New Molecular Descriptors Based on Local Properties at the Molecular Surface and a Boiling-Point Model Derived from Them

Bernd Ehresmann,[†] Marcel J. de Groot,[‡] Alexander Alex,[‡] and Timothy Clark*,[†]

Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25,
91052 Erlangen, Germany, and Pfizer Ltd., Global Research and Development, Sandwich Laboratories,
Sandwich, Kent CT13 9NJ, U.K.

Received September 28, 2003

New molecular descriptors based on statistical descriptions of the local ionization potential, local electron affinity, and the local polarizability at the surface of the molecule are proposed. The significance of these descriptors has been tested by calculating them for the Maybridge database in addition to our set of 26 descriptors reported previously. The new descriptors show little correlation with those already in use. Furthermore, the principal components of the extended set of descriptors for the Maybridge data show that especially the descriptors based on the local electron affinity extend the variance in our set of descriptors, which we have previously shown to be relevant to physical properties. The first nine principal components are shown to be most significant. As an example of the usefulness of the new descriptors, we have set up a QSPR model for boiling points using both the old and new descriptors.

INTRODUCTION

We[1] recently investigated the relationship between physical properties, molecular descriptors, and the drug-likeness of organic molecules. One of our major conclusions was that physical property space (as defined by our set of descriptors) is low-dimensional, as suggested by Lipinski et al.[2] Note that the fact that Lipinski et al.[2] based their conclusions on 2D-descriptors does not make them invalid in the present context. The dimensionality of physical-property space is independent of the set of descriptors used as long as they are able to describe molecules adequately. Our approach is analogous to the “chemical GPS” (global positioning system) concept of Oprea et al.[3,4] in that we try to map a large fraction of available chemistry in terms of physical property space. Interestingly, Oprea et al.[3,4] also identified nine principal components as being significant, although they used completely different descriptors, emphasizing the generality of Lipinski’s conclusions. Oprea[5] has also investigated the nature of the significant principal components for a large dataset and reaches conclusions very similar to those found in our work.[1]

One feature of our original work, however, must be regarded as preliminary. This is the use of element-specific descriptors. In our case, these are the sums of the molecular-electrostatic-potential derived charges[6] on all the atoms of a given element. These descriptors often function as pseudo-atom counts for the element in question. Their use in quantitative structure−property relationship (QSPR) models and their occurrence in some of the most significant principal components of the descriptor set for the Maybridge database[7] indicate that the remaining descriptors cannot describe some element-specific properties of the atoms in molecules. A further disadvantage of such element-specific descriptors is that one descriptor is required for each element present in the dataset, leading to a larger number of descriptors for a complete model (essentially one such descriptor of every type for each element) and to models that need to be extended when new elements are present in the dataset. Many whole-molecule descriptors exist, some of them based on quantum mechanical descriptions of molecules.[8] Our aim in this work, however, is to generate a set, complete as possible, of essentially orthogonal descriptors for describing physical properties. This aim contrasts strongly with the approach used by Karelson and Katrizky,[8] in which many descriptors were generated and the most appropriate were selected from among them. The use of descriptors based on local properties calculated at a molecular surface is part of a wider program to investigate the feasibility of modeling paradigms based on molecular surfaces rather than atomic structures.

To address this problem, we have now developed new surface-based molecular descriptors so that we can eliminate element-specific descriptors from those necessary to describe physical properties adequately. The goal of eliminating element-specific descriptors is not merely an intellectual exercise. The discovery of a new drug scaffold unrelated to the lead series or known analogues (scaffold hopping) depends heavily on the ability to establish quantitative structure−activity relationships (QSARs) based on descriptors calculated either for the entire molecule or for the molecular surface. Once this is possible, the structural formula disappears entirely from the descriptor space, so that scaffold hops become inherently more likely. The present paper reports new descriptors intended to contribute to the goal outlined above.

We emphasize that our approach is based on semiempirical molecular orbital (MO) theory. The local properties and their definitions are framed in terms appropriate to this theoretical framework and are not necessarily transferable to, for instance, density functional theory (DFT) or ab initio

[†] Friedrich-Alexander-Universität Erlangen-Nürnberg.
[‡] Pfizer Ltd.

10.1021/ci034215e CCC: $27.50
Published on Web 02/10/2004

calculations with large Gaussian basis sets. One reason for this limitation is that we are able to treat tens of thousands of molecules with semiempirical MO theory. Another is that this level of theory allows several simplifications that are not appropriate for DFT or ab initio calculations. We are, however, developing equivalent treatments for other levels of theory.

SURFACE-BASED PROPERTIES

Our current surface-based descriptors depend on the molecular electrostatic potential (MEP) calculated using semiempirical MO-theory at the solvent-excluded surface (SES).[9] These descriptors are based on original work by Murray, Politzer, et al.,[10] who, however, use density functional theory (DFT) to calculate the MEP at isodensity surfaces. Purely electrostatic descriptors, however, largely ignore the local donor/acceptor and polarizability properties at points on the surface. These properties, often described in terms of "hardness" and "softness" of the elements,[11] must also be captured in the descriptor set.

One solution to this problem is to use the local ionization energy, IE_L, which was first proposed by Sjoberg et al.[12] and made popular by Murray, Politzer et al.[13-16] The IE_L at a given point is defined using Koopmans' theorem as

$$IE_L = \frac{-\sum_{i=1}^{\mathrm{HOMO}} \rho_i(r)\epsilon_i}{\sum_{i=1}^{\mathrm{HOMO}} \rho_i(r)} \qquad (1)$$

where HOMO is the highest occupied molecular orbital (MO), $\rho_i(r)$ is the electron density attributable to the ith MO at the point (r) being considered, and ϵ_i is the Eigenvalue (i.e., ϵ_i is the Koopmans' theorem ionization potential associated with the ith MO).

We have recently[17] extended this definition to treat the acceptor properties by defining the local electron affinity, EA_L, as

$$EA_L = \frac{-\sum_{i=\mathrm{LUMO}}^{norbs} \rho_i(r)\epsilon_i}{\sum_{i=\mathrm{LUMO}}^{norbs} \rho_i(r)} \qquad (2)$$

where LUMO is the lowest unoccupied MO and *norbs* is the number of MOs. Note that this definition is only sensible for a very limited basis set and that it depends on the concept of meaningful virtual MOs. Clearly, a very extended basis set with a very large fraction of the orbitals unoccupied would give meaningless values for the local electron affinity calculated according to eq 2. However, eq 2 is extremely useful within the limits of semiempirical MO methods using minimal valence-only Slater basis sets—the methods used in this work. Similarly, the use of virtual orbitals within a minimal basis set approximation has a long tradition, although it cannot be justified theoretically as the virtual MOs are simply a consequence of the variational optimization of the occupied ones.

IE_L and EA_L can be combined to give a local Mulliken electronegativity,[18] χ_L and a local hardness,[11] η_L

$$\chi_L = \frac{(IE_L + EA_L)}{2} \qquad (3)$$

$$\eta_L = \frac{(IE_L - EA_L)}{2} \qquad (4)$$

The local electronegativity, if it were really the electronegativity, would have to be constant at all points on the surface of the molecule because of the principle of electronegativity equalization. Our "local electronegativity" is not constant because the local ionization energy and local electron affinity are not exactly equivalent to their nonlocal equivalents. However, the local electronegativity does not vary greatly. We therefore use only its average value at all surface points as a descriptor. As we have already described the local properties themselves,[17] we will not discuss them in detail here. Note, however, that our definition of the local hardness is an approximation appropriate to the semiempirical MO framework used.

We have recently defined an additive distributed polarizability model[19] based on Rivail's variational ansatz.[20] This procedure, which we originally defined for atom polarizabilities, can easily be extended to "atomic orbital polarizabilities". This allows us[17] to define a "local polarizability", α_L, at a point near the molecule as

$$\alpha_L = \frac{\sum_{j=1}^{norbs} \rho_j^1(r) q_j \bar{\alpha}_j}{\sum_{i=1}^{norbs} \rho_j^1(r) q_j} \qquad (5)$$

where q_j is the Coulson occupation and $\bar{\alpha}_j$ is the isotropic polarizability attributed to atomic orbital j. The density $\rho_j^1(r)$ is defined as the electron density at the point (r) in question due to an exactly singly occupied atomic orbital j.

