Suitability of Molecular Descriptors for Database Mining. A Comparative Analysis

Gabriele Cruciani,*,[†] Manuel Pastor,[‡] and Raimund Mannhold[§]

Dipartimento di Chimica, Laboratorio di Chemiometria, Universita di Perugia, Via Elce di Sotto, 10, 1-06123 Perugia, Italy, Department de Ciencies Experimentals i de la Salut, Universitat Pompeu Fabra, Doctor Aiguader 80, E-08003 Barcelona, Spain, and Institut für Lasermedizin, Arbeitsgruppe Molekulare Wirkstoff-Forschung, Heinrich-Heine-Universität Düsseldorf, Universitätsstrasse l, D-40225 Düsseldorf, Germany

Received December 12, 2000

Database mining methods rely on the molecular descriptors used to characterize a structural database. In the present investigation, five different types of descriptors (log *P*, UNITY fingerprints, ISIS keys, VolSurf, and GRIND) are applied to characterize various databases (n = 1007, 100, and 229) comprising drugs almost exclusively. The validity of the descriptors is comparatively analyzed via principal component analysis and its hierarchical variant, consensus principal component analysis. Both pharmacodynamic and pharmacokinetic aspects of database mining are treated. For pharmacodynamic aspects, clustering behavior achieved with the different descriptors is tested on the chemically homogeneous β-blockers, benzodiazepines, and penicillins and on the chemically more diverse class I antiarrhythmics. The following ranking is observed: UNITY fingerprints > ISIS keys and GRIND > VolSurf > log *P*. Regarding information content, the CPCA superweight plot indicates similarity between fingerprints and ISIS keys as well as between VolSurf and log *P*, while GRIND differs from all the remaining descriptors. Solubility data and blood/brain barrier penetrating behavior serve as test cases for pharmacokinetic aspects. Comparison of the descriptors applied to these data reveals that VolSurf has the most realistic and consistent behavior, GRIND shows intermediate behavior, while UNITY fingerprints and ISIS keys are not well suited for pharmacokinetic profiling. From this comparative analysis, we conclude that VolSurf descriptors exhibit particular advantages in treating pharmacokinetic aspects; UNITY fingerprints, ISIS keys, and GRIND descriptors are of special value for tackling pharmacodynamic aspects of database mining. The parameter log *P* is of limited applicability in database mining because of rather poor reliability and lack of completeness of data.

Introduction

Techniques from combinatorial chemistry gain increasing impact on lead generation and optimization in drug discovery. This involves the synthesis of sets of molecules containing large numbers of structurally related compounds that are rapidly generated by automated procedures. The joint development of high-throughput bioassays has meant that combinatorial chemistry provides a far more cost-effective approach to the discovery of bioactive compounds than traditional approaches that require the sequential synthesis and testing of individual molecules. In pharmaceutical research, the introduction of combinatorial chemistry implies the availability of large collections of compounds and consequently the availability of large databases collecting their structures, their measured biological activities, and other biological and physicochemical properties.

These databases are often regarded as a highly valuable source of information, and specific informatic manipulation is applied on them in order to evaluate the molecular diversity of their components, searching for compounds with particular pharmacodynamic or pharmacokinetic properties (pharmacophoric searches or pharmacokinetic profiling, respectively) and even trying to differentiate druglike from nondruglike compounds. In general, the informatic manipulation of the databases has been called “database mining”.

Irrespective of the purpose of database mining, its success is subordinated to the choice of an appropriate molecular characterization, producing a set of informative variables. Suitable characterizations must be, first of all, relevant to the properties to be studied, but some other characteristics should be also considered, such as the time needed for their computation and analysis and the storage requirements, to mention a few. Types of descriptors frequently used in database mining are reviewed by Brown[1,2] (see also Table 1).

Classically, structural descriptors were derived from global molecular properties. Physicochemical properties such as lipophilicity, electrostatics, steric shape, and bulk deserve mentioning. An advantage of these descriptors is their fast computation. A serious disadvantage, however, is their lack of completeness and accuracy, at least in some cases.[3] This holds in particular for some commercial software available for calculating log *P*. The problem of completeness, but not of accuracy, is solved with some of the newer whole-molecule approaches validated via neural networks.[3]

* To whom correspondence should be addressed. Phone: +39 075-45646. Fax: +39 075-45646. E-mail: gabri@chemiome.chm.unipg.it.
[†] Universita di Perugia.
[‡] Universitat Pompeu Fabra.
[§] Heinrich-Heine-Universitat Dusseldorf.

10.1021/jm0011326 CCC: $22.00
Published on Web 05/24/2002

Table 1. Descriptors Used for Database Mining[a]

	types of descriptors	type used in this study
I	descriptors derived from global molecular properties	log *P*
II	descriptors derived from 2D structure	
	fragment substructures	UNITY fingerprints, ISIS keys
	receptor recognition descriptors	
	topological indices	
III	descriptors derived from 3D structure	VolSurf, GRIND
IV	descriptors based on biological properties	
V	combination descriptors	

[a] Descriptors used for database mining according to the classification of Brown.[1,2] In the right-hand column, those descriptors are listed that are comparatively analyzed in this study.

Two-dimensional fragment-based descriptors, derived from substructure search systems, are frequently used in database mining. There are two main types. Structural keys make use of a predefined fragment dictionary and record the presence or absence of a number of small generic or specific fragments. Alternatives are hashed fingerprints.

Several 3D substructure search systems use a screening stage prior to geometric search. The screening methods used in these systems encode geometrical relationships between features such as atoms, ring centroids, and planes in terms of distances and angles.

The aim of the present work is to study the information given by different types of descriptors (Table 1) and to analyze their suitability for different practical purposes. With this aim, databases of 1007, 100, and 229 structures, mainly commercial drugs, have been characterized using different descriptors and then analyzed via simple chemometric tools such as PCA and CPCA. The analysis has been oriented to produce graphical representations of the databases where one can recognize the ability of the descriptors to discriminate between different families of compounds according to their different pharmacokinetic and pharmacodynamic properties. Special attention has been paid to calculated lipophilicity descriptors computed with four different commercial software packages.

