Descriptors, Physical Properties, and Drug-Likeness

Matthias Brüstle,[†] Bernd Beck,[†,‡] Torsten Schindler,[†] William King,[†] Timothy Mitchell,[§] and Timothy Clark*,[†]

Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25,
D-91052 Erlangen, Germany, and Millennium Pharmaceuticals Ltd., Granta Park, Great Abington,
Cambridge CB1 6ET, United Kingdom

Received August 23, 2001

We have investigated techniques for distinguishing between drugs and nondrugs using a set of molecular descriptors derived from semiempirical molecular orbital (AM1) calculations. The "drug" data set of 2105 compounds was derived from the World Drug Index (WDI) using a procedure designed to select real drugs. The "nondrug" data set was the Maybridge database. We have first investigated the dimensionality of physical properties space based on a set of 26 descriptors that we have used successfully to build absorption, distribution, metabolism, and excretion-related quantitative structure−property relationship models. We discuss the general nature of the descriptors for physical property space and the ability of these descriptors to distinguish between drugs and nondrugs. The third most significant principal component of this set of descriptors serves as a useful numerical index of drug-likeness, but no others are able to distinguish between drugs and nondrugs. We have therefore extended our set of descriptors to a total of 66 and have used recursive partitioning to identify the descriptors that can distinguish between drugs and nondrugs. This procedure pointed to two of the descriptors that play an important role in the principal component found above and one more from the set of 40 extra descriptors. These three descriptors were then used to train a Kohonen artificial neural net for the entire Maybridge data set. Projecting the drug database onto the map obtained resulted in a clear distinction not only between drugs and nondrugs but also, for instance, between hormones and other drugs. Projection of 42 131 compounds from the WDI onto the Kohonen map also revealed pronounced clustering in the regions of the map assigned as druglike.

Introduction

Recently, the emphasis in computational drug design has been extended from the traditional quantitative structure−activity relationship (QSAR) techniques used to find biologically active molecules to more quantitative structure−property relationship (QSPR)-oriented estimates of the absorption, distribution, metabolism, and excretion (ADME) properties of molecules. As part of this shift in emphasis, several groups have attempted to define the "drug-likeness" of molecules following the pioneering "rule of five" work by Lipinski.[1] The most common approach to this problem has been to use some type of molecular descriptors linked with a pattern recognition or interpolation technique, such as neural nets, to distinguish between a data set of drugs and one of nondrugs.[2−4] The problem often encountered is that the data sets are not orthogonal. There is no "nondrug" database. The failed molecules, those shown not to be drugs, are often not recorded or catalogued, and even so, there must be a significant proportion of molecules in a nondrug database that would make good drugs if they were biologically active. The only solution to this problem so far has been to assume that the number of drugs in a generic database is very small. Furthermore, the World Drug Index (WDI), a database often used as the "drug set", contains many chemicals that are not fully developed drugs. These may not have the required ADME properties necessary to be a drug. This problem was, however, addressed by Lipinski,[1] who used stringent selection criteria for his drug data set.

We now report a systematic investigation of possible approaches for distinguishing drugs from nondrugs. We have used the quantum mechanically derived descriptors that we have shown to be very successful in describing physical and partition properties of molecules[5] directly, rather than using the physical properties themselves. We have also investigated three different discrimination techniques.

Materials and Methods

Drugs Data Set. A drugs data set was selected from the WDI[6] as follows. The WDI 1997 contains 51 596 compounds. Of these, 7570 have been assigned a United States Adopted Name (USAN)[7] and 6307 have been assigned an International Nonproprietary Name (INN).[8] Combining these gives 8323 unique compounds, of which 3515 have an entry in the "Indication and Usage" (IU) field. Compounds with entries in the PT (activity) field were then excluded as follows: repellent, surfactant, sweetener, food-additive, radio protective, solubilizer, lubricant, synergistic, tonic, topical, skin, dermatological, vulneraries, flavor, insecticide, antiseptic, vitamin, chelator, cytostatic, emmollier, preservative, anaesthetic, diagnostic, dietary supplement, radiopaque, dye, emulsifier, laxative, radiosensitizer, rodenticide, solvent. Finally, 263 charged moieties were removed (because our treatment is at the moment not appropriate for charged moieties) and the remain-

* To whom correspondence should be addressed. Tel.: (+49)9131-8522948. Fax: (+49)9131-8526565. E-mail: clark@chemie.uni-erlangen.de.

[†] Friedrich-Alexander-Universität Erlangen-Nürnberg.

[‡] Current address: Boehringer Ingelheim Pharma KG, 88397 Biberach/Riss, Germany.

[§] Millennium Pharmaceuticals Ltd..

10.1021/jm011027b CCC: $22.00
Published on Web 07/03/2002

Table 1. Frequently Occurring Descriptors in QSPR Models[a]

descriptor	log*P*	boiling point	aqueous solubility	vapor pressure	vapor pressure (T-dependent)
ref	18, 19	21	22	19	20
molecular polarizability (α)	X	X	X	X	X
molecular surface area	X	X	X	X	
globularity	X	X			X
mean positive MEP	X	X	X	X	X
mean negative MEP	X	X		X	X
total variance (σ_{tot}^2)	X	X	X		
balance parameter (ν)	X	X	X		
positive variance (σ_+^2)		X	X		X
negative variance (σ_-^2)		X	X		
molecular weight		X	X	X	X
sum of MEP-derived charges on nitrogens	X	X	X	X	
sum of MEP-derived charges on oxygens	X	X	X	X	
sum of MEP-derived charges on phosphorus	X				
sum of MEP-derived charges on sulfur	X			X	

[a] Boxes marked with an X indicate that the descriptor is used in the given model. Those that are blank indicate that we would expect the descriptor to be used if the training data set contained larger numbers of phosphorus or sulfur compounds.

ing 2105 compounds were used as the "oral drug" data set. This procedure is slightly more extensive than the similar one used by Lipinski[1] but should give a very well-defined drug data set.

Nondrug Data Set. We also require a nondrug data set that is ideally orthogonal to the drug data set. It is commonly assumed that general chemical databases such as Maybridge[9] contain no drugs, although this is clearly not correct. We have, however, used the Maybridge database in its entirety and have assumed that it contains a significant proportion of compounds that would be suitable as orally available drugs. In our final approach, this data set plays the role of representing compounds in general (that is, chemical space) in order to set up a map in which drugs are localized in one or more areas. This approach is not unlike the "chemical GPS" technique used by Oprea et al.[10]

Processing Protocol. The compounds from the two data sets were converted from two-dimensional (2D) to single conformation three-dimensional (3D) structures using CORINA,[11] and these structures were used as starting geometries for AM1[12] geometry optimizations using VAMP 7.0.[13] This procedure corresponds exactly to that used earlier to optimize the structures of the Maybridge database.[14] The molecular electrostatics were stored using the natural atomic orbital-point charge (NAO-PC)[15,16] model, and these data were used to generate the descriptors with PROPGEN 1.0.[17]

Results and Discussion

Physical Properties and Descriptors. Lipinksi[1] has pointed out that the physical property space relevant to ADME property prediction is low-dimensional in comparison to, for instance, the descriptor space needed to estimate biological activity. Furthermore, extensive work by Murray and Politzer[18,19] has shown that descriptors based on the electrostatic properties at the surface of the molecule provide a good description of physical properties. We[20-24] have used such descriptors in conjunction with semiempirical (AM1[12] and PM3[25]) molecular orbital (MO) theory to develop a series of QSPR models for physical properties such as the logarithm of the octanol/water partition coefficient, log*P*,[20] the vapor pressure at 25° and at various temperatures,[21,22] the normal boiling point,[23] and aqueous solubility at 20 and 25°.[24] Although these models were developed independently of each other, a common subset of 14 descriptors appears in at least two of the models, as shown in Table 1.

We suggest that these descriptors provide a good description of physical property space, so that they and others that are similar can be used directly in defining drug fitness, rather than using measured physical properties directly or by using QSPR estimates of physical properties. The dimensionality of the information contained in our full set of 26 descriptors routinely used to build QSPR models can be tested by calculating their principal components (PCs)[26] for a database of general chemicals, such as the Maybridge database processed previously.[14] In this case, the fact that Maybridge contains a subset of molecules that would make good drugs is actually an advantage in investigating the dimensionality of physical property space.

