Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs

Christel A. S. Bergström,[†] Ulf Norinder,[†,§] Kristina Luthman,*[,‡] and Per Artursson[†]

Center of Pharmaceutical Informatics, Department of Pharmacy, Uppsala University,
Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden, Department of Medicinal Chemistry,
AstraZeneca Research and Development, SE-151 85 Södertälje, Sweden, and Department of Chemistry,
Medicinal Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden

Received December 18, 2002

The aim of the study was to investigate whether easily and rapidly calculated 2D and 3D molecular descriptors could predict the melting point of drug-like compounds, to allow a melting point classification of solid drugs. The melting points for 277 structurally diverse model drugs were extracted from the 12th edition of the Merck Index. 2D descriptors mainly representing electrotopology and electron accessibilities were calculated by Molconn-Z and the AstraZeneca in-house program Selma. 3D descriptors for molecular surface areas were generated using the programs MacroModel and Marea. Correlations between the calculated descriptors and the melting point values were established with partial least squares projection to latent structures (PLS) using training and test sets. Three different descriptor matrixes were studied, and the models obtained were used for consensus modeling. The calculated properties were shown to explain 63% of the melting point. Descriptors for hydrophilicity, polarity, partial atom charge, and molecular rigidity were found to be positively correlated with melting point, whereas nonpolar atoms and high flexibility within the molecule were negatively correlated to this solid-state characteristic. Moreover, the studied descriptors were successful in providing a qualitative ranking of compounds into classes displaying a low, intermediate, or high melting point. Finally, a mechanism for the relation between the molecular descriptors and their effect on the melting point and the aqueous solubility was proposed.

INTRODUCTION

The melting point is not only important in the screening for solid-state characteristics, it is also used as a descriptor for other properties, such as solubility,[1−5] the behavior of eutectic compositions,[6] and liquid viscosity.[7] Lately, great effort has been aimed at predicting the aqueous drug solubility to reduce the experimental settings needed in the preformulation studies.[8−12] Prediction models for solubility that include the melting point as a descriptor often result in acceptable accuracy of the solubility predictions.[2−5] Unfortunately, the melting point has not yet been predicted from theoretical descriptors, with the result that compounds have to be synthesized and experimentally analyzed for their melting point before other properties can be predicted. With the aim of reducing the time and cost of drug discovery and drug development, computational models for this solid-state characteristic based on calculated descriptors are warranted. If such computational protocols are developed, the compounds could be evaluated for their melting point as well as their solubility before they are synthesized.

We have recently worked with several different data sets composed of drug-like molecules to study if calculated molecular surface areas such as polar and nonpolar surface properties (PSA and NPSA, respectively) can be used to predict aqueous drug solubility. We found that mainly hydrophobic descriptors that are negatively correlated to solubility (i.e. NPSA and surface areas of atoms comprising NPSA) were selected in the variable selection performed with multivariate data analysis.[13,14] Interestingly, PSA, a descriptor highly correlated to the number of hydrogen bond forming groups,[15] was not included in the models. We speculated that this was a result of the positive influence that PSA has on the solid state and that molecules with a large PSA form more stable crystals than compounds with a small or no PSA.

In conducting this study, our aim was to investigate whether the melting point could be predicted from 2D descriptors for electrotopology or from 3D molecular surface area descriptors, or from these descriptors in concert, with an accuracy that would allow a theoretical melting point classification. Moreover, we wanted to clarify the role of PSA on the melting point and solubility characteristics of drug-like molecules.

METHODS

Data Set. The 277 compounds used for model building were extracted from the 12th edition of the Merck Index.[16] SKF105657 was determined in-house.[13] The data set showed a normal distribution with the majority of compounds displaying melting points between 140 and 160 °C (Figure 1). The selection criteria used were (i) the drugs should be structurally and physicochemically diverse (Figure 2a); (ii)

* Corresponding author phone: +46 31 772 2894; fax: +46 31 772 3840; e-mail: luthman@mc.gu.se.
[†] Uppsala University.
[‡] Göteborg University.
[§] AstraZeneca Research and Development.

10.1021/ci020280x CCC: $25.00
Published on Web 06/14/2003

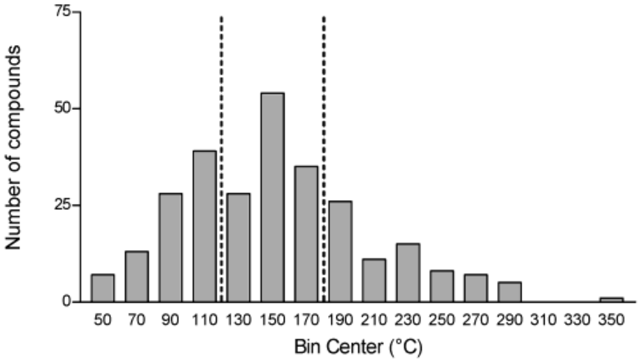

Figure 1. Frequency plot of the 277 compounds studied. A normal distribution is seen with the majority of the compounds displaying a melting point of 140−160 °C. The bars represent the bin centers ±10 °C. The dashed lines show the cutoff between low, intermediate, and high melting point values.

the melting point should be available for the pure compounds (i.e. no salt forms were included); (iii) it should be possible to analyze the conformational preferences of the molecules using molecular mechanics calculations; and (iv) the compounds should not display polymorphism or pseudopolymorphism (i.e. compounds displaying several melting points and compounds crystallized as hydrates and/or solvates were excluded).

Calculation of Descriptors. The 2D descriptors for electrotopology, PSA, and lipophilicity (calculated as logP_{oct}), were calculated with Molconn-Z[17] and the AstraZeneca in-house program Selma.[18] Molconn-Z[17] was used to calculate the electrotopological state indices. Briefly, the electrotopological state indices for a particular atom result from the topological and electronic environment. The indices will encode the electronegativity as well as the local topology of each atom by considering perturbation effects from its neighbors. The program Selma generates descriptors related to size, ring structure, flexibility, hydrogen bonds, polarity, BCUT parameters,[19] the connectivity indices,[20] electronic environment, partial atom charge, and lipophilicity. In total, Molconn-Z and Selma primarily generated 566 different descriptors.

