Molecular Characteristics for Solid-State Limited Solubility

Carola M. Wassvik,[†,§] Anders G. Holmén,[‡] Rieke Draheim,[†,#] Per Artursson,[†] and Christel A. S. Bergström*[,†]

Pharmaceutical Screening and Informatics, Department of Pharmacy, Uppsala University, Uppsala Biomedical Centre, Post Office Box 580, SE-75123 Uppsala, Sweden, and Lead Generation, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden

Received December 18, 2007

Solubility and solid-state characteristics were determined and multivariate data analysis was used to deduce structural features important for solid-state limited solubility of marketed drugs. Molecules with extended ring structures and large conjugated systems were less soluble, indicating that structural features related to rigidity and aromaticity result in solubility restricted by stable crystal structures. These descriptors successfully predicted the applied test set and can be useful for avoiding synthesis of compounds behaving like "brick dust".

Introduction

It is well recognized that, over the past few decades, drug leads and candidates have increased in size and lipophilicity, making it more likely that they will be poorly soluble in aqueous media.[1,2] If the solubility remains poor in the gastrointestinal fluids, poor absorption and a low bioavailability may result. This is further complicated by the fact that it is not always possible to find suitable formulations for these compounds.[3] It would be desirable to identify poorly soluble molecules at an early stage of drug discovery to avoid their selection for synthesis. This could potentially be achieved with computational models for drug solubility.

We have previously studied poorly soluble compounds for which the solubility was found to be restricted by poor solvation.[4] The lipophilicity interval of the compounds studied, as identified with the calculated octanol–water partition coefficient (ClogP[a]), was 3.5−6.8 with a mean ClogP value of 5.3. Our observation that the solvation process, i.e., the incorporation of the drug molecule into the water, was the most important factor contributing to the poor solubility of this data set is in agreement with the general solubility equation (GSE), eq 1, established by Yalkowsky and co-workers.[5] The GSE describes the influence of the solvation (represented by the octanol–water partition coefficient, log *P*) and the solid state (represented by the melting point; T_m) on solubility:

$$\log S_0 = 0.5 - 0.01(T_m - 25) - \log P \qquad (1)$$

where log S_0 is the intrinsic solubility in molar units. A closer examination of the predictions of this equation, using hypothetical compounds with low, intermediate, and high melting points and a large interval in lipophilicity is shown in Figure 1. It is clear that compounds with lower lipophilicity are more likely to display poor solubility resulting from a stable crystal structure (gray circle segment) than highly lipophilic compounds, which instead are solubility limited by poor solvation (white circle segment). For a compound with a T_m of 250 °C and a ClogP of 2, the GSE approximates the solubility to be governed by the solid state to 52%, whereas at a ClogP of 6 only 27% of the solubility of such high melting compounds is related to the solid state.

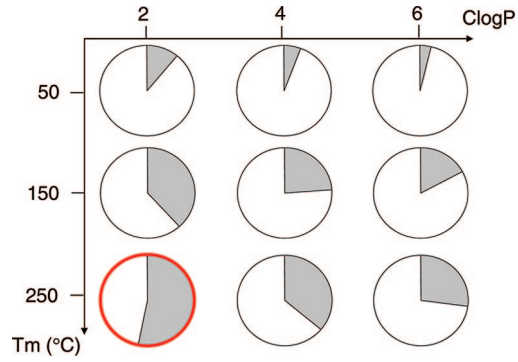

Figure 1. Influence of the solid state (gray) and the lipophilicity (white) on the solubility, as predicted by the GSE equation. At log $P \leq 2$ only, it is likely to find compounds whose solubility is mainly restricted by the solid state and not by solvation, identified with the red circle.

In this work, our aim was to identify structural features resulting in solid-state limited solubility. With the aid of the GSE equation (eq 1, Figure 1), we selected compounds with a ClogP value of ∼2. The intrinsic aqueous solubility (S_0) and solid-state characteristics were determined experimentally, and the dependence of the solubility on solvation and solid-state characteristics was investigated by multivariate statistical tools.