Our Maybridge single conformation data set contains 52 712 structures calculated with AM1 semiempirical MO theory[12] for which the descriptors given in Table 1 have been calculated. In addition to the 14 descriptors shown in Table 1, a further 12 related descriptors that we have used more recently than the published work were calculated to give a total of 26. These descriptors are defined in Table 2.

We have shown that the inclusion of correlated descriptors can be important in property prediction with neural nets.[20] This is because there is valuable information in the way these descriptors are not correlated.

Table 3 shows relatively few strong correlations. Interestingly, the molecular electronic polarizability, which we can now calculate accurately using the parametrized variational treatment[29,30] and which appears as a major descriptor in all of our current QSPR models, correlates very strongly with both the dipolar density[27] and the "size" descriptors molecular weight and volume. The other very strong correlation occurs between the molecular weight and the volume, suggesting that there is some redundancy in the size descriptors. The "electrostatic" descriptors are, however, pleasingly uncorrelated, suggesting that they provide a diverse description of the molecular electrostatic properties and, hence, to a large extent of the intermolecular interactions.

The total variance explained by the PCs calculated for these 26 descriptors is plotted against the number of PCs in Figure 1. About 90% of the variance is explained by the first 12 PCs, confirming the relatively low dimensionality of at least the space defined by these descriptors and by inference of physical property space.

Table 2. List of All Available Descriptors[a]

no.	acronym	description	ADME 26	FIRM 3	ref
1	μ	total molecular dipole moment	X		
2	dipden	dipolar density	X		27
3	α	total polarizability (original variational)			28
4	m0pol	total polarizability (parametrized model 0)	X		29
5	m2pol	total polarizability (parametrized model 2)			30
Sums of the Electrostatic Potential-Derived Atomic Charges on					
6	QsumH	H atoms	X	X	31
7	QsumN	N atoms	X		31
8	QsumO	O atoms	X		31
9	QsumP	P atoms	X		31
10	QsumS	S atoms	X		31
11	Qsumhal	halogen atoms			31
12	QsumF	F atoms	X		31
13	QsumCl	Cl atoms	X		31
14	QsumBr	Br atoms	X		31
15	QsumI	I atoms	X		31
16	npos	no. of triangles on the surface with a + MEP			18, 19
17	nneg	no. of triangles on the surface with a − MEP			18, 19
18	ESPmax	max MEP	X		18, 19
19	ESPmin	min MEP	X	X	18, 19
20	midpos	mean + MEP	X		18, 19
21	midneg	mean − MEP	X		
22	allmeanesp	total mean MEP			18, 19
23	midpos2	midpos^2			18, 19
24	midneg2	midneg^2			18, 19
25	σ^2_{tot}	total variance of the MEP	X		18, 19
26	ν	balance param	X		18, 19
27	$\sigma^2_{tot} \times \nu$	total variance * balance	X		18, 19
28	locpol	local polarity			18, 19
29	covHBac	covalent H bond acidity		X	18, 19
30	covHBbas	covalent H bond basicity			32
31	esHBac	electrostatic H bond acidity			32
32	esHBbas	electrostatic H bond basicity			32
33	nAcc	no. of H bond acceptor groups	X		32
34	nDon	no. of H bond donor groups	X		
35	nAryl	no. of aryl groups	X		
36	MW	mol wt	X		
37	Vol	molecular volume	X		33
38	Totsurface	total molecular surface area	X		33
39	Glob	globularity	X		34
Sum of E-States Based on QM-Calculated Bond Orders					
40	EstateN	N atoms			35
41	EstateO	O atoms			35
42	EstateP	P atoms			35
43	EstateS	S atoms			35
44	Estatehal	halogen atoms			35
45	EstateF	F atoms			35
46	EstateCl	Cl atoms			35
47	EstateBr	Br atoms			35
48	EstateI	I atoms			35
Sum of E-States Based on QM-Calculated Distance on					
49	Estate2N	N atoms			35
50	Estate2O	O atoms			35
51	Estate2P	P atoms			35
52	Estate2S	S atoms			35
53	Estate2hal	halogen atoms			35
54	Estate2F	F atoms			35
55	Estate2Cl	Cl atoms			35
56	Estate2Br	Br atoms			35
57	Estate2I	I atoms			35
Sum of Classical Kier and Hall E-States on					
58	EstateorgN	N atoms			36
59	EstateorgO	O atoms			36
60	EstateorgP	P atoms			36
61	EstateorgS	S atoms			36
62	Estateorghal	halogen atoms			36
63	EstateorgF	F atoms			36
64	EstateorgCl	Cl atoms			36
65	EstateorgBr	Br atoms			36
66	EstateorgI	I atoms			36

[a] The descriptors contained in the original 26 descriptor ADME set are marked under "ADME 26", and those selected by the FIRM analysis and used in the Kohonen net are marked under "FIRM 3".

Figure 2 shows a plot of the Eigenvalues of the PCs[26] against the number of the PC. Two tests have been proposed to determine the number of significant PCs for a given set of descriptors and data. The first, the Kaiser−Guttmann criterion,[38] is simply that all PCs with an Eigenvalue larger than one are significant. This test suggests that the first eight PCs are significant. The Scree test[39] proposes that Eigenvalue plots such as Figure 2 should show a kink between the significant and the less significant PCs. Figure 2 shows two such kinks, one at PC number five and one at seven. As the latter agrees relatively well with the Kaiser−Guttmann criterion, we conclude that the space described by our 26 descriptors for the Maybridge database is 7−8-dimensional. Our experience with QSPR models suggests that we can extrapolate this conclusion to give a rough idea of the dimensionality of physical property space in the context of ADME properties. What, however, is the nature of the PCs?

Nature of the PCs. The PCs[26] of the 26 descriptors calculated for the Maybridge database were calculated. The coefficients of the first nine PCs are shown in Table 4.

The first PC, which explains 23% of the total variance, consists mainly of the size descriptors such as the polarizability, molecular weight, volume, surface, and globularity. We interpret this factor as describing primarily the size and shape of the molecule.

The second PC, in contrast, consists almost entirely of the "Murray/Politzer" descriptors maximum and mean positive and negative electrostatic potentials, the total variance, σ_{tot}^2, and the product of the balance parameter and the total variance, $\sigma^2 \cdot \nu$. This PC, which accounts for 18% of the total variance, constitutes a general electrostatic description of the positive areas of the molecule.

The third PC consists of the total charge on fluorines, the minimum of the molecular electrostatic potential, the mean negative electrostatic potential, and the balance parameter, ν. The occurrence of the total charge on fluorines in this descriptor is a little puzzling, especially as it does not correlate with any of the others, but otherwise, this factor can be interpreted as the equivalent of PC2 for negative areas of the electrostatic potential. We have therefore labeled PC2 and PC3 "MEP+" and "MEP−" and suggest that the two together describe the surface electrostatics of the molecule quite effectively. Omitting the fluorine charge descriptor from the descriptors and recalculating the PCs result in an increase of the coefficients for the minimum MEP (to −0.475) and the sums of charges on hydrogens and nitrogens (to 0.341 and −0.254, respectively) but to a slight decrease in that for the balance parameter (to −0.438).