Computational Methods

Databases. The database for pharmacodynamic studies ($n = 1007$) comprises almost exclusively pharmaceuticals such as penicillins, β-blockers, benzodiazepines, and many more. It covers a broad spectrum of diverse chemistry ranging from simple structures such as acetyl salicylate to complex molecules such as erythromycin. Even if most compounds can be considered “druglike”, the database contains some molecules as lipophilic as β-caroten and others as hydrophilic as acarbose. Databases for pharmacokinetic studies comprise 100 molecules for the solubility example and 229 molecules for the blood/brain barrier permeation example.

Lipophilicity Descriptors. Methods for calculating lipophilicity (log *P*) can be divided into two classes: substructural approaches and whole-molecule approaches. The substructure approaches have in common that molecules are cut into atoms or groups; summing the single-atom or fragmental contributions (supplemented by applying correction rules in the latter case) results in the final log *P*. On the other hand, the whole-molecule approaches inspect the entire compound; they use, for instance, molecular lipophilicity potentials, topological indices, or molecular descriptors to quantify log *P*, and some reflect the impact of the three-dimensional structure on molecular lipophilicity.

In this study, we have obtained log *P* estimations with software based on both types of approaches: three substructure approaches (CLOGP version 4.34,[4,5] KOWWIN,[6] MOLCAD[7,8]) and one whole-molecule approach (HINT[9]). The fragmental system CLOGP[4,5] from Hansch and Leo is based on (in contrast to other fragmental methods) the principles of “constructionism”. The basic fragmental values were derived from accurately measured log *P* data of simple molecules such as hydrogen and methane, and then the remaining fragment set was constructed. A total of 200 fragment values and 25 correction factors are given. The KOWWIN software uses the atom/fragment contribution method by Meylan and Howard.[6] This is a reductionist approach derived by regression analysis. A total of 144 atom/fragment values and 235 correction factors are listed in version 1.54. A significant advantage of this approach is the fact that because of the simple atom/fragment methodology, so-called missing fragments only rarely occur. The MOLCAD[8] software uses the Ghose−Crippen method,[7] a purely atom-based procedure that exclusively applies atom contributions and avoids correction factors. Atoms (C, H, O, N, S, and halogens) are classified into 120 atom types. The program HINT[9] is a whole-molecule approach reflecting three-dimensionality by combining substructure contributions and conformational effects. The key parameter is the hydrophobic atom constant a_i, derived from Leo's fragment constants. HINT calculates hydrophobic atom constants using the following criteria: (1) the sum of atom constants within a fragment equals the fragment constant value; (2) bond, branching, or vicinal halogen factors are applied to all eligible atoms, while polar proximity factors are applied to the central atom of fragments; (3) superficial atoms are considered to be more important than central atoms.

Descriptors Derived from 2D Structure. Very often some form of fragment-based descriptors is used to characterize compounds in a database. Fingerprints are bit strings representing the answers to yes and no questions about the presence or absence of various substructural features within the molecular structure of a given compound. Fingerprints represent very high-dimensional chemistry spaces, typically from some hundreds up to thousands of bits. Representatives of fragment-based descriptors in this study are UNITY fingerprints and ISIS keys. UNITY fingerprints (e.g., similar to Daylight fingerprints used in many other papers) are based on a structural characterization using paths of connected terms and differ from ISIS keys by using a fixed number of defined substructures. UNITY fingerprints[10] were calculated with the standard definition rule file as implemented in UNITY 2.4. With ISIS keys,[11] molecular holograms were calculated via an SPL script (Sybyl, version 6.4).

Descriptors Derived from 3D Structure. Two recently developed descriptors included in our comparative study are the VolSurf descriptors[12] and the GRID independent descriptors (GRIND).[13] Both types of descriptors have in common the fact that they start from molecular interaction fields (MIF) computed with the program GRID. The large body of information contained in these fields (on the order of some hundred thousands of grid points) is then encoded using two different methods. VolSurf analyzes the MIF obtained with diverse probes (often water and the hydrophobic probe) and computes the volume and surface of the regions enclosing values of energies of interaction under certain cutoff limits, together with some other variables expressing their geometric distributions in the space. The result of such an analysis is a small number of variables describing the overall distribution of hydrophobic and hydrophilic regions around the molecule with a distinct physicochemical meaning. VolSurf descriptors are useful for describing pharmacokinetic and physicochemical properties.[12] On the other hand, GRIND descriptors have been designed mainly to represent pharmacodynamic properties. The computation of GRIND involves a preliminary simplification of a few MIF to extract the main pharmacophoric regions, followed by a particular type of autocorrelation transform. The results are a small set of alignment-independent descriptors representing the internal geometrical relationship of such pharmacophoric regions. These can be used directly for the chemometric analysis and can be interpreted with the appropriate software,[13] using graphical representations of the pharmacophoric regions and their interactions, together with the molecular structures, in interactive 3D plots.

Chemometric Analysis of the Descriptors. Since the aim of the study is to analyze the information given by the different types of descriptors, we have approached the problem using techniques generating simple 2D maps of the compounds. Of these, one can study the different ability of the descriptors to distinguish between different types of compounds, in terms of the distance between the different types of compound in the maps. The main technique used is the classic principal component analysis. To gain additional insight into the problem, a hierarchical variant of the PCA was also applied: the consensus principal component analysis.[14] In any case, the analysis starts by building an **X** matrix of data from the database. In this matrix, the rows represent the compounds and the columns contain the variables describing the compounds. The PCA is a well-known technique. To extract relevant information from a matrix, the PCA decomposes the matrix into a product of two smaller matrices: **T** (score matrix) and **P′** (loading matrix), which explain at best the overall variance of the original **X** matrix. The score matrix contains a few variables (Principal Components or PC) that are used to describe the objects (compounds) while the loading matrix again relates the original variables with the PC. From a practical point of view, the PCA analysis of the matrix allows us to obtain highly informative graphical representations of complex, high-dimensional matrices, in terms of a score plot and a loading plot. The score plot is a simplified representation of the objects where each compound is located in a particular position of the plot. The loading plot, on the contrary, is a representation of the participation of the original variables into the PC.