Results

Regression Analysis of log S_0 versus ClogP and Experimental Solid-State Properties. The data set studied in this work (Figure 2) displayed more than a 1000-fold range in S_0 (−1.75 to −4.83 on a log molar scale) and only a 10-fold range in ClogP (1.70–2.81). The R^2 of the correlation between log S_0 and ClogP of 0.54 for the 299 compounds that we investigated in a previous study[6] indicates that log S_0 is clearly dependent on the lipophilicity for a general druglike data set (Figure 3). In contrast, the subset of 20 compounds studied experimentally in this work shows no correlation to ClogP ($R^2 = 0.04$), demonstrating that for this data set the solubility was related to factors other than the lipophilicity.

* To whom correspondence should be addressed. Phone: +46 18 4714645. Fax: +46 18 4714223. E-mail address: christel.bergstrom@farmaci.uu.se.

† Uppsala University.

§ Current address: Carola Wassvik, Molecular Informatics, Johnson & Johnson PRD Spain, Janssen-Cilag S.A., Calle Rio Jarama 75, ES-45007 Toledo, Spain.

‡ AstraZeneca R&D Mölndal.

Current address: Rieke Draheim, Institute of Pharmaceutics and Biopharmaceutics, Heinrich-Heine-University Duesseldorf, D-40225 Duesseldorf, Germany.

[a] Abbreviations: ClogP, calculated octanol–water partition coefficient; GSE, general solubility equation; T_m, melting point; S_0, intrinsic solubility; ΔH_m, enthalpy of melting; ΔS_m, entropy of melting; PCA, principal component analysis; PLS, partial least-squares projection to latent structures.

10.1021/jm701587d CCC: $40.75
Published on Web 04/09/2008

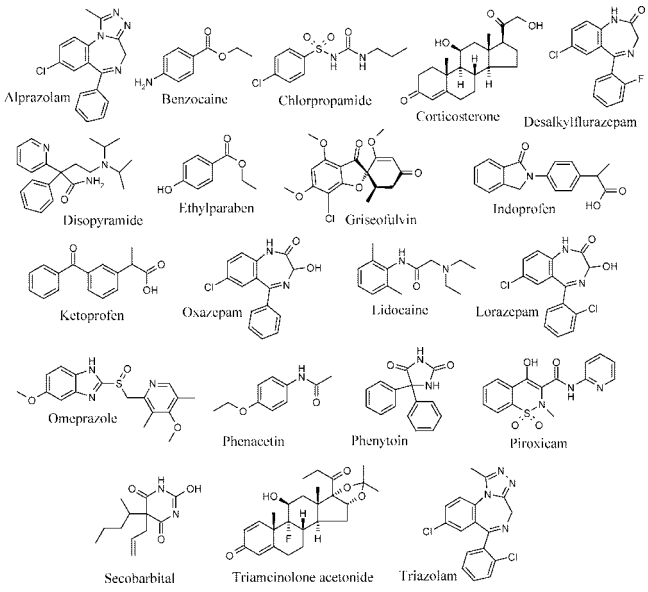

Figure 2. Chemical structures of the compounds studied.

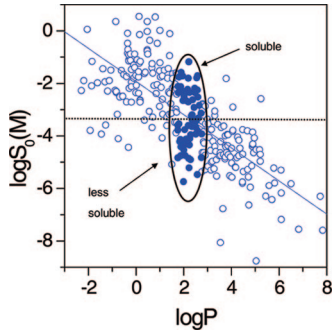

Figure 3. Correlation between log S_0 and ClogP for the 299 compounds compiled previously by Bergström et al.[13] The primary data set of 73 compounds with a ClogP value of ~2 is displayed as filled circles, and the cutoff value of log $S_0 = -3.31$, introduced to divide the compounds into two groups for the validation of the computational model, is shown.

The experimental values for all compounds in the data set are listed in Table 1. Regression analysis was conducted for log S_0 and each of the properties T_m, enthalpy of melting (ΔH_m), and entropy of melting (ΔS_m) with the intention of investigating which of them governed the solubility of these 20 compounds. Lorazepam and oxazepam showed unusually high values of ΔH_m and ΔS_m, and according to Grubb's test at the 95% confidence level, they were statistical outliers of the data set. A possible molecular explanation for the high values of lorazepam and oxazepam could be that the hydroxyl group in the C3 position and the unsubstituted nitrogen in position 1 in the diazepine ring can form dimers in the crystal through hydrogen bonding.[7] The log S_0 correlated well with both T_m ($R^2 = 0.70$, Figure 4) and ΔH_m, ($R^2 = 0.71$) but less so with ΔS_m ($R^2 = 0.31$). Thus, for the compounds investigated, the solubility values were independent of the lipophilicity, and instead, solid-state properties such as T_m and ΔH_m were important determinants for the solubility of these compounds.

