

# Articles

## Pharmacokinetically Based Mapping Device for Chemical Space Navigation

Tudor I. Oprea,<sup>\*,†</sup> Ismael Zamora,<sup>‡,§</sup> and Anna-Lena Ungell<sup>‡</sup>

*EST Chemical Computing, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden, and DMPK & Bioanalytical Chemistry, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden*

*Received December 20, 2001*

ChemGPS, the chemical global positioning system, is a tool that combines rules (equivalent to dimensions) and objects (chemical structures) to provide a consistent chemical space map (Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157–166.). Rules included, initially, general properties such as size, lipophilicity, and hydrogen bond capacity, while objects include “satellites”, intentionally placed outside the druglike space, as well as “core” objects, mostly orally available drugs. ChemGPS molecules (objects) were used in conjunction with the VolSurf (<http://www.moldiscovery.com>) descriptors (rules), which are relevant for ADME (absorption, distribution, metabolism, and excretion) properties. The combination of ChemGPS and VolSurf, GPSVS, was investigated with respect to the biopharmaceutics classification system, which is recommended by the Food and Drug Administration (FDA) ([http://www.fda.gov/cder/OPS/BCS\\_guidance.htm](http://www.fda.gov/cder/OPS/BCS_guidance.htm)), in particular with respect to permeability and solubility. The first GPSVS principal component correlates, with no further training, to passive transcellular permeability, as illustrated for the Caco-2, ghost erythrocyte, and blood–brain barrier datasets, respectively. The second GPSVS principal component correlates, without prior training, to solubility, as shown for the octanol–water partition and intrinsic solubility datasets, respectively. Although derived from principal component analysis, the two property axes rotate and form an angle of approximately 43°, thus being no longer orthogonal. GPSVS can be used to map the chemical space with respect to permeability and solubility, as recommended by FDA’s biopharmaceutics classification system.

### Introduction

Because of the increased number of compounds available from combinatorial and parallel synthesis,<sup>1</sup> there is a growing demand for methods that predict absorption, distribution, metabolism, and excretion (ADME) behavior in humans. The need to evaluate, early on, both permeability and solubility has recently been illustrated by Lipinski,<sup>2</sup> who analyzed the trends in two sets of compounds from Merck and Pfizer. The “rational design approach” at Merck seems to lead to clinical candidates with poorer permeability, whereas the “HTS approach” at Pfizer appears to result in clinical candidates with poorer solubility.<sup>2</sup> Since poor permeability and poor solubility are among the main reasons for failure in clinical trials, it has become apparent that awareness of these pitfalls should be introduced as early as possible in the lead discovery process.<sup>3</sup>

A decade ago, it was recognized that the fraction absorbed of a drug in the gastrointestinal tract could be predicted and estimated from in vitro systems such as the measurement of permeation through Caco-2 cell monolayer,<sup>4</sup> e.g., via permeability coefficients. However, this experimental technique requires high-purity soluble compounds in order to successfully assess permeability. Therefore, synthesis prioritization of compounds, or combinatorial libraries, needs to be performed not only by judging molecular diversity<sup>5–7</sup> but more often by considering ADME properties as well.<sup>8–10</sup> Thus, there is an increasing demand for fast and accurate predictions of pharmacokinetic properties.

However, the process of combinatorial library design and evaluation<sup>11</sup> is far from linear, and the decision of applying ADME property filters needs to be carefully balanced. Usually, a wide range of molecular descriptors<sup>12</sup> are evaluated in reactant and/or product space<sup>13–15</sup> prior to compound selection, with or without enumeration.<sup>14,16,17</sup> Other criteria, e.g., druglike<sup>18,19</sup> or leadlike<sup>20,21</sup> properties, followed by statistical analysis, e.g., via PCA<sup>22</sup> (principal component analysis) and design of experiments,<sup>23,24</sup> are used to limit the range of possibilities prior to synthesis. The choice of

\* To whom correspondence should be addressed. Phone: +46 31 776 2373. Fax: +46 31 776 3792. E-mail: Tudor.Oprea@astrazeneca.com.

† EST Chemical Computing.

‡ DMPK & Bioanalytical Chemistry.

§ Present address: Lead Molecular Discovery slt Francesc Cabanes i Alibau, 1–2 2<sup>o</sup>-1<sup>a</sup>, Sant Cugat del Valles, 08190 Barcelona, Spain. E-mail: Ismael.Zamora@telefonica.net.

**Table 1.** Description of VolSurf Descriptors

VolSurf code	definition
V	volume: total volume (computed at 0.25 kcal/mol)
S	surface: total surface (computed at 0.25 kcal/mol)
R	rugosity: volume/surface
G	globularity: surface of the compound divided by the surface of a sphere with the same volume
W1–W8	volume of interaction with the H <sub>2</sub> O probe at –0.2, –0.5, –1.0, –2.0, –3.0, –4.0, –5.0, and –6.0 kcal/mol levels
IW1–IW8	integy moment: proportional to the distance between the baricenter of the surface and the volume of interactions with the H <sub>2</sub> O probe at the different energy levels
CW1–CW8	capacity factor: volume of interaction with the H <sub>2</sub> O probe divided by the surface
Min1–Min3	energy minima: the first three energy minima interactions
D12, D13, D23	distance: the distances between the energy minima
D1–D8	volume of interaction with the dry probe at –0.2, –0.4, –0.6, –0.8, –1.0, –1.2, –1.4, and –1.6 kcal/mol levels
ID1–ID8	integy moment: proportional to the distance between the baricenter of the surface and the volume of interactions with the dry probe at the different energy levels
A	amphiphilic moment
CP	critical packing
HL1, HL2	balances of the hydrophilic–hydrophobic interactions
Wp1–Wp8	volume of interaction with the O probe at –0.2, –0.5, –1.0, –2.0, –3.0, –4.0, –5.0, and –6.0 kcal/mol levels
HB1–HB8	H-bond interaction at –0.2, –0.5, –1.0, –2.0, –3.0, –4.0, –5.0, and –6.0 kcal/mol levels
POL	molecular polarizability
MW	molecular weight

descriptors, reagents, and products is quite likely to have a significant influence on the compounds suggested for synthesis, and the introduction of ADME property filters will add to the complexity of the analyses; in most cases, local models will be developed for each particular problem, with a limited range of predictivity, unless one resorts to a uniform metric<sup>7,11</sup> for chemical space.

