QSAR Models for the Human H^+/Peptide Symporter, hPEPT1: Affinity Prediction Using Alignment-Independent Descriptors

Simon Birksø Larsen, Flemming Steen Jørgensen, and Lars Olsen*

Biostructural Research, Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, 2 Universitetsparken, DK-2100 Copenhagen, Denmark

Received September 17, 2007

A data set comprising the major known chemical classes of hPEPT1 ligands was compiled from the literature. For these compounds, alignment-independent descriptors (VolSurf, GRIND/Almond, and MOE) were computed. Using hierarchical partial least-squares projection to latent structures (H-PLS), a one-component model with $r^2 = 0.77$ and $q^2 = 0.75$ was obtained. The model satisfied a set of rigorous validation criteria and performed well in the prediction of an external test set. Mechanistic interpretation of the model reveals polarity properties to be the dominant factors in determining hPEPT1 affinity, with hydrophobic interactions contributing to a lesser extent. The model is superior to previously reported models due to its combination of quality and speed. Accordingly, it is suitable for ligand-based virtual screening, such as QSAR-based database mining.

INTRODUCTION

The human H^+/peptide symporter (hPEPT1), predominantly located in the brush border membrane of enterocytes, mediates the active uptake of di- and tripeptides in addition to a range of peptidomimetics, including pharmacologically important substances such as β-lactam antibiotics and angiotensin-converting enzyme (ACE) inhibitors.[1−4] Accordingly, hPEPT1 is able to transport diverse substrates differing both in size, net charge, and polarity. Due to its high capacity and low substrate specificity, hPEPT1 has been recognized as a promising route for oral drug delivery.[1]

In order to fully exploit hPEPT1 in drug design, a comprehensive understanding of structure−activity relationships (SAR) is required. Along these lines, computational methodologies provide an established strategy to probe the function and SAR of drug transporters.[5,6] With regard to affinity for hPEPT1, extensive three-dimensional quantitative structure−activity relationships (3D QSAR) have been reported by Gebauer et al.[7] and Biegel et al.,[8] where the CoMFA[9] and CoMSIA[10] methods were employed. Although CoMFA and CoMFA-like approaches are important 3D QSAR methodologies, they have several shortcomings. First, these methods are reported to be very sensitive to the chosen molecular alignment, which is carried out according to subjectively defined orientation rules.[11] Second, the alignment is time-consuming and not readily amenable to automated procedures. These characteristics preclude the use of CoMFA-like methods in a "high-throughput" fashion (e.g., QSAR-based database mining). To circumvent the alignment problem, Cruciani and co-workers have developed alignment-independent descriptors derived from the 3D molecular interaction field (MIF) of a molecule.[12] Two such types of descriptors are the VolSurf descriptors and GRid-INdependent Descriptors (GRIND). Other alignment-independent descriptors are the 2D descriptors, which use only information from the atom connectivity table but cannot capture information from the 3D conformation of a molecule.

Classification models for hPEPT1 inhibitors and noninhibitors have also been reported.[13] However, these models are only able to classify hPEPT1 ligands on a coarse, qualitative basis ("inhibitors", $K_i \leq 1$ mM; "unknown group", 1 mM $< K_i <$ 4.8 mM; "noninhibitors", $K_i \geq 4.8$ mM), and the protocol again requires molecular alignment.

Hence, existing models for hPEPT1 affinity prediction are incompatible with fast and quantitative applications in the pursuit of novel ligands with an affinity for hPEPT1. The strategy employed in this work addresses these limitations using computationally efficient, alignment-independent descriptors. Multivariate data analysis is carried out by means of partial least-squares projection to latent structures (PLS) and hierarchical PLS (H-PLS). In conclusion, we have developed a predictive QSAR model, which has been successfully challenged by a set of rigorous validation criteria. Key aspects of the mechanistic interpretation of our model are discussed and exemplified.

METHODS

Data Set. The data set consists of chemical structures and biological data previously described in the literature (see the Supporting Information).[8,14,15] For all compounds included in the data set, affinity for hPEPT1 was determined under comparable conditions using Caco-2 cell competition assays, preferably from the same laboratory, with [^{14}C]Gly-Sar as a reference substrate. This assay measures the ability of a test compound to inhibit the uptake of the reference substrate. Accordingly, the assay is not able to distinguish between transported compounds (substrates) and nontransported compounds (inhibitors).

The data set of Biegel et al.[8] contains 121 different compounds (90 di-/tripeptides, 31 β-lactam antibiotics). Di-/

* Corresponding author phone: +45 3533 6305; fax: +45 3533 6041; e-mail: lo@farma.ku.dk.

10.1021/ci700346y CCC: $40.75
Published on Web 12/20/2007

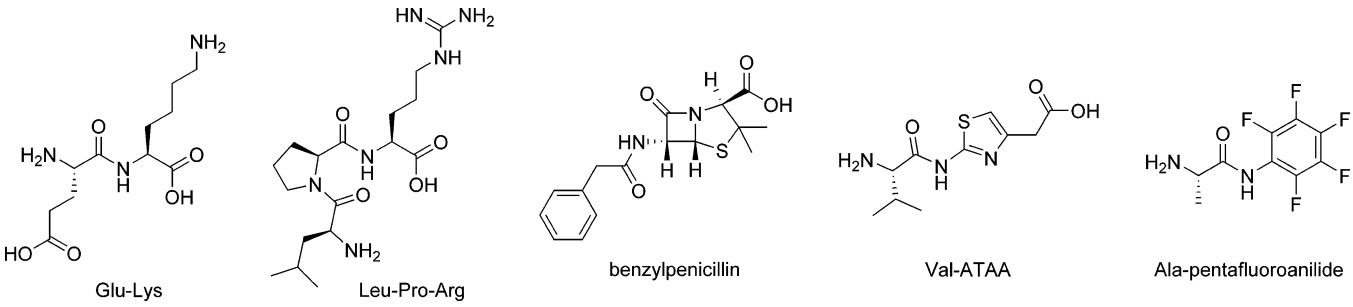

Figure 1. Representative structures of the data set.

tripeptides containing D-amino acids were excluded in the present study. Also, compounds subject to structural ambiguity in the references cited by Biegel et al.[8] were excluded. β-Lactam antibiotics from Luckner et al.,[14] some of which were not included by Biegel et al.,[8] were all included. Finally, a series of 2-aminothiazole-4-acetic acid (ATAA) derivatives recently reported by Biegel et al.[15] was included. The final data set thus contains 114 compounds (69 di-/tripeptides, 37 β-lactam antibiotics, and 8 ATAA derivatives). Representative structures of the data set are shown in Figure 1.