The local ionization energy, IE_L, has been discussed in detail by Murray, Politzer et al.[12-16] and will not be discussed further here. We have discussed the characteristics of EA_L, η_L, and α_L and their interpretation in terms of chemical reactivity in a separate article.[17]

Correlations Between Surface Properties. To investigate the independence of the surface properties and their relationship to the MEP, we calculated all five properties at points on the solvent-excluded surfaces of the top three prescription drug molecules[21] (nifedipine, **1**, quinapril, **2**, and zafirlukast, **3**, Chart 1) and combined the datasets to give the values of all five properties at 8633 points at the triangulation points on the solvent-excluded surfaces[9] of the three different molecules. The correlation matrix obtained for this dataset is shown in Table 1.

The local ionization energy and the three surface properties introduced above show very little correlation with the MEP, suggesting that they introduce new information not contained in our MEP-based descriptor set. Strong correlations among the other four properties are limited to the local ionization energy, IE_L, which correlates relatively strongly ($r^2 \approx 0.8$) with η_L. This correlation is not surprising because η_L is

Chart 1

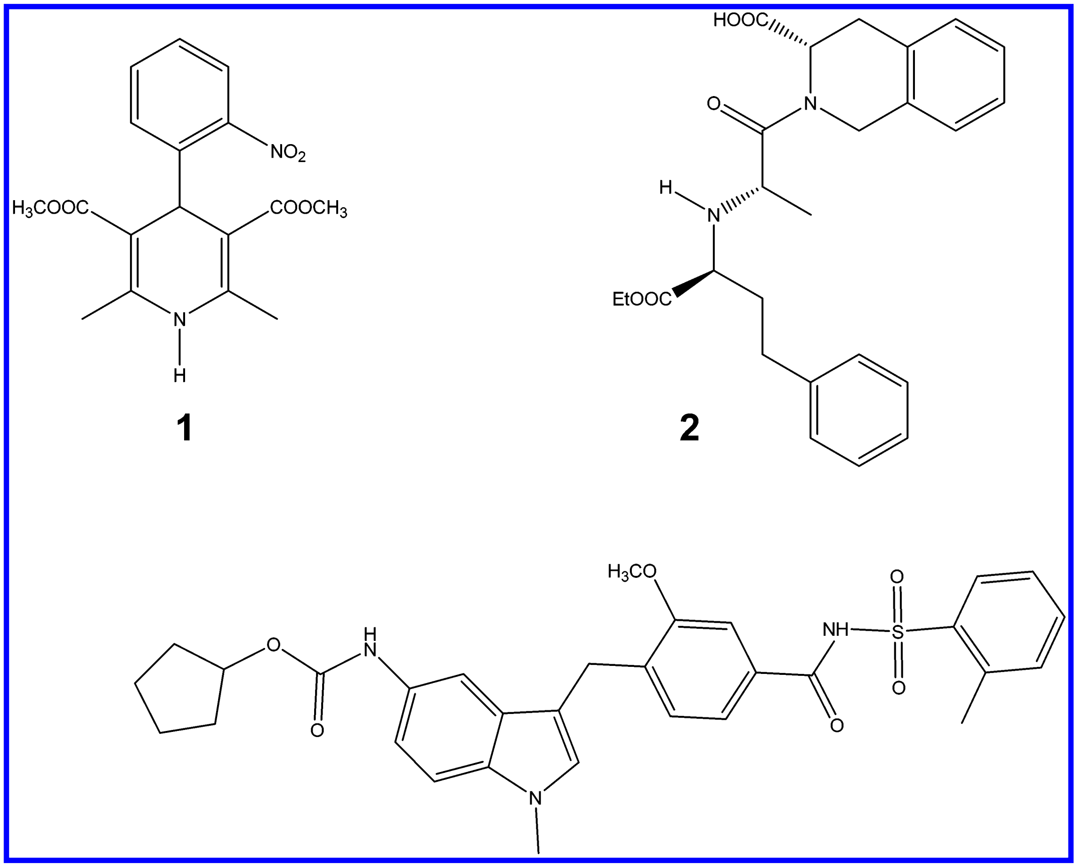

Table 1. Correlations between the Five Surface Properties for All Points Calculated at the Surfaces of the Three Drug Molecules (**1**−**3**, N = 8633)

	MEP	IE_L	EA_L	η_L	α_L
MEP	1				
IE_L	0.146	1			
EA_L	−0.122	0.178	1		
η_L	0.207	*0.807*	−0.438	1	
α_L	0.292	0.187	0.508	−0.139	1

Table 2. Principal Components of the Four Local Properties Calculated at the Surfaces of the Three Drugs (**1**−**3**, N = 8633)

	PC1	PC2	PC3	PC4
MEP	*0.481*	0.065	*−0.839*	−0.248
IE_L	*0.569*	0.154	*0.520*	*−0.618*
EA_L	*0.606*	0.244	0.147	*0.742*
α_L	−0.279	*0.955*	−0.065	−0.073
% variance	39.8	23.8	21.0	15.4
cumulative % variance	39.8	63.6	84.6	100
Eigenvalue	1.591	0.953	0.840	0.617

derived from IE_L. EA_L does not correlate strongly with η_L because its numerical variation is far smaller than that of IE_L, which therefore dominates the variation of η_L. Neither the MEP nor the donor/acceptor-related properties correlate strongly with the local polarizability, not even IE_L, which has been related conceptually to polarizability.[12−16] The correlation matrix suggests that MEP and α_L can be used with either EA_L and either IE_L or η_L to give two alternative sets of four independent surface properties.

However, we can also use the principal components (PCs) of the values of the properties on the surface of a test molecule to judge the amount and dimensionality of the information coded by the four properties used below to derive descriptors (MEP, IE_L, EA_L, and α_L). Table 2 and Figure 1 show the results of such a principal component analysis using the combined dataset calculated for the three drugs (i.e., the values of the four properties at all the surface points calculated for the three compounds). The eigenvalue test suggests that only the first PC (with an eigenvalue larger than 1.0) is significant. The eigenvalue for PC2 is, however 0.953 and the kink in the Scree plot (a plot of the Eigenvalues against the number of the principal component,[22] Figure 1) occurs at this factor. We conclude that the first two PCs are significant. The first PC comprises strong contributions from

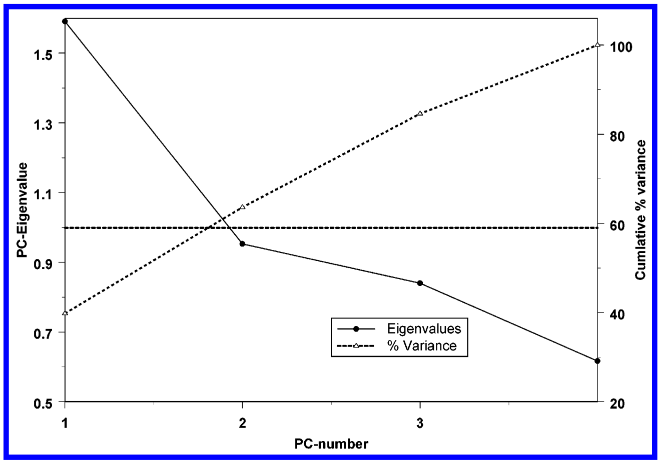

Figure 1. Scree-plot[22] of the principal components calculated for values of the MEP, IE_L, EA_L, and α_L at the surfaces of the three drugs **1**−**3** (N = 22 690).