Molecular Descriptors for log S_0 and Validation with an External Test Set. We were interested in deducing the structural features related to poor solubility in the narrow lipophilicity range around 2, which the GSE equation had identified as the lipophilicity value below which the solid state properties were likely to be major determinants of the solubility. Thus, we modeled our solubility data using 2D molecular descriptors and multivariate data analysis, which resulted in a model with $R^2 = 0.76$, $Q^2 = 0.75$, and RMSE = 0.45 log units. As shown in Figure 5, the solubility was related to the rigidity, captured by the number of rigid bonds, the Balaban index, and the number of rigid fragments and to the aromaticity of the molecule, identified with the Min eV #2 and Max eV #3 descriptors. For this data set, an increase in rigidity resulted in a decrease in solubility. This is in contrast to the solvation limited data set that we studied previously,[4] in which the rigidity of the molecule improved the solubility, most likely through the reduced demand of a large cavity formation in the water. In contrast, in the solid-state limited data set, the rigidity captures the stability of the solid state. It is well-known that flexible compounds do not form such stable crystal lattices as rigid ones because of the conformational freedom of the molecules. The rigidity of the molecule can further be linked to the Balaban index, a commonly encountered topological index.[8] This is a measure of molecular shape that is essentially independent of molecular size and the number of rings present in the molecule. Flat, extended ring structures (more rigid molecules) result in low values of the Balaban index, which in turn result in less soluble compounds. Finally, it was found that a high aromaticity will decrease the solubility. This identifies the importance of the nonspecific $\pi-\pi$-interactions for the stability of the crystal.

The model was tested on an external test set comprising the 53 compounds left after the primary selection of compounds with a narrow ClogP interval. The compounds, both training and test sets, were sorted into "soluble" or "less soluble" based on their position in the log S_0 vs log P plot (Figure 3). Compounds above the trend line were classified as "soluble", whereas compounds below the trend line were classified as "less soluble". The descriptors (Figure 5b) successfully predicted these groups, with 95% of the compounds in the training set being predicted accurately and a total of 79% of the test set (Table 2 and Supporting Information, Table S1). Notably, only 4 of the 29 "less soluble" compounds in the test set were falsely predicted as being "soluble".

Molecular Descriptors for T_m. The solubility of this data set was highly correlated to the T_m (Figure 4), making an investigation of which molecular descriptors were of importance for the T_m of interest. The resulting model ($R^2 = 0.74$, $Q^2 = 0.71$, and RMSE = 35.9 °C) captured structural features related to large ring structures, shape, and rigidity and were essentially the same as for the solubility model (Figure 6). Primarily, they encoded the lack of flexibility as a structural feature leading to a high T_m. Thus, a molecule like alprazolam that is flat and rigid because it has several interconnected rings is predicted to have a high T_m, while it is anticipated that a molecule like benzocaine that is small with flexible side chains will have a low T_m. These findings are in agreement with previous studies of structural features related to the melting point.[9,10]

Discussion

From the modeling results it becomes clear that molecules suffering from solid-state limited solubility are rigid and have a high aromaticity. These features are directly related to an increased stability of the solid state, i.e., a higher melting point and/or enthalpy of melting, since the molecules in our data set with extended ring systems and conjugated fragments display higher melting points and lower solubilities than molecules lacking these features. As an example of how the descriptors relate to the solubility, consider the structures presented in Figure 7. Compounds resulting in a low solubility and high melting

