0.3.

S. No	Descriptor Name	Descriptor populations with respect to bundle size (k) 2 to 10								
		2 ^a	3	4	5	6	7	8 ^b	9	10
1	ATCH1	28	211	688	1132	999	475	119	15	0
2	ATCH2	32	299	1090	1642	1014	227	0	0	0
3	ATCH3	28	210	642	913	641	217	30	0	0
4	ATCH4	46	648	3569	8853	10795	6869	2257	329	5
5	ATCH5	31	276	988	1629	1333	526	80	0	0
6	ATCH6	30	242	769	1156	880	328	48	0	0
7	ATCH7	31	259	859	1375	1153	510	112	10	0
8	ATCH8	34	323	1196	2071	1827	839	197	20	0
9	ATCH9	20	131	380	558	429	162	23	0	0
10	ATCH10	21	146	471	833	878	573	231	53	5
11	DIPVY	34	326	1142	1710	1122	274	*13	0	0
12	DIPVX	33	327	1246	2264	2204	1210	373	54	0
13	DIPVZ	51	780	4543	11502	13834	8409	2590	344	5
14	DIPMOM	26	174	381	366	162	29	1	0	0
15	ESDL1	32	288	1078	2010	2038	1144	336	41	0
16	ESDL2	32	288	1078	2010	2038	1144	336	41	0
17	ESDL3	21	130	339	445	316	125	28	3	0
18	ESDL4	33	317	1248	2395	2491	1452	455	62	0
19	ESDL5	38	420	1740	3128	2630	1020	172	8	0
20	ESDL6	42	506	2307	4757	4988	2798	847	127	5
21	ESDL7	33	311	1220	2379	2527	1495	468	62	0
22	ESDL8	31	247	772	1152	865	319	45	0	0
23	ESDL9	30	232	715	1067	811	307	45	0	0
24	ESDL10	33	295	1038	1679	1328	508	75	0	0
25	NSDL1	29	245	892	1708	1882	1212	425	62	0
26	NSDL2	29	233	731	1126	909	366	57	0	0
27	NSDL3	42	494	2128	4089	3973	2064	560	67	1
28	NSDL4	39	431	1792	3334	3148	1570	403	47	1
29	NSDL5	41	477	2036	3884	3794	1990	547	67	1
30	NSDL6	38	402	1670	3250	3320	1896	592	83	1
31	NSDL7	29	245	892	1708	1882	1212	425	62	0
32	NSDL8	35	319	1112	1749	1313	466	65	1	0
33	NSDL9	38	402	1670	3250	3320	1896	592	83	1
34	NSDL10	29	221	659	938	678	239	33	0	0
35	VDWVOL	25	149	339	327	133	23	0	0	0
36	SURFA	27	183	490	578	287	45	0	0	0
37	MOFIX	28	241	856	1488	1386	698	173	15	0
38	MOFIY	29	227	702	944	559	129	*8	0	0
39	MOFIZ	26	181	506	636	367	89	*8	0	0

40	PEAXX	27	193	544	678	382	89	*8	0	0
41	PEAXY	43	561	2860	6720	7985	4950	1534	195	5
42	PEAXZ	25	205	707	1071	658	158	*12	0	0
43	MOLWT	27	231	751	1073	695	173	*5	0	0
44	S81DX	39	467	2433	6259	8390	5952	2142	324	5
45	S81DY	32	316	1332	2742	2999	1803	575	79	0
46	S81DZ	34	357	1531	2979	2927	1503	393	45	0
47	S81CX	35	379	1652	3348	3496	1972	581	73	0
48	S81CY	35	386	1788	3943	4486	2724	864	130	5
49	S81CZ	42	561	3086	7797	9671	6115	1944	274	5
50	LogP	32	328	1337	2427	2104	857	128	0	0
51	MPNT	28	265	1083	2231	2491	1507	452	50	0
52	SUMF	33	304	1101	1881	1659	760	168	15	0
53	SUMR	40	477	2361	5706	7345	5164	1897	300	5

^a, the numbers in bold face correspond to high population group.

^b, The '*' marked descriptor populations are taken as zero in bundle 9 for the computation populations.

Table 6: A translation of models listed in **Table 3** in the form of descriptor population code; H- high population, L-low population.

Model No.	Model	No of OVB descriptors	Population code of the model
8	32s, 50	2	HL
9	33s, 50, 51	3	HLL
10	4, 7, 38s	3	HLL
11	4, 7, 39s	3	HLL
16	1s, 4, 38s	3	LHL
17	1s, 4, 39s	3	LHL
23	1s, 4, 45s	3	LHL
6	1, 4, 11, 12	4	LHHH
7	4, 7, 11, 12	4	HLHH
12	4, 7, 28, 39s	4	HLHL
13	4, 7, 28, 38s	4	HLLL
14	4, 7, 26, 39s	4	HLLL
15	4, 7, 26, 38s	4	HLLL
18	4, 11, 38s, 52	4	HHLH
19	4, 7, 26, 45s	4	HLLL
20	4, 11, 39s, 52	4	HHLH
21	4, 7, 12, 46s	4	HLHH
22	4, 7, 26, 40s	4	HLLL
24	4, 7, 11, 45s	4	HLHL
25	4, 7, 12, 41s	4	HLHH
26	1, 4, 11, 45s	4	LHHL
27	4, 32s, 41s, 50	4	HHHL
29	1s, 4, 11, 45s	4	LHHL
30	4, 11s, 45s, 52	4	HHLH
31	24, 32s, 46s, 51s	4	HHHL
28	4, 7, 11, 12, 46s	5	HLHHH