Table 3. Twenty-Five Molecular Descriptors Derived from the Local Ionization Energy, $\mathrm{IE_L}$, Electron Affinity, $\mathrm{EA_L}$, Electronegativity, χ_L, Hardness, η_L, and Polarizability, α_L[a]

descriptor	description	formula
σ^2_{IE}	variance in the local ionization energy	$\sigma^2_{\mathrm{IE}} = \frac{1}{N}\sum_{i=1}^{N}[\mathrm{IE}^{\,i}_{\mathrm{L}} - \overline{\mathrm{IE}_{\mathrm{L}}}]^2$
$\mathrm{IE}^{\max}_{\mathrm{L}}$	maximum value of the local ionization energy	
$\mathrm{IE}^{\min}_{\mathrm{L}}$	minimum value of the local ionization energy	
$\overline{\mathrm{IE}_{\mathrm{L}}}$	mean value of the local ionization energy	$\overline{\mathrm{IE}_{\mathrm{L}}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{IE}^{\,i}_{\mathrm{L}}$
$\Delta\mathrm{IE_L}$	range of the local ionization energy	$\Delta\mathrm{IE_L} = \mathrm{IE}^{\max}_{\mathrm{L}} - \mathrm{IE}^{\min}_{\mathrm{L}}$
$\sigma^2_{\mathrm{EA+}}$	variance in the local electron affinity for all positive values	$\sigma^2_{\mathrm{EA+}} = \frac{1}{m}\sum_{i=1}^{m}[\mathrm{EA}^{+}_{i} - \overline{\mathrm{EA}^{+}}]^2$
$\sigma^2_{\mathrm{EA-}}$	variance in the local electron affinity for all negative values	$\sigma^2_{\mathrm{EA-}} = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{EA}^{-}_{i} - \overline{\mathrm{EA}^{-}}]^2$
$\sigma^2_{\mathrm{EAtot}}$	sum of the positive and negative variances in the local electron affinity	$\sigma^2_{\mathrm{EAtot}} = \sigma^2_{\mathrm{EA+}} + \sigma^2_{\mathrm{EA-}}$
ν_{EA}	local electron affinity balance parameter	$\nu_{EA} = \frac{\sigma^2_{\mathrm{EA}} + {}^{\bullet}\sigma^2_{\mathrm{EA}}-}{[\sigma^2_{\mathrm{EA}}]^2}$
$\mathrm{EA}^{\max}_{\mathrm{L}}$	maximum of the local electron affinity	
$\mathrm{EA}^{\min}_{\mathrm{L}}$	minimum of the local electron affinity	
$\overline{\mathrm{EA}_{\mathrm{L}}}$	mean value of the local electron affinity	$\overline{\mathrm{EA}_{\mathrm{L}}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{EA}^{i}_{\mathrm{L}}$
$\Delta\mathrm{EA_L}$	range of the local ionization energy	$\Delta\mathrm{EA_L} = \mathrm{EA}^{\max}_{\mathrm{L}} - \mathrm{EA}^{\min}_{\mathrm{L}}$
$\delta\mathrm{A}^{+}_{\mathrm{EA}}$	fraction of the surface area with positive local electron affinity	$\delta\mathrm{A}^{+}_{\mathrm{EA}} = \frac{\mathrm{A}^{+}_{\mathrm{EA}}}{\mathrm{A}}$, A = total surface area
σ^2_{η}	*variance in the local hardness*	$\sigma^2_{\eta} = \frac{1}{N}\sum_{i=1}^{N}[\eta^{i}_{\mathrm{L}} - \overline{\eta_{\mathrm{L}}}]^2$
$\eta^{\max}_{\mathrm{L}}$	*maximum value of the local hardness*	
$\eta^{\min}_{\mathrm{L}}$	*minimum value of the local hardness*	
$\overline{\eta_{\mathrm{L}}}$	*mean value of the local hardness*	$\overline{\eta_{\mathrm{L}}} = \frac{1}{N}\sum_{i=1}^{N}\eta^{i}_{\mathrm{L}}$
$\Delta\eta_{\mathrm{L}}$	*range of the local hardness*	$\Delta\eta_{\mathrm{L}} = \eta^{\max}_{\mathrm{L}} - \eta^{\min}_{\mathrm{L}}$
$\overline{\chi_{\mathrm{L}}}$	mean value of the local electronegativity	$\overline{\chi_{\mathrm{L}}} = \frac{1}{N}\sum_{i=1}^{N}\chi^{\,i}_{\mathrm{L}}$
σ^2_{α}	variance in the local polarizability	$\sigma^2_{\alpha} = \frac{1}{N}\sum_{i=1}^{N}[\alpha^{i}_{\mathrm{L}} - \overline{\alpha_{L}}]^2$
$\alpha^{\max}_{\mathrm{L}}$	maximum value of the local polarizability	
$\alpha^{\min}_{\mathrm{L}}$	minimum value of the local polarizability	
$\overline{\alpha_{\mathrm{L}}}$	mean value of the local polarizability	$\overline{\alpha_{L}} = \frac{1}{N}\sum_{i=1}^{N}\alpha^{i}_{L}$
$\Delta\alpha_{\mathrm{L}}$	range of the local polarizability	$\Delta\alpha_{\mathrm{L}} = \alpha^{\max}_{\mathrm{L}} - \alpha^{\min}_{\mathrm{L}}$

[a] Variables not defined explicitly in the table are exactly analogous to those defined for the MEP in the text. The five italicized descriptors based on the local hardness were omitted from the descriptor set, as described in the text.

the MEP (23%), the local ionization energy (32%), and the local electron affinity (37%), whereas PC2 is almost exclusively (91%) made up of the local polarizability. Not surprisingly, adding the local electronegativity and hardness to the four properties does not add significant information (the most significant PCs are very similar to those without them except for correlation effects) but the local hardness is useful for visualizing molecular properties[17] and the mean value of the local electronegativity (see below) is a useful descriptor.

Molecular Descriptors from Surface Properties. Murray, Politzer, and co-workers[23−27] have pioneered the use of the statistics of the values of the MEP at points on the molecular surface to derive molecular descriptors.

Table 4. Correlation Matrix Obtained between the 14 "Old" Molecule-Based Descriptors from Our Original Set (Ref 1) with the "New" Set of 25 Descriptors Derived from the New Surface Properties (Table 3), All Calculated for the Maybridge Database (Positive and Negative Correlation Coefficients Greater than or Equal to 0.5 Are Shown in Italics)

25 "new" descriptors	14 "old" descriptors (ref 1)													
	μ	μ_D	α	V_{max}	V_{min}	$\bar{V}_+$	$\bar{V}_-$	σ^2_{tot}	ν	$\sigma^2_{tot}\nu$	MW	vol	A	G
σ^2_{IE}	0.06	0.15	−0.09	0.35	−0.12	0.41	−0.12	0.19	0.42	0.42	0.08	−0.15	−0.13	0.09
IE_L^{max}	0.22	0.12	0.09	0.42	0.07	0.45	0.00	0.08	*0.50*	0.37	0.32	0.16	0.19	−0.27
IE_L^{min}	−0.05	0.03	−0.22	−0.10	0.27	−0.03	−0.02	−0.05	−0.10	−0.10	−0.20	−0.11	−0.13	0.16
$\overline{IE_L}$	0.17	0.19	−0.15	0.24	0.13	0.35	−0.27	0.21	0.15	0.25	−0.09	−0.03	−0.01	−0.02
ΔIE_L	0.15	0.03	0.23	0.28	−0.20	0.24	0.01	0.09	0.33	0.27	0.33	0.18	0.20	−0.27
σ^2_{EA+}	0.05	0.10	−0.03	0.09	0.02	0.18	−0.10	0.08	0.06	0.11	0.00	−0.03	−0.03	0.04
σ^2_{EA-}	0.11	0.06	0.16	0.07	0.06	0.12	−0.02	−0.01	0.15	0.09	0.19	0.05	0.07	−0.13
$\sigma^2_{EA\mathit{tot}}$	0.11	0.07	0.16	0.08	0.06	0.14	−0.03	0.00	0.16	0.11	0.19	0.05	0.06	−0.12
ν_{EA}	0.09	0.12	−0.02	0.26	0.08	0.36	−0.13	0.12	0.18	0.23	0.07	−0.01	−0.01	0.02
EA_L^{max}	0.27	0.17	0.17	0.39	0.08	0.46	−0.14	0.14	0.35	0.33	0.32	0.14	0.17	−0.24
EA_L^{min}	−0.14	−0.07	−0.08	0.14	−0.12	0.06	0.11	−0.01	0.14	0.07	−0.01	−0.11	−0.11	0.12
$\overline{EA_L}$	0.15	0.13	0.08	0.35	0.07	0.49	0.04	0.03	0.46	0.32	0.30	−0.01	0.02	−0.12
ΔEA_L	0.30	0.18	0.18	0.18	0.14	0.30	−0.18	0.11	0.16	0.19	0.24	0.19	0.21	−0.26
δA^+_{EA}	0.15	0.18	−0.04	0.35	0.13	0.48	−0.18	0.13	0.31	0.30	0.13	−0.04	−0.04	0.03
σ^2_η	0.00	0.07	−0.02	0.23	−0.14	0.24	−0.03	0.10	0.30	0.28	0.09	−0.12	−0.11	0.07
η_L^{max}	0.25	0.15	0.04	0.33	0.11	0.39	−0.04	0.05	0.39	0.27	0.25	0.15	0.17	−0.21
η_L^{min}	−0.05	0.03	−0.22	−0.16	0.18	−0.08	−0.05	−0.03	−0.17	−0.13	−0.25	−0.11	−0.13	0.17
$\overline{\eta_L}$	0.01	0.03	−0.14	−0.07	0.03	−0.09	−0.19	0.11	−0.20	−0.04	−0.24	−0.01	−0.02	0.06
$\Delta\eta_L$	0.13	0.01	0.23	0.25	−0.14	0.20	0.03	0.04	0.29	0.21	0.32	0.15	0.18	−0.23
$\overline{\chi_L}$	0.26	0.27	−0.06	0.49	0.17	*0.70*	−0.19	0.20	*0.51*	0.47	0.17	−0.03	0.01	−0.11
σ^2_α	−0.05	0.00	−0.04	0.03	−0.01	0.07	0.01	0.00	0.06	0.04	0.13	−0.10	−0.10	0.11
α_L^{max}	−0.08	−0.08	0.09	−0.01	−0.10	−0.05	0.09	−0.04	0.02	−0.01	0.18	−0.02	−0.02	0.01
α_L^{min}	−0.24	−0.14	−0.02	−0.27	−0.23	−0.38	0.11	−0.05	−0.29	−0.20	−0.23	−0.15	−0.16	0.20
$\overline{\alpha_L}$	−0.17	−0.15	0.12	−0.18	−0.20	−0.29	0.14	−0.08	−0.16	−0.14	0.04	−0.05	−0.06	0.08
$\Delta\alpha_L$	−0.01	−0.04	0.10	0.07	−0.04	0.06	0.07	−0.02	0.10	0.05	0.26	0.03	0.03	−0.05