We have recently proposed ChemGPS (chemical global positioning system) as a tool<sup>25–27</sup> for providing a consistent map of the chemical space based on a set of rules (i.e., chemical space dimensions related to the medicinal chemistry space) and a set of objects (i.e., molecules of interest). The 423 ChemGPS objects consist of a set of “satellite” structures, intentionally placed outside the medicinal chemistry space, and a set of representative (“core”) structures, consisting mostly of orally available drugs. We have previously shown<sup>26</sup> the ability of ChemGPS to provide a global chemical space map by performing extensive comparisons with GRID-based<sup>28</sup> principal properties for heteroaromatic compounds<sup>29</sup> and principal properties (“z-scores”) of  $\alpha$ -amino acids,<sup>30</sup> as well as by comparison to locally derived PCA models. When 72 descriptors computed with the SaSA<sup>31</sup> and HYBOT<sup>32</sup> programs were used, a nine-dimensional ChemGPS map was derived from PCA *t*-scores, as previously described.<sup>26</sup>

Two of the ADME properties, passive permeability and solubility, have been used by the FDA (Food and Drug Administration) in their biopharmaceutics classification system, BCS,<sup>33</sup> as a guide for the *in vivo* bioavailability and bioequivalence studies for immediate-release solid oral dosage forms. Our work is therefore focused on the prediction of these two key properties, using the 423 ChemGPS objects as a reference system, in conjunction with the VolSurf<sup>34,35</sup> set of descriptors. On the basis of GRID<sup>28</sup>-calculated molecular interaction fields (MIFs), the VolSurf descriptors were previously shown to provide significant models for passive permeability in the intestine, as well as

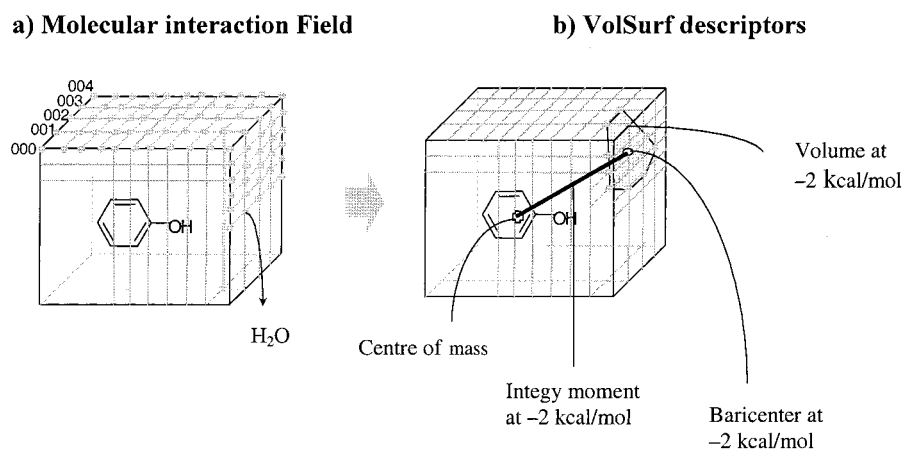
in the blood–brain barrier, BBB.<sup>36–38</sup> We show that GPSVS, the combination of ChemGPS objects<sup>26</sup> with VolSurf descriptors,<sup>35</sup> can be used to map the chemical space with respect to permeability and solubility, as recommended by FDA’s biopharmaceutics classification system. GPSVS could therefore become useful in the planning stages of combinatorial and medicinal chemistry synthesis to avoid those compounds that display poor permeability and poor solubility.

## Materials and Methods

**A. Data Sets.** VolSurf descriptors were calculated on six different sets: the ChemGPS training set,<sup>26</sup> permeability across the BBB,<sup>37</sup> passive transcellular permeability across the Caco-2 cell monolayer, passive transcellular permeability across erythrocyte ghost cells,<sup>35</sup> lipophilicity ( $\log P_{ow}$ ),<sup>39</sup> and solubility.<sup>40</sup>

**B. Molecular Modeling.** The ChemGPS training set (423 objects) was converted to 3D structures using CONCORD,<sup>41</sup> without further processing. When the 2D–3D conversions failed, molecules were built starting from X-ray structures or were modeled from parent compounds in SYBYL.<sup>42</sup> VolSurf descriptors, summarized in Table 1, were obtained in a two-step process (Figure 1): (a) The molecular interaction fields for the H<sub>2</sub>O, DRY and O probes were computed with GRID. (b) The resulting MIFs were analyzed, and the 72 VolSurf descriptors, related to the surface, volume, group distribution, and the relationships between them, were obtained (Table 1).

VolSurf descriptors provide directly interpretable maps for the hydrogen bond acceptor interactions (O probe), for the hydrogen bond donor interactions<sup>43</sup> (fields from the H<sub>2</sub>O probe, from which the fields derived with the O probe are subtracted), and for the hydrophobic interactions (DRY probe). Latent variables were extracted by applying PCA to the VolSurf descriptors set.



**Figure 1.** Two-step process of computing VolSurf descriptors: (a) GRID molecular interactions field (MIF) calculation; (b) descriptor derivation.

