J Comput Aided Mol Des (2008) 22:461–468
DOI 10.1007/s10822-008-9209-4

Size-intensive descriptors

George D. Purvis III

Received: 22 October 2007 / Accepted: 3 March 2008 / Published online: 21 March 2008

Abstract Non-linear effects in multi-linear structure–property (QSPR) models are sometimes included by using descriptors transformed by mathematical functions such as the square root or logarithm. Less commonly, products of two descriptors are used to account for cross dependencies. As described here, simple division of descriptors by chemical sample size (e.g. molecular weight, length, area or volume) creates *size-intensive descriptors* (alternatively, intrinsic descriptors) that are independent of the size of the chemical sample described, weakly correlated with the original descriptor, and important contributors to the best QSPR models. In our automated QSPRs, size-intensive descriptors in competition with their extensive descriptors are frequently selected as the best descriptors in the models with the highest r^2. Examples of QSPR models that use size-intensive descriptors are given, the lack of correlation of descriptors with their size-intensive version is demonstrated, and their physical significance is discussed.

Keywords CPSA · Connectivity indices · Descriptors · Size-intensive · Model · Non-linear models · QSAR · QSPR · Relationships · Structure–property · Size-extensitivity

Introduction

Central to the creation of a quantative property–structure relationship (QSPR) is the choice of structural descriptors. Decades of research have given us thousands [1, 2] of structural descriptors ranging from simple whole-molecule properties such as molecular weight, density and volume to quantum mechanically evaluated charge polar surface areas (CPSAs) [3]. Group counts, atom properties, topological connectivity indices [4, 5] and electrotoplogical indices [6] bridge the complexity range.

Today, creation of QSPRs from pools of hundreds of descriptors using linear regression is nearly fully automated in software [7–9]. Non-linear relationships are accommodated by transforming the data and/or descriptors using mathematical functions such as square, square root, logarithm, and inverse. Less frequently, combinations of descriptors such as products of two descriptors are made to account for non-linear cross-dependencies.

In the process of developing automated QSAR/QSPR software that exhaustively searches for the best combination of a few descriptors from a pool of hundreds, experiments with single-descriptor non-linear mathematical transformations including squares, square roots, logarithms, reciprocals and others were performed. While descriptors transformed by mathematical functions often do appear in the best models, cross-descriptors created by dividing a descriptor by chemical sample [10] size appear in the best models more frequently.

In the following, it is important to understand the term *chemical sample* [10]. A chemical sample is the part of the universe that is the subject of chemical study. The chemical sample may contain one or more molecules and undergo chemical or physical transformation from its current state to a new state. In modeling chemistry, we are most often concerned with modeling chemical samples.

Originally presented at the American Chemical Society Fall 2007 National Meeting, Boston, MA, Phil Magee Memorial Symposium: QSAR Reborn, 1:30 PM–4:50 PM, Wednesday, August 22, 2007 under the title "Intrinsic Descriptors".

G. D. Purvis III (✉)
Fujitsu Computer Systems, Biosciences Group, 15244 NW Greenbrier Pkwy, Beaverton, OR 97006, USA
e-mail: gpurvis@us.fujitsu.com

In analogy to extensive and intensive properties, descriptors that scale with sample size are called *extensive descriptors*. Descriptors that are independent of size could be called *intensive descriptors*. However in this paper, descriptors that are independent of size are called *size-intensive descriptors* to avoid confusion with the frequently-used expression "computationally-intensive descriptors".

Extensive descriptors, such as group and atom counts, depend up the size of the chemical sample. To understand the difference between extensive and size-intensive descriptors, compare the hydroxyl (–OH) group count descriptor for chemical sample A containing a single methanol to the hydroxyl count for a chemical sample B containing two methanols. The hydroxyl group count for A is one and that for B is two. By contrast, the size-intensive descriptor created by dividing hydroxyl counts by the sample molecular weights are the equal for A and B. Descriptor scaling with sample size is appropriate for extensive properties such as total system energy, but intuitively less appropriate for intensive properties such as boiling point.

A great number of choices are available to define the chemical sample size and thereby different size-intensive descriptors. Measures of chemical sample size include molecular weight, surface area, volume, longest path, length, width, depth, containing box size, repeat unit count, certain connectivity indices, and others that are extensive. Quite generally, a size-intensive descriptor is created when any extensive descriptor is divided by a different extensive descriptor.