These descriptors vary from simple MEP values such as the maximum, minimum, mean, and the means of the positive and negative values[23] to statistical descriptions[24] such as the positive and negative variances, σ^2_+ and σ^1_-, which are defined as

$$\sigma^2_+ = \frac{1}{m}\sum_{i=1}^{m}[V_i^+ - \bar{V}^+]^2 \qquad (6)$$

and

$$\sigma^2_- = \frac{1}{n}\sum_{i=1}^{n}[V_i^- - \bar{V}^-]^2 \qquad (7)$$

where m and n are the numbers of surface points with positive and negative MEPs, respectively, V_i^+ and V_i^- are the MEP values for point i on the surface (with positive or negative MEP, respectively), and $\overline{V^+}$ and $\overline{V^-}$ are the mean positive and negative MEPs, respectively. The total variance, σ^2_{tot} is defined as the sum of σ^2_+ and σ^2_-. A so-called balance parameter, ν, is defined as

$$\nu = \frac{\sigma^2_+\,\sigma^1_-}{[\sigma^2_{tot}]^2} \qquad (8)$$

and the simple product of the balance parameter and the total variance is also used as a descriptor. Finally, a so-called local polarity (which is actually a nonlocal property), Π, is defined[25] as

$$\Pi = \frac{1}{N}\sum_{i=1}^{N}|V_i - \bar{V}| \qquad (9)$$

where N is the total number of surface points and $\bar{V}$ is the overall mean value of the MEP.

We have used the same definitions applied to the new set of surface properties to define the set of 25 molecular descriptors shown in Table 3. Note that, for properties that can only have either positive or negative values, the simple variance, σ^2, is used, rather than the definitions given above, and, as discussed above, only the mean value of the local electronegativity is used.

Correlations Between and Significance of the New Descriptors. To test the significance and interdependences of our new set of descriptors, we calculated them for all of the molecules contained in the Maybridge database[7] processed using our standard protocol[1,28] using AM1[29] with VAMP 8.1.[30] We then calculated the correlation matrix between the new descriptors described in Table 3 and the 14 of our standard set of 26 descriptors used previously[1] that are derived from purely molecular properties (see below). The results are shown in Tables 4 and 5.

Table 4 shows the correlations between the older and the new set of descriptors. There are pleasingly few even moderately correlated pairs of descriptors. The mean positive

Table 5. Correlation Matrix Obtained for the 25 New Descriptors Calculated for the Maybridge Database (Positive and Negative Correlation Coefficients Greater than or Equal to 0.7 Are Shown in Italics)

	σ^2_{IE}	$\mathrm{IE}_{\mathrm{L}}^{\mathrm{max}}$	$\mathrm{IE}_{\mathrm{L}}^{\mathrm{min}}$	$\overline{\mathrm{IE}_{\mathrm{L}}}$	$\Delta\mathrm{IE}_{\mathrm{L}}$	$\sigma^2_{\mathrm{EA}+}$	$\sigma^2_{\mathrm{EA}-}$	$\sigma^2_{\mathrm{EAtot}}$	ν_{EA}	$\mathrm{EA}_{\mathrm{L}}^{\mathrm{max}}$	$\mathrm{EA}_{\mathrm{L}}^{\mathrm{min}}$	$\overline{\mathrm{EA}_{\mathrm{L}}}$	$\Delta\mathrm{EA}_{\mathrm{L}}$	$\delta\mathrm{A}_{\mathrm{EA}}^{+}$
σ^2_{IE}	1.00													
$\mathrm{IE}_{\mathrm{L}}^{\mathrm{max}}$	0.24	1.00												
$\mathrm{IE}_{\mathrm{L}}^{\mathrm{min}}$	−0.67	−0.02	1.00											
$\overline{\mathrm{IE}_{\mathrm{L}}}$	−0.48	0.33	0.60	1.00										
$\Delta\mathrm{IE}_{\mathrm{L}}$	*0.70*	0.50	*−0.88*	−0.36	1.00									
$\sigma^2_{\mathrm{EA}+}$	0.08	0.03	−0.01	−0.01	0.02	1.00								
$\sigma^2_{\mathrm{EA}-}$	0.36	0.01	−0.42	−0.31	0.36	0.08	1.00							
$\sigma^2_{\mathrm{EAtot}}$	0.36	0.01	−0.41	−0.30	0.36	0.22	0.99	1.00						
ν_{EA}	0.15	0.13	0.00	0.03	0.07	0.69	0.11	0.20	1.00					
$\mathrm{EA}_{\mathrm{L}}^{\mathrm{max}}$	0.31	0.36	−0.23	0.00	0.37	0.33	0.61	0.64	0.52	1.00				
$\mathrm{EA}_{\mathrm{L}}^{\mathrm{min}}$	0.27	0.15	−0.14	−0.20	0.20	0.05	−0.18	−0.17	0.14	0.08	1.00			
$\overline{\mathrm{EA}_{\mathrm{L}}}$	0.59	0.36	−0.37	−0.28	0.50	0.17	0.55	0.56	0.32	0.67	0.31	1.00		
$\Delta\mathrm{EA}_{\mathrm{L}}$	0.03	0.15	−0.07	0.14	0.13	0.20	0.58	0.59	0.28	0.68	−0.67	0.27	1.00	
$\delta\mathrm{A}_{\mathrm{EA}}^{+}$	0.32	0.22	−0.11	−0.06	0.20	0.36	0.35	0.39	0.67	0.62	0.16	0.57	0.34	1.00
σ^2_{η}	*0.94*	0.08	*−0.75*	−0.64	0.69	0.12	0.50	0.50	0.15	0.33	0.29	0.59	0.03	0.33
$\eta_{\mathrm{L}}^{\mathrm{max}}$	0.09	*0.80*	0.11	0.46	0.29	0.00	−0.11	−0.10	0.08	0.20	0.06	0.15	0.10	0.16
$\eta_{\mathrm{L}}^{\mathrm{min}}$	−0.68	−0.07	*0.95*	0.60	*−0.86*	−0.06	−0.52	−0.52	−0.10	−0.38	−0.19	−0.50	−0.14	−0.23
$\overline{\eta_{\mathrm{L}}}$	−0.67	−0.02	0.60	*0.80*	−0.53	−0.11	−0.54	−0.54	−0.19	−0.42	−0.32	*−0.80*	−0.08	−0.40
$\Delta\eta_{\mathrm{L}}$	0.69	0.31	*−0.90*	−0.45	*0.93*	0.06	0.47	0.47	0.12	0.43	0.21	0.53	0.17	0.28
$\overline{\chi_{\mathrm{L}}}$	0.09	0.58	0.18	0.60	0.12	0.13	0.20	0.22	0.29	0.56	0.10	0.61	0.34	0.42
σ^2_{α}	0.43	−0.02	−0.21	−0.40	0.17	0.02	0.28	0.28	0.05	0.10	0.15	0.34	−0.04	0.19
$\alpha_{\mathrm{L}}^{\mathrm{max}}$	0.53	−0.12	−0.49	−0.62	0.37	0.01	0.39	0.39	0.01	0.12	0.18	0.40	−0.04	0.11
$\alpha_{\mathrm{L}}^{\mathrm{min}}$	0.08	−0.46	−0.23	−0.49	−0.02	−0.02	0.02	0.01	−0.09	−0.23	0.26	−0.08	−0.36	−0.18
$\overline{\alpha_{\mathrm{L}}}$	0.42	−0.39	−0.48	*−0.81*	0.22	0.00	0.37	0.36	−0.05	−0.01	0.20	0.29	−0.15	0.03
$\Delta\alpha_{\mathrm{L}}$	0.53	0.02	−0.44	−0.50	0.39	0.02	0.41	0.40	0.04	0.20	0.11	0.44	0.06	0.17