SaSA is an in-house program that calculates 72 descriptors starting from the 2D representation of the molecule. Size-related descriptors included molecular weight (MW), the number of heavy atoms, the number of carbons, and the calculated molecular refractivity (CMR).<sup>44</sup> Polarizability is estimated by CMR and by an atom-based polarizability scheme.<sup>45</sup> Flexibility and rigidity are estimated by counting the total number of bonds and rings (RNG), the number of rotatable bonds (RTB), and the number of rigid bonds (RGB)<sup>46</sup> and by several topological indices that estimate other properties<sup>47</sup> as well (e.g., size). The Wiener, Balaban, Randic, and Motoc indices, as well as the Kier and Hall suite of topological descriptors,<sup>48</sup> are used in SaSA. Hydrogen-bonding capacity is estimated using four HYBOT<sup>32</sup> descriptors: the maximum free energy H-bond donor factor ( $C_d$ ), the sum of  $C_d$  values, the maximum free energy H-bond acceptor factor ( $C_a$ ), and the sum of  $C_a$  values. All  $C_d$  values were given a positive sign, as previously suggested.<sup>49</sup> In addition, we use the simple count of oxygens, nitrogens, H-bond donors (HDO), and H-bond acceptors (HAC), as implemented in SaSA. Charge is estimated by counting the positive (N\_POS) and negative (N\_NEG) ionization centers, as well as the maximum positive and negative charge, as calculated using the Gasteiger–Marsili method.<sup>50</sup> Lipophilicity is estimated with two methods, CLOGP<sup>51</sup> and ACD-LogP.<sup>52</sup> Both calculate the logarithm of the octanol–water partition coefficient,  $\log P$ .<sup>53</sup>

**C. Statistics.** All statistical analyses, including PCA and PLS (partial least squares),<sup>54,55</sup> were performed using the SIMCA package.<sup>56</sup> To yield descriptor columns with 0 average and 1 as standard deviation, an autoscaling pretreatment of the variables was performed prior to any analysis. PCA models were obtained for the ChemGPS training set, as well as for the different validation sets. The ChemGPS model was then used to estimate PCA  $t$ -scores on the basis of the coordinate transformation matrix derived from the reference set. This procedure, also termed<sup>56</sup> PCA prediction, was previously described for ChemGPS score estimation starting from 2D descriptors.<sup>26</sup>

ChemGPS scores obtained using the VolSurf descriptors were compared to those obtained from the SaSA and HYBOT for the training set (423 compounds). Furthermore,

the  $t$ -scores derived via PCA prediction were compared to the ones derived from local PCA models for the same sets. PLS models were derived whenever experimental data were available, with the same pretreated VolSurf descriptors. PLS coefficients obtained for each of the datasets were then compared to the PCA loadings predicted using the ChemGPS training set.

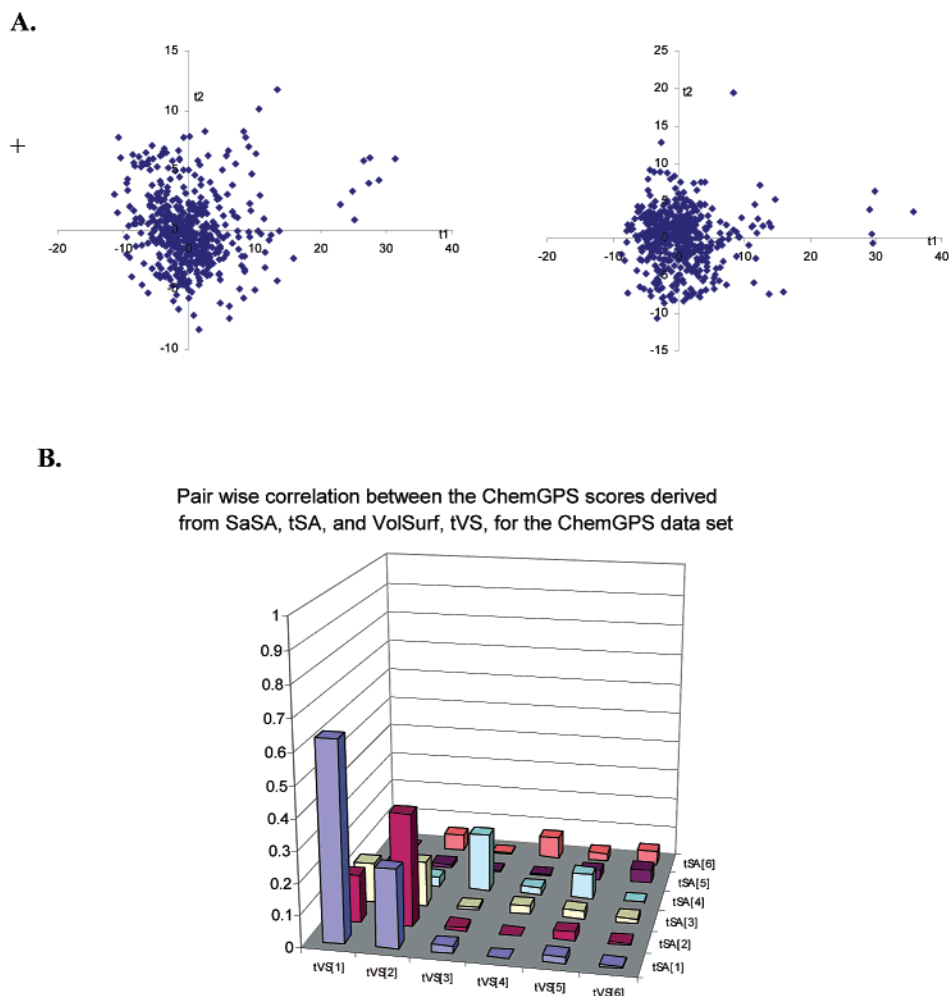
**D. Comparative Analysis of the PCA  $t$ -Scores.** Principal component axes derived from local models are often rotated when compared to other PC axes from local or global models, as previously discussed.<sup>26</sup> To ascertain that the maximum correlation coefficient has been found between the local and global PCA  $t$ -scores, we used the following procedure:

$$t_1 = t'_1 \cos \alpha + t'_2 \sin \alpha \quad (1)$$

When  $t_1$  is completely aligned to  $t'_1$ ,  $\alpha$  equals  $0^\circ$  and the  $t$ -scores are well correlated. Whenever rotation occurs,  $\alpha$  is not  $0^\circ$ , implying that the information extracted by the two models, though it may be identical, needs to be reoriented for better comparison, as performed in, for example, factor analysis.<sup>57</sup>