When the descriptors are not single variables but blocks of variables and when there is an interest in evaluating the discriminating power of the variables themselves, hierarchical variants of PCA are superior to regular PCA. Among the available methods, we decided to apply consensus principal component analysis (CPCA), as implemented in GOLPE.[15] Details of the algorithm are described elsewhere.[16] Basically, CPCA uses exactly the same objective function of PCA and tries to best explain the overall variance of the **X** matrix, but the analysis is made at two levels: the block level, which expresses the “opinion” of each of the blocks of variables, and the superlevel, which express the “consensus” of all blocks. As implemented in GOLPE, CPCA provides a solution in the superlevel that is identical to the solution found in regular PCA. The same **T** and **P′** matrices are obtained. Additionally, the method produces block scores T_b and block loadings P_b for each of the probes used and a weight matrix that expresses the participation of each block in the overall scores. The block loadings are essentially identical to the “piece” of the loading corresponding to a certain block except for the use of a different normalization. On the contrary, block scores represent a peculiar point of view of the model given by a certain probe and provide unique information not present in regular PCA. Object distances in the block scores are used in GOLPE to assess the relative importance of the different blocks of variables in their discrimination.

Chemometric analysis was performed on SGI 02 workstations, using the GOLPE software, version 4.5.[15]

Results and Discussion

Database mining strictly depends on the selection of appropriate molecular descriptors, guided by the specific aim of a study. Both pharmacodynamic and pharmacokinetic aspects of database mining are treated here. The following descriptors are comparatively analyzed to cover most different types: log *P* data represent descriptors derived from global molecular properties. UNITY fingerprints and ISIS keys are examples of descriptors derived from 2D structure. Finally, VolSurf descriptors and GRIND, calculated with the new software ALMOND, represent descriptors derived from 3D structure. Descriptors are compared by using the consensus principal component analysis (CPCA) option of the software GOLPE. Criteria for comparison are aspects of information content, reliability, completeness, and rapidity in deriving the individual descriptors.

Pharmacodynamic Aspects of Database Mining. One main goal of database mining is the search for pharmacodynamic similarity. The database ($n = 1007$), used for this part of our study, comprises drugs almost exclusively. The chemically homogeneous classes of β-blockers (29 structures), benzodiazepines (15 structures), and penicillins (26 structures) as well as the chemically more diverse class I antiarrhythmics (9 structures) are used as test cases to exemplify their clustering behavior. Application of CPCA to the entire database using 1771 variables, derived with five descriptors (for details, see Computational Methods), yields a two-component model explaining about 20% of the variance. The four chemical classes are relatively well separated and nicely clustered (see Figure 1). Figure 2 elucidates which descriptor block is responsible for this separation. Separation obtained with the UNITY fingerprints is documented in Figure 2a, clearly indicating a nice clustering of all four pharmacological classes. Thus, UNITY fingerprints exhibit the highest contribution to the global separation as shown in Figure 1. GRID independent descriptors GRIND, calculated with the ALMOND software, also cluster penicillins, 1-blockers, and benzodiazepines, but class I antiarrhythmics are mixed between the other clusters (Figure 2b). ISIS keys show a less evident separation (Figure 2c). Penicillins are well clustered and separated from the other groups, but β-blockers and benzodiazepines are less well separated and exhibit some overlap with the class I antiarrhythmics. Only the penicillins are separated when using the VolSurf descriptors, while the remaining groups overlap (Figure 2d). The log *P* descriptors completely fail to cluster the four groups, which all are mixed, as can be derived from Figure 2e. From this comparison one can conclude the following ranking of descriptor suitability for pharmacodynamic aspects of database mining:

$$\text{UNITY fingerprints} > \text{GRIND} \gtrsim \text{ISIS keys} > \text{VolSurf descriptors} > \log P(\text{s})$$

Results of another interesting aspect of descriptor comparison, i.e., information content, are summarized in Figure 3 via the loading plots, which represent the original variables in the space of the principal components. The loading of a variable indicates how much this variable participates in defining the principal component. Similarity in projection indicates similarity in information content. A comparative view in Figure 3 elucidates, for example, a certain similarity between log *P* and VolSurf descriptors and between UNITY fingerprints and ISIS keys; GRIND descriptors differ from the other descriptors in a peculiar way. The CPCA superweight plot, shown in Figure 4, summarizes the above observations: the log *P* block is closest to the VolSurf block, UNITY fingerprints are closest to ISIS keys, while the isolation of GRIND descriptors under-

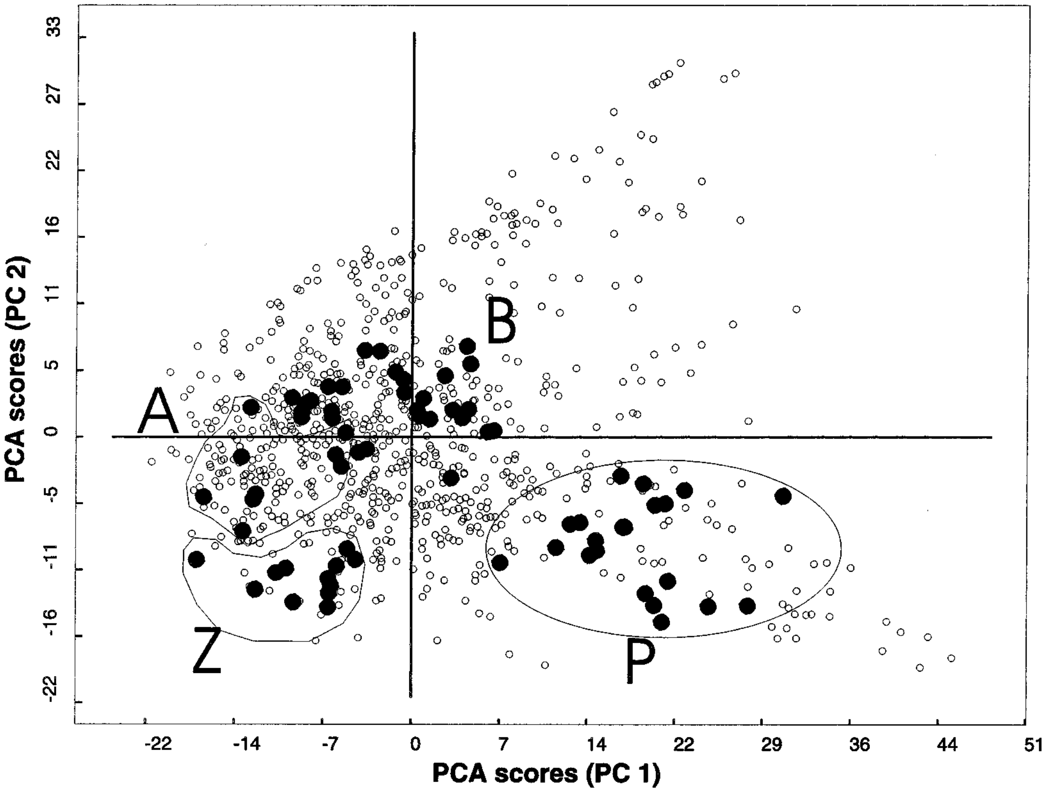

Figure 1. Application of consensus principal component analysis to the entire database using 1771 variables, derived with five descriptor blocks, yielding a global two-component model explaining 20% of the variance. The four chemical classes are nicely clustered. Coding is the following: β-blockers (B), penicillins (P), benzodiazepines (Z), and class I antiarrhythmics (A).