	σ^2_{η}	$\eta_{\mathrm{L}}^{\mathrm{max}}$	$\eta_{\mathrm{L}}^{\mathrm{min}}$	$\overline{\eta_{\mathrm{L}}}$	$\Delta\eta_{\mathrm{L}}$	$\overline{\chi_{\mathrm{L}}}$	σ^2_{α}	$\alpha_{\mathrm{L}}^{\mathrm{max}}$	$\alpha_{\mathrm{L}}^{\mathrm{min}}$	$\overline{\alpha_{\mathrm{L}}}$	$\Delta\alpha_{\mathrm{L}}$
σ^2_{η}	1.00										
$\eta_{\mathrm{L}}^{\mathrm{max}}$	−0.06	1.00									
$\eta_{\mathrm{L}}^{\mathrm{min}}$	*−0.79*	0.08	1.00								
$\overline{\eta_{\mathrm{L}}}$	*−0.77*	0.19	0.68	1.00							
$\Delta\eta_{\mathrm{L}}$	*0.75*	0.22	*−0.95*	−0.61	1.00						
$\overline{\chi_{\mathrm{L}}}$	−0.03	0.50	0.08	−0.01	0.07	1.00					
σ^2_{α}	0.47	−0.07	−0.26	−0.46	0.24	−0.05	1.00				
$\alpha_{\mathrm{L}}^{\mathrm{max}}$	0.62	−0.23	−0.54	−0.63	0.45	−0.18	*0.86*	1.00			
$\alpha_{\mathrm{L}}^{\mathrm{min}}$	0.21	−0.58	−0.19	−0.25	0.01	−0.47	0.10	0.30	1.00		
$\overline{\alpha_{\mathrm{L}}}$	0.58	−0.53	−0.49	−0.69	0.32	−0.43	0.65	*0.79*	0.64	1.00	
$\Delta\alpha_{\mathrm{L}}$	0.58	−0.07	−0.50	−0.59	0.47	−0.05	*0.87*	*0.96*	0.01	0.64	1.00

MEP, $\bar{V}$, and the MEP-balance parameter, ν, correlate moderately with the mean of the local electronegativity. Otherwise, all correlation coefficients are 0.5 or smaller, suggesting that the new descriptors add significant information to our data.

Table 5 shows the intercorrelations among the set of 25 descriptors described in Table 3. All correlation coefficients larger than or equal to 0.7 are highlighted by italics. Correlations greater than 0.9 are only found between σ^2_{IE} and σ^2_{η} and for IE_L^{min} with both η_L^{min} and $\Delta\eta_L$, which also correlates strongly with ΔIE_L. Perhaps not surprisingly, more moderate correlations are found between the balance parameter for the local electron affinity and the variance in the negative values of the same property. The minimum value and the range of the local hardness also correlate moderately, as do the maximum and range of the local polarizability. We can conclude, however, that most significant correlations can be removed from our data by removing all descriptors based on the local hardness. These five descriptors were therefore omitted from our descriptor set for the remainder of this work to give a set of 20 new descriptors based on the local ionization energy and electron affinity and also including the mean value of the local electronegativity on the molecular surface.

Principal Components of the Descriptor Set. As in our earlier work on descriptors and physical properties,[1] we have investigated the description of the molecules of the Maybridge dataset. In the present case, however, we have used the set of 20 local ionization energy, electron affinity, and polarizability based descriptors shown in Table 3 along with

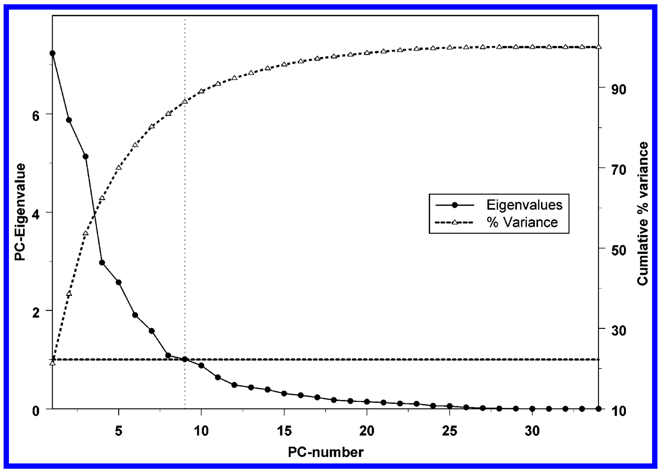

Figure 2. Scree-plot[22] of the principal components calculated for the set of 34 descriptors described in the text for the entire Maybridge dataset.

the following 14 electrostatic, shape, and size descriptors. These describe molecular properties not covered by the new surface properties: dipole moment, μ; dipolar density,[31] μ_D; molecular electronic polarizabilty, α[32]; maximum (most positive) MEP, V_{max}; minimum (most negative) MEP, V_{min}; mean of the positive MEP values, $\bar{V}_+$; mean of the negative MEP values, $\bar{V}_-$; total variance in the MEP, σ^2_{tot} [10]; MEP balance parameter, ν[10]; product of the total variance in the MEP and the balance parameter, $\sigma^2_{tot}\nu$ [10]; molecular weight, MW; molecular volume, VOL; molecular surface area, A; and globularity, G[33].

These 14 descriptors together with the 20 defined in Table 3 comprise the full set of 34 calculated for the Maybridge dataset. The principal components of the descriptors thus obtained were then calculated. The resulting Scree-plot[22] is shown in Figure 2.

The Eigenvalue test[34] suggests that nine PCs are significant. The Eigenvalue plot shows a significant kink[35] between PCs 8 and 9, so we have chosen to include nine PCs in our analysis. Note that our earlier set of 26 descriptors,[1] which also included element-specific descriptors, and the descriptor set used by Oprea[5] also gave a Scree-plot that suggested that nine PCs were significant. The vectors corresponding to the nine most significant PCs are shown in Table 6.

The individual PCs can, to some extent, be interpreted in terms of familiar concepts, although they are not as transparent as we might like.

PC1 is a combination of many descriptors derived from the MEP and the new surface properties. The descriptors derived from the MEP, local ionization energy, and local electron affinity contribute 23.4%, 16.8%, and 39.5%, respectively, to this vector to give a total of almost 80% for these three properties. This vector, which describes 22% of the variance in the dataset, is related to both donor/acceptor and Coulomb interactions. This is perhaps surprising as we instinctively expect Coulomb interactions to be far stronger than donor/acceptor. This descriptor can, however, be interpreted as describing predominantly Coulomb and acceptor interactions of the molecule.

PC2 is also composed predominantly of descriptors derived from the four properties found in PC1. However, the local polarizability plays a very significant role in this factor, with a contribution of over 33%. The MEP-derived descriptors are also important (19%), but in this factor the local ionization energy (21%) plays a more significant role than the local electron affinity (6%). We suggest that this factor is the donor equivalent of PC1. We noted in our earlier paper[1] that factors tend to occur in pairs.