Preparation of Structures. Molecules were built in SYBYL (v.7.2, Tripos Associates Inc., St. Louis, MO) or in the Molecular Operating Environment (MOE) (v.2006.08, Chemical Computing Group Inc., Montreal, Canada). Chemical groups were modeled in their dominating charge state at physiological pH. The MMFF94 force field was used for short energy relaxations (dielectric constant set to $\epsilon = 4$, max. iterations = 100). In the case of MOE descriptors, PEOE partial charges were included prior to descriptor calculation. All calculations were run on a personal computer (the total time required, for calculations described in this work, was <1 hour).

Molecular Descriptors. Three sets of molecular descriptors were considered: VolSurf descriptors, GRid-INdependent Descriptors (GRIND), and descriptors implemented in the MOE software.

VolSurf Descriptors. The principles behind the VolSurf procedure have been described in detail elsewhere.[16−19] In brief, the 3D MIF of a molecule is used to calculate volume and surface descriptors (size and shape as well as hydrophilic and hydrophobic regions and the balance between them). Although VolSurf has mainly been used in the modeling of pharmacokinetic and physicochemical properties, its applicability in protein−ligand affinity predictions has also been reported.[20] In the present study, the water (OH2), hydrophobic (DRY), carbonyl oxygen (O), and amide NH (N1) probes were selected, resulting in 110 descriptors. The descriptors were calculated using the VolSurf program (v.4.1.4.1, Molecular Discovery Ltd., Middlesex, United Kingdom). A grid spacing of 0.5 Å and eight energy levels were used.

GRid-INdependent Descriptors (GRIND). A detailed account of the principles behind GRIND can be found elsewhere.[21−23] The procedure for obtaining GRIND involves three automated steps: (i) 3D MIF calculation, (ii) extraction of relevant MIF regions, and (iii) encoding of geometrical relationships into GRIND. The latter step works by computing the product of interaction energies between extracted nodes of the MIF. For a given distance range, only the numerically largest product is kept. GRIND were calculated using the Almond program (v.3.3.0, Molecular Discovery Ltd., Middlesex, United Kingdom). The hydrophobic (DRY), carbonyl oxygen (O), amide NH (N1), and shape (TIP) probes were selected. The program was run with the following settings: grid spacing 0.5 Å, relative field weight 50%, number of filtered nodes 100, and correlogram size 50. The number of descriptors resulting from these settings is 500.

MOE Descriptors. These descriptors include 2D (physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and κ shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors) and 3D (potential energy descriptors, surface area, volume, and shape descriptors, and conformation-dependent charge descriptors) molecular descriptors. Detailed information can be found in the MOE manual (v.2006.08, Chemical Computing Group Inc., Montreal, Canada), and references cited therein. Descriptors dependent on an external reference frame or computationally demanding, semiempirical calculations were omitted, resulting in the calculation of 204 MOE descriptors.

Training and Test Set Selection. The data set was partitioned into training and test sets using a D-optimal onion design (DOOD), the principles of which are outlined under the *Multivariate Data Analysis* section. The DOOD was based on a principal component analysis (PCA) of the entire data set including binding affinities (data not shown). In this way, two-thirds of the data set, corresponding to 76 compounds (44 di-/tripeptides, 27 β-lactam antibiotics, and 5 ATAA derivatives), were assigned to the training set. The remaining 38 compounds were used as a test set. The training and test set span pK_i ranges of -2.2 to $+2.0$ and -1.7 to $+1.7$, respectively. Thus, both sets cover more than 3 orders of magnitude of biological activity.

Multivariate Data Analysis. PCA and PLS are well-established methods of multivariate data analysis.[24,25] PCA scores, being orthogonal and few in numbers, are useful design variables for statistical molecular design (SMD), such as the D-optimal designs for representative (diverse) training set selection.[26] In DOOD, the data set is partitioned into subsets (layers), and a D-optimal design is applied to each layer. QSAR models based on rationally selected training sets are more likely to result in predictive models.[27]

In PLS, X scores are extracted to summarize variation in the **X** matrix as well as to maximize the covariance between the **X** matrix and a matrix of dependent variables (**Y** matrix). H-PLS is an extension to traditional PLS regression.[28,29] With H-PLS, variables are first divided into conceptually mean-

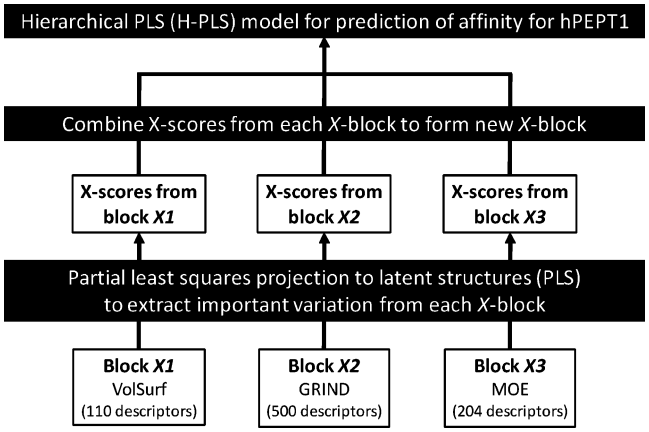

Figure 2. Schematic overview of the workflow and hierarchical PLS (H-PLS) modeling. On the lower (base) level, the X variables are divided into conceptually meaningful blocks (VolSurf, GRIND, and MOE descriptors, respectively). Each X block is then modeled locally by PLS to extract a few X scores ("super variables") that are combined at the upper (top) level to form a new X block. This new X block is finally modeled with PLS.

ingful X blocks (in our case, VolSurf, GRIND, and MOE descriptors, respectively; cf. Figure 2).

Using PCA or PLS, each block of variables is then summarized by a few X scores ("super variables") that are concatenated into a new X block. The new X block is used for final PLS modeling at the top level (Figure 2). The advantage of using PLS, rather than PCA, at the base level is that the projection of each X block is oriented in a way relevant to the description of the Y block. This has the effect of stabilizing the PLS model at the top level, resulting in models with increased predictive ability.[28] It is emphasized that the objective of the base-level modeling is to summarize the original variables, not to optimize the predictive power.