lines the peculiar information inherent in this descriptor type. GRIND descriptors show a compromise between wholistic descriptors (such as VolSurf) and fingerprint descriptors (such as UNITY). It should be pointed out that GRIND descriptors are the only descriptors reported here that depend on the molecular conformation. Although the latter can be a problem in the modeling phase, the results clearly demonstrate that GRIND descriptors are still able to clusterize compounds on the basis of pharmacodynamic properties. Figure 5 shows the molecular interaction field maps obtained with an acceptor and a donor probe for amoxicillin and cyclacillin molecules. The two penicillins differ mainly in the phenol moiety replaced by a cyclohexane moiety in cyclacillin. UNITY and ISIS keys cluster amoxicillin and cyclacillin really close in the PCA space. Conversely, when GRIND descriptors are used, the two penicillins are more separated. The molecular interaction fields (from which the GRIND descriptors are calculated) are really similar for the two molecules. However, the regions produced by the phenolic hydroxyl are not present in cyclacillin and produce a larger potential interaction with a virtual receptor site. Thus, GRIND descriptors tend to favor 3D potential interaction while fingerprint descriptors tend to favor similar patterns in the one- or two-dimensional structure.

log *P* Fragmentation. The rather disappointing results with log *P* as a descriptor initiated us to investigate this parameter in more detail. The idea to perform database mining by PCA of a set of log *P* descriptors stems from a previous paper.[17] The log *P* parameter can be seen as a "latent variable" comprising a variety of molecular properties, and indeed, different calculation methods are able to highlight different molecular properties. Important features of chemical libraries could thus be discovered by combining a few different calculation methods (in our case KOWWIN, CLOGP, HINT, and MOLCAD) as descriptors for the **X** space. The log *P* data, calculated with these four programs, are extremely different in a lot of cases. To investigate the discrepancies between the four calculation methods in a systematic way, we used the following approach. For the entire database of 1107 structures, we calculated a Δ log *P* value by subtracting the average of the remaining three log *P* values from the one individual log *P*, as described in the following equations:

$$\Delta \log P\ (\mathrm{KOWWIN}) = \log P\ (\mathrm{KOWWIN}) - \text{average of } (\log P\ (\mathrm{CLOGP}) + \log P\ (\mathrm{HINT}) + \log P\ (\mathrm{MOLCAD})) \quad (1)$$

$$\Delta \log P\ (\mathrm{CLOGP}) = \log P\ (\mathrm{CLOGP}) - \text{average of } (\log P\ (\mathrm{KOWWIN}) + \log P\ (\mathrm{HINT}) + \log P\ (\mathrm{MOLCAD})) \quad (2)$$

$$\Delta \log P\ (\mathrm{HINT}) = \log P\ (\mathrm{HINT}) - \text{average of } (\log P\ (\mathrm{KOWWIN}) + \log P\ (\mathrm{CLOGP}) + \log P\ (\mathrm{MOLCAD})) \quad (3)$$

$$\Delta \log P\ (\mathrm{MOLCAD}) = \log P\ (\mathrm{MOLCAD}) - \text{average of } (\log P\ (\mathrm{KOWWIN}) + \log P\ (\mathrm{CLOGP}) + \log P\ (\mathrm{HINT})) \quad (4)$$

In a second step, the compounds of the database were fragmented and represented by a sort of fingerprint indicating the presence of a fragment with "1" and its