PC number four, which accounts for 8% of the variance, has strong contributions from the total charges on nitrogen and the H bond donor count. These two descriptors are relatively strongly negatively correlated and have opposite signs in the PC, so that we can

Table 3. Correlation Matrix Obtained for the 26 Descriptors Calculated for the Maybridge Database[a]

Section 1

descriptor	acronym	m	dipden	m0pol	QsumH	QsumN	QsumO	QsumP	QsumS	QsumF	QsumCl	QsumBr	QsumI	ESPmax	ESPmin
total dipole	μ	1.00													
total dipole/volume	Dipden	**0.79**	1.00												
mean polarizability	m0pol	0.15	−0.37	1.00											
total H charges	QsumH	0.04	−0.24	0.48	1.00										
total N charges	QsumN	0.06	0.13	−0.14	−0.25	1.00									
total O charges	QsumO	−0.25	−0.05	−0.25	−0.21	−0.22	1.00								
total P charges	QsumP	0.04	0.01	0.06	0.06	0.02	−0.09	1.00							
total S charges	QsumS	0.21	0.08	0.14	0.00	−0.05	−**0.62**	−0.06	1.00						
total F charges	QsumF	−0.09	0.02	−0.05	0.09	0.06	−0.02	0.01	−0.06	1.00					
total Cl charges	QsumCl	−0.03	−0.12	0.19	−0.21	−0.07	0.03	−0.01	0.02	0.05	1.00				
total Br charges	QsumBr	−0.03	−0.02	0.00	−0.13	0.06	0.04	0.00	−0.02	0.04	−0.05	1.00			
total I charges	QsumI	−0.03	−0.01	0.01	−0.09	0.02	0.01	0.00	0.00	0.02	−0.02	−0.01	1.00		
max MEP	ESPmax	0.34	0.27	0.04	0.00	−0.12	−0.44	0.16	**0.56**	−0.20	0.00	0.00	0.00	1.00	
min MEP	ESPmin	−0.13	−0.06	−0.17	−0.18	0.34	−0.05	−0.24	−0.13	−0.17	0.03	0.01	−0.06	−0.10	1.00
mean + MEP	midpos	0.42	**0.50**	−0.24	−0.18	0.02	−0.26	0.05	0.23	−0.21	−0.04	0.06	0.00	**0.67**	0.14
Mean − MEP	midneg	−0.37	−0.49	0.19	−0.03	0.00	0.41	−0.08	−0.29	−0.30	0.17	0.01	0.00	−0.30	0.36
total variance	σ_{tot}^2	0.42	0.46	−0.11	0.05	−0.17	−0.37	0.24	0.49	0.15	−0.10	0.00	0.01	**0.59**	−**0.52**
balance param	$\boldsymbol{\nu}$	0.11	0.14	−0.06	−0.17	−0.05	−0.04	−0.10	0.01	−0.37	0.08	−0.01	−0.02	**0.52**	0.27
variance * balance	$\sigma_{tot}^2 \times \nu$	0.37	0.43	−0.12	−0.07	−0.17	−0.31	0.06	0.39	−0.09	−0.03	−0.01	0.00	**0.78**	−0.22
no. of acceptor dipoles	NAcc	0.13	−0.08	0.35	0.26	−0.22	−0.34	−0.02	−0.14	0.01	−0.05	0.05	−0.01	0.10	−0.06
no. of donor dipoles	NDon	0.00	0.02	−0.06	0.41	−0.48	−0.02	−0.02	−0.07	0.01	−0.02	−0.04	0.00	0.26	−0.13
no. of aryl rings	NAryl	0.10	−0.22	0.66	0.02	−0.14	0.02	0.02	0.09	−0.11	0.14	−0.01	−0.01	0.04	−0.07
mol wt	MW	0.22	−0.29	**0.90**	0.33	−0.12	−0.38	0.06	0.25	−0.32	0.26	0.13	0.09	0.24	−0.07
volume	Vol	0.19	−0.35	**0.96**	**0.57**	−0.12	−0.35	0.08	0.17	−0.17	0.15	−0.03	−0.02	0.10	−0.11
total surface	Totsurface	0.19	−0.34	**0.95**	**0.55**	−0.15	−0.34	0.07	0.16	−0.19	0.15	−0.03	−0.02	0.11	−0.11
globularity	Glob	−0.19	0.32	−**0.87**	−0.44	0.21	0.28	−0.03	−0.12	0.22	−0.15	0.05	0.03	−0.14	0.12

Section 2

descriptor	acronym	midpos	modneg	σ_{tot}^2	ν	$\sigma_{tot}^2 \times \nu$	nAcc	nDon	nAryl	MW	Vol	Totsurface	Glob
mean + MEP	midpos	1.00											
mean − MEP	midneg	−0.32	1.00										
total variance	σ_{tot}^2	0.44	−**0.71**	1.00									
balance param	ν	0.44	0.10	−0.05	1.00								
variance * balance	$\sigma_{tot}^2 \times \nu$	**0.60**	−0.48	**0.70**	**0.62**	1.00							
no. of acceptor dipoles	NAcc	0.15	−0.21	0.05	0.13	0.14	1.00						
no. of donor dipoles	NDon	0.17	−0.14	0.20	0.32	0.37	0.21	1.00					
no. of aryl rings	NAryl	−0.07	0.26	−0.17	0.08	−0.08	0.20	−0.13	1.00				
mol wt	MW	0.03	0.17	−0.05	0.11	0.03	0.42	−0.04	**0.60**	1.00			
volume	Vol	−0.13	0.18	−0.08	−0.06	−0.11	0.38	−0.04	**0.56**	**0.92**	1.00		
total surface	Totsurface	−0.12	0.18	−0.08	−0.03	−0.09	0.41	−0.02	**0.57**	**0.93**	**0.99**	1.00	
globularity	Glob	0.08	−0.19	0.08	−0.05	0.03	−0.45	0.00	−**0.59**	−**0.86**	−**0.89**	−**0.93**	1.00

[a] Positive and negative correlation coefficients greater than or equal to 0.5 are shown in boldface.

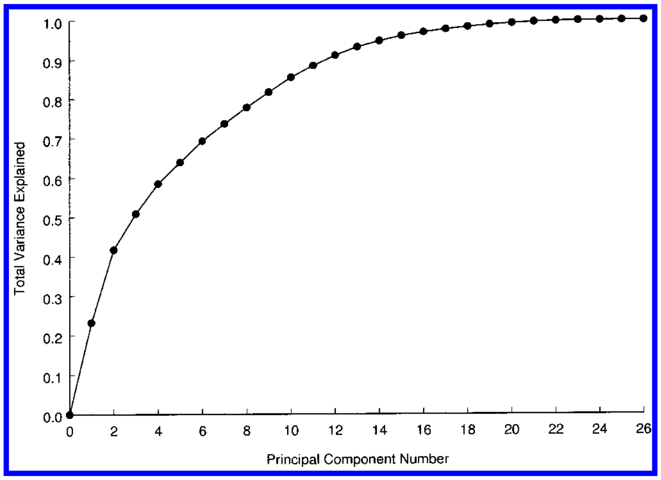


Figure 1. Total variance explained by the PCs of the 26 descriptors calculated for the Maybridge database.

interpret this factor as describing the H bond donor ability of the molecule.

Although PC4 consists of contributions from five different descriptors, they are all related to H bond acceptor properties, so that this factor can be interpreted as being the hydrogen bond acceptor equivalent of the hydrogen bond donor PC number 5.

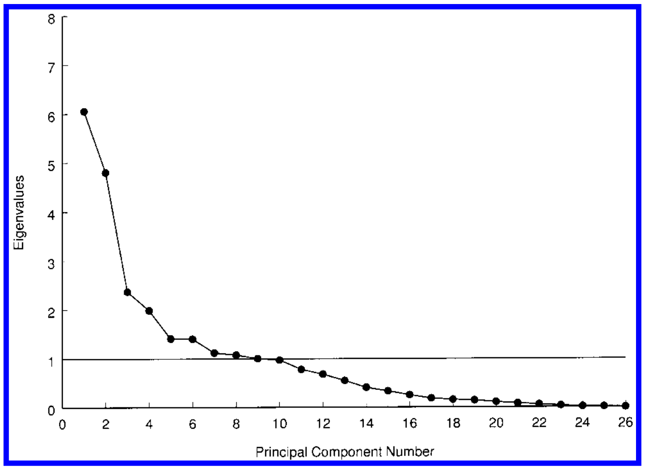


Figure 2. Eigenvalues of the PCs of the 26 descriptors calculated for the Maybridge database.

PC6, which explains about 5% of the total variance, consists of large contributions from the "dipole" descriptors (dipole moment and dipolar density) and from the total charge on sulfurs. As with the fluorine descriptor in PC3, the contribution from the total charge on sulfurs is difficult to explain, but otherwise, this descriptor can be interpreted as a simple dipolar (polarity) factor.