The following sections examine *size-intensive descriptors* as a class of descriptors and demonstrate that

(1) conceptually, this class can be thought of as a "density or concentration of" class;
(2) the size-intensive form of a descriptor is usually only weakly correlated with either the descriptor or the chemical sample size;
(3) size-intensive descriptors appear in the best QSPR/QSAR models;
(4) size-intensive descriptors can produce more compact models;
(5) and the best models are not the result of chance.

Of course, size-intensive descriptors have been used previously in QSPR model building. For example, the fractional CPSA descriptors [3] are size-intensive as are molecular density and the count of fluorine atoms divided by the total number of atoms [11, 12]. In automated QSPR programs such as CODESSA [9], 20% or more of the descriptors could be classified as size-intensive descriptors. Undoubtedly many more examples could be enumerated.

In the development of quantum methods of extensive chemical sample properties such as total energy, *size extensivity* [13, 14] has proven important. Similarly, *size intensivity* can be important in development of QSPR models of intensive properties. While an obvious use of size-intensive descriptors is in modeling chemical samples that contain multiple molecules such as mixtures, in this paper, the focus is on size-intensive descriptors of chemical samples that contain a single molecule because size-intensive descriptors can also be important for models built with single molecule descriptors. Here the question "what do we learn when standard descriptors compete to build the best QSPR model with their size-intensive version created by dividing by chemical sample size?" is explored.

A simple illustration: hydroxyl counts

Consider the count of hydroxyl groups. As a descriptor of sugars it increases with the size of the sugar. On the other hand, the size-intensive descriptor created by dividing by the sample's molecular weight remains nearly constant as shown in Table 1.

By contrast in Table 2, as a descriptor of primary alcohols the hydroxyl count remains constant while the size-intensive descriptor decreases significantly.

The size-intensive hydroxyl count maps sugars to a similar value and primary alcohols to a range of values.

While the size-intensive hydroxyl count is derived from the hydroxyl count, it is only weakly correlated with hydroxyl count. To see how size-intensive descriptors can be only weakly correlated to the descriptor from which they are derived, examine Table 3. When the descriptors are centered and normalized to one significant figure so that the unique value is 0 and the magnitude of the non unique value is 1, the lack of correlation between descriptors is apparent. As confirmation, the largest pairwise correlation

Table 1 Sugars

Sugar	–OH count	–OH/MW
Threose	3	0.0266
Ribose	4	0.0278
Manose	5	0.0285

Table 2 Primary alcohols

Alcohol	–OH count	–OH/MW
Methanol	1	0.0321
Ethanol	1	0.0217
Propanol	1	0.0166

Table 3 Descriptor independence

Chemical	Descriptors			Centered and normalized		
	# –OH	MW	# –OH/MW	# –OH	MW	# –OH/MW
Methanol	1	32.4	0.031	0	−1	0
Ethylene glycol	2	62.1	0.032	1	0	0
Propanol	1	61.0	0.017	0	0	−1

between the original unscaled descriptors in Table 3 is 0.55.

Look at Fig. 1 to visualize the difference between the count of hydroxyl groups in sugars and the count divided by size.

The black box in Fig. 1 outlines a uniform length in the extended conformation of three sugars to emphasize that sugars as a class tend to have one hydroxyl per carbon atom (except for the end aldehyde). A size-intensive hydroxyl descriptor captures this characteristic of sugars by assigning all sugars nearly the same numerical value, 0.027 in Table 1. Physically, a size-intensive hydroxyl descriptor provides a “density” or “concentration” of hydroxyl’s per atom, per angstrom, per area or per volume depending upon the chemical sample size used in the definition.

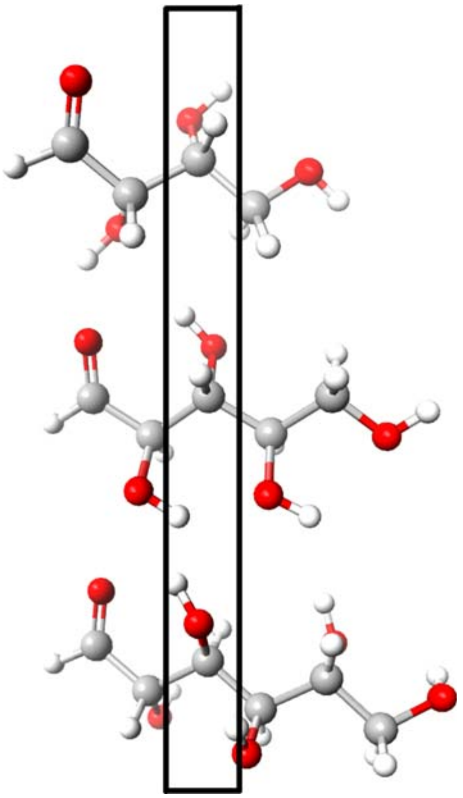

Fig. 1 Threose, ribose and manose

A larger example: 50 HIA inhibitors

It is informative to see how the *hydrogen bond donor count*, molecular weight and the size-intensive descriptor *hydrogen bond donor count divided by molecular weight* interplay in a larger data set. Figure 2, shows three pairwise plots of these descriptors for fifty inhibitors used to build a QSAR model of human intestinal absorption [15].