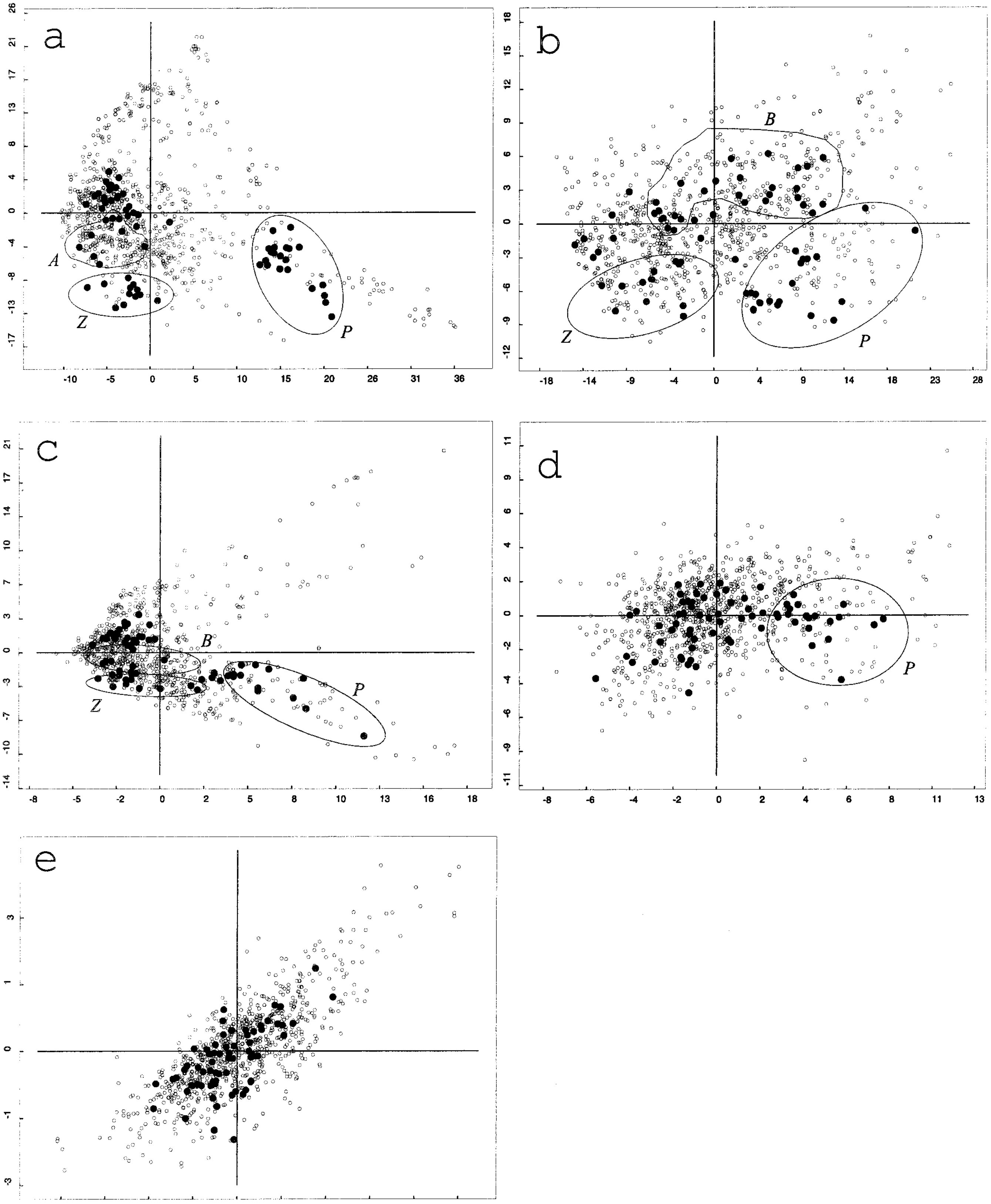

Figure 2. Individual CPCA models with the five descriptors used in this study. Filled points represent members of the chemical classes of β-blockers (B), penicillins (P), benzodiazepines (Z), and class I antiarrhythmics (A). (a) CPCA model with UNITY fingerprints. All four pharmacological classes are well separated. For the sake of graphical clarity, only three classes (A, P, Z) are shown. Thus, UNITY fingerprints mostly contribute to the global separation as shown in Figure 1. (b) CPCA model with GRIND. GRIND descriptors also cluster penicillins, β-blockers, and benzodiazepines, while class I antiarrhythmics are mixed. (c) CPCA model with ISIS keys showing a less evident separation. Penicillins are well clustered, and β-blockers are less well separated and exhibit some overlap with the other classes of compounds. (d) CPCA model with VolSurf descriptors. Only the penicillins are separated, while all the remaining groups are overlapping. (e) CPCA model with log *P* as descriptor. The log *P* parameter completely fails to cluster the four test groups, which are all mixed.