Table 4. Nine Most Significant PCs of the 26 Descriptors for the Maybridge Database[a]

descriptor	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
dipole moment	−0.0629	−0.2777	−0.0130	−0.2250	−0.1288	**−0.4941**	−0.1672	0.0834	−0.0442
dipolar density	0.1569	−0.2783	−0.0186	−0.1734	−0.1244	**−0.4621**	−0.1216	0.0864	−0.0387
polarizability	**−0.3880**	0.0448	0.0528	−0.0620	−0.0710	−0.0355	0.0144	0.0439	0.0011
Sums of the Potential Derived Charges on									
H	−0.2114	−0.0013	0.2837	0.2392	**0.3416**	−0.0785	−0.2390	−0.1017	−0.0131
N	0.0881	0.0530	−0.1263	**0.5018**	0.2597	−0.0408	0.0334	−0.1274	0.0271
O	0.1443	0.2353	−0.0507	0.2575	**−0.4222**	−0.2452	−0.0338	−0.0516	−0.0154
P	−0.0295	−0.0626	0.1507	−0.0395	−0.2080	0.0526	0.0032	**−0.6601**	0.2312
S	−0.0815	−0.2428	0.0246	−0.2338	0.0139	**0.4947**	−0.1613	0.1314	−0.0516
F	0.0812	0.0295	**0.3908**	−0.0676	0.0536	0.0245	0.2684	**0.3675**	0.1342
Cl	−0.0725	0.0427	−0.1323	−0.0264	**−0.3608**	0.1822	0.0940	**0.4915**	0.2282
Br	0.0071	0.0037	−0.0376	−0.0640	−0.0196	−0.0162	**0.7007**	−0.2220	0.2612
I	0.0034	0.0001	0.0069	−0.0315	−0.1054	0.0806	**0.3053**	−0.0697	**−0.8909**
max MEP	−0.0636	**−0.3740**	−0.1726	0.0561	−0.0250	0.2358	−0.0209	−0.1113	0.0356
min MEP	0.0560	0.1188	**−0.4189**	−0.1134	**0.4081**	−0.0311	−0.0235	0.0966	0.0431
mean + MEP	0.0460	**−0.3280**	−0.2571	−0.0029	0.0456	−0.1054	0.0861	−0.0684	0.0269
mean − MEP	−0.0764	**0.3080**	**−0.3062**	0.0921	−0.1095	0.0812	−0.1317	−0.1355	−0.0020
variance (σ_{tot}^2)	0.0297	**−0.3774**	0.2627	−0.0276	−0.1380	0.1052	0.0024	−0.0433	0.0181
balance param (ν)	−0.0077	−0.1715	**−0.4740**	0.2571	0.0290	0.0056	0.0429	0.0103	0.0202
$\sigma_{tot}^2 \times \nu$	0.0190	**−0.3988**	−0.1071	0.1596	−0.0646	0.0941	0.0460	0.0228	0.0187
no. of H bond acceptors	−0.1879	−0.0891	0.0381	0.1592	0.2820	−0.2757	**0.3900**	0.1076	0.0185
no. of H bond donors	−0.0120	−0.1411	0.0510	**0.5623**	0.1421	−0.0122	−0.0004	0.0715	−0.0078
no. of aromatic rings	−0.2593	0.0525	−0.1194	−0.0581	**−0.3189**	−0.0679	0.0480	0.0676	0.0179
mol wt	**−0.3833**	−0.0371	−0.1093	−0.0767	−0.0565	0.0184	0.1275	−0.0277	−0.0388
mol volume	**−0.3963**	0.0147	0.0318	−0.0634	0.0381	−0.0362	−0.0575	−0.0272	0.0009
total surface area	**−0.4001**	0.0100	0.0150	−0.0432	0.0281	−0.0492	−0.0495	−0.0192	−0.0013
globularity	**0.3808**	0.0003	0.0301	−0.0074	0.0168	0.0771	0.0232	−0.0119	0.0023
Eigenvalue	6.0581	4.8052	2.3689	1.9850	1.4139	1.4065	1.1263	1.0803	1.0076
% variance explained	23.30	18.48	9.11	7.63	5.44	5.41	4.33	4.15	3.88
total % variance explained	23.30	41.78	50.89	58.53	63.97	69.38	73.71	77.86	81.74
qualitative description	size, shape	MEP +	MEP −	H bond donor	H bond acceptor	dipolar (polarity)	Br	O, F, Cl	I

[a] Coefficients larger than 0.3 are shown in boldface.

The remaining three PCs shown in Table 4 are dominated by the sums of potential-derived charges on different elements and thus can probably be interpreted as describing the chemical diversity of the compounds. PC7 is dominated by bromine and iodine, PC8 is dominated by oxygen, fluorine, and chlorine, and PC9 is dominated by iodine.

The above results provide a pleasingly consistent picture of the factors describing physical properties. These are, in order of descending importance, the size and shape of the molecule, its electrostatic properties summarized in two complementary descriptors, the hydrogen bond donor and acceptor properties, the dipolar polarity, and a series of descriptors describing the chemical constitution.

Can the Individual Descriptors Discriminate between Drugs and Nondrugs? We first plotted histograms (shown in the Supporting Information) of the percentage frequencies of the individual PCs and compared them for the Maybridge and drug data sets. Such histogram comparisons should reveal the extent to which the PCs can discriminate between drugs and nondrugs. The results are disappointing for all but one PC. PCs 5−8 show little discrimination because they describe relatively little variance. Of the four most significant PCs, PC3 (MEP−) is the only one that can discriminate between the drugs and the nondrugs. A 70.6% amount of the drugs have a value higher than 1.15 and only 21.6% of the nondrugs, as shown in Figure 3. Thus, this single factor can distinguish between drugs and nondrugs as well as many published procedures. It is remarkable that this ability to discriminate is only found for one of the PCs calculated. It is also important to note that the similarity found for PC1 between the two data sets rules out a simple size discrimination between drugs and nondrugs between our two data sets. On the basis of these results, we can propose a numerical index of drug-likeness, Δ, defined by:

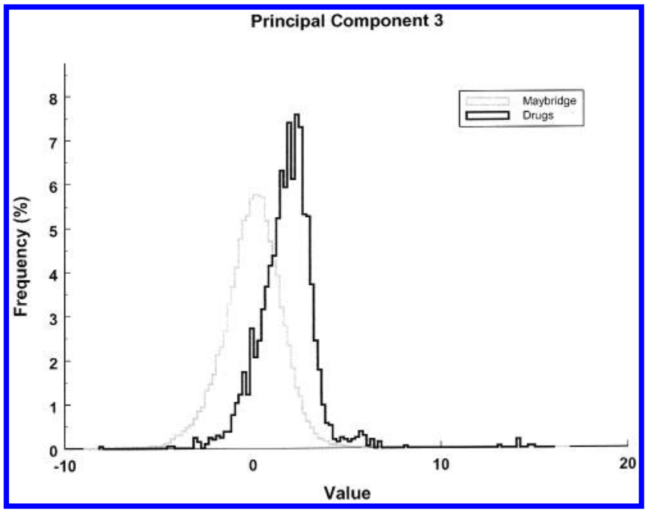


Figure 3. Frequency histogram for PC number 3 (MEP-) for drugs and nondrugs.

$$\Delta = \text{PC3} - 1.15 \qquad (1)$$

which would give positive values for most drugs and negative values for most nondrugs.