Visual examination of the first three graphs in Fig. 2 reveals no excessively strong cross correlation among the three descriptors. This is consistent with the pairwise correlation values of 0.74, 0.00 and 0.58 respectively. The fourth graph in Fig. 2 plots the size-intensive descriptor and the approximation of the size-intensive descriptor that results from the linear regression with H-bond donor count and molecular weight. The r^2 for the regression of the size-intensive descriptor against donor count and molecular weight was 0.85. Consequently, the variance inflation factor for the size-intensive descriptor if the other two are included in a regression model is less than 4 and all three descriptors could be used in the same three-descriptor model.

In summary, the size-intensive descriptor is not strongly correlated with its component base descriptors and it is not well described by a linear combination of the base descriptors. Including the size-intensive descriptor in a QSPR expands descriptive space in a non-linear and unique direction.

Although it is possible to imagine trivial counter examples[1], this observation for the H-bond donor size-intensive descriptor holds generally for other descriptors, even in the non-trivial worst-case scenario involving size threshold descriptors. Consider the case of the threshold descriptor MW > 260 [16]. This descriptor is zero until the chemical sample has a molecular weight of 260 or greater. For molecular weights 260 greater, the descriptor is equal to the sample’s molecular weight minus 260.

Figure 3 shows that even in a worst-case scenario, the size-intensive descriptor is significantly different from the base descriptor and the regression equation that relates the two has an r^2 of 0.85. The correlation of the MW > 260 with its size-intensive descriptor is 0.92, a value that is on the threshold of being unacceptably large if both appear in the same model. In practice, as demonstrated in the next sections, a descriptor and its size-intensive variants do not appear in the same best model.

[1] A trivial example of a useless size-intensive descriptor is molecular weight divided by molecular weight which is intensive, but also devoid of all information.

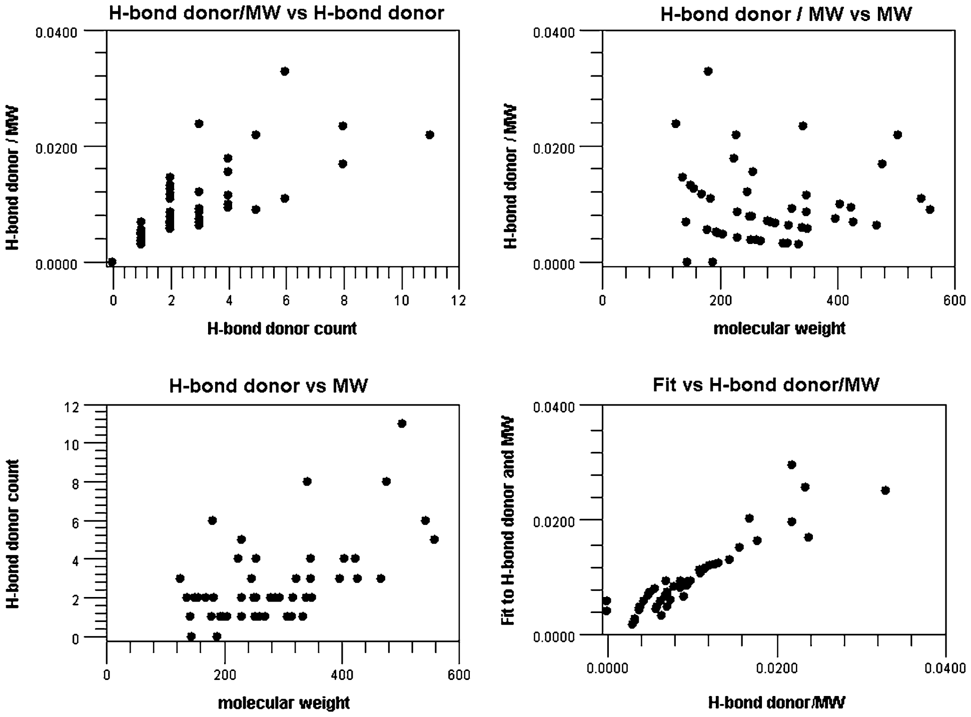

Fig. 2 Pairwise relationships between H-bond donor count, molecular weight and H-bond donor count divided by molecular weight and the fit of H-bond donor/MW to H-bond donor and MW