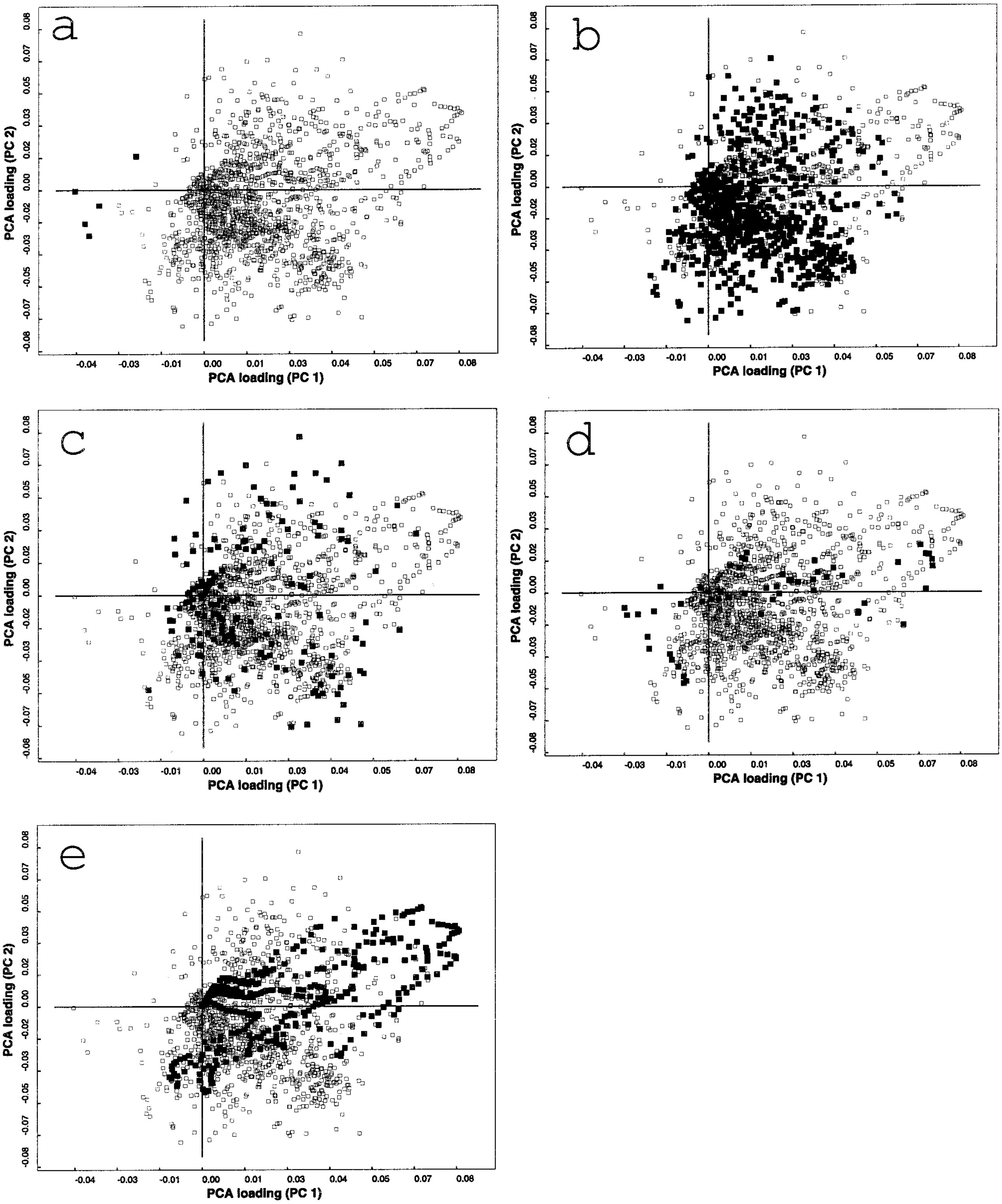

Figure 3. Comparison of the information content of the five descriptors via loading plots. Filled points represent the loadings for the individual descriptor under consideration. (a) Loadings plot for log *P* as descriptors. (b) Loadings plot for ISIS keys. (c) Loadings plot for UNITY fingerprints. (d) Loadings plot for VolSurf descriptors. (e) Loadings plot for GRIND. Comparison of parts a−e elucidates a similarity (corresponding to the pattern between the filled points) between log *P* and VolSurf and between UNITY fingerprints and ISIS keys; GRIND descriptors differ from the other descriptors in a peculiar way.

absence with “0”. In the present work, the fragments considered were extracted from the analysis of the structures included in the database. Via this procedure, a matrix with 900 variables and 1107 objects (database molecules) was obtained. When Δ log *P* is included as a dependent variable, the PLS analysis yields a model explaining 20% of the variance. It means that 20% of the big intermethod variation in log *P* is due to the presence of some fragments that are peculiar. The remaining 80% is due to a misparametrization that has no “structure”. Such an approach allows for identifying single fragments responsible for intermethod log *P*

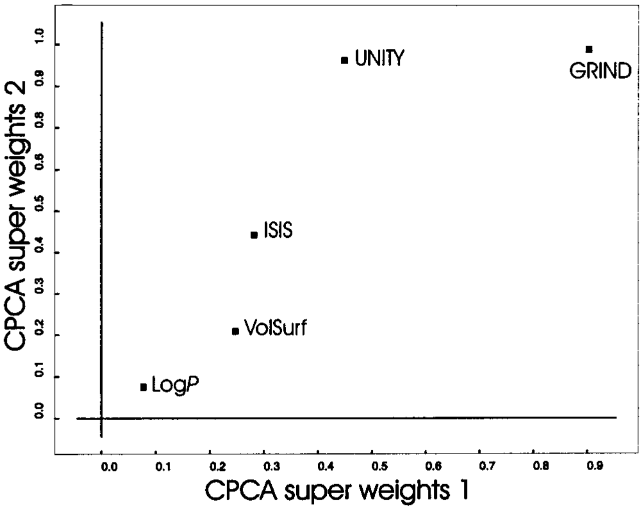

Figure 4. CPCA superweight plot summarizing the information content of the five descriptor blocks. The VolSurf block is located between log *P* and ISIS blocks, UNITY fingerprints and ISIS keys exhibit some similarity, while the peculiar information inherent in the GRIND descriptor type is reflected by its outlier behavior. For evaluating similarity, one has to consider the projection of the block descriptors on both weight axes.

differences. For instance, the parametrization of the diphenylmethyl moiety is responsible for the large differences between HINT and the other three methods in calculating cinnarizine or flunarizine. The presence of a $CO-NH-CH_2-CH(OH)-CH_2OH$ fragment as, for example, in iohexol or iopromide causes large differences ($\Delta \log P \approx 4$) between MOLCAD and the remaining software. An even more dramatic example is bleomycin, for which KOWWIN and CLOGP differ by 6−8 log units from HINT and MOLCAD. Our fragmentation approach allows us to attribute this discrepancy to the presence of a thiazole moiety. Taken together, the big intermethod variations in log *P* cast doubt on the applicability of this descriptor in database mining.

Pharmacokinetic Aspects of Database Mining. Solubility data and blood/brain barrier penetrating behavior serve as examples to investigate pharmacokinetic aspects of database mining.

Aqueous solubility is one of the key factors for the bioavailability of pharmaceutical agents, and approaches to calculate this property are highly demanded. In the dataset used for investigating pharmacodynamic aspects of database mining, solubility data are known to us for only 40 compounds. Thus, we used our own dataset of 100 structures (mixed drugs and nondrugs) with known solubilities for a comparative PLS analysis applying ISIS keys, VolSurf parameters, and GRIND as molecular descriptors. When using the ISIS keys, the PLS analysis yields the following model: with five components the squared correlation coefficient r^2 amounts to 0.84; model validation with the LOO approach gives a cross-validated correlation coefficient q^2 of 0.60. Even if the model is rather good, the standard deviation of the error of prediction (SDEP) is too high (1.40). VolSurf descriptors produce a significantly better model; the corresponding squared correlation coefficient for the five-component model is 0.91, while model validation results in a cross-validated correlation coefficient of 0.83, again applying the LOO approach. It should be noted that the standard deviation of the error of prediction is 0.90 log units in this case, which is half a log unit smaller than in the previous model. To underline the superiority of VolSurf descriptors over ISIS keys in solubility models, score plots from principal component analysis are comparatively shown in Figure 6. In the case of ISIS keys, no discrimination between soluble (black circles) and insoluble compounds (open circles) can be found; soluble and insoluble compounds are evenly distributed in the plot of Figure 6a. In contrast, a significant clustering regarding compound solubility is observed, when using VolSurf descriptors (Figure 6b). Soluble and insoluble compounds are nicely separated, and even the compounds with intermediate solubility (gray circles) are clustered between the highly soluble and the insoluble structures. GRIND descriptors produce a model with a squared correlation coefficient for the five-component model of 0.77, while model validation results in a cross-validated correlation coefficient of 0.65. In this latter case, the standard deviation of the error of prediction is 1.20 log units. GRIND descriptors are closer to ISIS keys than to VolSurf in the solubility dataset.