PC3 is composed predominantly (60%) of the size and shape descriptors molecular electronic polarizability, molecular weight, molecular volume, molecular surface area, and globularity. This vector corresponds closely (see below) to the most significant PC found using our older set of descriptors.[1]

PC4 is dominated by MEP- (51%) and EA_L- (29%) derived descriptors.

PC5 is a very "delocalized" factor comprising roughly 30% each of the descriptors derived from the MEP and the local electron affinity. The local ionization energy contributes only 12%, but the dipole moment (4.4%) and the dipolar density (4%) are also significant.

PC6 is dominated (76%) by descriptors derived from EA_L. Thus, this factor is a classical Lewis-acid descriptor.

PC7 consists mostly (53%) of descriptors derived from the local polarizability. As these descriptors describe the variation of the polarizability, PC7 can be considered to be describing chemical diversity.

PC8 has strong contributions from the molecular dipole (32%) and the dipolar density (27%). This factor corresponds very loosely to the dipolar polarity PC found using our original descriptor set,[1] although the two do not correlate significantly (see below).

PC9 is another descriptor in which the local electron affinity (55%) dominates. The mean local electronegativity (8%) also plays a significant role.

As suggested above, the principal components derived from the new set of descriptors are not as easily interpreted as those found previously.[1] PC3 is clearly the size and shape descriptor found previously and PC8 corresponds very loosely to the dipolar polarity descriptor also found previously. The remaining seven significant descriptors all represent a combination of Coulomb, donor, acceptor, and polarization descriptors. These combined effects describe intermolecular interactions from the purely electrostatic through hydrogen bonds to dispersion.

To test the correspondence between the PCs calculated from the new descriptor set and the original one, we correlated the top nine original principal components reported in ref 1 with the nine most significant calculated using the new set of 34 descriptors. The results are shown in Table 7.

The strongest correlation (0.88) is found between the most significant principal component calculated using the old set of descriptors (size and shape[1]) with PC3 calculated with the new set. This is not surprising as the descriptors involved are all found in the old set of descriptors and are largely independent of the Coulomb, donor, or acceptor descriptors. Most significantly, however, this factor, which is largely unchanged between the two sets of descriptors, explains 23.3% of the variance using the old descriptors, but only 15% using the new set. As the number of significant principal components remains the same, this must mean that the new descriptors are adding variance to the data. Of the remaining descriptors, only PC2 obtained with the old set of descriptors (designated MEP+)[1] correlates moderately ($R^2 = 0.63$) with

Table 6. The Eleven Most Significant Principal Component Vectors Calculated Using the 34-Descriptor Set Described in the Text for the Entire Maybridge Dataset

descriptor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
μ	0.15	−0.13	−0.07	−0.21	0.21	−0.09	0.06	−0.57	−0.17
μ_D	0.11	−0.21	0.15	−0.14	0.20	−0.10	0.02	−0.52	−0.14
α	0.08	0.18	−0.36	−0.10	0.04	0.08	0.03	−0.01	−0.08
V_{max}	0.23	−0.16	−0.02	−0.19	−0.13	0.04	0.06	0.28	0.03
V_{min}	−0.02	−0.04	−0.01	0.43	−0.13	−0.16	0.10	−0.17	0.06
$\bar{V}_+$	0.25	−0.23	0.07	−0.06	−0.07	−0.03	0.07	0.00	0.01
$\bar{V}_-$	−0.09	0.19	−0.10	0.27	−0.27	−0.06	−0.08	−0.08	−0.02
σ^2_{tot}	0.13	−0.17	0.07	−0.39	0.18	0.09	0.05	0.24	0.00
ν	0.20	−0.10	0.02	0.10	−0.31	−0.16	−0.05	0.07	0.08
$\sigma^2_{tot}\nu$	0.23	−0.19	0.07	−0.23	−0.06	−0.02	0.01	0.25	0.05
MW	0.16	0.14	−0.34	−0.07	−0.06	0.04	0.15	−0.02	−0.03
vol	0.07	0.13	−0.39	−0.11	0.02	0.08	0.07	−0.03	0.00
A	0.08	0.13	−0.40	−0.10	0.01	0.06	0.06	−0.02	0.00
G	−0.10	−0.11	0.38	0.07	0.01	−0.01	0.00	−0.01	0.02
σ^2_{IE}	0.23	0.12	0.19	−0.10	−0.15	−0.07	−0.16	−0.02	0.19
IE_L^{max}	0.18	−0.12	−0.11	0.01	−0.30	−0.13	−0.05	−0.13	0.09
IE_L^{min}	−0.17	−0.23	−0.07	0.15	0.01	0.07	0.40	0.08	−0.21
$\overline{IE_L}$	−0.03	−0.33	−0.13	0.04	−0.03	−0.08	0.14	0.13	−0.05
ΔIE_L	0.23	0.14	0.00	−0.12	−0.16	−0.12	−0.37	−0.13	0.23
σ^2_{EA+}	0.10	−0.04	0.04	0.15	0.10	0.49	0.01	−0.13	0.34
σ^2_{EA-}	0.22	0.14	0.04	0.20	0.27	−0.16	−0.14	0.15	−0.22
σ^2_{EAtot}	0.23	0.13	0.04	0.21	0.28	−0.09	−0.14	0.13	−0.17
ν_{EA}	0.16	−0.07	0.03	0.18	0.04	0.52	0.05	−0.09	0.23
EA_L^{max}	0.29	−0.02	−0.04	0.20	0.10	0.12	−0.09	0.03	−0.15
EA_L^{min}	0.03	0.06	0.11	−0.10	−0.41	0.31	−0.03	−0.11	−0.40
$\overline{EA_L}$	0.29	0.07	0.07	0.14	−0.13	0.02	−0.03	−0.03	−0.30
ΔEA_L	0.19	−0.06	−0.11	0.22	0.38	−0.14	−0.05	0.11	0.17
δA^+_{EA}	0.23	−0.05	0.06	0.19	0.01	0.29	0.05	−0.04	−0.03
$\overline{\chi_L}$	0.21	−0.22	−0.05	0.15	−0.13	−0.05	0.09	0.08	−0.29
σ^2_α	0.12	0.20	0.18	−0.01	−0.03	−0.08	0.50	−0.01	0.07
α_L^{max}	0.13	0.29	0.17	−0.05	−0.01	−0.07	0.31	−0.01	0.06
α_L^{min}	−0.10	0.19	0.16	−0.12	0.05	0.22	−0.18	0.09	−0.31
$\overline{\alpha_L}$	0.04	0.33	0.18	−0.08	0.08	0.07	0.09	0.05	−0.14
$\Delta\alpha_L$	0.17	0.25	0.13	−0.02	−0.03	−0.14	0.38	−0.03	0.15
cumulative % variance explained	22.1	39.7	54.8	63.3	70.6	76.0	80.6	83.7	86.7
Eigenvalue	7.51	5.98	5.14	2.89	2.49	1.83	1.55	1.07	1.00

Table 7. Correlation Matrix Calculated for the Maybridge Database for the Descriptor Set Introduced in This Paper ("new descriptors") and that Used in Ref 1 ("old descriptors")

	old descriptors								
new descriptors	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
PC1	0.25	0.63	−0.41	0.07	−0.22	−0.02	0.31	0.02	−0.04
PC2	0.33	−0.57	0.10	0.01	−0.21	−0.21	0.41	0.03	−0.07
PC3	*−0.88*	0.07	0.02	−0.05	−0.12	−0.14	0.28	0.03	−0.04
PC4	−0.16	−0.47	−0.56	0.15	0.23	0.12	0.06	0.04	0.01
PC5	−0.04	0.09	0.56	0.34	−0.09	0.27	0.08	0.15	−0.02
PC6	0.08	0.00	0.21	0.11	0.05	−0.23	−0.07	−0.04	0.01
PC7	0.06	0.10	−0.02	0.24	0.19	−0.04	0.41	−0.21	0.17
PC8	0.05	0.09	0.13	−0.34	0.41	−0.56	0.19	−0.05	−0.02
PC9	−0.03	0.00	0.02	−0.22	0.31	0.10	0.03	−0.18	−0.14

PC1 obtained with the new descriptors, whereas the other original PC designated as being electrostatic in nature (PC3, designated MEP−)[1] correlates moderately (−0.56 and 0.56, respectively) with PCs 4 and 5 obtained with the new set of descriptors. PC8 obtained with the new descriptor set correlates only moderately (−0.56) with the dipolar descriptor (PC6) obtained with the old descriptors.