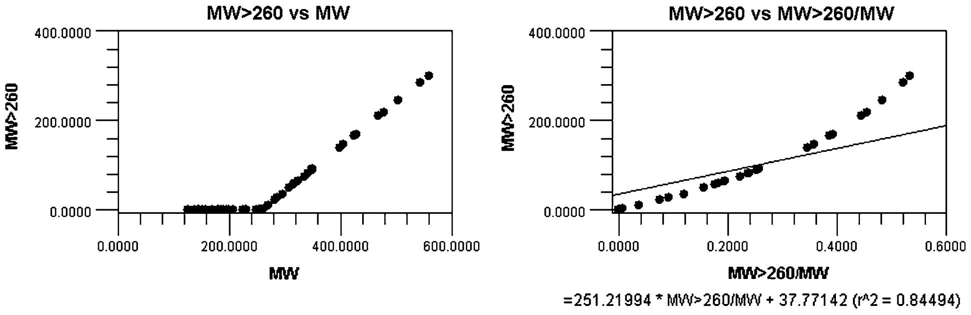

Fig. 3 Threshold descriptor MW > 260

Do size-intensive descriptors appear in the best models?

Expansion of descriptor space is useful in QSPR when it results in better models with smaller errors and more compact equations that include the new descriptors. Otherwise, addition of new descriptors just because they expand the space can be a hindrance to discovery of the best models by unnecessarily expanding the space of possible models.

Methodology

To explore whether size-intensive descriptors appear often in the best models, illustrative QSPR models for human intestinal absorption, boiling points and water solubility using data sets of 50, 158 and 259 compounds respectively were developed. The results of models developed using all combinations of two, three and four descriptors from pools of 100 to 600 descriptors for each data set are reported. Models for these properties have been reported previously too many times to fully enumerate. Compared to the best reported models, the data sets used here are smaller and the models developed have fewer descriptors. Consequently, these illustrative models should not be used predictively.

Screened pools of descriptors used in model building are created by selecting from a large set of computed descriptors those that have variance and at least two descriptor values are different from the other values. Unlike common practice, descriptors are not screened for high pairwise correlation during the process of pool creation. Instead, correlation screening is delayed until model building. If during the process of creating a model from all combinations of descriptors, two descriptors used in that model are correlated by more than a specified threshold (0.95), the model is skipped. The delay increases the chance of finding the best combination of descriptors which otherwise could be lost.

The number of descriptors is limited to four or fewer so that models from all combinations of descriptors can be compared and the best model within the pool of all descriptors could be found, thereby avoiding problems with search algorithms that can not guarantee that the best solution is found. The appearance of size-intensive descriptors in the best models with few descriptors is important. When choosing between two models of the same quality but differing number of descriptors, the model with the fewest number of descriptors is preferred.

HIA models

The human intestinal absorption data of 50 chemicals [15] optimized in extended conformations with MM2 [17] were modeled using a screened pool of 349 descriptors that included both linear and non-linear versions of topological descriptors (connectivity indices, group counts, atom properties, and so forth). Of the 349 descriptors 97 (1 in 4) were size-intensive descriptors created by dividing a topological descriptor by molecular weight. The resulting best single, double, triple and quadruple descriptor models are:

$$\begin{aligned}&\textit{Experimental HIA}(\%) = -47.5285\\&\quad * \mathrm{sqrt}(\textit{H - bond acceptor count}) + 176.1413;\\&\mathrm{r}^2 = 0.6832 \text{ adj } \mathrm{r}^2 = 0.6832 \ \mathrm{q}^2 = 0.6623 \text{ avErr}\\&\quad = 16.4234 \text{ SD} = 20.9735 \text{ cvSD} = 21.6543\end{aligned} \tag{1}$$

$$\begin{aligned}&\textit{Experimental HIA}(\%) = -2268.6511\\&\quad * \boldsymbol{H - bond\ donor\ count/MW} - 156.2767\\&\quad * \boldsymbol{molecular\ weight > 260/MW} + 109.6942;\\&\mathrm{r}^2 = 0.7714 \text{ adj } \mathrm{r}^2 = 0.7617 \ \mathrm{q}^2 = 0.7434 \text{ avErr}\\&\quad = 12.7566 \text{ SD} = 17.8141 \text{ cvSD} = 18.8738\end{aligned} \tag{2}$$