Why should PC3 be able to distinguish between drugs and nondrugs? Superficially, PC2 and PC3 form a complementary set of electrostatic surface descriptors that together describe the electrostatic binding characteristic electrostatics of the surface of the molecule. Our interpretation of PC2 and PC3 is that the former describes the total polarity of the molecular surface, as evidenced by the importance of the total variance, σ_{tot}^2, and both the mean positive and the mean negative MEPs (with opposite signs, indicating that the MEP range is important). However, in PC3, the balance parameter, ν, plays the most important role, indicating that the juxtaposition of positive and negative binding areas is important. The large coefficient of the minimum MEP descriptor in PC3 and the maximum MEP in PC2 may indicate some emphasis on opposite polarities for these two PCs, so that the acronyms MEP+ and MEP−, respectively, seem appropriate. For a more complete discussion of the meanings and purposes of MEP surface descriptors, see Murray and Politzer.[18,19]

Generally, however, we can conclude that, perhaps surprisingly, all other descriptors that have been found to be very well-suited for QSPR models of ADME-related properties do not do a very good job of distinguishing between our two data sets. We have therefore extended our descriptor set to the entire 66 calculated routinely by our software[17] in the hope that we can introduce more ability to recognize drugs. To develop a further calculational technique for recognizing drugs, we have first used recursive partitioning[40] to partition the drugs and nondrugs on the basis of the 66 descriptors. We have not, however, used this technique for discrimination because the Maybridge data set does not only contain nondrugs. Instead, we have used recursive partitioning to select suitable descriptors for an unsupervised learning (Kohonen net) approach in which the decision as to whether a Maybridge molecule would be a suitable drug or not is never made. We have used this approach previously to select descriptors for training back-propagation neural nets for physical[23] and spectroscopic[41] properties.

Recursive Partitioning. We have used the Formal Inference-based Recursive Modeling (FIRM) algorithm[42] to partition the drugs and nondrugs on the basis of our total set of 66 descriptors. FIRM partitions the data set recursively on the basis of the descriptors used. The first step is to partition the data set for each available descriptor into a maximum of 20 categories for ordinal data types and 10 categories for real data types. Grouping together similar adjacent categories reduces the number of these categories. The similarity is calculated using Student's *t* distribution for ordinal values and the χ^2 distribution for real values. For each descriptor, a probability value is calculated using Pearson's test or the *F* test. The data set is then split using the most significant descriptor. This is repeated until a user-supplied threshold is reached.

Using the FIRM software with the standard parameters suggested by the documentation resulted in three descriptors being used to partition the data: the sum of the potential-derived charges on hydrogens (QsumH), the minimum electrostatic potential (ESPmin), and the covalent hydrogen bond acidity[32] (covHBac) (see Table 2). Note that QsumH and ESPmin are both represented quite strongly (coefficients of 0.28 and 0.42, respectively) in the MEP− factor discussed above as being able to distinguish between drugs and nondrugs. The covHBac descriptor, which was not included in the original 26 descriptors and was introduced by Cronce et al.,[32] is defined as the difference in the orbital energy of the lowest unoccupied orbital (E_{LUMO}) of the molecule in question and that of the highest occupied orbital (E_{HOMO}) of water (−12.464 eV at AM1). ESPmin can be interpreted as being related to the strength of the strongest H bond acceptor.

To judge the ability of the FIRM analysis to discriminate using the three descriptors, we first normalized the results in order to take the differences in the numbers of compounds in the data sets into account. After this normalization, 77.4% of all drugs are found in end-nodes with a relative majority of drugs and 80.2% of all nondrugs are found in end-nodes with a relative majority of nondrugs. Thus, the three descriptors seem well-suited for a nonlinear mapping technique that should cluster the drugs.

Artificial Neural Networks. We and others have previously used supervised learning for the back-propagation training technique used to set up our QSPR models.[20−24] Gasteiger and Zupan[43] have pioneered the use of unsupervised learning in the form of Kohonen nets[44] in chemistry and have demonstrated that they can be remarkably predictive in a variety of applications. Kohonen nets are also often known as self-organizing maps (SOMs), and this description best describes our intention in applying them to this problem. The unsupervised learning process removes the need for a nondrug data set. However, we need to select descriptors for the mapping that can achieve our goal; otherwise, we cannot expect the neural network to cluster drugs away from nondrugs. However, the FIRM-based selection procedure outlined above should provide us with the best descriptors for our purpose.

A Kohonen net with a 2D organization of the network nodes (neurons) was used. To prevent any border effects, the neurons were organized toroidally, so that every neuron is equivalent to the others. The principles of Kohonen nets have been described many times[43,44] and will not be discussed here. All Kohonen net calculations reported here used the SOM_PAK program.[45]

A 200 × 200 node architecture was chosen for the first tests in order to give the 50 000 molecules enough space to distribute. Other network dimensions were used to investigate the effect of size on performance, and we eventually settled for a 100 × 100 node architecture. The nodes were arranged in a rectangular grid. The so-called "bubble" function was used as the radial adjustment function.[46] The area and rate of adjustment decrease with training cycles, but the rate of adjustment does not depend on the distance to the matching neuron.

The projection of the drug data set onto the 100 × 100 Kohonen map obtained from using the three descriptors selected by the FIRM analysis descriptors is shown in Figure 4. Both show the results for neural networks. Both networks show much better separation of drugs and nondrugs than all of our previous approaches. The drugs are well-clustered in an irregularly shaped island, suggesting that our approach can recognize potential drugs. However, before we analyze the performance of the net, we should investigate the relationship between the individual descriptors and the position of the drugs on the map.

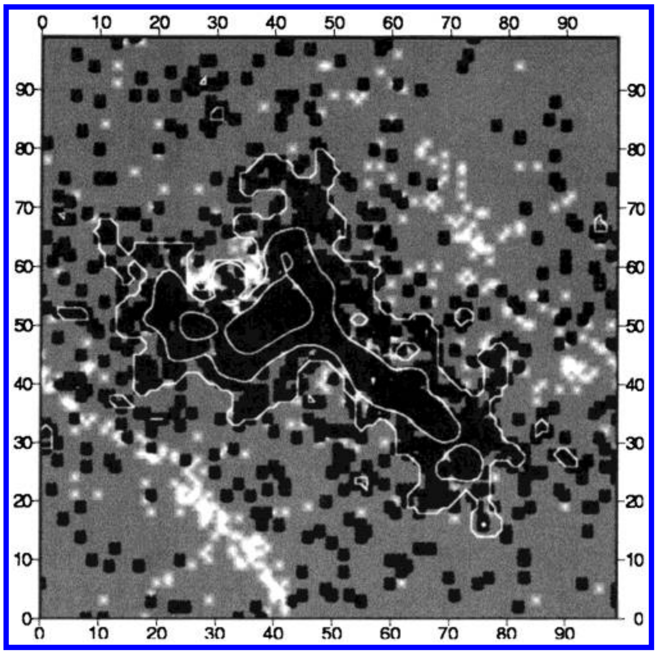


Figure 4. 100 × 100 Kohonen network trained with the three FIRM-selected descriptors (black, drugs; gray, nondrugs). The training parameters were as follows: number of iterations for the first training run, 10 000; starting adjustment value (α) for the first training run, 0.03; starting adjustment radius for the first training run, 200; number of iterations for the second training run, 100 000; starting adjustment value (α) for the second training run, 0.01; and starting adjustment radius for the second training run, 40.

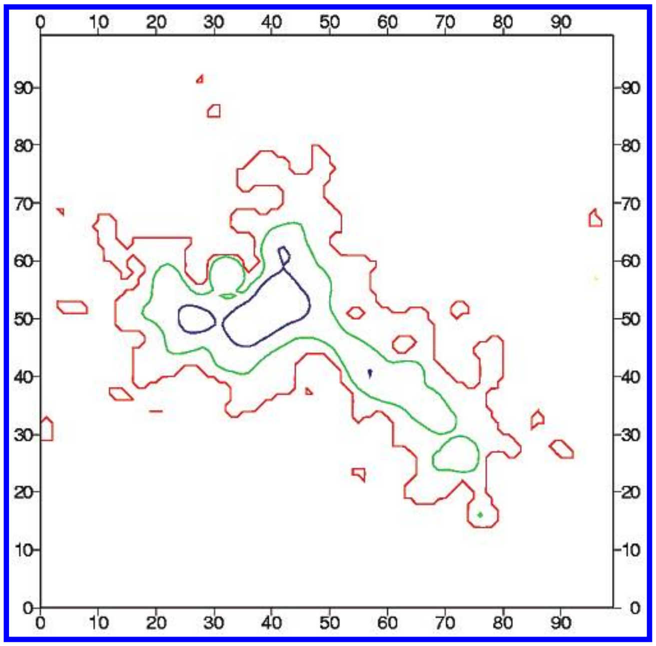


Figure 6. Contour plot of the occurrence of the drug molecules within the Kohonen map. The data have been smoothed as described in the text. The contours correspond to 0.2 (red), 0.67 (green), and 0.8 drugs per node (blue). The red contour corresponds to the value expected for a uniform distribution of the drugs over all of the nodes.