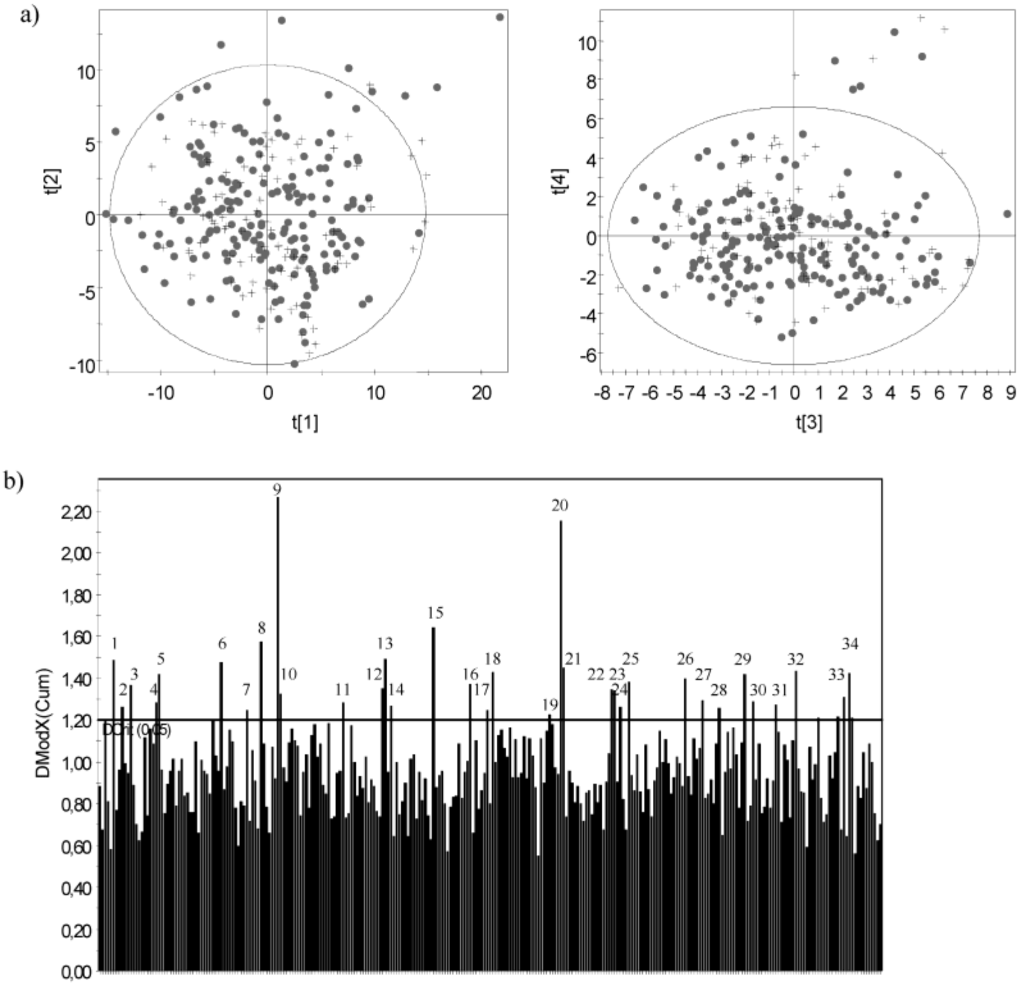

Figure 2. Heterogeneity of the selected data set investigated by principal component analysis (PCA). (a) The scores of the four first principal components (t1−t4) describing 60% of the diversity in the descriptor space are shown. Both 2D and 3D molecular descriptors were used as input matrixes. The training set (filled circles) and test set (crosses) covered all four quadrants of the PCA plot, showing that the selected training and test sets were heterogeneous. Moreover, the test set was well represented by the training set. (b) The cumulative Distance to Model with a 95% CI, explaining 96% of the descriptor space, identified 34 compounds as outliers of the descriptor space: 1. promethazine, 2. ranitidine, 3. novonal, 4. chloroquine, 5. noxythiolin, 6. acecarbromal, 7. vabartan, 8. pirozadil, 9. benzoic acid, 10. nadolol, 11. sertaconazole, 12. florfenicol, 13. taurolidine, 14. metizoline, 15. famotidine, 16. piposulfan, 17. irbesartan, 18. letrozole, 19. albutoin, 20. delavirdine, 21. acitretin, 22. moxesterol, 23. prazosin, 24. orotic acid, 25. isosorbide, 26. valnoctamide, 27. terbutaline, 28. xylometazolin, 29. pyrinoline, 30. astemizole, 31. zileuton, 32. fosfosal, 33. vigabatrin, and 34. tirofiban. None of these were considered to be large outliers, and, therefore, all 277 compounds were included in the analysis.

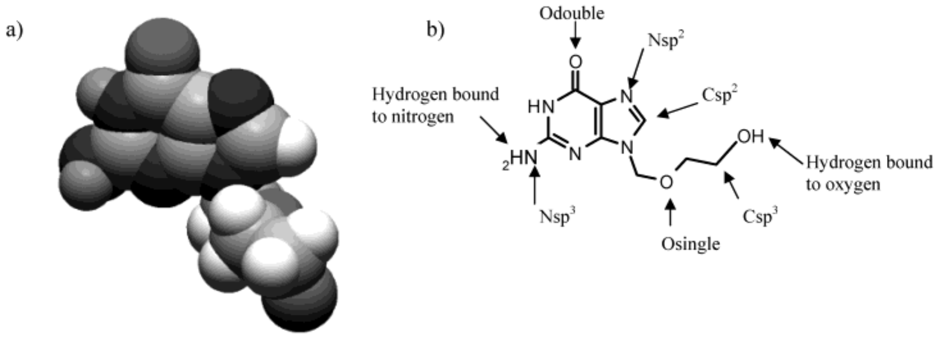

Figure 3. Molecular surface areas. (a) 3D conformation of acyclovir displaying polar atoms in dark gray. (b) Examples of partitioned total surface areas (PTSAs) included in the multivariate data analysis. The PTSAs represent the surface areas of each type of atom calculated with Marea.[21] The polar surface area (PSA) is composed of the PTSAs of oxygen atoms, nitrogen atoms, and hydrogen atoms bound to these heteroatoms. All other atom types are included in the nonpolar surface area (NPSA).