## Results and Discussion

**Comparison of ChemGPS Scores Derived from SaSA and VolSurf Descriptors.** The PCA score plots for the first and second components, based on the two sets of descriptors and the ChemGPS set of objects (423 compounds), are shown in Figure 2A. These two reference systems contain different maps and clusters. The first component has a correlation index close to 0.6 in the pairwise comparison (Figure 2B), a reflection of the importance of molecular size in both descriptor sets. The remaining latent variables do not show a significant correlation. This is not surprising, since VolSurf is designed to cover ADME-related properties and is computed from 3D structures, whereas SaSA features a rather diverse set of 2D-based molecular descriptors designed to cover chemical variability in property and topology space. These results indicate that the information extracted by ChemGPS/SaSA and ChemGPS/VolSurf (GPSVS) is complementary, not redundant, and is potentially useful when planning combinatorial libraries.



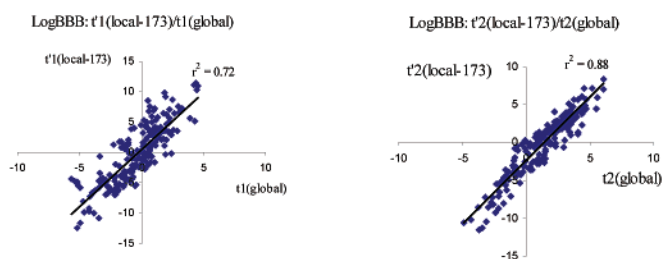
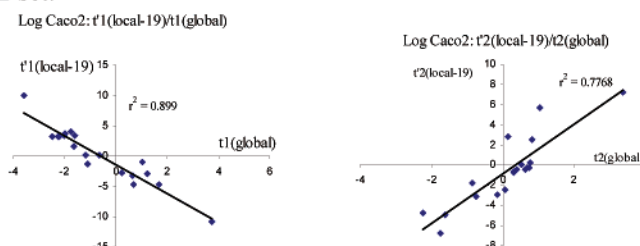
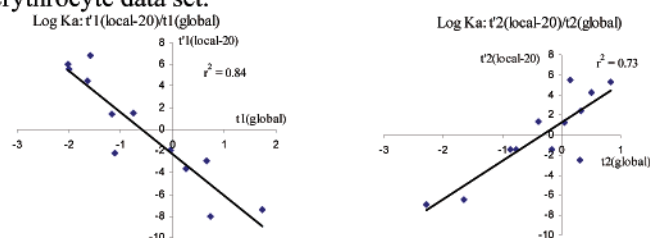
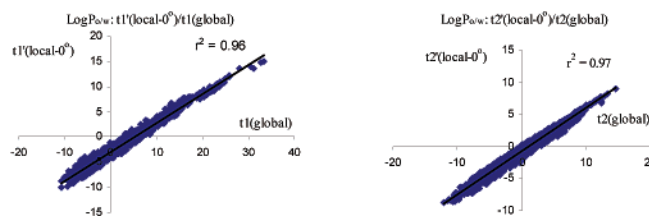
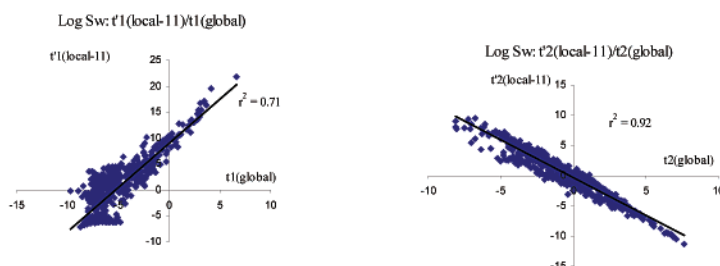
**Figure 2.** (A) PCA score plots derived from SaSA (left) and VolSurf (right) descriptors. (B) Pairwise correlation between SaSA and VolSurf principal components.

**GPSVS, a Global Model.** One of the primary objectives when designing the ChemGPS system was to avoid extrapolation<sup>26</sup> in the positioning of a novel compound in the druglike or leadlike chemical space. When combining ChemGPS with VolSurf, we aimed at preserving the ability to provide a global model geared for pharmacokinetically relevant properties. To assess the predictive power of the GPSVS system, *t*-scores derived from the ChemGPS training set via PCA-prediction, referred to as GPSVS scores, were compared to the *t*-scores obtained from the local models of the five validation sets. The advantage of deriving a global model (applicable across chemistries) is evident if one considers the pitfalls of local models; if all the molecules were monocarboxylic acids, this information would not be captured by the local model but would be highlighted in a global (chemically diverse) model. This often manifests itself as a principal component perturbation, or rotation, and it may even lead to interdimensional swapping. For example, PC1 in model A (global) may well correlate with PC3, but not with PC1, in model B (local), while PC2 in both models could be directly correlated.