H-PLS is an alternative to potentially risky variable selection (i.e., deletion of unimportant variables) as well as to the challenge of block-scaling, both of which can be carried out in a multitude of ways. Another advantage of H-PLS is enhanced model interpretability. Since parameters are distributed over the levels of the model hierarchy, there are fewer parameters to interpret at each level. Statistics and diagnostics that apply to conventional PLS also apply to H-PLS.[28]

To determine the optimal dimensionality of a PLS model, cross-validation (CV) is often employed.[25] From CV, one obtains q^2, the cross-validated r^2. Root-mean-square error of the fit for observations in the training set (RMSEE) and root-mean-square error of the prediction for observations in the test set (RMSEP) are calculated as follows, eqs 1 and 2:

$$\mathrm{RMSEE} = \sqrt{\frac{\sum (y_{i,\mathrm{obs}} - y_{i,\mathrm{calc}})^2}{N - c - 1}} \quad (1)$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum (y_{i,\mathrm{obs}} - y_{i,pred})^2}{N}} \quad (2)$$

where N is the number of compounds and c is the number of model components.

Attention has been drawn to the fact that a high q^2 does not by itself guarantee a QSAR model to have high predictive ability.[30] To estimate the predictive ability of a QSAR model, prediction of an external test set is indispensable. For the test set, we use a capital R for the correlation coefficient between the measured affinities ($y_{i,\mathrm{obs}}$) and the predicted affinities ($y_{i,\mathrm{pred}}$). Furthermore, R_0 and R_0' denote corresponding correlation coefficients for regressions through the origin (i.e., for $y_{i,\mathrm{obs}} = k y_{i,\mathrm{pred}}$ and $y_{i,\mathrm{pred}} = k' y_{i,\mathrm{obs}}$, respectively). As proposed by Tropsha and co-workers,[30,31] a QSAR model can be considered predictive, if the following set of validation criteria are satisfied:

$$q^2 > 0.5 \quad (3)$$

$$R^2 > 0.6 \quad (4)$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \text{ and } 0.85 \le k \le 1.15$$

or

$$\frac{R^2 - R_0'^2}{R^2} < 0.1 \text{ and } 0.85 \le k' \le 1.15 \quad (5)$$

In the present work, PCA and PLS analyses were performed with the SIMCA-P software (v.10.5, Umetrics, www.umetrics.com) using default settings. Variables were centered and scaled to unit variance (autoscaling), except for GRIND, which were centered only. In the statistical analysis, K_i values were modeled as pK_i. CV (seven rounds) was used to identify the optimum number of components in hierarchical PLS models. The DOOD was generated with the MODDE software (v.7, Umetrics, www.umetrics.com).

RESULTS

Model Development and Validation. The training set selected by DOOD was used for base-level PLS modeling with VolSurf, GRIND, and MOE descriptors comprising distinct X blocks (Figure 2). For the base-level models, CV returned two (GRIND) or three (VolSurf, MOE) significant model components. To ensure adequate description of the X blocks, three components were included in all three cases (accounting for 47−63% of the variation in the descriptor blocks). A single compound (ceftibuten) was excluded in the derivation of the models (identified as a moderate outlier in the residuals' normal probability plots of each base-level model). A summary of model statistics pertaining to the base-level models can be found in Table 1E−G.

Transferring the X scores from the base-level models to the top level resulted in a one-component model with $r^2 = 0.77$ and $q^2 = 0.75$ (Table 1A). The calculated affinities for compounds in the training set are shown in Figure 3 together with predictions for the test set. Affinity estimation for three (cephaloridine, Asp-Ala-NH_2, and 6-aminopenicillanic acid) and two (Leu-Pro-Arg and Val-ATAA) compounds, respectively, deviate more than one logarithmic unit from the measured values. Inspection of the structures did not suggest reasons for the deviations. However, for cephaloridine and 6-aminopenicillanic acid, determination of the K_i values required extensive extrapolation beyond the measured range of concentrations in the assay.[14] This fact provides an explanation for the deviation observed for these two com-

Table 1. Summary of Model Statistics

	descriptors[a]			training set			test set						
model	V	G	M	r^2	q^2	RMSEE	R^2	RMSEP	R_0^2	R'^2_0	k	k'	valid[b]
A	X	X	X	0.77	0.75	0.49	0.72	0.51	0.71	0.71	0.96	0.75	yes
B		X	X	0.75	0.74	0.51	0.67	0.58	0.66	0.66	0.86	0.76	yes
C	X		X	0.75	0.72	0.52	0.70	0.53	0.68	0.69	1.00	0.70	yes
D	X	X		0.74	0.72	0.52	0.71	0.52	0.70	0.71	0.99	0.72	yes
E	X			0.68	0.42	0.59	0.73	0.52	0.72	0.72	1.21	0.60	no
F		X		0.67	0.35	0.60	0.57	0.66	0.55	0.56	0.81	0.70	no
G			X	0.70	0.60	0.58	0.59	0.66	0.55	0.56	0.83	0.67	no

[a] For base-level models (i.e., E−G), V, G, and M are the original VolSurf, GRIND, and MOE descriptors, respectively. For hierarchical models (i.e., A−D), V, G, and M are the X scores ("super variables") from base-level models E−G. [b] "Yes" indicates that the model satisfies conditions 3−5 discussed under the *Multivariate Data Analysis* section.

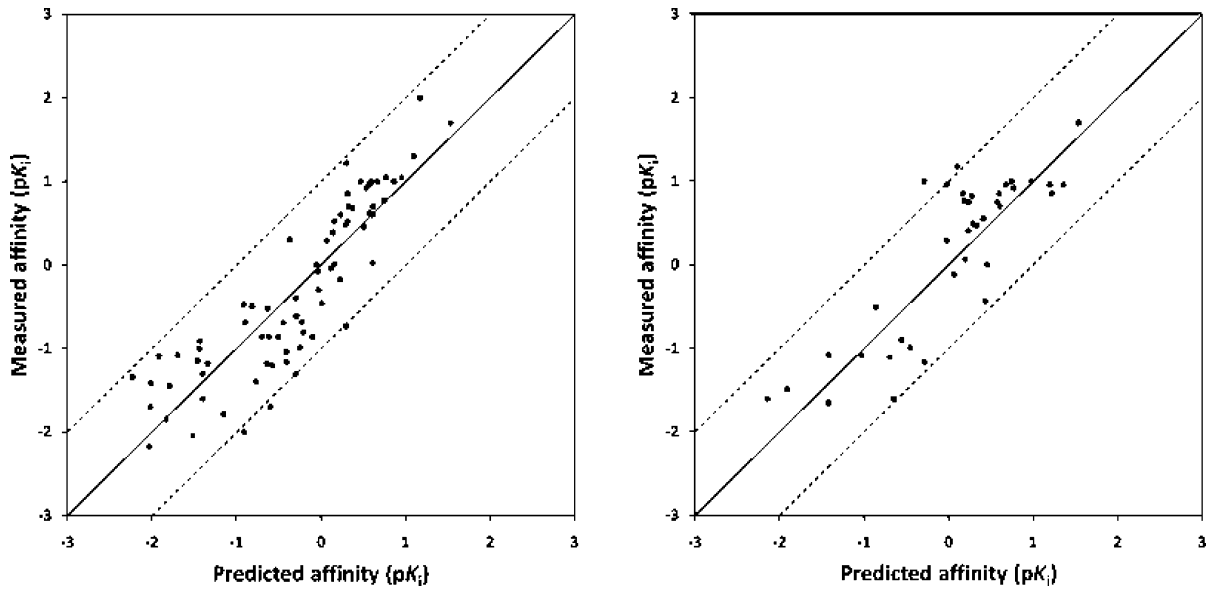

Figure 3. Hierarchical PLS model performance. Measured versus predicted affinities for compounds in the training set ($N = 75$; left) and test set ($N = 38$; *right*). Solid lines represent perfect correlations between measured and predicted affinities, and dashed lines indicate a deviation of one logarithmic unit.