The 3D descriptors for molecular surface areas, such as PSA, NPSA, and partitioned total surface areas (PTSA) (Figure 3), were obtained from conformations generated using the BatchMin program and the MMFF force field in MacroModel version 6.5. The conformational analysis was performed in vacuum with the compounds in their un-ionized state. The in-house computer program Marea[21] was used to calculate the free surface area of each atom and the molecular volume using the van der Waals radii, as follows: sp- and sp^2-hybridized carbons 1.94 Å; sp^3-hybridized carbons 1.90 Å; oxygen 1.74 Å; nitrogen 1.82 Å; sulfur 2.11 Å; chloride 2.03 Å; electroneutral hydrogen 1.50 Å; hydrogen bound to oxygen 1.10 Å; and hydrogen bound to nitrogen 1.125 Å (obtained from PCMODEL v. 4.0, see ref 22). The surface areas were defined as previously described.[13,15] Briefly, composite properties, such as NPSA and the PSA, were calculated as well as PTSA descriptors. The PSA was defined as the surface area occupied by oxygen and nitrogen atoms, and hydrogen atoms bound to these heteroatoms, whereas the NPSA was defined as the total surface area (SA) minus the PSA. The calculated PTSA descriptors correspond to the surface area of a certain type of atom. For example, the NPSA attributable to carbon atoms can be partitioned into the surface areas of sp-, sp^2-, and sp^3-hybridized carbon atoms and the hydrogen atoms bound to these carbon atoms. In a similar way, the PSA attributable to oxygen atoms can be partitioned into the surface areas of single-bonded oxygen, double-bonded oxygen, and hydrogen atoms bound to single-bonded oxygen atoms (Figure 3). Both the absolute static surface areas and the static surface areas relative to the SA were calculated for the global minimum conformation found after a conformational search of 500 steps. In total, 46 different 3D descriptors were calculated.

Table 1. Statistics of Models[a]

model	R^2	Q^2	RMSE$_{tr}$ (°C)	RMSE$_{te}$ (°C)
I	0.56 (0.60)	0.53	36.9 (34.6)	51.7 (40.0)
II	0.31 (0.33)	0.30	46.1 (44.8)	50.3 (46.5)
III	0.57 (0.63)	0.54	36.6 (32.5)	49.8 (40.7)
IV	0.63 (0.64)		35.1 (33.9)	44.6 (39.1)

[a] Coefficient of determination (R^2), the cross-validated coefficient of determination (Q^2), and root-mean-square error of training (RMSE$_{tr}$) and test (RMSE$_{te}$) sets are presented. Values in parantheses are after exclusion of statistical outliers (residuals of ±2.5 standard deviations between experimental and observed value) in each model. Model I is based on 2D descriptors, model II on 3D descriptors, model III on both 2D and 3D descriptors, and model IV is the averaged consensus of the three developed models.

Data Analysis. Melting point values were predicted by principal component analysis (PCA)[23] and partial least squares projection to latent structures (PLS)[24] using Simca.[25] Skewed descriptors were cubic root-transformed prior to the multivariate data analysis. Variables which did not obtain a skewness within ±1.5 were excluded from further data analysis to avoid them obtaining a too heavy weighting in the models. Thus, 121 of the 612 descriptors were included in the PLS variable selection. The training set was selected to include two-thirds of the data set, and the test set was composed of every third compound of the data set when listed in ascending melting point order (see Table 1 in Supporting Information). A PCA plot of the compounds revealed that the test set was well represented by the training set and that a large range in the descriptor space was covered (Figure 2a).

Three different matrixes were used for the PLS predictions: I. 2D descriptors generated by Molconn-Z and Selma; II. 3D descriptors generated by MacroModel and Marea; and III. a combined matrix of both 2D and 3D descriptors. The number of PLS components computed was assessed by Q^2, the leave-one-out cross-validated R^2, using seven cross-validation rounds. Applying this protocol, the training set was divided into seven groups, and each group was left out once to assess the general predictivity of the model. Only PLS components resulting in a positive Q^2 were computed. The models were refined through stepwise selection of the descriptors. Initially all the descriptors were included in the PLS model, and then after the first round, the descriptor with the least influence on the prediction was deleted and the PLS repeated. If the exclusion of the least important descriptor resulted in a more predictive model (as assessed by a higher Q^2), that descriptor was permanently left out of the model. This procedure was repeated until no further improvement of the model could be achieved. The predictive power of the models was assessed by root-mean square error (RMSE) of the test set (RMSE$_{te}$). The three models obtained from matrixes I−III were subjected to consensus modeling, where the average value of the three predicted values was used. Finally, the resulting four models were used in an attempt to classify the compounds into three classes representing low, intermediate, and high melting points as follows: class 1. melting point below 120 °C ($n = 87$); class 2. melting point of 120−180 °C ($n = 116$); and class 3. melting point above

180 °C (n = 73). The accuracy of these predictions is reported as the percentage correct classification.

RESULTS

Data Set. The compounds were considered to be structurally diverse as identified in the PCA plots of the training and test sets, and both training and test sets were distributed in all four quadrants of the PCA plot (Figure 2a). Moreover, the distance to model analysis of the PCA model obtained from the investigated descriptors identified 34 outliers using a 95% confidence interval (Figure 2b). The majority of these compounds were only small outliers, but two larger outliers were identified (benzoic acid and delavirdine). However, we did not consider these two compounds to be large enough outliers to exclude them from further analysis.

The classification of melting point used two cutoff values: 120 °C as the cutoff value between a low and an intermediate melting point and 180 °C as the cutoff value between an intermediate and a high melting point. These cutoff values were based on the frequency plot of the 277 compounds investigated in this work (Figure 1), but the choice was also supported by empirical knowledge from organic chemistry. The frequency plot showed that the compounds followed a normal distribution with the majority of drug-like molecules being found in the interval 140−160 °C. The compounds were distributed as follows: class 1 31.5%, class 2 42%, and class 3 26.5%.

2D Descriptors. The variable selection of the 2D descriptors resulted in the possibility of theoretically describing 56% of the melting point (Figure 4 and Table 1). The most important descriptors for the melting point mainly reflected partial atom charges, polar and nonpolar properties, and the flexibility of the molecule. The multivariate data analysis showed that the descriptor representing the average of all positive partial atom charges was positively correlated to the melting point, whereas the average of all negative partial atom charges was negatively correlated to the response parameter. This can be mathematically explained by the fact that the values of the positive charge are positive and the values of negative charge are negative. Thus, both the positive and negative partial charge will have a positive influence on the melting point, i.e. these properties will result in the formation of more stable crystals with higher melting point values. Moreover, hydrophilic and polar measures, i.e. PSA and hydrogen bond acceptors, were positively correlated to the melting point, whereas nonpolar measures such as NPSA and the electron accessibility between sp^2-and sp^3-carbon atoms (e1C1C2) were negatively correlated resulting in decreased melting point values. Flexibility descriptors showed the largest negative influence on the melting point, while rigidity measures such as ring structures were found to have a positive influence (Figure 4 and Table 3).