$$\begin{aligned}&\textit{Experimental HIA}(\%) = -2383.7550\\&\quad * \boldsymbol{H - bond\ donor\ count/MW} - 141.7792\\&\quad * \boldsymbol{molecular\ weight > 260/MW} - 7.2663\\&\quad * \textit{carboxyl count}^2 + 112.8166;\\&\mathrm{r}^2 = 0.8043 \text{ adj } \mathrm{r}^2 = 0.7915 \ \mathrm{q}^2 = 0.7685 \text{ avErr}\\&\quad = 12.0598 \text{ SD} = 16.4844 \text{ cvSD} = 17.9281\end{aligned} \tag{3}$$

$$\begin{aligned}&\textit{Experimental HIA}(\%) = -2020.9795\\&\quad * \boldsymbol{Csp^3\ bonded\ to\ 2\ C/MW} + 1082.4441/\textit{length}\\&\quad - 458.5202/\textit{hydrogen count} - 33.6139\\&\quad * \textit{sqrt}(\textit{H - bond acceptor count}) + 114.0506;\\&\mathrm{r}^2 = 0.8464 \text{ adj } \mathrm{r}^2 = 0.8327 \ \mathrm{q}^2 = 0.8165 \text{ avErr}\\&\quad = 10.9371 \text{ SD} = 14.6040 \text{ cvSD} = 15.9634\end{aligned} \tag{4}$$

Size-intensive descriptors are bold for emphasis. Of the 10 descriptors appearing in the four equations, size-intensive descriptors appear five times. The temptation to classify as size-intensive descriptors the reciprocal of the *length* and *hydrogen count* that appear in the best quadruple has been resisted.

The descriptors in all four models have variance inflation factors less than 2 and no two descriptors are correlated by more than 0.95. The usual statistical measures appear below each equation. The r^2 adjusted for the degrees of freedom is called adj r^2, q^2 is the leave-one-out cross-validation result, avErr is the unsigned average error, SD is the standard deviation of the error and cvSD is the standard deviation of the errors in the leave-one-out cross validation.

Comparison to random

The best quadruple model from 389 descriptors is the best model from over 600 million possible models. Except for models that would contain highly correlated (>0.95) descriptors, each possible model was built and compared to find the models with the highest r^2. Whenever a QSPR model is built, there is a possibility that the best model is chance correlation. When the best model from over 600 million is selected, the possibility of a chance correlation is increased.

To measure the probability that the HIA models were random, 21 best single, double, triple and quadruple models were created from 21 different sets of 389 random descriptors each. The average and standard deviation of the random models were computed and 99% confidence limits were established at the average plus three times the standard deviation. For three descriptor models, the average r^2 of the best models built with random descriptors was 0.45 with a standard deviation of 0.03. The three-sigma limit for triples is 0.54 and the chance that a three descriptor models for the 50 sample HIA data set with r^2 of 0.55 or greater is random is less than 1%. For four descriptor models, the average r^2 of 21 best models was 0.53 with a standard deviation of 0.04. The probability that models with r^2 larger than 0.65 are random is less than 1%. All four of the models for HIA were well above their three sigma threshold and all have less than a 1% probability of being correlations of random descriptors.

Boiling point models

Size-intensive descriptors are created by dividing by chemical sample size. Innumerable measures of chemical sample size can be imagined and the natural question is “Which size measurement is best?” To explore answers to this question, size-intensive topological descriptors were created using molecular weight, containing box volume, containing box surface area, molecule length, accessible volume and accessible surface area for sample size. All descriptors were placed into the descriptor pool, greatly increasing its size, so that they could compete for inclusion in the best model.

Boiling point models were the first reported by Randic in his work on connectivity indices [18]. Here, a dataset of 259 boiling points of diverse chemicals [19–23] was modeled using a selected pool of 689 descriptors. The chemical samples included these elements (lowest: highest: number of molecules): H(0:26:255), C(1:14:259), O(0:4:181), F(0:6:12), S(0:2:26), Cl(0:6:36), Br(0:4:8), and I(0:1:3). The first two numbers in parentheses list the lowest and highest counts of the element in any molecule and the last number is the number of molecules in which the element appeared. The measured boiling points were distributed normally over the range from 187.5 to 653.1 K.