The local vs global relationships were scrutinized for the five external data sets, stressing the highest correlation between GPSVS predicted scores and the values obtained from local models via eq 1. In all data sets, the fraction of

explained variance,  $r^2$ , is higher than 0.7 for both PC1 and PC2 and higher than 0.8 for at least one of the two components (see Figure 3). This indicates a good predictive power for the GPSVS model, as well as some particularities of the local models. A rotation of the initial GPSVS scores was required for all models to achieve the maximum correlation. The predictivity of GPSVS is explained in part by the good chemical structural span of the ChemGPS training set, which appears to cover the space described by the local test sets.

GPSVS does not have the directionality (sign) problem that occurs in local PCA models, since the rotations present in Figure 3, i.e., positive or negative slopes, do not occur. This is due to the conventional design of the ChemGPS system that derives PCA scores in a uniform manner, i.e., via prediction. Furthermore, the property correlations discussed below, which establish the relationship between the GPSVS PC1 and permeability on one hand and between the GPSVS PC2 and solubility on the other hand, were obtained with no a priori input of either permeability or solubility data. Rather, the relationships to experimental and theoretical data for these two properties were established a posteriori. This makes GPSVS a general model suitable for direct

**A. Blood-brain barrier set.****B. Caco-2 set.****C. Ghost erythrocyte data set.****D. Octanol/water partition (LogPo/w) set.****E. Solubility (LogSw) set.**

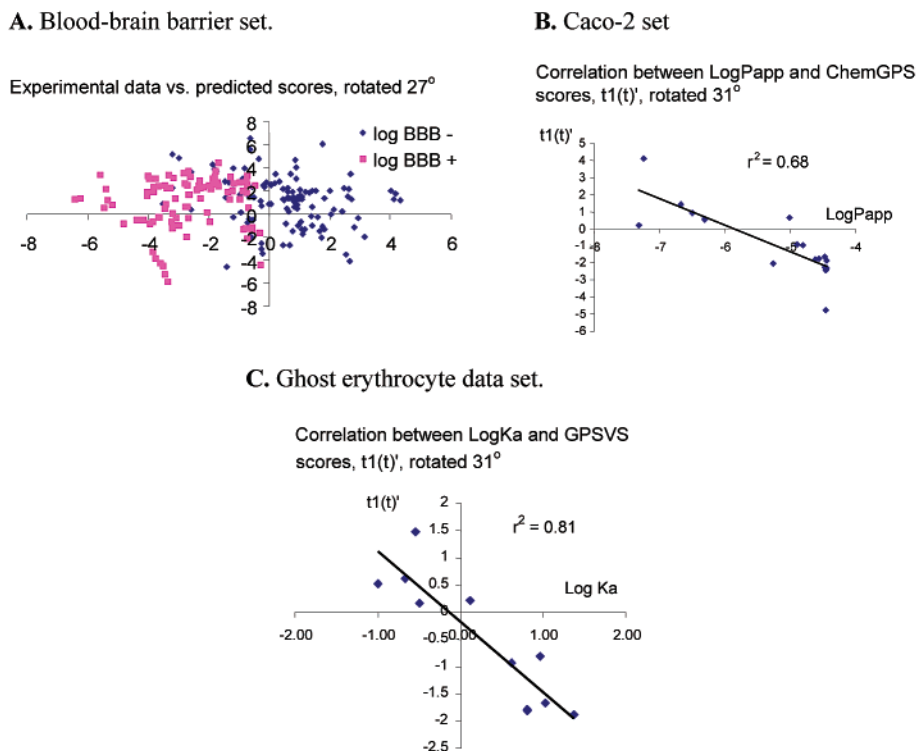
**Figure 3.** Intercomponent relationships between the GPSVS scores (global model) and the local  $t$ -scores for the first two components in each validation set.

interpretation, since the chemical space maps derived with this system apply the same convention in a consistent manner.

**GPSVS and Passive Permeability.** The passive transcellular mechanism of absorption across the epithelial tissue in the gastrointestinal tract or across the BBB was modeled using several different techniques. Three data sets were analyzed in this case: permeability across the BBB, perme-

ability across a Caco-2 cell monolayer, and permeability across erythrocyte ghost cells.

VolSurf descriptors have successfully been used in predicting the brain penetration of 227 compounds<sup>37</sup> by using PLS discriminant analyses (PLS-DA). A local PCA model was shown to produce results similar to results from PLS-DA.<sup>37</sup> PCA scores for the same 227 compounds were obtained using GPSVS. A correlation analysis was performed



**Figure 4.** Correlation between the biological activity and the scores predicted by the GPSVS system: (A) for the blood–brain barrier; (B) for the permeability across the Caco-2 cell monolayer; (C) for the permeability in the ghost erythrocytes cells.

between the PLS pseudocoefficients and the loadings for the different components. The rotated PC1 correlates well with the PLS pseudocoefficients ( $r^2 = 0.86$ ), indicating that both models extract the same information (data not shown). Furthermore, the two-dimensional GPSVS score plot shown in Figure 4A illustrates a direct relationship to the BBB classification scheme,<sup>37</sup> suggesting that GPSVS scores can serve as a simple classifier for passive BBB permeability.

We have previously described<sup>38</sup> a PLS-based VolSurf model for passive transcellular permeability across a Caco-2 cells monolayer. PLS pseudocoefficients obtained from this model were compared to the GPSVS loadings. The resulting correlation ( $r^2 = 0.79$ ) indicates that both methods extract similar information (data not shown). GPSVS scores have a direct relationship to Caco-2 permeability ( $r^2 = 0.67$ ), as illustrated in Figure 4B. Thus, in the absence of any a priori input, GPSVS scores correlate with Caco-2 permeability data in a manner that is comparable to those of the PLS model based on VolSurf descriptors for the same 22 compounds<sup>38</sup> ( $q^2 = 0.79$ ). Similar analyses were performed for the permeability data obtained from experiments across ghost erythrocytes. The correlation between the PLS pseudocoefficients and the loading for the GPSVS PC1 component ( $r^2 = 0.61$ , data not shown) indicates that GPSVS scores are correlated to this measure of passive permeability as well ( $r^2 = 0.81$ ; see Figure 4C).