The descriptors identified were able to sort low melting point from high melting point, and only two of the 185 training set compounds, i.e. pirozadil and dextromoramide, and four of the 92 test set compounds, i.e. isosorbide, zidovudine, tirofiban and probenecid, were falsely predicted between the low (class 1) and high (class 3) melting point classes, respectively (Table 2). Of these compounds, PCA had identified pirozadil, isosorbide, and tirofiban as outliers of the descriptor space investigated, which may be a reason

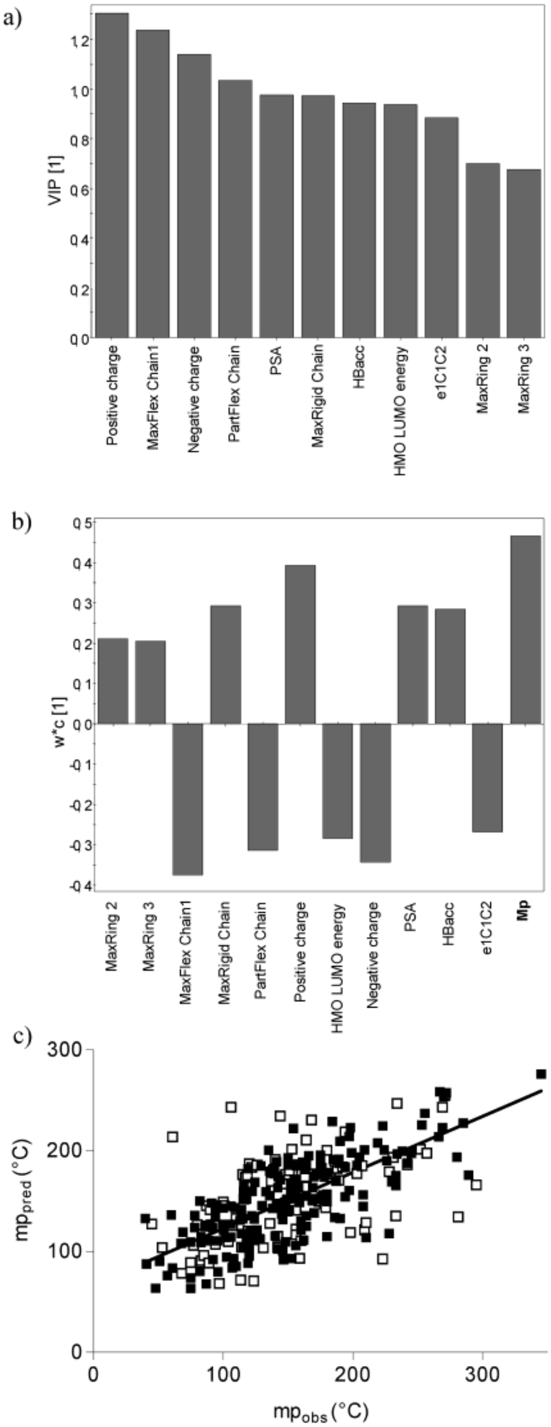

Figure 4. PLS model for in silico prediction of the melting point using 2D descriptors only. The following descriptors were used: averaged positive and negative partial atom charge calculated from all charged atoms within the molecule; length of the longest chain with only rotatable bonds (MaxFlex Chain1); length of the longest partially flexible chain (PartFlex Chain); 2D PSA; lengths of the three longest chains containing only rigid bonds (MaxRigid Chain); number of hydrogen bond acceptors (HBacc); energy of the lowest unoccupied molecular orbital (HMO LUMO energy); electron accessibility between sp^2- and sp^3-hybridized carbon atoms (e1C1C2); size of the second and third largest ring (MaxRing 2 and MaxRing 3) (a) Variable Influence on Projection (VIP) plot showing the importance of each descriptor in the prediction of the melting point. (b) Loading plot showing the interrelationship of the 2D descriptors and the melting point. Only one principal component was extracted in the PLS analysis. (c) The melting point estimated from 2D descriptors for electrotopology, bond energies and PSA. ■ = training set and □ = test set.

Table 2. Classification Results from the Different Models[a]

	I_{tr}	I_{te}	II_{tr}	II_{te}	III_{tr}	III_{te}	IV_{tr}	IV_{te}
% correct class 1	48.3	41.4	37.9	37.9	50.0	44.8	46.6	37.9
% correct class 2	62.8	51.3	79.5	82.0	65.4	56.4	79.5	71.8
% correct class 3	63.3	54.2	40.8	20.8	67.3	45.8	61.2	29.3
% class 1 predicted as 2	50.0	51.7	62.1	58.6	50.0	48.3	53.4	58.6
% class 1 predicted as 3	1.7	6.9	0	3.4	0	6.9	0	3.4
% class 2 predicted as 1	16.7	20.5	5.1	5.1	18.0	12.8	6.4	7.7
% class 2 predicted as 3	20.5	28.2	15.4	12.8	16.7	30.8	14.1	20.5
% class 3 predicted as 1	2.0	8.3	10.2	0	6.1	8.3	4.1	0
% class 3 predicted as 2	34.7	37.5	49.0	79.2	26.5	45.8	34.7	70.8

[a] Classification based on Models I (2D descriptors), II (3D descriptors), III (2D and 3D descriptors), and IV (consensus model) presented for training (tr) and test (te) sets. Class 1 corresponds to low melting point values which includes compounds displaying melting points less than 120 °C, class 2 corresponds to intermediate melting point values with compounds displaying melting points between 120 and 180 °C, and class 3 corresponds to high melting point with compounds displaying melting points higher than 180 °C.