pounds. Asp-Ala-NH_2 is somewhat atypical due to its C-terminal carboxamide, which may account for the observed deviation. No simple explanation could be found for the poor prediction of the two compounds in the test set.

The residual standard deviation for the test set (RMSEP = 0.51) was found to be very similar to that of the training set (RMSEE = 0.49); 0.5 log units corresponds to a factor of 3.

A summary of model statistics for the final hierarchical model (A) and for models where X scores from one of the base-level models have been left out of the model development (B−D) is also shown in Table 1. It should be noted that the inclusion of more than three components from each base-level model improved r^2 and q^2 (eventually exceeding 0.9) of the hierarchical model. This improvement was, however, not mirrored by an increase in predictive ability, as judged from predictions of the test set.

We also performed a response permutation test to assess the robustness of the final hierarchical model. In response permutation, several models are derived with the responses (i.e., affinities) randomized. The results are shown in Figure 4. Intercepts of the r^2- and q^2-regression lines are below limits of 0.3 and 0.05, respectively, reinforcing the validity of the model.[24]

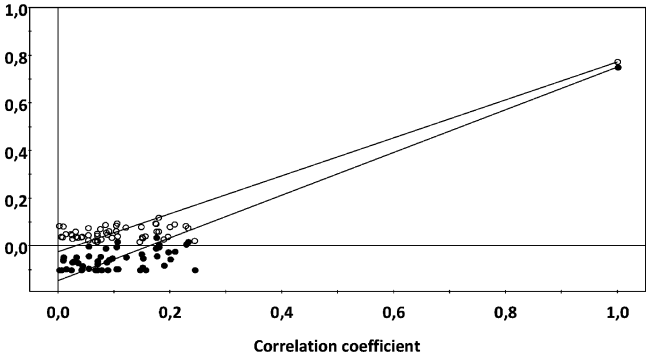

Figure 4. Response permutation testing for the hierarchical model. The abscissa axis is the correlation coefficient between the original and permuted pK_i values. The ordinate axis represents r^2 (circles) and q^2 (black dots) for 50 random models and the original model (the latter with correlation coefficient 1). The intercepts of the r^2- and q^2-regression lines with the ordinate axis are −0.02 and −0.15, respectively.

Model Parameters. PLS coefficients for the final hierarchical model are shown in Figure 5 (top) along with X weights for the first component (t1) of the base-level VolSurf (middle) and GRIND model (bottom). X weights for MOE descriptors have been omitted for clarity. In Table 2, a correlation matrix, representing the pair-wise correlation