Comparison of the Descriptors for a QSPR. To test the effectivity of the new set of descriptors, we used our boiling-point dataset reported previously[36] to compare a model

Table 8. Descriptors Sets Used to Train the Two Boiling Point Models Based on the New Descriptor Set without Element-Specific Descriptors (Descriptors 1−10 Were Used for the First Model and 1−13 Were Used for the Second Model)

no.	symbol	descriptor
1	α	molecular electronic polarizability
2	MW	molecular weight
3	$\bar{V}_+$	mean of the positive values of the MEP
4	$\bar{V}_-$	mean of the negative values of the MEP
5	σ^2_{tot}	MEP total variance
6	$\sigma^2_{\mathrm{tot}}\nu$	product of the MEP total variance and the MEP balance parameter
7	$\mathrm{EA}_{\mathrm{L}}^{\mathrm{min}}$	minimum of the local electron affinity
8	$\overline{\mathrm{EA}_{\mathrm{L}}}$	mean value of the local electron affinity
9	$\overline{\chi_{\mathrm{L}}}$	mean value of the local electronegativity
10	$\overline{\alpha_{\mathrm{L}}}$	mean value of the local polarizability
11	$\alpha^2_{\mathrm{EAtot}}$	sum of the positive and negative variances in the local electron affinity
12	$\sigma^2_{\mathrm{EA-}}$	variance in the local electron affinity for all negative values
13	σ^2_{α}	variance in the local polarizability

constructed with the 18 descriptors used previously, which include element-specific sums of MEP-derived charges,[36] with the new set of exclusively whole-molecule descriptors. Note that the dataset is fairly noisy,[36] so that the advantage we hope to gain from the new descriptors is not necessarily an improvement in the overall accuracy of the model, which would bring the danger of overfitting to poor experimental datapoints, but rather to provide a more robust and general model by, for instance, requiring less descriptors. Descriptors for the new model were first selected by using the boiling point as the lead variable for recursive partitioning[37] based on the set of 34 descriptors. The first selection run resulted in 10 descriptors (shown in Table 8, descriptors 1−10) six of which are based on the MEP or the geometry and four on the new properties introduced in this work. Note that, in contrast to the results found by Oprea et al.,[38] we did not find any significant correlation between significant PCs of the descriptors and the boiling point.

Table 9. Results Obtained for the Boiling-Point Model Using the 18 Descriptors Reported Previously[34] and the 10- and 13-Descriptor Sets Given in Table 8[a]

descriptor:	ref 34 (18)	Table 8 (1−10)	Table 8 (1−13)
architecture:	18:10:1	10:9:1	13:10:1
training set			
MUE	17.3	20.1	18.0
largest error	118.4	118.7	128.3
cross-validation			
MUE	20.0	22.8	20.2
largest error	158.8	229.8	170.8
validation set			
MUE	20.6	21.3	21.6
largest error	156.8	141.9	184.3

[a] The training set consisted of 5455 compounds divided into 10 sets. An independent validation set of 608 compounds was used. All errors are given in °C. MUE = mean unsigned error.

The model using the previous set of 18 descriptors was trained using an 18:10:1 feedforward neural net architecture within our committee-machine approach[39] using 10 separate networks. The model using the 10 descriptors from the new set used the same techniques with a 10:9:1 network architecture. The new model therefore uses roughly half as many weights as the one reported previously.[36] Despite this, the results (shown in Figure 3(a) and Table 9) are only slightly worse than those obtained with the old descriptor set. Adding another three descriptors to the new set (11−13 in Table 8) to give a 13:10:1 network architecture improves the performance of the net with the new descriptors considerably, giving almost the same mean unsigned errors both for the training set, cross-validation on the training set[39] and for the validation set. The results are shown in Figure 3b and Table 9.

Thus, the new descriptors set is able to describe the chemical diversity previously treated using the element-

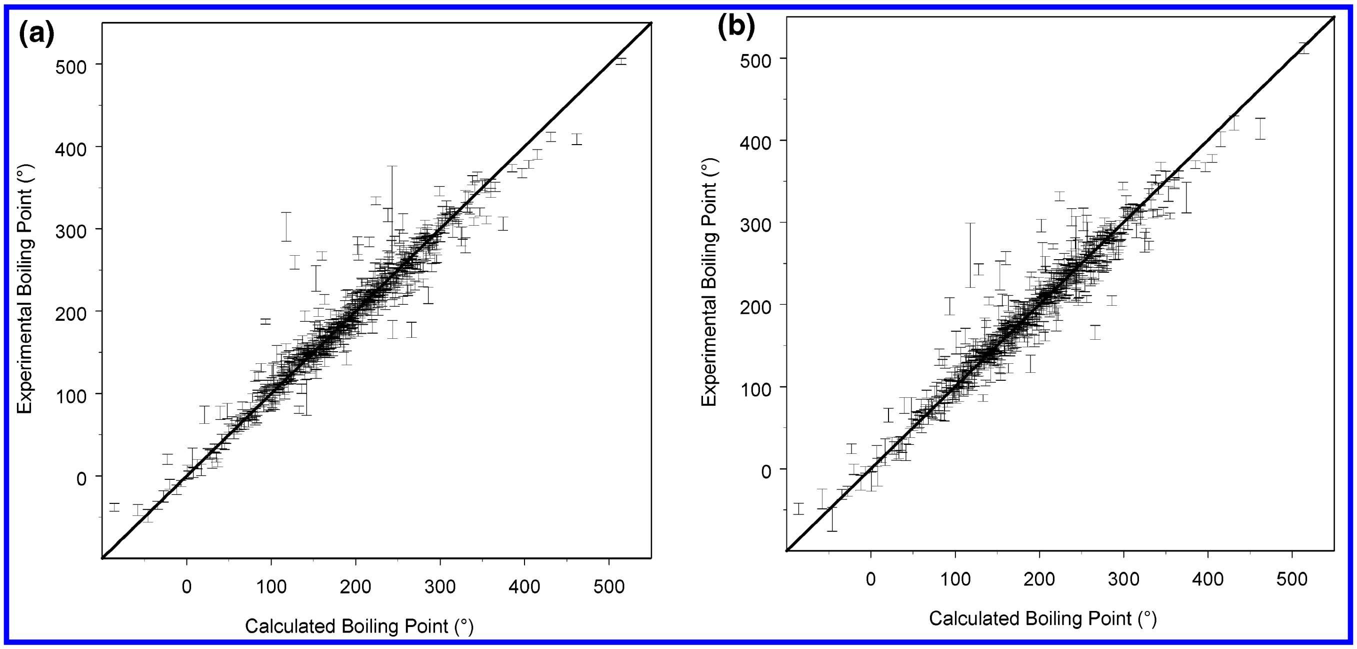

Figure 3. Scatter plots of the calculated boiling points against the experimental values for (a) the 10:9:1 net and (b) the 13:10:1 net. The results are shown for the validation set of 606 compounds. The error bars indicate ± one standard deviation of the results of the 10 individual nets for each compound.

specific descriptors. The reduction in the number of descriptors necessary to achieve essentially the same performance for the model is gratifying. The descriptors given in Table 8 also confirm some conclusions suggested by the principal component analysis. Of the seven descriptors listed in Table 8 that are derived from the new local properties, four (EA_L^{min}, $\overline{EA_L}$, σ^2_{EAtot} and σ^2_{EA-}) are derived directly from the local electron affinity and one ($\overline{\chi_L}$) indirectly. The other two ($\overline{\alpha_L}$ and σ^2_{α}) are derived from the local polarizability. The four MEP-derived descriptors, the molecular electronic polarizability and the molecular weight are familiar descriptors from almost all of our published QSPR-models.