Table 3. Descriptor Coefficients Obtained by PLS Analysis[a]

descriptor	model I	model II	model III
constant	2.73	2.73	2.73
# rings	n.i.	n.i.	0.20
MaxRing 2	0.10	n.i.	n.i.
MaxRing 3	0.10	n.i.	n.i.
MaxFlex Chain1	−0.17	n.i.	−0.20
MaxRigid Chain	0.14	n.i.	0.23
PartFlex Chain	−0.15	n.i.	n.i.
Positive partial charge	0.18	n.i.	0.13
Negative partial charge	−0.16	n.i.	−0.09
HMO LUMO energy	−0.13	n.i.	−0.18
PSA 2D	0.14	n.i.	n.i.
HBacc	0.13	n.i.	n.i.
e1C1C2	−0.12	n.i.	n.i.
PSA 3D	n.i.	0.20	0.13
%Hneutral 3D	n.i.	−0.23	n.i.
%NPSAsat 3D	n.i.	−0.23	−0.15
%PSA 3D	n.i.	n.i.	0.13

[a] The following abbreviations are used: number of rings (# rings); size of the second and third largest ring (MaxRing 2 and MaxRing 3); length of the longest chain with only rotatable bonds (MaxFlex Chain1); lengths of three longest chains containing only rigid bonds (MaxRigid Chain); length of the longest partially flexible chain (PartFlex Chain); averaged positive and negative partial charge calculated from all charged atoms within the molecule; energy of lowest unoccupied molecular orbital (HMO LUMO energy); number of hydrogen bond acceptors (HBacc); electron accessibility between sp^2- and sp^3-hybridized carbon atoms (e1C1C2); fraction of the surface area occupied by hydrogen atoms bound to carbon atoms (%Hneutral), fraction saturated NPSA (%NPSAsat) and fraction PSA (%PSA). n.i. = not included in the final model.

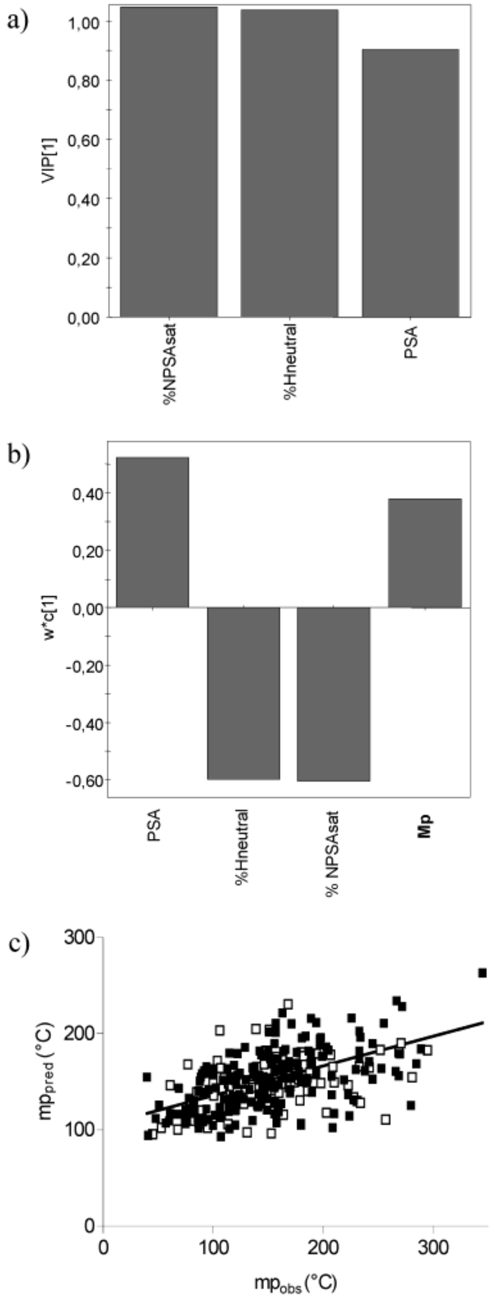

Figure 5. PLS model for in silico prediction of the melting point using 3D descriptors only. The following descriptors were used: PSA, fraction saturated NPSA (%NPSAsat) and fraction of the surface area occupied by hydrogen atoms bound to carbon atoms (%Hneutral). (a) Variable Influence on Projection (VIP) plot showing the importance of each descriptor in the prediction of the melting point. (b) Loading plot showing the interrelationship of the 3D descriptors and the melting point. Only one principal component was extracted in the PLS analysis. (c) The melting point estimated from the 3D descriptors. ■ = training set and □ = test set.

for the misclassification of these drugs (Figure 2b). Moreover, the 2D generated descriptors failed to describe the azide function present in zidovudine, which explains the false prediction for this compound. The selected 2D descriptors could better discriminate between intermediate and high melting point values, than between low and intermediate (Table 2). However, several of the misclassified compounds were falsely predicted within the range of ±10 °C of the borders used for discrimination between low, intermediate, and high melting points. For the training set, 17.2% of the class 1 compounds were falsely predicted as class 2 with melting points <130 °C, 20.5% of the class 2 compounds were falsely predicted as belonging to either class 1 with melting points >110 °C or class 3 with melting points <190 °C, and 6.1% of the class 3 compounds were falsely predicted as class 2 with melting points >170 °C. Thus, we can conclude that a large proportion of the falsely predicted compounds was predicted within ±10 °C of the cutoff points between low, intermediate, and high melting points.

3D Descriptors. The variable selection of matrix II containing descriptors for surface areas was not successful in the prediction of the melting point (Figure 5 and Table 1). However, the PSA proved to be one of the most important descriptors of the surface areas included and, with the fraction of NPSA (%NPSA) and the fraction of neutral hydrogen atoms bound to carbon atoms (%Hneutral), the PSA explained 31% of the melting point values (Figure 5 and Table 1). The multivariate analysis also showed that the PSA is positively correlated to the melting point and that the nonpolar descriptors NPSA and neutral hydrogen atoms have a negative influence on the melting point (Table 3).

Unfortunately, the surface areas estimate the melting point of most of the compounds to be intermediate (class 2), resulting in a large number of false predictions whereby the low and high melting points were estimated to be intermediate. Furthermore, with this model a rather large proportion of false predictions occurred within ±10 °C of the borders set. For the training set, 15.5% of the class 1 compounds were falsely predicted as class 2 with melting points <130 °C, 10.2% of the class 2 compounds were falsely predicted as being either class 1 with melting points >110 °C or class 3 with melting points <190 °C, and 10.2% of the class 3 compounds were falsely predicted as class 2 with melting points >170 °C.