Direct Comparison of the Drug Molecule Distribution with the Descriptor Values. Figure 5 shows the distribution of the descriptor values within the Kohonen map. Maximum values are depicted in red, and minimum values are shown in magenta. The areas with the highest drug molecule concentration are outlined in white. The distribution of the QsumH values (Figure 5a) is very similar to that of the drug molecules. Druglike molecules have values around the maximum. This descriptor discriminates best between drugs and nondrugs and was therefore selected by FIRM as the first partitioning descriptor and also has a coefficient of 0.28 in the MEP+ PC.

The distribution for the ESPmin values is shown in Figure 5b. The minimum values for this descriptor appear close to one end of the drug molecule cluster and the maximum close to the other. As will be shown below, this behavior corresponds to an additional selection within the drug data set.

The distribution for the third descriptor (covHBac) is shown in Figure 5c. The distribution is similar to that found for QsumH but shifted upward in the diagram. The drug cluster lies parallel and close to the area of maximum values for this descriptor.

Thus, two of the descriptors, QsumH and covHBac, are very similarly distributed to the drug cluster and a combination of the two can clearly define the position of the cluster fairly well. The third descriptor, ESPmin, provides some discrimination within the cluster that we had not expected (see below).

Quantitative Analysis. Figure 6 shows a contour plot of the numbers of molecules from the drug data set per neuron. The data have been smoothed using a

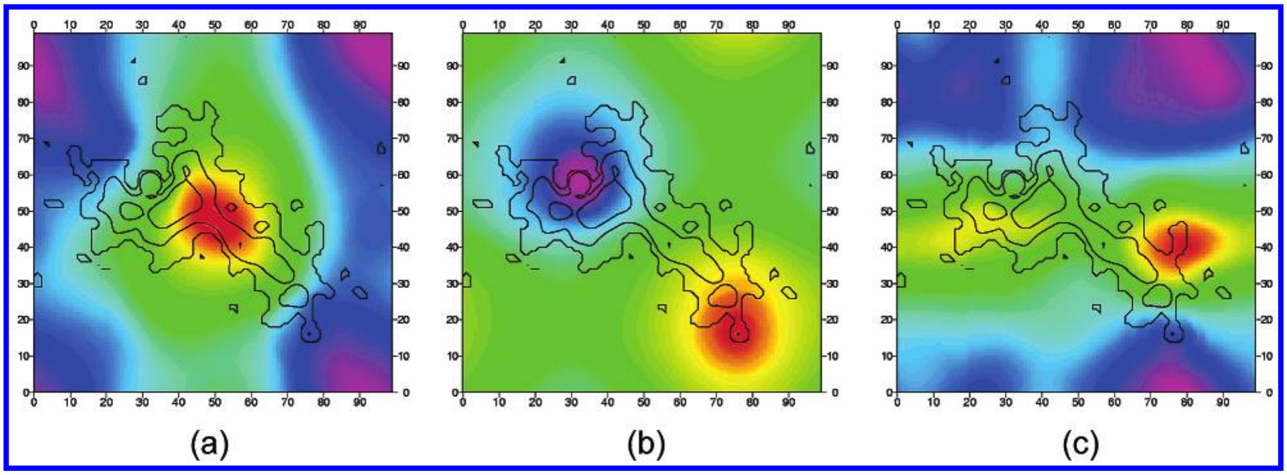


Figure 5. Color-coded (red, minimum; magenta, maximum) Kohonen maps of the values of the three descriptors selected by the FIRM analysis: (a) QsumH, (b) Espmin, and (c) covHBac. The contours shown correspond to those given in Figure 6.

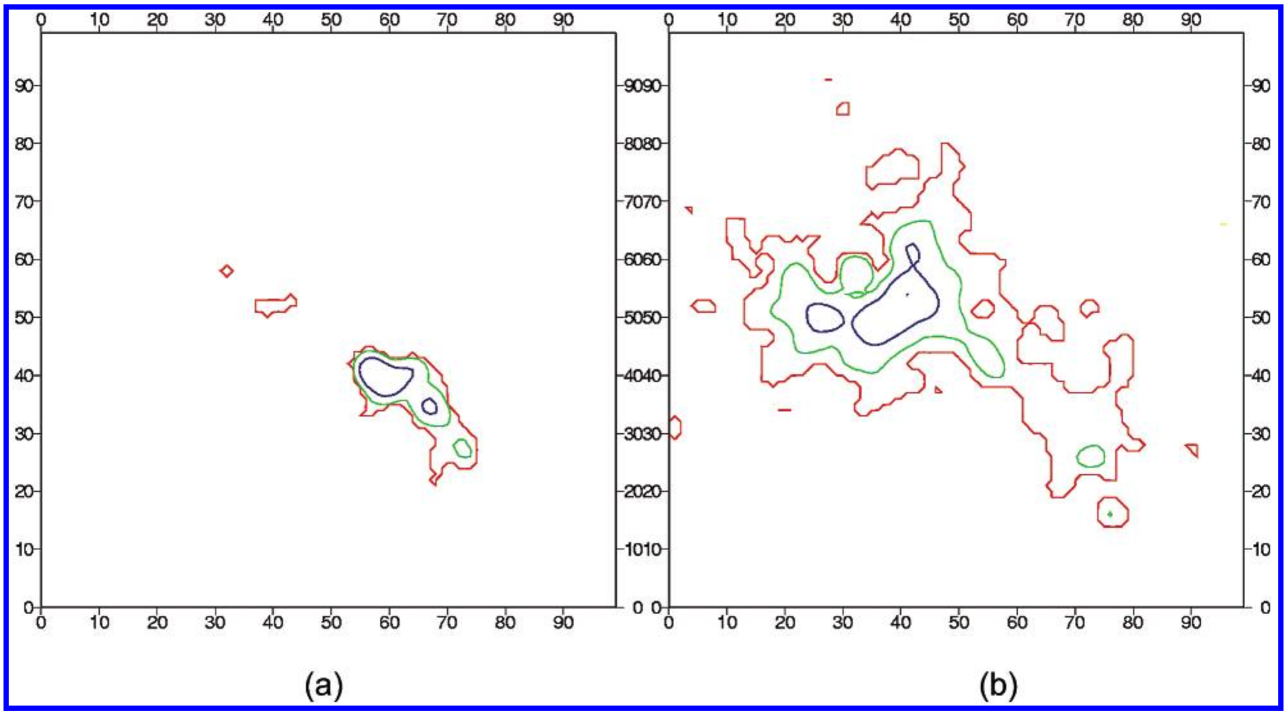


Figure 7. Smoothed contour plots for (a) the 252 hormones from the drugs data set and (b) the remaining compounds.

simple linear distance-dependent function extending out to the fourth nearest neighbors. The lowest contour level shown corresponds to a smoothed occupancy of 0.2 drugs per node, the value expected for a uniform distribution of the molecules over the 10 000 bins. Thus, the area within the red contour is that in which the drugs are proportionally overrepresented.

The drugs are found in 20% of the neurons, suggesting that the Kohonen map provides a practical and useful tool for identifying drug candidates. Before assessing the performance of this technique, however, we will discuss the occurrence of different types of drug within the map and whether we can actually identify the best areas for different applications or targets.