Table 7: Models identified in CP-MLR along with statistical parameters.^a

Model No.	Model Origin ^b	Model ^c	Normal stat. ^d				Cross-validation stat.		
			<i>r</i>	<i>r</i> -bar	<i>s</i>	F	<i>Q</i> ²	SPRES S	SDE P
40	G,M	38,50,52	0.84 9	0.83 1	0.46 0	23.2 7	0.64 7	0.518	0.483
41	G,M	17,36,50	0.84 8	0.82 9	0.46 2	23.0 4	0.64 4	0.520	0.485
42	M	39,50,52	0.84 7	0.82 9	0.46 2	22.9 3	0.64 3	0.520	0.485
43	G,M	17,38,50	0.83 4	0.81 7	0.47 6	21.1 5	0.60 4	0.548	0.511
44		17,39,50	0.83 5	0.81 4	0.47 9	20.7 1	0.60 0	0.550	0.514
45		4,5,11	0.82 9	0.80 8	0.48 7	19.8 3	0.61 2	0.543	0.506
46	G,M	4,17,40,50	0.88 0	0.86 0	0.42 2	22.3 1	0.63 6	0.532	0.490
47		17,35,50,51	0.86 2	0.83 9	0.44 9	18.8 5	0.64 2	0.531	0.486
48		17,35,43,50	0.86 1	0.83 8	0.45 1	18.6 7	0.64 5	0.529	0.484
49		17,35,37,50	0.86 1	0.83 8	0.45 1	18.6 6	0.63 9	0.533	0.488
50		8,36,50,52	0.85 4	0.83 0	0.46 1	17.5 8	0.64 1	0.532	0.487
51	M	4,5,11,39,50	0.90 9	0.89 0	0.37 7	23.8 2	0.69 6	0.499	0.448
52	M	4,5,11,38,50	0.90 9	0.89 0	0.37 7	23.7 8	0.69 6	0.499	0.448
53	M	4,17,35,37,50	0.90 5	0.88 5	0.38 5	22.7 1	0.67 6	0.515	0.462
54		4,11,39,50,52	0.90 4	0.88 4	0.38 6	22.4 8	0.69 9	0.496	0.446
55		4,11,38,50,52	0.90 2	0.88 1	0.39 1	21.7 5	0.69 2	0.502	0.451
56	M	4,11,38,48,50,52	0.92 4	0.90 4	0.35 4	23.2 4	0.75 4	0.458	0.403
57	M	4,11,39,48,50,52	0.92 4	0.90 4	0.35 4	23.2 3	0.75 1	0.461	0.405
58	M	4,11,38,47,50,52	0.92 4	0.90 3	0.35 4	23.1 9	0.74 9	0.463	0.407
59	M	4,11,39,47,50,52	0.92 3	0.90 3	0.35 5	23.0 9	0.74 6	0.466	0.410
60		4,11,12,38,50,52	0.91 8	0.89 6	0.36 7	21.2 5	0.74 3	0.469	0.412

^a, filter-1 as 0.79; filter-2 as 2.0; filter-3 as 0.80; filter-4 as $0.36 \leq Q^2 \leq 1.0$.

^b, G for **GFA** (ref.8) M for **MUSEUM** (ref.9).

^{c,d}, see footnotes b and c, respectively, of **Table 3**.

Table 8: Models identified in **CP-MLR** along with statistical parameters.^a

Model No.	Model ^b	Normal stat. ^c				Cross-validation stat.		
		<i>r</i>	<i>r</i> -bar	<i>s</i>	F	Q^2	SPRESS	SDEP
54	4,11,39,50,52	0.904	0.884	0.386	22.48	0.699	0.496	0.446
55	4,11,38,50,52	0.902	0.881	0.391	21.75	0.692	0.502	0.451
61	38,50,51s,52	0.893	0.870	0.407	19.70	0.708	0.489	0.439
62	1,*4s,12,39,50	0.924	0.903	0.362	23.02	0.769	0.444	0.390
63	1,*4s,12,38,50	0.922	0.901	0.365	22.69	0.765	0.448	0.394
64	4,11,39,50s,52	0.922	0.901	0.361	22.52	0.738	0.473	0.416
65	4,11,38,50s,52	0.921	0.900	0.361	22.37	0.739	0.472	0.416
60	4,11,12,38,50,52	0.918	0.896	0.367	21.35	0.743	0.469	0.412
66	4,32s,39,50,51s	0.923	0.905	0.352	20.37	0.727	0.493	0.425
67	1s,12,38,50,51s	0.925	0.901	0.361	19.49	0.729	0.491	0.423
68	1,4,12,38,50,51s	0.922	0.897	0.366	18.62	0.717	0.502	0.433
69	1,4,12,39,50,51s	0.921	0.895	0.369	18.24	0.710	0.508	0.438
70	1,4,11,12,41s,50	0.919	0.893	0.374	17.87	0.731	0.489	0.421
71	4,11,39,50s,51s,52	0.941	0.918	0.327	21.14	0.782	0.450	0.379
72	4,11,38,50s,51s,52	0.939	0.916	0.331	20.66	0.778	0.454	0.383
73	4,11,40,50s,51s,52	0.934	0.909	0.343	18.96	0.748	0.484	0.408

^a, filter-1 as 0.79; filter-2 as 2.0; filter-3 as 0.80; filter-4 as $0.63 \leq Q^2 \leq 1.0$.

^{b,c}, see footnotes b and c, respectively, of **Table 3**.