Combined Matrix. The variable selection of the combined matrix, which included both 2D and 3D descriptors, resulted in an accuracy equal to that of the model developed using 2D descriptors only (Figure 6 and Table 1). The most important descriptors in this model for the melting point were almost identical to the ones obtained in model I, with the exception that the 2D descriptor for the PSA was replaced with the 3D descriptor for this property. The same trend was shown in this model as in model I for the degree of influence on the melting point: hydrophilic and polar measures were positively correlated to the melting point, and nonpolar descriptors were negatively correlated to it. Negative and positive partial atom charge strongly affected the melting point values, with the argument given in the explanation for the 2D model being valid again here. Moreover, the larger the flexibility, the larger the negative influence on the melting point will be, whereas descriptors of rigidity (i.e. ring structures of the molecule) have a large positive influence on the melting point (Figure 6 and Table 3).

A model which explained 57% of the melting point was obtained. Model I based on 2D descriptors alone and model III based on both 2D and 3D descriptors proved to be of similar accuracy regarding the prediction of the melting point class, even though the absolute RMSE values ($\mathrm{RMSE_{tr}}$ and $\mathrm{RMSE_{te}}$) were better for model III than model I (Table 1). The 2D and 3D descriptors calculated were able to separate the low melting points from the high ones, and only three compounds in the training set, i.e. dextromoramide, albutoin, and acitretin, and four compounds in the test set, i.e. isosorbide, zidovudine, vigabatrin, and tirofiban, were falsely predicted between the low (class 1) and high (class 3) melting point categories (Table 2). Of these compounds, PCA had identified albutoin, acitretin, isosorbide, vigabatrin, and tirofiban as outliers of the descriptor space investigated (Figure 2b), which may be a reason for the misclassification of these drugs, whereas the parametrization of the azide function in the 2D generated descriptors explains the false prediction of zidovudine in this model too. This model also misclassified a rather large proportion of compounds within ±10 °C of the borders taken for discrimination between the three classes. For the training set, 12.1% of the class 1 compounds were falsely predicted as being class 2 with

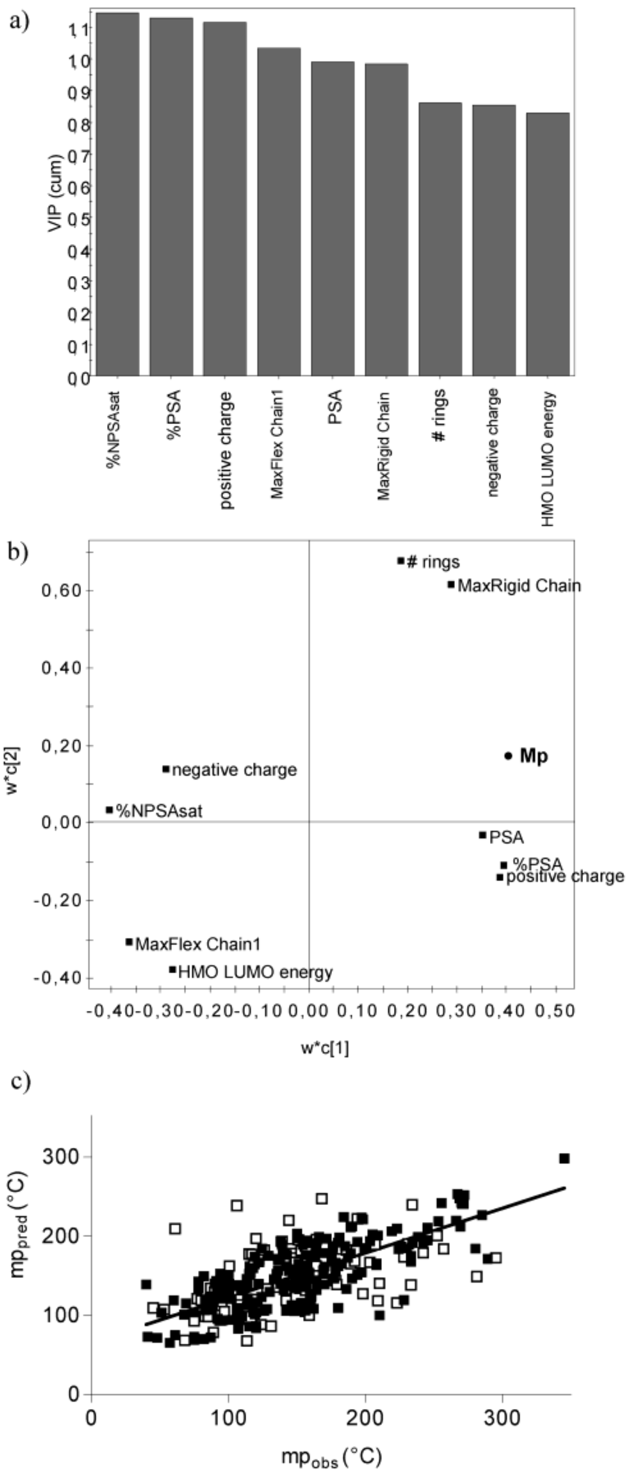

Figure 6. PLS model for in silico prediction of the melting point using a combination of 2D and 3D descriptors. The following descriptors were used: fraction saturated NPSA (%NPSAsat), fraction PSA (%PSA), and PSA, averaged positive and negative partial atom charge calculated from all charged atoms within the molecule; length of the longest chain with only rotatable bonds (MaxFlex Chain1); lengths of three longest chains containing only rigid bonds (MaxRigid Chain); number of rings (# rings); and the energy of lowest unoccupied molecular orbital (HMO LUMO energy); (a) Variable Influence on Projection (VIP) plot showing the importance of each descriptor in the prediction of the melting point. (b) Loading plot showing the interrelationship of the 2D and 3D descriptors and the melting point. Two principal components were extracted in the PLS analysis. (c) The melting point estimated from the combined 2D and 3D matrix. ■ = training set and □ = test set.

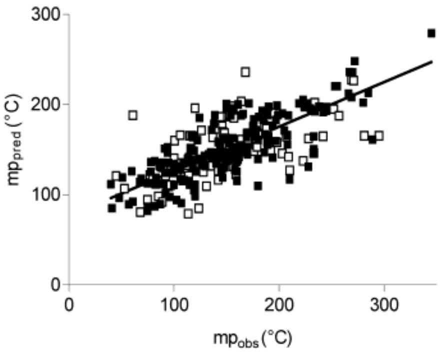

Figure 7. Consensus model of all three obtained models improved R^2 with 0.06 (10%) and resulted in a more stable model with regard to outliers. ■ = training set and □ = test set.

melting points <130 °C, 11.5% of the class 2 compounds were falsely predicted as either class 1 with melting points >110 °C or class 3 with melting points <190 °C, and 8.2% of the class 3 compounds were falsely predicted as class 2 with melting points >170 °C.