Table 1. Characteristics of the Compounds Studied

compd	CAS no.[a]	MW (g/mol)	log S_0 ± SD[b] (M)	pK_a[c]	acid/base[d]	ClogP[e]	T_m ± SD[f] (°C)	ΔH_m ± SD[g] (kJ/mol)	ΔS_m ± SD[h] (J/(mol·K))
alprazolam	28981-97-7	308.8	−3.60 ± 0.00	2.4	b	2.56	228.6 ± 0.2	32.0 ± 0.61	63.9 ± 1.22
benzocaine	94-09-7	165.2	−2.34 ± 0.03	2.5	b	1.92	89.4 ± 0.2	24.6 ± 0.30	67.8 ± 0.86
chlorpropamide	94-20-2	276.7	−3.30 ± 0.00	4.8	a	2.35	128.0 ± 0.1	25.7 ± 0.41	64.0 ± 1.02
corticosterone	50-22-6	346.5	−3.28 ± 0.02	n	n	2.32	185.3 ± 0.1	35.5 ± 0.50	77.5 ± 1.10
desalkylflurazepam	2886-65-9	288.7	−3.72 ± 0.02	11.6	a	2.81	208.0 ± 0.1	30.7 ± 1.03	63.9 ± 2.14
disopyramide	3737-09-5	339.5	−2.38 ± 0.03	10.1	b	2.58	96.5 ± 0.1	26.7 ± 0.54	72.3 ± 1.47
ethylparaben	120-47-8	166.2	−2.38 ± 0.01	8.3	a	2.51	116.0 ± 0.1	27.9 ± 0.75	71.6 ± 1.92
griseofulvin	126-07-8	352.8	−4.83 ± 0.08	n	n	1.91	218.2 ± 0.0	44.7 ± 0.78	90.8 ± 1.59
(±)-indoprofen	31842-01-0	281.3	−4.72 ± 0.12	4.6	a	2.74	211.6 ± 0.5	40.3 ± 2.38	83.2 ± 4.48
(±)-ketoprofen	22071-15-4	254.3	−3.52 ± 0.01	4	a	2.76	95.0 ± 0.1	37.3 ± 0.33	101.2 ± 0.12
lidocaine	137-58-6	234.3	−1.75 ± 0.00	8.5	b	1.95	67.8 ± 0.2	18.8 ± 0.53	55.1 ± 1.57
(±)-lorazepam	846-49-1	321.2	−3.74 ± 0.07	11.5	a	2.37	180.0 ± 0.3	75.2 ± 2.19	165.9 ± 4.71
(±)-omeprazole	7359-58-6	345.4	−3.40 ± 0.02	8.9/4.1	a/b	1.70	164.5 ± 0.3	na	na
(±)-oxazepam	604-75-1	286.7	−4.19 ± 0.02	n	n	2.31	205.6 ± 0.4	86.4 ± 0.01	180.5 ± 0.13
phenacetin	62-44-2	179.2	−2.48 ± 0.00	n	n	1.77	134.4 ± 0.1	34.1 ± 0.92	83.7 ± 2.27
phenytoin	57-41-0	252.3	−4.15 ± 0.04	8.3	a	2.09	295.8 ± 0.3	40.1 ± 0.75	70.4 ± 2.97
piroxicam	36322-90-4	331.4	−4.03 ± 0.01	4.5/3.6	a/b	1.89	200.5 ± 0.5	36.3 ± 0.15	76.7 ± 0.25
secobarbital	76-73-3	238.3	−2.36 ± 0.01	7.8	a	2.16	98.6 ± 0.1	22.9 ± 0.86	61.7 ± 2.31
triamcinolone acetonide	76-25-5	434.5	−4.46 ± 0.01	n	n	2.21	302.3 ± 1.1	na	na
triazolam	28911-01-5	343.2	−4.04 ± 0.01	2.3	b	2.62	241.3 ± 0.2	41.0 ± 1.00	79.7 ± 1.96
min		165.2	−4.83			1.70	67.8	18.8	55.1
max		434.5	−1.75			2.81	302.3	86.4	180.5

[a] Chemistry Abstracts Service registry number (CAS no.). [b] Intrinsic solubility (log S_0) expressed as the average log molar concentration ± standard deviation (SD). SD values of 0.00 have the experimental error in the third decimal. [c] n = neutral compound (no proteolytic function in the pH range 2–12). Values for pK_a were experimental, obtained from ref 17, or calculated from ref 18. [d] a = acid, *b* = base, n = neutral, a/b = ampholyte. [e] The calculated octanol–water partition coefficient (ClogP) using the Daylight software, version 4.9. [f] Melting point (T_m) expressed as the average value ± standard deviation (SD). [g] Enthalpy of melting (ΔH_m) expressed as the average value ± standard deviation (SD). na = not applicable (decomposing compound). [h] Entropy of melting (ΔS_m) expressed as the average value ± standard deviation (SD). na = not applicable (decomposing compound).