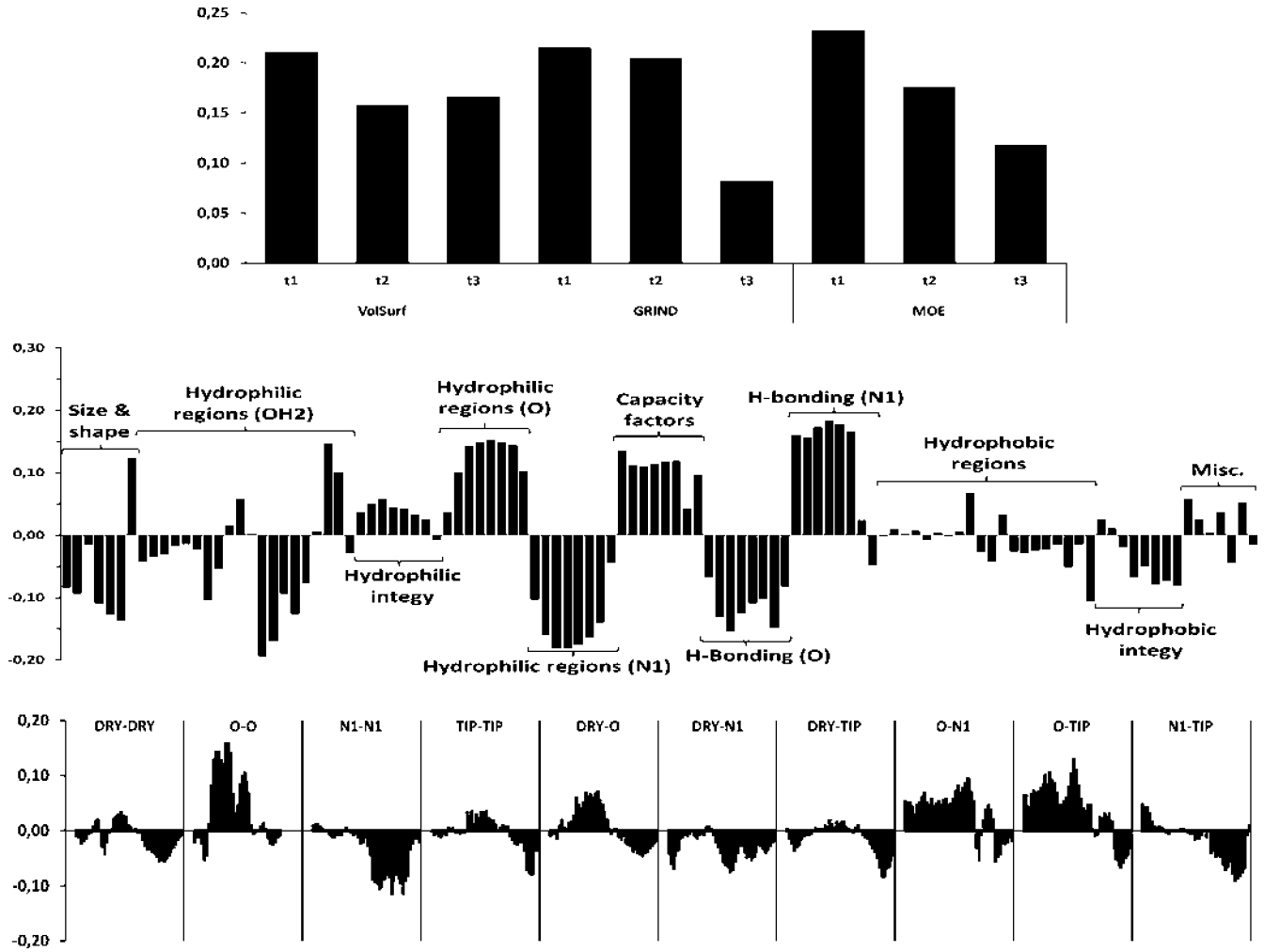

Figure 5. Top: PLS coefficients for the hierarchical model. Middle: PLS X weights for the first component (t1) of the base-level VolSurf model. Bottom: PLS X weights for the first component (t1) of the base-level GRIND model. PLS X weights for MOE descriptors have been omitted for clarity. In all cases, vertical bars indicate the degree of importance for the coefficient or X weight. For further discussion, see the text.

Table 2. Pair-Wise Correlation (r^2 Values) between X Scores for Base-Level Models (Values Pertaining to the Training Set)

VolSurf	t1	1								
	t2	0	1							
	t3	0	0	1						
GRIND	t1	0.72	0.01	0	1					
	t2	0	0.33	0.20	0	1				
	t3	0.09	0.23	0.27	0	0	1			
MOE	t1	0.90	0	0.03	0.72	0	0.15	1		
	t2	0	0.74	0	0.03	0.25	0.28	0	1	
	t3	0	0.01	0.14	0.06	0.28	0.25	0	0	1
		t1	t2	t3	t1	t2	t3	t1	t2	t3
		VolSurf			GRIND			MOE		

between X scores originating from different base-level models, is shown.

DISCUSSION

Model Characteristics. The training set used for model development consists of compounds from different structural classes: dipeptides, tripeptides, β-lactam antibiotics, and aminothiazole derivatives, as well as amino acid amides (Figure 1). Accordingly, the major chemotypes of known hPEPT1 ligands have been included in the development of the models discussed below.

By examining the statistics and diagnostics of the hierarchical models (Table 1A−D), we see that these all satisfy the set of validation criteria outlined under the *Multivariate Data Analysis* section. Furthermore, the model including X scores from all three base-level models (A) is only marginally superior in terms of fit and predictive ability. That satisfactory hierarchical models are attainable by inclusion of X scores from just two base-level models suggests some redundancy among the scores. As illustrated in Table 2, this can indeed be shown to be the case by calculating the pair-wise correlation. Particularly, the first components (t1) are all mutually correlated (r^2 values 72−90%) across the base-level models. The same information is thus conveyed to a considerable extent by the first component, and this fact explains why satisfactory models are possible, even when

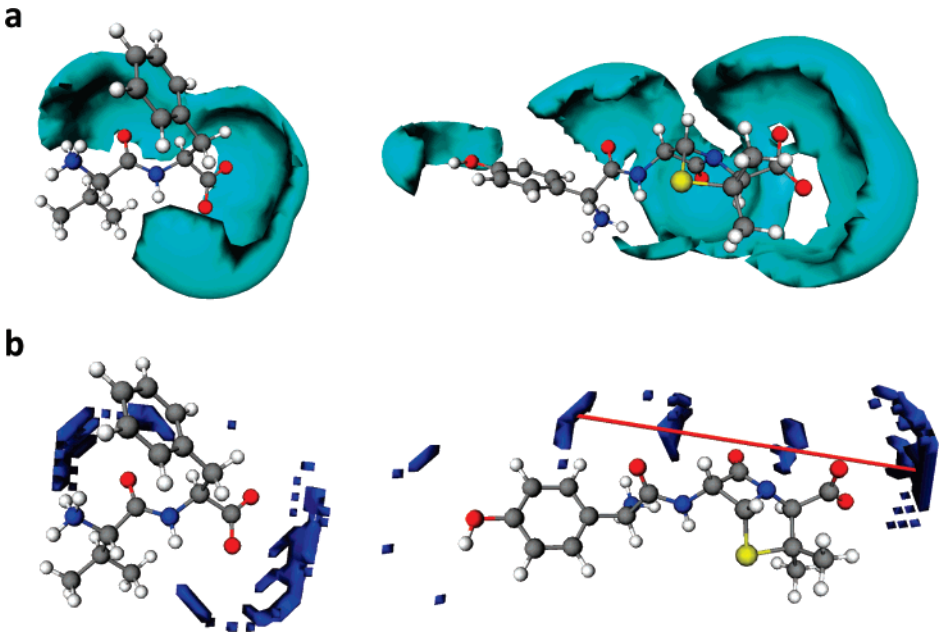

Figure 6. Comparison of the high-affinity dipeptide Val-Phe (K_i = 0.05 mM) and the low-affinity β-lactam antibiotic amoxicillin ($K_i \approx$ 25 mM). (a) MIF derived using the H-bond donor probe (N1) at the −3.0 kcal/mol level of interaction energy (shown in cyan). Amoxicillin (right) shows the largest field, and as this field is inversely related to affinity, it is a contributing factor to the lower affinity relative to Val-Phe (left). (b) The filtered MIF derived from the H-bond donor probe (N1) is shown as blue regions. In the case of amoxicillin (right), the product of interaction energies at the points separated by a distance defined by the red line (approximately 13.6 Å) is detrimental to affinity. This distance is not defined in Val-Phe (left).

excluding X scores from a single base-level model. Conversely, the third component (t3) from the VolSurf base-level model is only sparsely correlated with X scores from the GRIND and MOE model and thus exemplifies an X score carrying considerable unique information. Further discussions of the information content of 2D and 3D descriptors can be found in the literature.[32] Here, it suffices to point out that X scores from the base-level models are significantly correlated, and as a consequence, calculation of one of the descriptor blocks could possibly be omitted in future QSAR studies without serious model deterioration.

The base-level models (Table 1E−G) also deserve some comments. Notably, none of these models satisfies the set of validation criteria considered in this work. The VolSurf model (E) achieves a remarkably high R^2 (0.73) for the test set despite a moderate q^2 (0.42). In addition, r^2 (0.68) for the training set is less than R^2 (0.73) for the test set, and RMSEE (0.59) is larger than RMSEP (0.52). That is, the predictions of the test set are better than the fit of the training set compounds. Hence, at least some of the apparent predictive ability of this model is attributable to chance as we cannot expect a model to predict unknown compounds better than compounds used for model calibration. Finally, for model E, neither k nor k' is close to unity, signifying that the high test set R^2 values are obtained as a result of a correlation between predicted and actual affinities that deviates from the proper 1:1 relationship. Although the base-level VolSurf model has some attractive properties, more confidence should be assigned to the hierarchical models, overall showing favorable statistics for both training and test sets.