Consensus Modeling. In an attempt to improve the predictions and classifications, the three models obtained in this study were used for consensus modeling. This was performed in two steps: first by averaging the results obtained from models I and III as these showed a predictive power of greater than 50%. Thereafter, we also included model II in the consensus modeling. Surprisingly, the inclusion of model II in the consensus modeling proved to be the most successful method of determining the melting point, resulting in improvements both in the R^2- and the RMSE values in comparison to models I−III (Table 1). However, the classification using the consensus model resulted in equal accuracy to both models I and III (Table 2).

DISCUSSION

When starting up this project we searched the Internet for guidelines for computational predictions of the melting point. Surprisingly, we did not find a single publication containing a theoretical analysis of the melting point, although molecular modeling has been used lately to try to predict crystal lattice energies.[26,27] Even though the techniques for performing experimental melting point determinations are relatively fast and straightforward, the clear disadvantage of experimental determination is that the compound has to be synthesized and isolated in its solid state before one can measure the melting point. Computational models generating melting point predictions would result in early estimations of characteristics such as solubility,[1−5] the behavior of eutectic compositions,[6] and liquid viscosity[7] from the 2D structure drawn in the computer. Thus, high-throughput in vitro methods used for pharmacological screening would not only screen large chemical libraries for pharmacological potency, but also result in high-output screening if combined with reliable virtual filters for properties such as the melting point, absorption characteristics, distribution, metabolism, and toxicity (ADMET).

When constructing the model for the melting point, it became apparent from the analyses that the descriptors for hydrophilic and polar fragments were positively correlated to the melting point, whereas descriptors for nonpolar properties were negatively correlated to it. Furthermore, ring structures were found to increase the melting point, whereas a large degree of molecular flexibility resulted in a lowering of the melting point. These results led us to the conclusion that molecules which are polar and rigid generally form more stable and better ordered crystal structures than do molecules in which nonpolar fragments and aliphatic chains with high flexibility constitute the dominant part of the molecule.

Even though the multivariate data analysis did not succeed in predicting the melting point from the surface area descriptors alone (model II), the result from the analysis was in accordance with the findings from the analysis using the 2D descriptors (model I). Thus some fundamental conclusions can be drawn regarding the relationship between polar and nonpolar fragments and the melting point. The analysis showed that the PSA is positively related to the melting point. We interpreted this as the PSA being a driving force for the formation of stable crystal structures, since the presence of polar functional groups allows the formation of energetically favorable intermolecular hydrogen bonds within the solid structure. Nonpolar measures proved to be negatively correlated to the melting point (Figure 5), and this was interpreted as follows: crystals of molecules with large nonpolar surface areas will be less stable, and, thus, the melting point will be lower for such compounds than for compounds incorporating polar fragments in the structure. These results are supported by the differences in strength between the intermolecular interactions between polar fragments (hydrogen bonds) and the van der Waals interactions between nonpolar fragments. Interestingly, the calculated lipophilicity value (ClogP_{oct}) was not selected as an important descriptor for any of the models. This might partly be explained by the selection of several other measures of nonpolarity and hydrophobicity, i.e. the NPSA, the relative surface area of neutral hydrogen atoms bound to carbon atoms and the electron accessibility between carbon atoms, which overlap with the properties incorporated in ClogP_{oct}.

The information obtained from the combined matrix using both 2D and 3D descriptors in the analysis of the melting point (model III) showed comparable results as observed in the two separate analyses (models I and II). However, the 2D descriptor for the PSA selected in model I was now replaced by the corresponding 3D descriptor. This 3D descriptor is calculated for a low-energy conformation of the molecules, which improves the accuracy of the surface area calculations, and could explain the improvement obtained when the 3D generated PSA is chosen rather than the 2D one. When using the combined matrix we obtained a model that was slightly better than model I and fewer variables were required.

The results obtained in models I−III will also influence the selection of variables when modeling aqueous drug solubility as the solid state is an important property affecting the water solubility. In previous publications,[13,14] we have successfully predicted solubility from the calculated molecular surface properties. These PLS analyses resulted in the selection of descriptors restricting solubility such as the NPSA and size descriptors. Only two hydrogen bonding descriptors were identified as being important for the solubility (the surface areas of double bonded oxygen atoms and hydrogen atoms bound to nitrogen). However, the selection of these polar measures seemed to be more data set dependent than the selection of nonpolar descriptors, which we considered as more general descriptors for the solubility prediction. We have so far worked with three different data sets, ranging from 17 compounds in a small data set[13] up to 300 compounds in the largest data set (unpublished data). In none of these analyses has the PSA proved to be an important descriptor for solubility. We speculated that this was dependent on the solid state and that the PSA contributes to the stabilizing forces in the crystal structure rather than being a driving force in the solvatization process.[13] In contrast, molecules with large nonpolar surface areas do not form as strong intermolecular interactions as molecules also containing polar fragments and the crystals can, therefore, be expected to dissolve faster in a proper, hydrophobic solvent in accordance with the "like dissolves like" theory. This interpretation is in agreement with the results obtained in the present study. Moreover, in agreement with the "like dissolves like" theory, compounds with large lipophilic/nonpolar parts will not be highly soluble in water, resulting in that nonpolar atoms are negatively correlated to aqueous solubility in our solubility models.[13,14] Thus these results have led us to the following conclusion: the solubility of nonpolar compounds is not restricted by the formation of strong crystal structures, but rather, by the energy needed in the solvatization process. Moreover, the solubility of compounds with large PSAs is mainly restricted by the formation of strong crystal structures. To efficiently stabilize crystal structures, the compounds should contain functional groups with the capability to both donate and accept hydrogen bonds. The conclusions drawn from our models regarding the influence of hydrophobic and hydrophilic descriptors on melting point and solubility are supported by experimental and theoretical studies performed by others.[26−29] For instance, Stella and co-workers found the melting point of phenytoin prodrugs to decrease and the solubility in glycerol esters to increase as the hydrophobicity of the prodrug increased.[28]

With the intention of improving the accuracy of the melting point prediction, an averaged consensus approach was taken. The three predicted melting point values of each compound obtained from PLS models I−III were averaged in model IV (Table 1). This resulted in a more robust model with a higher R^2 (0.63) and lower RMSE values (RMSE$_{tr}$ of 35.1 °C and RMSE$_{te}$ of 44.6 °C) than those obtained in models I and III. However, the RMSE values were still rather high, and, most importantly, the accuracy of the classification did not improve. Thus, for the data set investigated in this work and for the 2D and 3D descriptors used, we conclude that model I, based on the most rapidly generated descriptors, can be used with an accuracy equal to that of the more elaborate models III and IV.