The best single, double, triple and quadruple descriptor models are:

$$\begin{aligned}\mathit{boiling\ points} &= 174.2863 \\ &\quad * \mathit{sqrt}(\mathit{connectivity\ index}\ 1) + 95.8668; \\ \mathrm{r}^2 &= 0.6127 \text{ adj } \mathrm{r}^2 = 0.6127 \text{ cvr}^2 = 0.6070 \text{ avErr} \\ &= 35.0581 \text{ SD} = 48.4180 \text{ cvSD} = 48.7704\end{aligned} \tag{5}$$

$$\begin{aligned}\mathit{boiling\ points} &= 1.91764\mathrm{e} + 04 \\ &\quad * \boldsymbol{H - bond\ donor\ count/box\ volume} + 277.3106 \\ &\quad * \ln(\mathit{total\ accessible\ surface\ area}) - 1119.3457; \\ \mathrm{r}^2 &= 0.7745 \text{ adj } \mathrm{r}^2 = 0.7728 \text{ q}^2 = 0.7697 \text{ avErr} \\ &= 28.0882 \text{ SD} = 36.9404 \text{ cvSD} = 37.3353\end{aligned} \tag{6}$$

$$\begin{aligned}\mathit{boiling\ points} &= 22.7807 \\ &\quad * \mathit{double\ bond\ count} + 8247.7480 * \boldsymbol{H - bond} \\ &\quad \boldsymbol{donor\ count/total\ accessible\ surface\ volume} \\ &\quad + 247.3029 * \ln(\mathit{total\ accessible\ surface\ area}) \\ &\quad - 975.9711;\ \mathrm{r}^2 = 0.8841 \text{ dof adjusted} \\ \mathrm{r}^2 &= 0.8828 \text{ q}^2 = 0.8807 \text{ avErr} \\ &= 18.4911 \text{ SD} = 26.4806 \text{ cvSD} = 26.8688\end{aligned} \tag{7}$$

$$\begin{aligned}\mathit{boiling\ points} &= 1.16070\mathrm{e} + 04 \\ &\quad * \boldsymbol{H - bond\ donor\ count/total\ accessible\ surface\ area} \\ &\quad + 845.5113 * \boldsymbol{single\ bond\ count/total\ accessible} \\ &\quad \boldsymbol{surface\ volume} + 260.3694 * \boldsymbol{Fluorine\ count/length} \\ &\quad + 195.4066 * \mathrm{sqrt}(\mathit{connectivity\ index}\ 1) + 131.2526; \\ \mathrm{r}^2 &= 0.9262 \text{ adj } \mathrm{r}^2 = 0.9251 \text{ q}^2 = 0.9228 \\ &\text{avErr} = 15.7835 \text{ SD} = 21.1314 \text{ cvSD} = 21.6140\end{aligned} \tag{8}$$

As before, the five size-intensive descriptors are bold. Consistent with Randic's work, *connectivity index 1* appears in the best single and quadruple models. The quadruple descriptor model also includes a fluorine size-intensive descriptor similar to the fluorine size-intensive descriptor reported by Katritsky et al. [11] in their model of boiling points.

As seen in the equations, no single size is preferred for creating the size-intensive descriptors used in the best models. *Box volume*, *accessible surface volume*, *accessible surface area* and *length* all appear. This observation is consistent with that reported in Ref. [11] where the best eight descriptor model for 584 diverse organic compounds ($r^2 = 0.96$) contains size-intensive descriptors where the size is measured by the number of atoms in the compound (N_F/N) and the total surface area (HDSA2).

The lack of a strongly preferred chemical sample size measure would create a problem if all size-intensive descriptors based on all possible size measures were required. Fortunately, the quality of the model is not often strongly dependent upon the size measure chosen. For example, the second best quadruples model replaces the *Fluorine count/length* descriptor with *Fluorine count/accessible surface area* but keeps the same 0.9262 r^2 with a slightly lower q^2 of 0.9226. The fourth best quadruples model replaces the *length* with *accessible surface volume* yielding r^2 and q^2 of 0.9244 and 0.9204.

Building a model which includes only those size-intensive descriptors that use molecular weight as the sample size results in this best model:

$$\begin{aligned}\mathit{boiling\ points} = &- 38.8977 * \mathit{Fluorine\ count} \\ &+ 4849.6101 * \boldsymbol{H - bond\ donor\ count/MW} \\ &- 470.4260 * \boldsymbol{single\ bond\ count/MW} \\ &+ 193.1372 * \mathrm{sqrt}(\mathit{connectivity\ index}\ 1) \\ &+ 114.9292\end{aligned} \tag{9}$$

Since r^2 is 0.9189 for Eq. 9, there is an increase in the error of the model, but it is much smaller than the change in going from a three-descriptor model to a four descrip tor model. For this descriptor pool, the 17th best four-descriptor model is the first that contains no size-intensive descriptors. It has an r^2 of 0.9045.