For the base-level models built from GRIND and MOE descriptors, statistics for the test set are altogether unsatisfactory, even though, in the case of MOE, training set statistics appear promising. In summary, only the hierarchical models satisfy the validation criteria considered.

Model Interpretation: Features Influencing Binding Affinity. In the previous section, the hierarchical model built from all base-level X scores (Table 1A) was established as the best model. Accordingly, all mechanistic interpretations apply to this model. The basis for interpretation is summarized in Figure 5, where PLS coefficients are shown along with X weights for the first component (t1) of the VolSurf and GRIND base-level models. X weights for MOE descriptors have been omitted for clarity since several MOE descriptors inherently are difficult to interpret (e.g., the graph-theoretical topological descriptors). Also, X weights for subsequent components, t2 and t3, are not shown. Thus, some aspects, generally of a more subtle character, will not be discussed. For illustrative purposes, two compounds have been selected to aid interpretation: a compound of high affinity (the dipeptide Val-Phe, K_i = 0.05 mM) and a compound of low affinity (the β-lactam amoxicillin, $K_i \approx$ 25 mM).[33] For these molecules, MIFs relating to important GRIND and VolSurf descriptors are depicted in Figure 6 and will be referred to where appropriate.

The PLS coefficients reveal that all X scores contribute to binding affinity in a positive manner. Inspection of VolSurf X weights pertaining to t1 (Figure 5, middle) reveals several trends. In essence, hydrophilic regions originating from the H-bond acceptor probe (O), capacity factors (hydrophilic surface per surface unit), and H-bonding parameters (difference between hydrophilic volume obtained with the water probe and one of the polar probes) originating from the H-bond donor probe (N1) all affect affinity in a positive way. In contrast, hydrophilic regions derived from the water and H-bond donor (N1) probes, in addition to H-bonding parameters derived from the H-bond acceptor probe (O), have

a negative impact on affinity. In the case of OH2 hydrophilic regions, two noteworthy exceptions, with a marked positive impact on affinity, are descriptors representing spatial separation (distance) between the best three local minima of interaction energy. Size and shape descriptors are mainly inversely correlated with affinity. To some extent, this reflects simple, chemical subclass differences. For example, β-lactams are larger and have lower affinity than the dipeptides, causing the regression to recognize size as a property detrimental to affinity. With respect to hydrophobic interactions, especially the hydrophobic integy moments (imbalance between the center of mass and the center of hydrophobic regions) have a negative impact on affinity. That is, concentration of hydrophobic regions in one part of the molecular surface will result in decreased affinity. On the other hand, hydrophilic integy moments have a moderate positive impact on affinity.

A major difference between Val-Phe and amoxicillin, in terms of VolSurf descriptors, is the hydrophilic volumes obtained with the H-bond donor probe (N1). At every level of interaction energy evaluated, amoxicillin has the greater volume. This is shown graphically in Figure 6a at the −3.0 kcal/mol level.

As mentioned above, and as shown in Figure 5, the hydrophilic regions derived from the N1 probe decrease affinity for hPEPT1, partly explaining the lower affinity of amoxicillin as compared to Val-Phe.

Comparing H-bonding capabilities obtained from the O and N1 probes, Val-Phe has the smaller values in the case of the O probe, and the larger values in the case of N1 (graphics not shown). This corresponds to a less negative and a more positive effect on affinity (cf. Figure 5), respectively. Capacity factors are also greater for Val-Phe, again augmenting the higher affinity of this compound, as these descriptors contribute positively to affinity.

To summarize the importance of different VolSurf descriptors, it can be concluded that polarity descriptors (hydrophilic regions, H-bonding parameters, and capacity factors) contribute relatively more than do hydrophobic interactions. More specifically, the H-bond donor capabilities of a ligand generally appear to be beneficial for affinity, whereas the H-bond acceptor properties are inversely correlated with affinity.

These findings are in reasonable agreement with those of Andersen et al.[34] who employed the VolSurf approach to develop a QSAR model for the binding of 25 tripeptides to hPEPT1. In contrast to the present work, Andersen et al.[34] found hydrophilic integy moments to have a pronounced negative influence on affinity. However, one should keep in mind that any QSAR merely reflects the differences between compounds included in the analysis. This could account for the observed discrepancy. Interestingly, for the 3D QSAR model reported by Biegel et al.,[8] polar interactions, notably H-bond donor capabilities of the ligand, were found to dominate, whereas hydrophobic properties contributed to a smaller extent. These results are in good agreement with those of our study.

Examining the X weights for GRIND t1 (Figure 5, bottom), several trends are again apparent. For X weights relating to the carbonyl oxygen autocorrelogram (O−O), large positive contributions to affinity are observed. Other correlograms containing descriptors with a considerable positive influence on affinity are DRY−O, O−N1, and O−TIP. On the other hand, X weights pertaining to the N1−N1 correlogram have a distinct, negative impact on affinity, particularly at longer distances (corresponding to the right-hand side of the correlogram). Similar comments apply to X weights representing the N1−TIP correlogram. Additional descriptors with negative influence on affinity are found in other correlograms as well.

The characteristics can again be visualized by comparing MIFs for Val-Phe and amoxicillin. Shown in Figure 6b are the filtered MIFs obtained with the H-bond donor probe (N1). For amoxicillin, the product of interaction energies for positions separated by approximately 13.6 Å in the MIF has an important, negative influence on the binding affinity of this compound. This distance is not defined for the more active compound, Val-Phe.

As for the VolSurf descriptors, the importance of the various GRIND variables can be regarded as being dominated by polar interactions, again with limited contribution from hydrophobic properties.

In terms of drug design, it is crucial to translate the QSAR interpretation into actual chemical modifications required to achieve biologically active compounds. Particularly, balance between polarity properties (e.g., number and position of H-bond donors and acceptors) is of critical importance in the design of new ligands for hPEPT1. Although highly desirable, more unambiguous directions for ligand design are beyond the capabilities of the present QSAR model.

Overall, the SAR of hPEPT1, as outlined in this study, seems to be governed largely by polarity features. Nonetheless, the nature of the SAR appears somewhat blurred. Taking into account that hPEPT1 displays pronounced promiscuous binding (e.g., the binding pocket accommodates and binds ligands ranging from the smallest dipeptide, Gly-Gly, to the largest tripeptide, Trp-Trp-Trp),[8] it does not come as a surprise that a clear-cut pattern of affinity-enhancing, molecular attributes cannot be defined, disregarding common pharmacophoric elements such as charged N and C termini.[3] Elucidation of the mechanism for the promiscuity of hPEPT1 and proteins with similar function might have to await the appearance of an X-ray structure. Yet, the aspects described in this study provide some general insight into the SAR of hPEPT1.