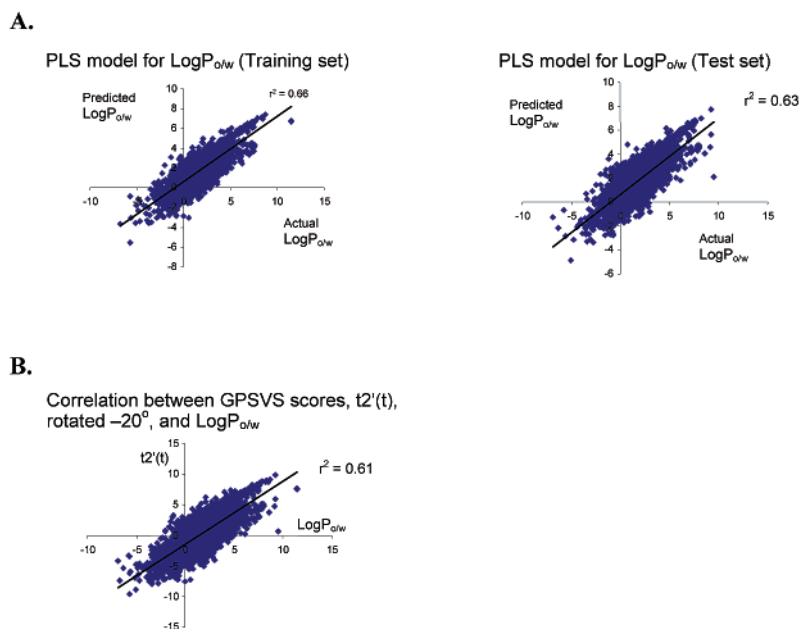
**Solubility-Related Models.** Most of the efforts to predict solubility have been focused on its relationship with  $\log P_{o/w}$  (the logarithm of the octanol–water partition coefficient).<sup>53</sup> We have used the LogPstar dataset, a collection of 7954 compounds from the Pomona Masterfile.<sup>39</sup> A significant PLS model, judged by its external predictivity ( $q^2_{\text{cross-validated}} = 0.66$ ;  $r^2_{\text{external}} = 0.63$ ), was derived (see Figure 5A). The PLS

pseudocoefficients for this model correlate well ( $r^2 = 0.97$ , data not shown) with the PC2 loadings from GPSVS. In the absence of any input related to  $\log P_{o/w}$ , GPSVS PC2 scores correlate with the 7954 measured  $\log P_{o/w}$  values ( $r^2 = 0.61$ ) (see Figure 5B). The GPSVS predictivity,  $r^2 = 0.61$ , is quite similar to the external predictivity of the PLS model,  $r^2_{\text{external}} = 0.63$ . While this may indicate the limited ability of VolSurf descriptors to derive better  $\log P_{o/w}$  models, it also illustrates the general predictivity of GPSVS with respect to this property.

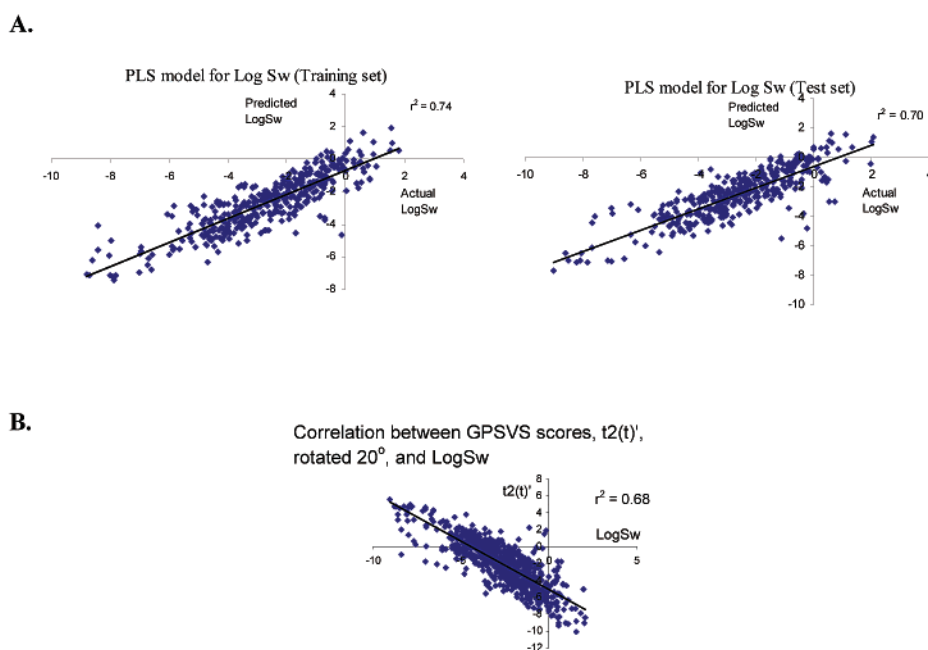
Intrinsic solubility,  $\log S_w$ , values<sup>40</sup> were also compared to GPSVS scores for 794 compounds. The data set was divided into a training set (404 compounds) and a test set (390 compounds). A significant PLS model, judged by its external predictivity ( $q^2_{\text{cross-validated}} = 0.70$ ;  $r^2_{\text{external}} = 0.70$ ), was derived (see Figure 6A). The PLS pseudocoefficients correlate well with the PC2 loadings from GPSVS ( $r^2 = 0.85$ ; data not shown). Without any input related to aqueous solubility, GPSVS PC2 scores correlate directly with the 794  $\log S_w$  values ( $r^2 = 0.68$ ) (see Figure 6B). The GPSVS predictivity,  $r^2 = 0.68$ , is quite similar to the external predictivity of the PLS model,  $r^2_{\text{external}} = 0.70$ , indicating perhaps not only the limited ability of VolSurf descriptors to derive better  $\log S_w$  models but also the ability of GPSVS to estimate solubility trends in a general manner.