The second pharmacokinetic example is dedicated to blood/brain barrier penetrating behavior. Central nervous system (CNS) active compounds need to be able to permeate the blood/brain barrier, whereas for peripherally acting drugs blood/brain barrier penetration should be kept as low as possible to avoid CNS-related side effects. In any case, permeability of the blood/brain barrier is always a key parameter in pharmacokinetic drug profiling. Experimental approaches for this parameter are complicated and not well suited in the investigation of large databases. A fast and reliable computational method for predicting blood/brain barrier

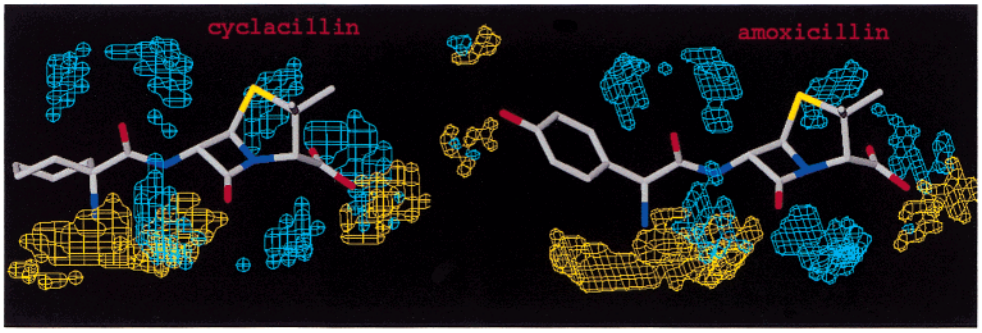

Figure 5. Molecular interaction field maps for amoxicillin (right part) and cyclacillin (left part) obtained with a donor NH (blue regions) and an acceptor carbonyl CO (yellow regions) probe. The 3D pattern of the molecular interactions is really similar, but amoxicillin shows longer interaction distances because of the hydroxy phenol group.

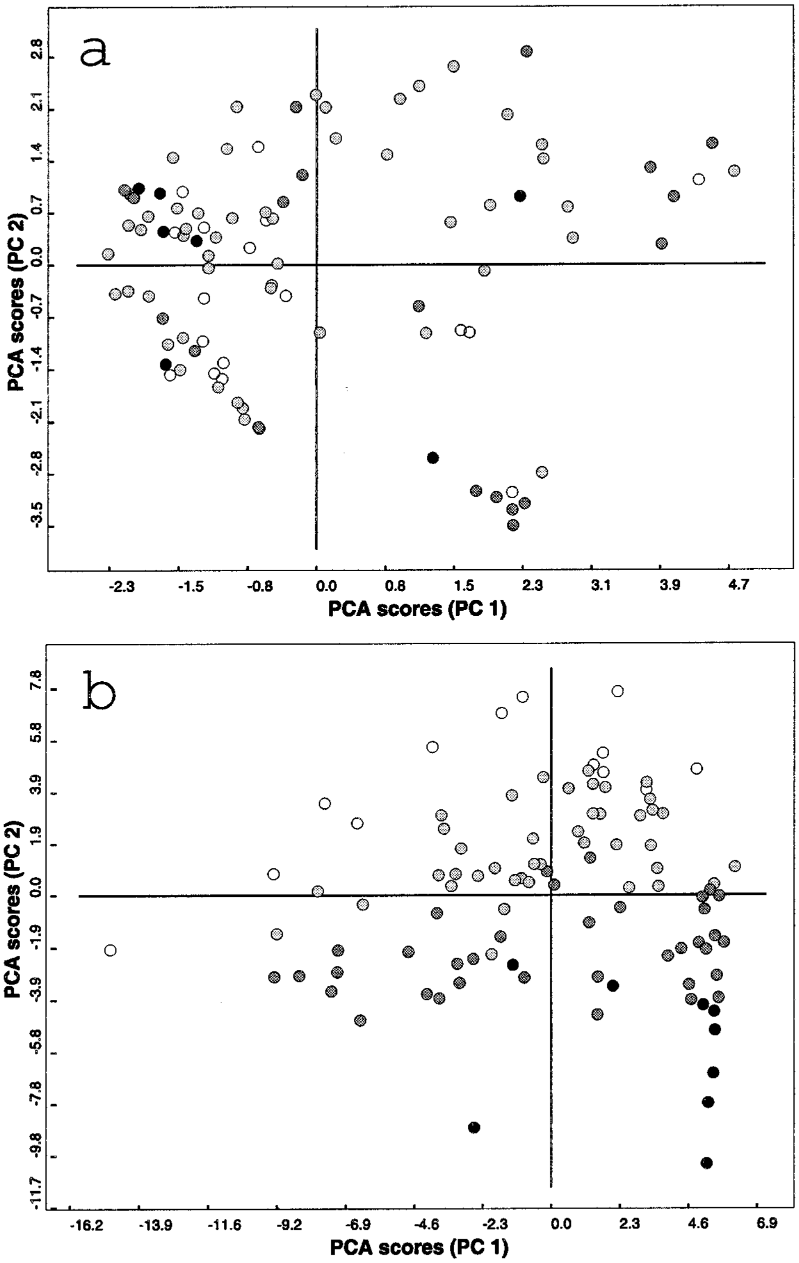

Figure 6. Comparison of score plots from principal component analysis for VolSurf descriptors and ISIS keys in solubility models. Black circles indicate soluble compounds, open circles indicate insoluble compounds, and gray circles correspond to a solubility between the two extremes. Whereas in the case of ISIS keys soluble and insoluble compounds are mixed (a), a clear separation between the two groups is observed when using VolSurf descriptors (b).

permeability would be a significant progress for drug development.

A step in this direction was published by Cruciani et al.[18] These authors used VolSurf descriptors to produce a simple model suitable for the external prediction of blood/brain barrier permeability. The database used includes 229 drugs with a well-defined brain penetration profile. For these compounds an **X** matrix comprising 72 VolSurf descriptors was calculated and discriminant PLS was used for chemometric analysis, thus assigning a score of 1 to the blood/brain barrier permeating drugs and a score of -1 to the nonpermeating drugs. From PLS modeling and cross-validation, two significant latent variables were obtained and the model assigned the correct permeation profile to more than 90% of the compounds in the database. Figure 7a demonstrates clearly the significant separation between blood/brain barrier permeating and nonpermeating drug via the scores plot of principal component analysis. Taken together, the results summarized here demonstrate that it is possible to predict the blood/brain barrier permeation of putative drugs from their three-dimensional structures using VolSurf descriptors. A further advantage of this approach is that it allows for quantifying the favorable and unfavorable contributions of physicochemical and structural properties to blood/brain barrier permeation. Therefore, the method allows for manipulating the structures in a rational way in order to improve the brain permeation properties of drug candidates under development.