To improve the accuracy of melting point predictions, two main approaches need to be taken. First, an in-house database with a large number of high quality experimentally determined melting points of drug-like molecules should be generated. Preferably, these measurements should be performed with the same experimental settings, i.e. DSC. Next, the descriptor space has to be increased. It is expected that other response parameters influence the melting point, such as crystal packing forces, heat capacity, and the enthalpy and entropy of fusion, and these need to be studied theoretically to generate calculated descriptors which would be useful to explain these phenomena. If these two criteria are fulfilled, we believe that in silico predictions of several solid-state characteristics will be possible.

CONCLUSION

In conclusion, we have performed a theoretical analysis of molecular properties determining the melting point, a solid state characteristic commonly used in drug discovery and drug development. Our results show that rapidly generated molecular descriptors can explain 63% of the melting point variation of drug-like molecules, and descriptors generated from the 2D representation of the molecule were more successful in the prediction of melting point than were descriptors generated from the 3D conformation. Descriptors for hydrophilicity/polarity, partial atom charge, and rigidity were found to increase the melting point, whereas nonpolar descriptors and descriptors for molecular flexibility lowered the melting point. The accuracy of the models allows a qualitative classification of the melting point to be made, with good separation between compounds with low and high melting points. However, to improve the classification and the accuracy of the predictions, we expect that more specific descriptors for solid-state characteristics need to be incorporated. Finally, we could conclude from the present work that solubility predictions based on molecular surface areas[13,14] can be explained partly from the properties of the solid state.

ACKNOWLEDGMENT

This work was supported by grant No. 9478 from the Swedish Research Council, the Swedish Foundation for Strategic Research, the Knut and Alice Wallenberg Foundation, and the Swedish Fund for Research without Animal Experiments. We are grateful to Helena Hansson for assistance in collecting the data.

Supporting Information Available: A table of the observed values taken from the Merck Index and the predicted values obtained from the different models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Yalkowsky, S. H.; Valvani, S. C. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **1980**, *69*, 912−922.

(2) Yalkowsky, S. H. Solubility and partitioning V: Dependence of solubility on melting point. *J. Pharm. Sci.* **1981**, *70*, 971−973.

(3) Yalkowsky, S. H.; Pinal, R. Estimation of the aqueous solubility of complex organic compounds. *Chemosphere* **1993**, *26*, 1239−1261.

(4) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.

(5) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(6) Law, D.; Wang, W.; Schmitt, E. A.; Long, M. A. Prediction of poly(ethylene) glycol-drug eutectic compositions using an index based on the van't Hoff equation. *Pharm. Res.* **2002**, *19*, 315−321.

(7) Nikmo, J.; Kukkonen, J.; Riikonen, K. A model for evaluating physicochemical substance properties required by consequence analysis models. *J. Hazard. Mater.* **2002**, *91*, 43−61.

(8) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(9) 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.

(10) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

(11) Huuskonen, J. Estimation of aqueous solubility in drug design. *Comb. Chem. High Throughput Screening* **2001**, *4*, 311−316.

(12) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* **2002**, *19*, 497−503.

(13) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of aqueous drug solubility. *Pharm. Res.* **2002**, 182−188.

(14) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption classification of oral drugs from molecular surface properties. *J. Med. Chem.* **2003**, *46*, 558−570.

(15) Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.* **2001**, *44*, 1927−1937.

(16) *The Merck Index*, 12th ed.; Budavari, S., Ed.; Merck & Co, Inc: Whitehouse Station, 1996.

(17) Molconn-Z v.3 15S, Hall Associates Consulting, Quincy, MA.

(18) Olsson, T.; Sherbukhin, V. Synthesis and Structure Administration (SaSA), AstraZeneca R&D Mölndal.

(19) Pearlman, R. S.; Smith, K. M. *New software tools for chemical diversity*, in *3D-QSAR and drug design*; Kubinyi, H., Martin, Y., Folkers, G., Eds.; Kluwer Academic: Dordrecht, 1997.

(20) *Connectivity in structure−activity analysis;* Kier, L. B., Hall, L. H., Eds.; Research Studies Press: John Wiley and Sons: Letchwort, 1986.

(21) The program MAREA is available upon request from the authors. The program is provided free of charge for academic users. Contact Johan Gråsjö (e-mail johan.grasjo@farmaci.uu.se).

(22) Gajewski, J. J.; Gilbert, K. E.; McKelvey, J. MMX an enhanced version of MM2. *Adv. Mol. Model.* **1990**, *2*, 65−92.

(23) Jackson, E. J. *A users guide to principal components;* Wiley: New York, 1991.

(24) Höskuldsson, A. PLS regression methods. *J. Chemometrics* **1988**, *2*, 211−228.

(25) Simca-P v. 8.0; Umetrics AB, Box 7960, SE-907 19 Umeå, Sweden.

(26) Li, Z. J.; Ojala, W. H.; Grant, D. J. Molecular modeling study of chiral drug crystals: lattice energy calculations. *J. Pharm. Sci.* **2001**, *90*, 1523−1539.

(27) Adsmond, D. A.; Grant, D. J. Hydrogen bonding in sulfonamides. *J. Pharm. Sci.* **2001**, *90*, 2058−2077.

(28) Yamaoka, Y.; Roberts, R. D.; Stella, V. J. Low-melting phenytoin prodrugs as alternative oral delivery modes for phenytoin: a model for other high-melting sparingly water-soluble drugs. *J. Pharm. Sci.* **1983**, *72*, 400−405.

(29) Rollinger, J. M.; Gstrein, E. M.; Burger, A. Crystal forms of torasemide: new insights. *Eur. J. Pharm. Biopharm.* **2002**, *53*, 75−86.

CI020280X