SUMMARY

We have presented new descriptors based on the values of the local ionization energy, the local electron affinity, and the local polarizability. These descriptors are intended to replace the element-specific descriptors and feature counts used in our previous models in order to make our descriptor set independent of properties that are not derived from the entire molecule. In this way, we are able to eliminate reference to the 2D-structure or individual groups in the molecules from our descriptors. This should result in (a) an increased likelihood of scaffold hopping in QSAR or virtual screening applications and (b) more robust and general QSPR models. Although the current descriptors are not suited for activity prediction because they do not address specific binding sites in molecules, the local properties can be used analogously to our MEP-based alignment-free QSAR technique.[40] We have, for instance, been able to define a local hydration energy on the molecular surface[41] that is an essential component of future surface-based scoring functions. Neither the values of the surface properties on the SES of drug molecules or the descriptors derived from them show strong correlations among themselves or with the MEP or descriptors derived from it. A new set of 34 molecular descriptors based on whole-molecule properties, the MEP, local ionization energy, local electron affinity, and local polarizability has been proposed that we believe treats size, Coulomb, donor/acceptor, and van der Waals interactions well. Currently, the molecular shape is represented only by the globularity, so our descriptor set would probably benefit from more detailed shape descriptors.[42]

ACKNOWLEDGMENT

We thank Pfizer Ltd. for financial support.

REFERENCES AND NOTES

(1) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Propertes and Drug-Likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(3) Oprea, T. I.; Gottfries, J. ChemGPS: A Chemical Space Navigation Tool. In *Rational Approaches to Drug Design: 13th European Symposium on QSAR*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, 2001; pp 437−446.

(4) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(5) Oprea, T. I. On the Information Content of 2D and 3D Descriptors for QSAR. *J. Braz. Chem. Soc.* **2002**, *13*, 811−815.

(6) Beck, B.; Clark, T.; Glen, R. C. A Detailed Study of VESPA Electrostatic Potential-Derived Atomic Charges. *J. Mol. Model.* **1995**, *1*, 176−187.

(7) Maybridge Chemicals Company Ltd.: Trevillet, Tintangel, Cornwall PL34 OHW, England.

(8) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(9) Pascual-Ahuir, J. L.; Silla, E.; Tuñon, I. GEPOL: An improved Description of Molecular Surfaces III. A New Algorithm for the Computation of a Solvent-Excluded Surface. *J. Comput. Chem.* **1994**, *15*, 1127−1138.

(10) Murray, J. S.; Politzer, P. Statistical Analysis of the Molecular Surface Electrostatic Potential: An Approach to Describing Noncovalent Interaction in Condensed Phases. *J. Mol. Struct (THEOCHEM)* **1998**, *425*, 107−114. Murray, J. S.; Lane, P.; Brinck, T.; Paulsen, K.; Grince, M. E.; Politzer, P. Relationships of Critical Constants and Boiling Points to Computed Molecular Surface Properties. *J. Phys. Chem.* **1993**, *97*, 9369−9373.

(11) Pearson, R. G. Absolute electronegativity and hardness: application to inorganic chemistry. *Inorg. Chem.* **1988**, *27*, 734−740.

(12) Sjoberg P.; Murray J. S.; Brinck T.; Politzer P. A. Average local ionization energies on the molecular surfaces of aromatic systems as guides to chemical reactivity. *Can. J. Chem.* **1990**, *68*, 1440−1443.

(13) Politzer, P.; Murray, J. S.; Grice, M. E.; Brinck, T.; Ranganathan, S. Radial behavior of the average local ionization energies of atoms. *J. Chem. Phys.* **1991**, *95*, 6699−6704.

(14) Politzer, P.; Murray, J. S.; Concha, M. C. The complementary roles of molecular surface electrostatic potentials and average local ionization energies with respect to electrophilic processes. *Int. J. Quantum Chem.* **2002**, *88*, 19−27, and references therein.

(15) Hussein, W.; Walker, C. G.; Peralta-Inga, Z.; Murray, J. S. Computed electrostatic potentials and average local ionization energies on the molecular surfaces of some tetracyclines. *Int. J. Quantum Chem.* **2001**, *82*, 160−169.

(16) Murray, J. S.; Abu-Awwad, F.; Politzer, P. Characterization of aromatic hydrocarbons by means of average local ionization energies on their molecular surfaces. *J. Mol. Struct. (THEOCHEM)* **2000**, *501*−*502*, 241−250.

(17) Ehresmann, B.; Martin, B.; Horn, A. H. C.; Clark, T. Local molecular properties and their use in predicting reactivity *J. Mol. Model.* **2003**, *9*, 342−347.

(18) Mulliken, R. S. New electroaffinity scale; together with data on valence states and on valence ionization potentials and electron affinities. *J. Chem. Phys.* **1934**, *2*, 782−793.

(19) Martin, B.; Gedeck, P.; Clark, T. An Additive NDDO−Based Atomic Polarizability Model. *Int. J. Quantum Chem.* **2000**, *77*, 473−497.

(20) Rinaldi, D.; Rivail, J. L. Calculation of molecular electronic polarizabilities. Comparison of different methods. *Theor. Chim. Acta* **1974**, *32*, 243−251. Rinaldi, D.; Rivail, J. L. Molecular polarisability and dielectric effect of medium in the liquid phase. Theoretical study of the water molecule and its dimers. *Theor. Chim. Acta* **1973**, *32*, 57−70.

(21) http://www.rxusa.com/top100+.htm.

(22) See Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley-Interscience: New York, 2002, and references therein.

(23) Murray, J. S.; Politzer, P. Statistical analysis of the molecular surface electrostatic potential: an approach to describing noncovalent interactions in condensed phases. *J. Mol. Struct. (THEOCHEM)* **1998**, *425*, 107−114.

(24) Murray, J.S.; Ranganathan, S.; Politzer, P. Correlations between the solvent hydrogen bond acceptor parameter β and the calculated molecular electrostatic potential. *J. Org. Chem.* **1991**, *56*, 3734−3737.

(25) Politzer, P.; Lane, P.; Murray, J. S.; Brinck, T. Investigation of Relationships between Solute Molecule Surface Electrostatic Potentials and Solubilities in Supercritical Fluids. *J. Phys. Chem.* **1992**, *96*, 7938−7943.

(26) Murray, J. S.; Lane, P.; Brinck, T.; Paulsen, K.; Grice, M. E.; Politzer, P. Relationships of Critical Constants and Boiling Points to Computed Molecular Surface Properties. *J. Phys. Chem.* **1993**, *97*, 9369−9373.

(27) Brinck, T.; Murray, J. S.; Politzer, P. Quantative determination of the total local polarity (charge separation) in molecules. *Mol. Phys.* **1992**, *76*, 609−617.

(28) Beck, B.; Horn, A. H. C.; Carpenter, J. E.; Clark, T. Enhanced 3D-Databases: A Fully Electrostatic Database of AM1-Optimized Structures. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1214−1217.

(29) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909. Holder, A. J. AM1. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, 1998; pp 8−11.

(30) Clark, T.; Alex, A.; Beck, B.; Burkhardt, F.; Chandrasekhar, J.; Gedeck, P.; Horn, A. H. C.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. VAMP 8.1, Erlangen, 2002.

(31) Mu, L.; Drago, R. S.; Richardson, D. E. A model based QSPR analysis of the unified nonspecific solvent polarity scale. *J. Chem. Soc., Perkin Trans. 2* **1998**, 159−167.

(32) Schürer, G.; Gedeck, P.; Gottschalk, M.; Clark, T. Accurate Parametrized Variational Calculations of the Molecular Electronic Polarizability by NDDO−Based Methods. *Int. J. Quantum Chem.* **1999**, *75*, 17−31.

(33) Meyer, A. Y. The Size of Molecules. *Chem. Soc. Rev.* **1986**, *15*, 449−475.

(34) Kaiser, H. F. *Educ. Psychol. Meas.* **1960**, *20, 141−151.*

(35) *Cattell, R. B. Multivariate Behav. Res. B* **1966**, *1*, 245−252.

(36) Chalk, A. J.; Beck B.; Clark, T.A. Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457−462.

(37) Zhang, H.; Singer, B. *Recursive Partitioning in the Health Sciences*; Springer-Verlag: Telos, 1999.

(38) Oprea, T. I.; Zamora, I.; Ungell, A.-L. Pharmacokinetically Based Mapping Device for Chemical Space Navigation. *J. Comb. Chem.* **2002**, *4*, 258−266.

(39) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and logP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.

(40) Clark, T. Quantum Cheminformatics: An Oxymoron? (Part 2). In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, 2001; pp 29−40

(41) Ehresmann, B.; Clark, T. unpublished data.

(42) Mezey, P. G. *Shape in Chemistry*; VCH Publishers: New York, 1993.

CI034215E