LogW models

The improvement in r^2 coming from size-intensive descriptors in the four-descriptor boiling point model was a little over 1%. To demonstrate the importance of small improvements, the best double, triple and quadruple models for water solubility with and without size-intensive descriptors are compared. The results for 158 water solubilities [24] using a pool of 610 linear and nonlinear descriptors are

$$\begin{aligned}\log W &= -1.3750 * \log P - 0.9876 \\ &\quad * \mathit{ring\ count\ all\ nonaromatic} + 1.0009; \\ \mathrm{r}^2 &= 0.9210 \text{ adj } \mathrm{r}^2 = 0.9210 \text{ q}^2 = 0.9183 \text{ avErr} \\ &= 0.4863 \text{ SD} = 0.6305 \text{ cvSD} = 0.6412\end{aligned} \tag{10}$$

$$\begin{aligned}\log W = {} & -183.4463 * (\boldsymbol{\log P/total\ accessible\ surface}\\ & \boldsymbol{volume}) - 0.0702 * (\text{connectivity index } 1)^2\\ & + 1.6749;\\ \mathrm{r}^2 = {} & 0.9306 \text{ adj } \mathrm{r}^2 = 0.9306 \text{ q}^2 = 0.9280 \text{ avErr}\\ & = 0.4650 \text{ SD} = 0.5910 \text{ cvSD} = 0.6017\end{aligned} \tag{11}$$

$$\begin{aligned}\log W = {} & -1.2789 * \log P + 0.8843\\ & * \textit{donatable hydrogen count} - 1.200 * \textit{ring}\\ & \textit{count nonaromatic } 6 + 0.5892;\\ \mathrm{r}^2 = {} & 0.9419 \text{ adj } \mathrm{r}^2 = 0.9419 \text{ q}^2 = 0.9392 \text{ avErr}\\ & = 0.4157 \text{ SD} = 0.5407 \text{ cvSD} = 0.5530\end{aligned} \tag{12}$$

$$\begin{aligned}\log W = {} & 79.2011 * (\boldsymbol{ring\ count\ aromatic\ 6/}\\ & \boldsymbol{molecular\ weight}) - 195.8089\\ & * (\boldsymbol{\log P/total\ accessible\ surface\ volume})\\ & - 0.0785 * (\textit{connectivity index } 1)^2 + 1.7210;\\ \mathrm{r}^2 = {} & 0.9507 \text{ adj } \mathrm{r}^2 = 0.9507 \text{ q}^2 = 0.9480 \text{ avErr}\\ & = 0.3945 \text{ SD} = 0.4981 \text{ cvSD} = 0.5115\end{aligned} \tag{13}$$

$$\begin{aligned}\log W = {} & 1.2721 * \log P + 1.0098 * \textit{carbonyl count}\\ & + 0.5682 * \textit{donotable hydrogen count}\\ & - 1.3650 * \textit{ring count nonaromatic}6 + 0.3397;\\ \mathrm{r}^2 = {} & 0.9428 \text{ adj } \mathrm{r}^2 = 0.9413 \text{ q}^2 = 0.9388 \text{ avErr}\\ & = 0.4165 \text{ SD} = 0.5365 \text{ cvSD} = 0.5550\end{aligned} \tag{14}$$

$$\begin{aligned}\log W = {} & 0.6271 * \textit{donotable hydrogen count}\\ & + 65.3322 * (\boldsymbol{ring\ count\ 6\ member/}\\ & \boldsymbol{molecular\ weight}) - 178.7146\\ & * (\boldsymbol{\log P/total\ accessible\ surface\ volume})\\ & - 0.0784 * (\textit{connectivity index } 1)^2 + 1.3770;\\ \mathrm{r}^2 = {} & 0.9570 \text{ adj } \mathrm{r}^2 = 0.9559 \text{ q}^2 = 0.9540 \text{ avErr}\\ & = 0.3715 \text{ SD} = 0.4649 \text{ cvSD} = 0.4811\end{aligned} \tag{15}$$

Once again, of the 10 possible times that a size-intensive descriptor could appear in a model, it appears 5 times.

The water solubility models were tested against 32 known solubilities and the test results were consistent with the expected errors. For example, the average error in Eq. 15 is 0.37 and the average error of the tested compounds was 0.44.

While the differences in r^2 between the non-size-intensive models and the non-linear size-intensive models are between 1% and 2%, it is important to note that the three-descriptor non-linear model has a higher r^2 (0.9507) than the four descriptor non-size-intensive model (0.9428).