Perspectives. It is interesting to compare the predictive performance of our hierarchical model to that of published QSAR models for hPEPT1. Gebauer et al.[7] provide test set residuals from which the following test set statistics can be obtained: $R^2 = 0.77$, RMSEP = 0.36, $R_0^2 = 0.74$, $R_0'^2 = 0.76$, $k = 0.93$, and $k' = 0.85$ for 26 compounds. These statistics are slightly better than, though still comparable with, those obtained for our model. The two models are, however, derived from different compound sets, with Gebauer et al.[7] considering only dipeptide-type ligands. Thus, what our hierarchical model sacrifices in terms of predictive ability is gained from a much wider applicability domain, encompassing several chemotypes. Biegel et al.[8] provide no detailed test set statistics, and only a qualitative comparison based on graphical presentations of predictive performance is possible. Again, a similar performance is observed.

It should be emphasized that neither of the two models used for comparison are able to compete with the speed and simplicity of the automated protocol used in the present work.

It is well-documented that hPEPT1 has a strong preference for peptides with amino acids in the L configuration.[3,7,8] A limitation of the descriptors in this work is that they do not distinguish between enantiomers (e.g., L-Ala-L-Ala and D-Ala-D-Ala will result in the same descriptor values). Although differences between diastereomers might be captured, we chose not to include any peptides containing D-amino acids. The problem of chirality is, however, fundamental in QSAR modeling, a notable exception being the CoMFA-like methods, which handle chirality by default. Although this chirality issue to some extent poses a limitation on our model, chirality-sensitive descriptors have been developed in recent years.[35,36] Future work should thus be directed toward incorporation of such descriptors. Until our model has been extended to include chirality-sensitive descriptors, care should be taken in predicting the affinity of compounds stereochemically dissimilar to compounds in the training set. Indeed, careful consideration is required whenever a QSAR model is used to predict properties of compounds from a part of the chemical space differing from that of the training set.

With respect to ligand conformations, peptides were built with trans peptide bonds, in accord with the postulated stereospecificity of hPEPT1,[3,37] but no conformational analysis was carried out for any of the compounds included in the data set. It is thus appropriate to address the possible conformational influence on predictions. VolSurf descriptors have been shown to be relatively insensitive to conformational sampling and averaging,[17−19] whereas GRIND show a low-to-moderate conformational dependence.[21,38] Still, GRIND is claimed to be more robust than other methods.[22] For the descriptors implemented in MOE, the majority are 2D descriptors, which are independent of conformation. To investigate whether this relative conformational independence applies to the present model, we selected the conformationally most flexible compounds of the test set (i.e., di- and tripeptides without Pro residues; $N = 16$) and generated 10 random conformations for each compound, many of which were quite unreasonable in terms of conformational energy (no short energy relaxation was applied). Visual inspection verified that backbone and side chains covered a considerable conformational space. Descriptor calculation followed by affinity prediction demonstrated (except for Ile-Val-Tyr) that the standard deviation (SD) for the pK_i values spanned by any single compound was at most 0.25 log units (average SD was 0.14 log units); 0.14 and 0.25 log units correspond to a factor of 1.4 and less than 2, respectively. For Ile-Val-Tyr, the pK_i SD was more than 1.1 log units (predicted K_i spanned a range of almost 3 orders of magnitude), mainly owing to three high-energy conformations. In the next step, all random conformations were submitted to the short energy relaxation used in our protocol, followed by descriptor calculation and affinity prediction. This caused the average SD for the pK_i values to remain constant at 0.14 log units, now including Ile-Val-Tyr. Remarkably, the pK_i SD for Ile-Val-Tyr dropped to 0.26 log units (highest observed) with a pK_i range < 1 log unit, and all predictions for this compound were within 1 log unit of the experimentally observed affinity (max. deviation found was 0.7 log units). These investigations confirm that conformational selection is relatively unimportant with regard to the descriptors used, and only subordinate contributions to the prediction error (RMSEP) can be ascribed to the conformation. However, short energy relaxation seems to be an effective measure to ensure this relative conformational independence. For these reasons, it is clear that significant improvement cannot be achieved by inclusion of a conformational search, as long as at least a reasonable conformation (e.g., an extended conformation) is provided as input for descriptor calculation.

Finally, we are attentive to the fact that affinity for hPEPT1 does not by itself guarantee that a given compound is in fact also transported by the protein. In other words, compounds with high affinity for hPEPT1 can be inhibitors or substrates. However, translocation data in a consistent assay format are at this time sparse, complicating the QSAR modeling of substrate transport. On the other hand, there is strong evidence suggesting that "high" affinity is a necessary, although not sufficient, prerequisite to ensure translocation of a compound.[39,40] Consideration of hPEPT1 binding affinity is thus still highly relevant, and until further translocation data become available, affinity is an acceptable surrogate parameter to guide the selection of compounds for *in vitro* testing.

CONCLUSIONS

On the basis of hierarchical PLS modeling, we have developed a predictive QSAR model, which allows a prediction of affinity for diverse hPEPT1 ligand chemotypes. The model was derived from alignment-independent descriptors and operates on a computationally efficient and fast protocol, requiring neither time-consuming molecular superimposition nor conformational sampling and averaging. Mechanistically, polarity properties of the ligands were disclosed to influence affinity relatively more than hydrophobic properties.

The model described in this work may find applications in ligand-based virtual screening, such as QSAR-based database mining, where fast and automated procedures are of paramount importance. This area of application is currently being actively pursued in our laboratory.

ABBREVIATIONS

ATAA, 2-aminothiazole-4-acetic acid; CoMFA, comparative molecular field analysis; CoMSIA, comparative molecular similarity indices analysis; CV, cross-validation; DOOD, D-optimal onion design; GRIND, GRid-INdependent Descriptors; hPEPT1, human H^+/peptide symporter; H-PLS, hierarchical PLS; MIF, molecular interaction field; MOE, Molecular Operating Environment; PCA, principal component analysis; PEOE, partial equalization of orbital electronegativities; PLS, partial least-squares projection to latent structures; (Q)SAR, (quantitative) structure-activity relationships; RMSEE, root mean square error of the fit for observations in the training set; RMSEP, root mean square error of the prediction for observations in the test set; SMD, statistical molecular design.

ACKNOWLEDGMENT

Prof. G. Cruciani is acknowledged for access to VolSurf and Almond. L.O. acknowledges financial support from the Carlsberg Foundation.