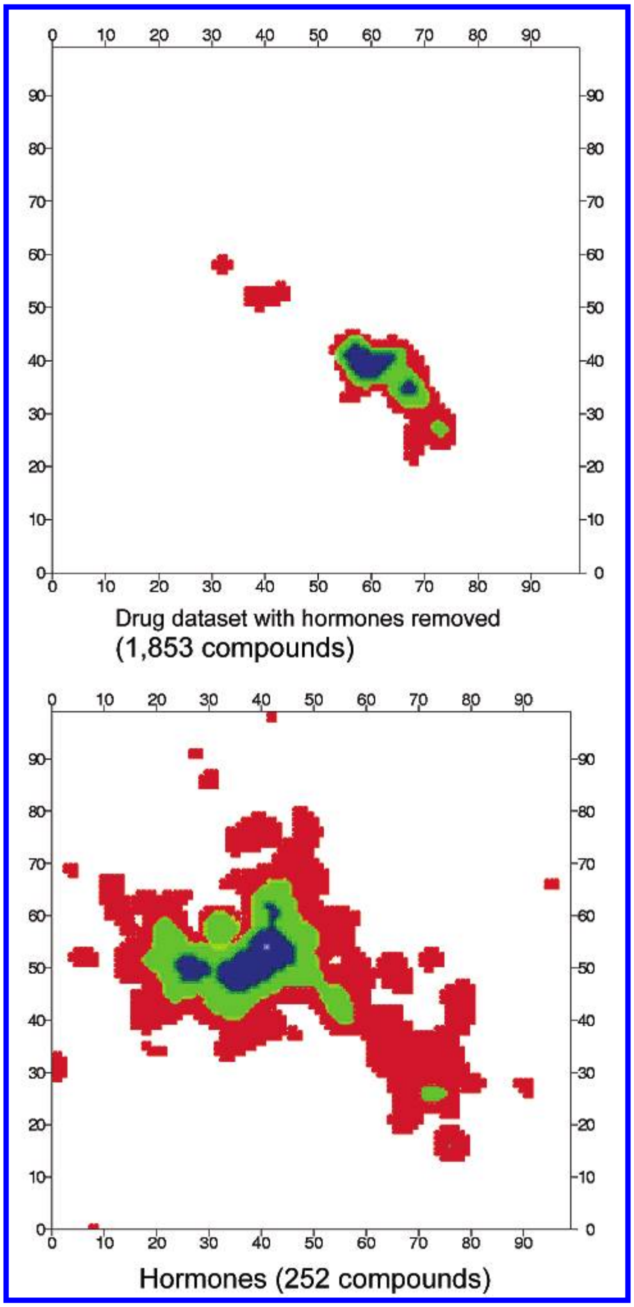
Which Drugs Are Where? To investigate possible partitioning within the drug data set, the drugs were classified into 10 different classes according to their application. The distribution of each class within the Kohonen map was then plotted. Interestingly, the different classes of drugs are somewhat discriminated by the ESPmin descriptor. Antiinfectives, central nervous system (CNS) drugs, antiallergics, cardiants, and bronchial drugs are concentrated within the 0.2 molecule/node contour found for the total drug set but at the end of the distribution close to the minimum values of ESPmin. Immunomodulants, enzyme−inhibitors/enhancers, and digestion-related drugs are, if at all, only very loosely clustered. Gratifyingly, dermatics (a classification that was accidentally not excluded in our drug selection procedure) are also fairly evenly distributed over the map. However, the hormones are concentrated at the opposite end of the total drug cluster (that with average values of ESPmin adjacent to the ESPmin maximum) to the majority of drugs. We therefore separated the hormones (252 compounds) from the remainder of our data set and plotted separate contour plots for the two data sets. The results are shown in Figure 7.

The two plots show that by separating the predominantly steroidal (151 of 268 compounds in the hormone data set are steroids) hormones from the remaining drugs, we can not only identify hormonelike compounds but also use a significantly smaller cluster for the remaining drug compounds. We have therefore set up two binning schemes in order to allocate a drug-likeness score in four levels.

Drug-Likeness Classification. We have divided the frequency with which drugs are found in a given bin in the above smoothed Kohonen maps into three ranges based on the frequency and plotted the results for the hormones and the drugs−hormones data sets in Figure 8.

The red area (class C) indicates only that at least one member of the data set was assigned to this node and is therefore only a very weak indication that such a compound may be suitable as a drug. The green (class B) and blue (class A) areas indicate a far higher concentration of drugs and are therefore a strong indication that the compound has the correct physical properties. For the drug data set, there are 389 class B nodes and 143 class A nodes. The two highest classes together therefore account for 5.3% (3.9 and 1.4%, respectively) of the 10 000 available nodes. The corresponding data for the hormone data set are 80 class B and 49 class A nodes (0.8 and 0.5% of the total number of nodes, respectively). Nodes in which no drugs appear are assigned to class D.

Analysis of the Performance of the Models for the Data Sets. Figure 9 shows a histogram of the percent frequency of the occurrence in classes A−D of the molecules in the Maybridge and total drugs data sets (the numerical data are shown in Table 5 in the Supporting Information). The distinction between the two data sets is impressive. The ratios drugs:nondrugs for classes A−D are 17.7:1, 4.9:1, 1.7:1, and 0.2:1, respectively. This selection is mainly caused by a very sharp falloff in the proportion of Maybridge molecules in the druglike classes. Only 1.7% of the Maybridge molecules, for instance, are found in class A. However,



Drug dataset with hormones removed
(1,853 compounds)

Hormones (252 compounds)

Figure 8. Map of the binning used to classify compounds according to the neuron to which they are assigned in the trained Kohonen net.

the simple numerical index, Δ, also provides a useful distinction between the two data sets. The mean Δ for the Maybridge data set is −1.15, with a standard deviation of 1.54, whereas the corresponding values for the drugs data set are +0.62 and 1.56. Thus, the widths of the two distributions are very similar, as shown in Figure 3, but the maxima are shifted relative to each other.

We also processed a large subset (42 131 compounds) of the WDI selected using a size limit of 150 atoms and projected these onto the Kohnen map obtained by the above procedure. The resulting distribution is shown in Figure 10. The large data set clusters were remarkably similar to the selected drug data set, as also shown in Table 5 (Supporting Information). The results suggest that the large data set, which we will call WDI for simplicity, is significantly more druglike than Maybridge (73.5% of the compounds are found in classes A−C) but less so than our small selected drug data set. This conclusion is significant because the drug data set was combined with Maybridge to train the Kohonen net, although the distinction between the two data sets was used to select the descriptors. The ratios of the frequency of occurrence of the drug data set as compared to WDI in classes A−D are 1.4:1, 1:1, 1:1, and 1:1.6, respectively. The corresponding ratios between Maybridge and WDI are 1:12.8, 1:4.9, 1:1.6, and 3:1. Thus, the drug data set is enriched in class A and poorer in class D compounds than the unselected WDI data set. The frequencies of occurrence of class B and C compounds are very similar in the two data sets. This result is both consistent with our expectations and a useful validation of the techniques used. We should perhaps note here that there are many nondrugs in the WDI data set.

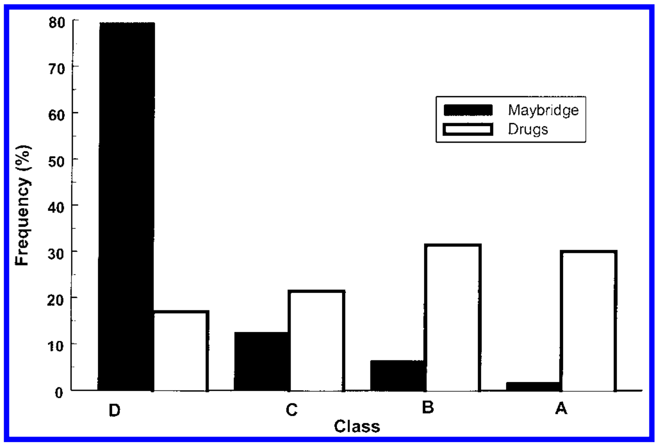


Figure 9. Performance of the Kohonen map-based classification scheme for the Maybridge and drug data sets.