## Conclusions

The importance of early drug discovery awareness regarding passive transcellular permeability and solubility has been previously stressed.<sup>2</sup> This paper documents, with multiple datasets derived from independent experiments, that GPSVS scores relate to observed permeability (PC1) and solubility (PC2) values, without prior input of experimental data.



**Figure 5.** (A) PLS model for the LogPstar dataset: actual vs estimated values for the 3867 compounds in the training set (left) and actual vs predicted values for the 4069 compounds in the test set. Eighteen compounds were excluded because of problems in the VolSurf descriptor calculation step. (B) The correlation between GPSVS PC2 scores and measured  $\log P_{\text{o/w}}$  values.

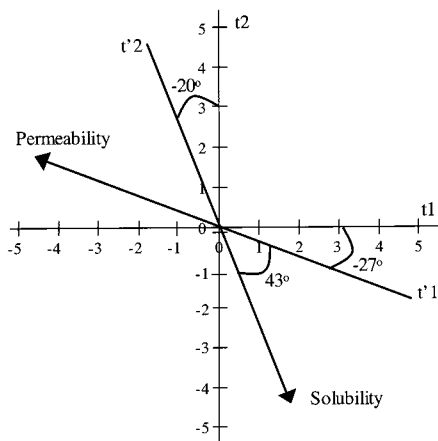


**Figure 6.** (A) PLS model for the solubility dataset: actual vs estimated values for the 404 compounds in the training set (left) and actual vs predicted values for the 390 compounds in the test set. (B) The correlation between the GPSVS PC2 scores and measured  $\log S_w$ .

Usually, PCA models are interpreted by comparing the loadings of various descriptors in the X (descriptor) block, which then relate latent variables to physical meaning. However, in GPSVS, the first two PCs are directly related to measured properties (e.g., Caco-2 permeability and solubility), an observation that is further substantiated by comparison to PLS models derived for the same datasets. While useful in estimating trends, GPSVS is not suitable for exact estimates of permeability and solubility; even though GPSVS PC1 correlates with passive permeability, other routes of absorption, e.g., active transport and efflux mechanisms, are not likely to be related to this score. And while GPSVS PC2 can generally be regarded as a trend for

(intrinsic) aqueous solubility, one does not expect an exact relationship between this score and other measures of solubility, e.g., those related to different crystallization forms of the same compound.

Whereas principal components are orthogonal by definition, the GPSVS scores relating to permeability and solubility are not. Their rotation amounts to an angle of approximately  $43^\circ$  when PC1 and PC2 are plotted together (see Figure 7). Even though the angle between the two GPSVS axes is close to  $45^\circ$ , implying a 1:1 relationship in PCA score units, we found this scaling was not reflected by the comparison of permeability and solubility. Thus, a two-log-unit increase in intrinsic solubility ( $r^2$  in Figure 7) is likely to result in a



**Figure 7.** Permeability scores  $t_1$  are rotated  $-27^\circ$  from orthogonality whereas the solubility scores  $t_2$  are rotated  $-20^\circ$ , when compared to the GPSVS scores. The permeability and solubility axes form an angle of  $43^\circ$ .

one-log-unit drop in passive transcellular permeability ( $t_1$  in Figure 7), whereas a two-log-unit increase in permeability is likely to result in a one-log-unit decrease in solubility (see Figure 7). In relationship to the biopharmaceutical classification system (BCS), this implies that solubility is more restrictive compared to permeability. In terms of library design, this indicates that compounds with high solubility/bad permeability can be easier brought into the medium solubility/medium permeability range, in comparison with compounds having low solubility/good permeability. Further consequences of this axial rearrangement and its relationship to BCS are currently under investigation in our group.

The ChemGPS set of objects provides different chemical space maps, when used in conjunction with 2D-based descriptors computed by SaSA and HYBOT, vs the 3D-based VolSurf descriptors; all correlation indices among the first six principal components are lower than 0.2 except for PC1 ( $r^2 = 0.61$ ; see Figure 2B). This advocates the use of 3D-based descriptors (i.e., VolSurf), not 2D-based descriptors, in combination with ChemGPS when mapping the medicinal chemistry space with respect to permeability and solubility (BCS). In conclusion, the GPSVS system, a training-set free model for permeability and solubility, can be used as a BCS-related global mapping device for large sets of compounds to assist combinatorial and medicinal chemists in synthesis planning, thus accelerating the drug discovery process.

**Acknowledgment.** We thank Drs. Johan Gottfries and Per-Olof Eriksson (AstraZeneca R&D Mölndal) and Dr. Ulf Norinder (AstraZeneca R&D Södertälje) for valuable input.

## References and Notes

- Lehn, J. M. Dynamic combinatorial chemistry and virtual combinatorial libraries. *Chem.—Eur. J.* **1999**, *5*, 2455–2463.
- Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
- 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.