Supporting Information Available: Tables with data set compounds and hPEPT1 affinity. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rubio-Aliaga, I.; Daniel, H. Mammalian peptide transporters as targets for drug delivery. *Trends Pharmacol. Sci.* **2002**, *23*, 434−440.

(2) Nielsen, C. U.; Brodin, B.; Jørgensen, F. S.; Frokjaer, S.; Steffansen, B. Human peptide transporters: therapeutic applications. *Expert Opin. Ther. Pat.* **2002**, *12*, 1329−1350.

(3) Brandsch, M.; Knutter, I.; Leibach, F. H. The intestinal H+/peptide symporter PEPT1: structure-affinity relationships. *Eur. J. Pharm. Sci.* **2004**, *21*, 53−60.

(4) Daniel, H.; Kottra, G. The proton oligopeptide cotransporter family SLC15 in physiology and pharmacology. *Pflügers Arch.* **2004**, *447*, 610−618.

(5) Chang, C.; Swaan, P. W. Computational approaches to modeling drug transporters. *Eur. J. Pharm. Sci.* **2006**, *27*, 411−424.

(6) Chang, C.; Ekins, S.; Bahadduri, P.; Swaan, P. W. Pharmacophore-based discovery of ligands for drug transporters. *Adv. Drug Delivery Rev.* **2006**, *58*, 1431−1450.

(7) Gebauer, S.; Knutter, I.; Hartrodt, B.; Brandsch, M.; Neubert, K.; Thondorf, I. Three-dimensional quantitative structure-activity relationship analyses of peptide substrates of the mammalian H+/peptide cotransporter PEPT1. *J. Med. Chem.* **2003**, *46*, 5725−5734.

(8) Biegel, A.; Gebauer, S.; Hartrodt, B.; Brandsch, M.; Neubert, K.; Thondorf, I. Three-dimensional quantitative structure-activity relationship analyses of beta-lactam antibiotics and tripeptides as substrates of the mammalian H+/peptide cotransporter PEPT1. *J. Med. Chem.* **2005**, *48*, 4410−4419.

(9) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular-Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(10) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(11) Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, Netherlands, 1998; pp 583−618.

(12) *Molecular Interaction Fields. Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; WILEY-VCH: Weinheim, Germany, 2006.

(13) Kamphorst, J.; Cucurull-Sanchez, L.; Jones, B. A performance evaluation of multiple classification models of human PEPT1 inhibitors and non-inhibitors. *QSAR Comb. Sci.* **2007**, *26*, 220−226.

(14) Luckner, P.; Brandsch, M. Interaction of 31 beta-lactam antibiotics with the H+/peptide symporter PEPT2: analysis of affinity constants and comparison with PEPT1. *Eur. J. Pharm. Biopharm.* **2005**, *59*, 17−24.

(15) Biegel, A.; Gebauer, S.; Hartrodt, B.; Knütter, I.; Neubert, K.; Brandsch, M.; Thondorf, I. Recognition of 2-aminothiazole-4-acetic acid derivatives by the peptide transporters PEPT1 and PEPT2. *Eur. J. Pharm. Sci.* **2007**, *32*, 69−76.

(16) Mannhold, R.; Berellini, G.; Carosati, E.; Benedetti, P. Use of MIF-based VolSurf Descriptors in Physicochemical and Pharmacokinetic Studies. In *Molecular Interaction Fields. Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; WILEY-VCH: Weinheim, Germany, 2006; pp 173−196.

(17) Cruciani, C.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *THEOCHEM* **2000**, *503*, 17−30.

(18) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.

(19) 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.

(20) Zamora, I.; Oprea, T.; Cruciani, G.; Pastor, M.; Ungell, A. L. Surface descriptors for protein-ligand affinity prediction. *J. Med. Chem.* **2003**, *46*, 25−33.

(21) Pastor, M. Alignment-independent Descriptors from Molecular Interaction Fields. In *Molecular Interaction Fields. Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006; pp 117−143.

(22) Pastor, M.; Cruciani, G.; Mclay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(23) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating molecular shape into the alignment-free GRid-INdependent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805−2815.

(24) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis. Principles and Applications*; Umetrics AB: Umeå, Sweden, 2001.

(25) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(26) Olsson, I. M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 37−46.

(27) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.

(28) Eriksson, L.; Johansson, E.; Lindgren, F.; Sjostrom, M.; Wold, S. Megavariate analysis of hierarchical QSAR data. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 711−726.

(29) Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemom.* **1996**, *10*, 463−482.

(30) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(31) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

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

(33) For hPEPT1, the following classification is often used: $K_i \leq 0.5$ mM, high affinity; 0.5 mM $< K_i \leq 5.0$ mM, medium affinity; $K_i > 5$ mM, low affinity. See ref 3.

(34) Andersen, R.; Jørgensen, F. S.; Olsen, L.; Vabeno, J.; Thorn, K.; Nielsen, C. U.; Steffansen, B. Development of a QSAR model for binding of tripeptides and tripeptidomimetics to the human intestinal di-/tripeptide transporter hPEPT1. *Pharm. Res.* **2006**, *23*, 483−492.

(35) Golbraikh, A.; Tropsha, A. QSAR Modeling using chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 144−154.

(36) Natarajan, R.; Basak, S. C.; Neumann, T. S. Novel approach for the numerical characterization of molecular chirality. *J. Chem. Inf. Model.* **2007**, *47*, 771−775.

(37) Brandsch, M.; Thunecke, F.; Kullertz, G.; Schutkowski, M.; Fischer, G.; Neubert, K. Evidence for the absolute conformational specificity of the intestinal H+/peptide symporter, PEPT1. *J. Biol. Chem.* **1998**, *273*, 3861−3864.

(38) Benedetti, P.; Mannhold, R.; Cruciani, G.; Ottaviani, G. GRIND/ALMOND investigations on CysLT(1) receptor antagonists of the quinolinyl(bridged)aryl type. *Bioorg. Med. Chem.* **2004**, *12*, 3607−3617.

(39) Bretschneider, B.; Brandsch, M.; Neubert, R. Intestinal transport of beta-lactam antibiotics: Analysis of the affinity at the H+/peptide symporter (PEPT1), the uptake into Caco-2 cell monolayers and the transepithelial flux. *Pharm. Res.* **1999**, *16*, 55−61.

(40) Vig, B. S.; Stouch, T. R.; Timoszyk, J. K.; Quan, Y.; Wall, D. A.; Smith, R. L.; Faria, T. N. Human PEPT1 pharmacophore distinguishes between dipeptide transport and binding. *J. Med. Chem.* **2006**, *49*, 3636−3644.

CI700346Y