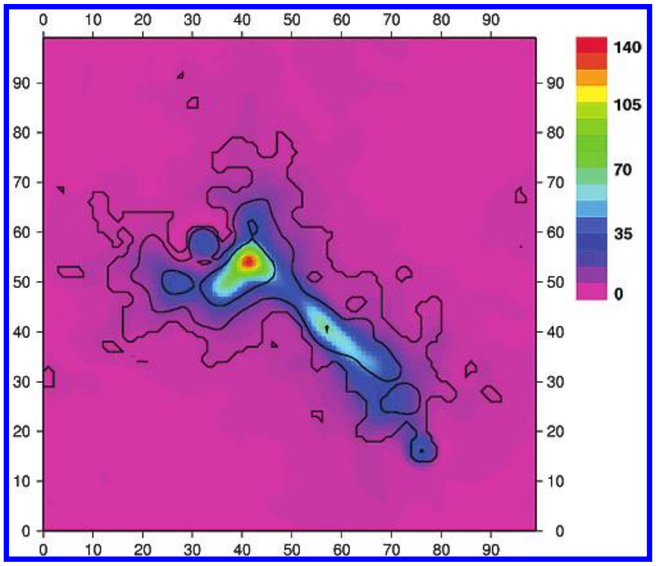


Figure 10. Distribution of the 42 131 compounds with 150 atoms or less taken from the WDI on the Kohonen map obtained for the Maybridge data set.

Summary and Conclusions

The techniques investigated here provide considerable new information about the relationship between descriptors and physical properties as well as how drugs can be distinguished from nondrugs. The assignment of the PCs of the 26 descriptor set to the individual factors shape/size, MEP+, MEP−, H bond donor, H bond acceptor, dipolar polarity, and chemical diversity descriptors (in order of decreasing importance) provides a simple and instinctive framework for further QSPR work.

The fact that only one of these descriptors, MEP−, is able to discriminate between drugs and nondrugs is particularly interesting. The simple drug-likeness index, Δ, defined in eq 1, provides a simple, one-dimensional estimate of the suitability of a given compound as an oral drug. Its performance is comparable with many of the methods described in the literature.

The Kohonen map, trained using only three descriptors identified by recursive partitioning, provides more information and is even able to discriminate, for instance, between hormones and other drugs. The neuron to which a given compound is assigned allows a qualitative classification of the compound as a potential drug or not. It is both remarkable and gratifying (from our point of view) that the classical 2D descriptors included in the full 66 descriptor set, such as the counts of hydrogen bond donors and acceptors, apparently do not have the resolution to distinguish as well as the three quantum mechanically derived descriptors QsumH, ESPmin, and covHBac. This is perhaps surprising as the three descriptors used are clearly connected to the hydrogen-bonding properties of the molecule. They do not, however, correlate strongly with their 2D equivalents.

Above all, however, the fact that our Kohonen map-based classification does not depend on the definition of a nondrug data set encourages us to believe that the techniques described here can form the basis for an objective and routine screening of drug candidates.

Acknowledgment. We thank Dr. J. Einsiedel and Apothekerin A. Moll (Institut für Pharmazeutische Chemie der Universität Erlangen-Nürnberg) for classifying our drug data set. We also thank Johnny Gasteiger for introducing us to Kohonen nets. This work was supported by the Fonds der Chemischen Industrie and by the Alexander-von-Humboldt Foundation (W.K.).

Supporting Information Available: Eight figures showing histograms of the percentage frequencies of the individual principal components (not shown in the text) for the Maybridge and drug data sets plus a table showing the classification of the different types of molecules using the drug-likeness index Δ and the Kohonen mapping. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) 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.

(2) Sadowski, J.; Kubinyi, H. A. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(3) Wagener, M.; van Geersestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.

(4) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(5) (a) Clark, T. Quantum Cheminformatics: An Oxymoron? Proceedings of the Beilstein Institute Workshop: Chemical Data Analysis in the Large, Bozen, Italy, May 22−26, 2000; p 88−99. (b) Clark, T. Quantum Cheminformatics: An Oxymoron? In *Rational Approaches to Drug Design*, 13th European Symposium on QSAR; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, 2001; Part II, p 29−40.

(6) Derwent World Drug Index (Derwent WDI), 1997, Derwent Information, 14 Great Queen Street, London, WC2B 5DF, U.K.

(7) United States Adopted Names (USAN) Council, American Medical Association, P.O. Box 10970, Chicago, IL 60610.

(8) World Health Organisation (WHO), http://www.who.int/medicines/organization/qsm/activities/qualityassurance/inn/innguide.shtml.

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

(10) (a) 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; p 437. (b) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(11) Sadowski, J.; Gasteiger, J. *Corina, version 1.8*; Oxford Molecular: Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, England.

(12) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(13) Clark, T.; Alex, A.; Beck, B.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Rauhut, G.; Sauer, W.; Steinke, T. *VAMP 7.0*; Oxford Molecular Ltd.: Medawar Centre, Oxford Science Park, Standford-on-Thames, Oxford, OX4 4GA, U.K., 1998.

(14) Beck, B.; Horn, A.; 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.

(15) Rauhut, G.; Clark, T. Multicenter Point Charge Model for High Quality Molecular Electrostatic Potentials from AM1 Calculations. *J. Comput. Chem.* **1993**, *14*, 503−509.

(16) Beck, B.; Rauhut, G.; Clark, T. The Natural Atomic Orbital Point Charge Model for PM3: Multiple Moments and Molecular Electrostatic Potential. *J. Comput. Chem.* **1995**, *15*, 1064−1073.

(17) Beck, B. *Propgen 1.0*; Oxford Molecular Ltd.: Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, U.K., 2000.

(18) 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.

(19) 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.

(20) Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. Prediction of the *n*-Octanol/Water Partition Coefficient, logP, using a Combination of Semiempirical MO−Calcualtions and a Neural Network. *J. Mol. Model.* **1997**, *3*, 142−155.

(21) 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.

(22) Chalk, A. J.; Beck, B.; Clark, T. A Temperature-Dependent Quantum Mechanical/Neural Net Model for Vapor Pressure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1053−1059.

(23) 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.

(24) Beck, B.; Clark, T. Manuscript in preparation.

(25) (a) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method *J. Comput. Chem.* **1989**, *10*, 209−220. (b) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.

(26) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

(27) 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.

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

(29) 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.

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

(31) Beck, B.; Glen, R. C.; Clark, T. VESPA: A New, Fast Approach to Electrostatic Potential Derived Atomic Charges from Semiempirical Methods. *J. Comput. Chem.* **1997**, *18*, 744−756.

(32) Cronce, D. T.; Famini, G. R.; DeSoto, J. A.; Wilson, L. Y. Using Theoretical Descriptors in Quantitative Structure−Property Relationships: Some Distribution Equilibria. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293−1301.

(33) 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.

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

(35) Beck, B. Unpublished results.

(36) Kier, L. B.; Hall, L. H. *Molecular Structure Description*; Academic Press: New York, 1999.

(37) 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**, *2*, 159−167.

(38) (a) Guttmann, L. Some necessary conditions for common factor analysis. *Psychometrika* **1954**, *19*, 149−162. (b) Kaiser, H. F.; Dickmann, K. Analytic determination of common factors. *Am. Psychol.* **1959**, *14*, 425−439.

(39) (a) Catell, R. B.; Vogelmann, S. A comprehensive trial of the scree and KG-criteria for determining the number of factors. *Multi. Behav. Res.* **1977**, *12*, 289−325. (b) Catell, R. B. The scree test for the number of factors. *Multi. Behav. Res.* **1966**, *1*, 245−276.

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

(41) Brüstle, M. M. Sc. Thesis, Universität Erlangen-Nürnberg, 2000.

(42) Hawkins, D. M. FIRM, http://www.stat.umn.edu/users/FIRM/index.html.

(43) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists, An Introduction*; VCH: Weinheim, 1993.

(44) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer-Verlag: Berlin, 1989.

(45) Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. *SOM PAK: The Self-Organizing Map Program Package*; Technical Report A31; Helsinki University of Technology, Laboratory of Computer and Information Science: FIN-02150 Espoo, Finland, 1996.

(46) Honkela, T. *Comparisons of Self-Organized Word Category Maps*, Proceedings of WSOM ‘97, Workshop on Self-Organizing Maps, Espoo, Finland, 1997; http://citeseer.nj.nec.com/427658.html.

JM011027B