Also note that the standard deviation in the error decreases 9% from 0.5407 in the nonsize-intensive three-descriptor model to 0.4981 in the size-intensive descriptor model. It decreases 15% from 0.5365 in the non size-intensive four-descriptor model to 0.4649 in the size-intensive model. Small improvements in r^2 of good models make much larger reductions in the error.

Conclusion

Size-intensive descriptors are a class of descriptors that are divided by the chemical sample size to remove dependence on size. The chemical sample size used to create a size-intensive descriptor can be measured by the number of atoms, molecular weight, length, width, depth, surface area, volume, longest path, containing box size, repeat unit count, certain connectivity indices, and so forth.

Size-intensive descriptors account for non-linear cross-dependencies between the extensive descriptor and the size descriptor. The original extensive descriptor and its size-intensive form are not highly correlated. Furthermore, size-intensive descriptors expand descriptor space in a way that cannot be represented as a linear combination of the original descriptor with sample size.

Conceptually, a size-intensive descriptor has a simple physical interpretation as a "density" or "concentration" of the base descriptor in a chemical sample. For example, the count of hydroxyl groups becomes the concentration of hydroxyl groups per square angstrom when the count is divided by the surface area.

Size-intensive descriptors have been demonstrated to appear frequently in the best models for human intestinal absorption, boiling points and water solubility where they were 5 of the 10 possible descriptors in each model set of best single, double, triple and quadruple descriptors.

Size-intensive descriptors have been shown to provide significant reduction in model error when compared to models that do not use size-intensive descriptors. Finally, size-intensive descriptors were shown to lead to models with fewer total descriptors for the same quality. This is important, models with fewer descriptors are easier to discover, usually more robust and more easily interpreted.

References

1. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH
2. Karleson M (2000) Molecular descriptors in QSAR/QSPR. Wiley-Interscience, New York
3. Stanton DT, Jurs PC (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. Anal Chem 62:2323–2329
4. Kier LB (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
5. Kier LB, Hall L (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
6. Kier LB (1999) Molecular structure description: the electrotopological state. Academic Press, New York

7. Stuper AJ, Jurs PC (1976) J Chem Inf Comput Sci 2:99
8. Jurs PC, Chou JT, Yuan M (1979) In: Olson RC, Christoffersen RE (eds) Computer-assisted drug design. American Chemical Society, Washington, DC, pp 103–129
9. CODESSA, www.codessa-pro.com/descriptors/. Accessed 13 Aug 2007
10. Purvis GD III (1994) The chemical sample: a fundamental object for molecular modeling. J Chem Inf Comput Sci 34:17–21
11. Katritzky AR, Lobanov VS, Karelson M (1998) Normal boiling points for organic compounds: correlation and prediction by a quantitative structure-property relationship. J Chem Inf Comput Sci 38:28–41
12. Katritzky AR, Mu L, Lobanov VS, Karelson M (1996) Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and test set of 9 simple inorganics. J Phys Chem 100:10400–10407
13. Bartlett RJ, Purvis GD (1978) Int J Quantum Chem 14:561
14. Bartlett RJ, Purvis GD III (1980) Molecular applications of coupled cluster and many-body perturbation methods. Phys Scr 21:255–265
15. Wessel MD, Jurs PC, Tolan JW, Muskal SM (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. J Chem Inf Comput Sci 38:726–735
16. Hou TJ, Xu XJ (2003) ADME: evaluation in drug discovery. 3. modeling blood-brain barrier partitioning using simple molecular descriptors. J Chem Inf Comput Sci 43:2137–2152
17. CAChe Worksystem Pro 6.1, Fujitsu Computer Systems, Beaverton, OR, 97007, (2007)
18. Randic M (1975) J Am Chem Soc 97:6606–6615
19. Guha R (2007) Chemical informatics functionality in R. J Stat Softw 18(5), supplemental material, http://www.jstatsoft.org/. Accessed 20 July 2007
20. Stanton DT, Jurs PC, Hicks MG (1991) Computer-assisted prediction of normal boiling points of furans, tetrahydrofurans, and thophenes. J Chem Inf Comput Sci 31:301–310
21. Egolf LM, Wessel MD, Jurs PC (1994) Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. J Chem Inf Comput Sci 34:947–956
22. Goll E, Jurs P (1999) Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model. J Chem Inf Comput Sci 39(6):974–983
23. Lowell H Hall LH, Story CT (1996) Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. J Chem Inf Comput Sci 36:1004–1014
24. Liang C, Gallagher D (1997) Prediction of physical & chemical properties by quantitative structure-property relationships: water solubility prediction. Am Lab March:34–40