For the sake of comparison, results of a corresponding score plot from principal component analysis with the

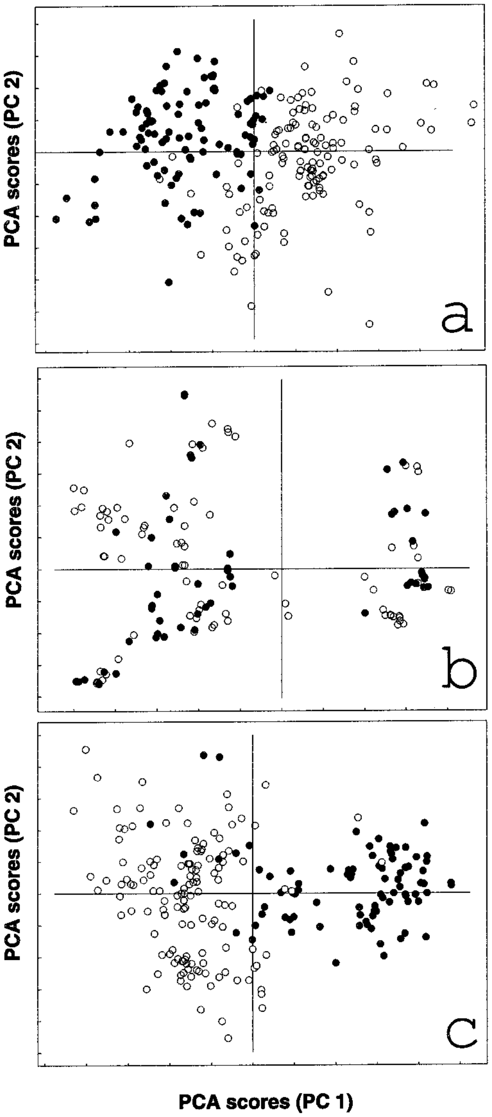

Figure 7. Comparison of score plots from principal component analysis for VolSurf descriptors, UNITY fingerprints, and GRIND descriptors in models of blood/brain barrier permeating behavior. Black circles indicate permeating compounds, and open circles indicate nonpermeating compounds. A clear clustering of the groups of permeating and nonpermeating compounds is observed when using VolSurf descriptors (a). In the case of UNITY fingerprints, the groups of permeating and nonpermeating compounds are mixed (b). A clear clustering is also obtained using GRIND descriptors (c).

same dataset, but using UNITY instead of VolSurf descriptors, are given in Figure 7b. The mixed pattern of blood/brain barrier permeating and nonpermeating drugs underlines the superiority of VolSurf descriptors for pharmacokinetic aspects of database mining. When GRIND descriptors are used, the PCA model, reported in Figure 7c, clearly demonstrates the nice performance of these kinds of descriptors.

Concluding Remarks

From this comparative analysis, we conclude that UNITY fingerprints, ISIS keys, and GRIND descriptors are of special value for tackling pharmacodynamic aspects of database mining. For the test classes of β-blockers, benzodiazepines, penicillins, and class I antiarrhythmics, the following ranking of descriptor suitability was observed: UNITY fingerprints > GRIND ≥ ISIS keys > VolSurf descriptors > log *P*. VolSurf descriptors exhibit particular advantages in treating pharmacokinetic aspects. For such important aspects of pharmacokinetic profiling as solubility and blood/brain barrier permeating behavior, we could show a significant superiority of VolSurf descriptors over ISIS keys. The log *P* parameter is of limited applicability in database mining because of rather poor reliability and lack of completeness of data.

Acknowledgment. Raimund Mannhold expresses his gratitude for cooperation and hospitality to Sergio Clementi and Gabriele Cruciani, Laboratory of Chemometrics, University of Perugia, Italy. The database with 1007 compounds was generously supplied by Joachim März. We thank the following colleagues for providing the log *P* data: Glenn Kellogg (HINT); Robert Jäger and Jochen Brickmann (MOLCAD); Ute Abraham and Hugo Kubinyi (CLOGP). Wolfgang Guba and Hans Matter are thanked for the supply of ISIS keys and UNITY fingerprints. Massimo Baroni is thanked for the work on fragmentation of log *P*.

Supporting Information Available: Composition of the databases together with the SMILES codes of the individual compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Brown, R. D.; Martin, Y. C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
(2) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31–49.
(3) Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule approaches for calculating log *P*. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337–354.
(4) Leo, A. J.; Jow, P. Y. C.; Silipo, C.; Hansch, C. Calculation of hydrophobic constant (log *P*) from *x*- and *f*-constants. *J. Med. Chem.* **1975**, *18*, 865–868.
(5) Hansch, C.; Leo, A. J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley: New York, 1979.
(6) Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol–water partition coefficients. *J. Pharm. Sci* **1995**, *84*, 83–92.
(7) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
(8) Brickmann, J.; Waldherr-Teschner, M. Interaktive Computergraphik und Molekülmodellierung (Interactive Computer Graphics and Molecular Modeling). *Informationstechnik* **1991**, *33*, 83–90.
(9) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545–552.
(10) *UNITY Chemical Information Software*, version 2.4; Tripos (1699 S. Hanley Road, St. Louis, MO 63144).
(11) *ISIS/Base*, version 2.1.3; Molecular Design Ltd. (14600 Catalina Street, Irvine, CA 92714).
(12) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular Fields in Quantitative Structure–Permeation Relationships: The VolSurf Approach. *THEOCHEM* **2000**, *503, 17–30.*
(13) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid Independent Descriptors (GRIND). A Novel Class of Alignment-Independent Three-Dimensional Descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3242.

(14) Westerhuis, J. A.; Kourti, T.; Macgregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **1998**, *12*, 301−321.

(15) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.−Act. Relat.* **1993**, *12*, 9−20.

(16) Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. J.; Fox, T. GRID/PCA: A New Computational Tool To Design Selective Ligands. *J. Med. Chem.* **2000**, *43*, 3033−3044.

(17) Mannhold, R.; Cruciani, G.; Dross, K.; Rekker, R. Multivariate analysis of experimental and computational descriptors of molecular lipophilicity. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 573−581.

(18) Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting Blood−Brain Barrier Permeation from Three-Dimensional Molecular Structure. *J. Med. Chem.* **2000**, *43*, 2204−2216.

JM0011326