- Artursson, P.; Karlsson, J. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2 cells). *Biochem. Biophys. Commun.* **1991**, *175*, 880–885.
- Warr, W. A. Combinatorial chemistry and molecular diversity. An overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 132–140.
- Martin, Y. C.; Brown, D.; Bures, M. G. Quantifying diversity. In *Combinatorial chemistry and molecular diversity in drug discovery*; Gordon, E. M., Kerwin, J. F., Eds.; Wiley-Liss: New York, 1998; pp 369–385.
- Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 339–353.
- Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263–272.
- Darvas, F.; Dorman, G. Early integration of ADME/Tox parameters into the design process of combinatorial libraries. *Chim. Oggi* **1999**, *17*, 10–13.
- Oprea, T. I.; Zamora, I.; Svensson, P. Quo Vadis, scoring functions? Toward an integrated pharmacokinetic and binding affinity prediction framework. In *Combinatorial library design and evaluation for drug design*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker, Inc.: New York, 2001; pp 233–266.
- Willett, P. Chemoinformatics—similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **2000**, *11*, 85–88.
- Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
- Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63–70.
- Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161–1172.
- Lobanov, V. S.; Agrafiotis, D. K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 460–470.
- Oprea, T. I. Rapid estimation of hydrophobicity for virtual combinatorial library analysis. *SAR QSAR Environ. Res.* **2001**, *12*, 129–141.
- Shi, S.; Peng, Z.; Kostrowicki, J.; Paderes, G.; Kuki, A. Efficient combinatorial filtering for desired molecular properties of reaction products. *J. Mol. Graphics Modell.* **2000**, *18*, 478–496.
- Sadowski, J.; Kubinyi, H. A Scoring Scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.
- 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.
- Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. I. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 3743–3748.
- Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.
- Jackson, J. E. *A Users Guide to Principal Components*; Wiley: New York, 1991.
- Austel, V. A manual method for systematic drug design. *Eur. J. Med. Chem.* **1982**, *17*, 9–16.
- Johnson, M. E.; Nachtsheim, C. J. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics* **1983**, *25*, 271–277.

- (25) Oprea, T. I.; Gottfries, J.; Sherbukhin, V.; Svensson, P.; Kühler, T. C. Chemical information management in drug discovery: Optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graphics Modell.* **2000**, *18*, 512–524.
- (26) Oprea, T. I.; Gottfries, J. Chemography: The Art of Chemical Space Navigation. *J. Comb. Chem.* **2001**, *3*, 157–446.
- (27) Oprea, T. I.; Gottfries, J. ChemGPS: A chemical space navigation tool. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science Press: Barcelona, Spain 2001; pp 437–446.
- (28) Goodford, P. J. Computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (29) Clementi, S.; Cruciani, G.; Fifi, P.; Riganelli, D.; Valigi, R.; Musumarra, G. A new set of principal properties for heteroaromatics obtained by GRID. *Quant. Struct.–Act. Relat.* **1996**, *15*, 108–120.
- (30) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
- (31) Olsson, T.; Sherbukhin, V. Synthesis and Structure Administration (SaSA), 1997–2001. AstraZeneca. <http://www.astrazeneca.com>.
- (32) Raevsky, O. A.; Grigor'ev, V. Yu.; Kireev, D.; Zefirov, N. S. Complete thermodynamic description of H-Bonding in the framework of multiplicative approach. *Quant. Struct.–Act. Relat.* **1992**, *11*, 49–64. HYBOT is available from Pion Inc., Cambridge, Massachusetts, <http://www.pion-inc.com>.
- (33) Waiver of in vivo bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on a biopharmaceutics classification system. U.S. Department of Health and Human Services, Food and Drug Administration. [http://www.fda.gov/cder/OPS/BCS\\_guidance.htm](http://www.fda.gov/cder/OPS/BCS_guidance.htm) (accessed 2000).
- (34) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17–30. VolSurf is available from Molecular Discovery Ltd., <http://www.moldiscovery.com>.
- (35) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A new tool for pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11* (Suppl. 2), S29–S39.
- (36) Guba, W.; Cruciani, G. Molecular field-derived descriptors for the multivariate modeling of pharmacokinetic data. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jørgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 89–94.
- (37) Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood–brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43* (11), 2204–2216.
- (38) Zamora, I.; Oprea, T. I.; Ungell, A.-L. Prediction of oral drug permeability. In *Rational Approaches to Drug Design*; Höltje, H.-D., Sippl, W., Eds.; Prous Science Press: Barcelona, Spain, 2001; pp 271–280.
- (39) The Pomona Masterfile is an extensive collection of experimental log *P* values, available from Albert Leo, Pomona College, California, <http://clogp.pomona.edu>.
- (40) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868–880.
- (41) CONCORD; Tripos, Inc.: St. Louis, MO, 2000; <http://www.tripos.com>.
- (42) SYBYL, version 6.2; Tripos, Inc.; St. Louis, MO, 2000.
- (43) Bobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M. New hydrogen-bond potential for use in determining energetically favourable binding sites of molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.
- (44) Leo, A.; Weininger, A. *CMR3 Reference Manual*; Daylight Chemical Information Systems: Santa Fe, NM, 1995; <http://www.daylight.com/>.
- (45) Glen, R. C. A Fast Empirical Method for the Calculation of Molecular Polarizability. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 457–466.
- (46) Oprea, T. I. Property Distribution of Drug-related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.
- (47) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891–898.
- (48) Balaban, A. T. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res.* **1998**, *8*, 1–21.
- (49) Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct.–Act. Relat.* **1996**, *15*, 480–490.
- (50) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity: A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3222.
- (51) CLOGP; Biobyte, Inc.: Claremont, CA; <http://clogp.pomona.edu/>.
- (52) ACDLogP; ACD Labs: Toronto, Canada; <http://www.acdlabs.com/>.
- (53) Leo, A. Estimating Log $P_{oct}$  from Structures. *Chem. Rev.* **1993**, *5*, 1281–1306.
- (54) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III The collinearity problem in linear regression. The partial least squares approach to generalised inverses. *J. Sci. Stat. Comput.* **1984**, *5*, 735–743.
- (55) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211–228.
- (56) SIMCA-P, version 8.0; Umetrics: Umeå, Sweden; <http://www.umetrics.com/>.
- (57) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esverse, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. Multivariate Data Analysis in Chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*; Kowalski, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1983; pp 17–96.