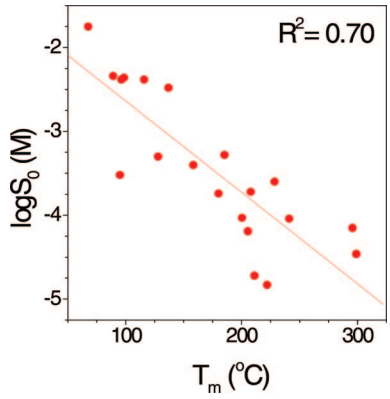

Figure 4. Regression analysis of the solubility versus T_m. A more detailed analysis of the solid state showed that the melting enthalpy was highly correlated to the solubility (R^2 = 0.71), whereas the correlation between solubility and melting entropy only was R^2 = 0.31.

point all have a large exposed flat area and are rigid in comparison to the highly soluble, low melting compounds. In contrast, the molecules with flexible side chains can only form weak nonspecific interactions in the solid state, making the whole crystal lattice weak.

Neither the solubility model nor the melting point model identified any descriptor related to the hydrogen bond properties of the molecules as an important determinant, a feature that has been identified previously as being important for the formation of a stable crystal lattice.[9,11,12] However, compounds with a large number of intermolecular hydrogen bonds in the crystal, e.g., phenytoin as identified in the Cambridge Structural Database, version 5.27 (Cambridge Crystallographic Data Center, U.K.), were correctly sorted using these rigidity and aromaticity descriptors. We note that the number of hydrogen bond donors is essentially equal for the compounds in the data set, varying between 0 and 2, whereas the number of hydrogen bond acceptors varies between 2

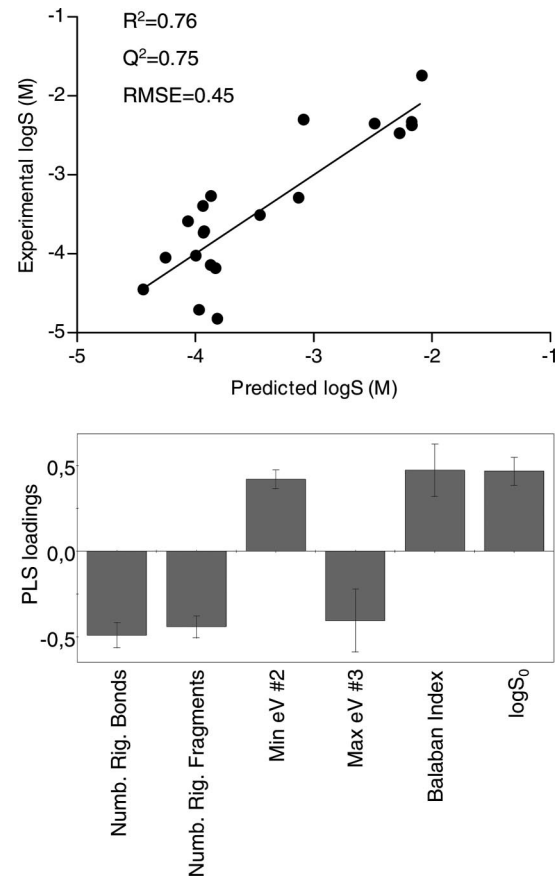

Figure 5. (a, top) Prediction obtained from the PLS solubility model. (b, bottom) Loadings of molecular descriptors: the number of rigid bonds (Numb. Rig. Bonds), the Balaban index, the number of rigid fragments (Numb. Rig. Fragments), the second smallest eigenvalue (Min eV #2), and the third largest eigenvalue (Max eV #3). The error bars correspond to the limits of the 95% confidence interval.

Table 2. Correct Classification of the Compounds Using Molecular Descriptors[a]

class[b]	less soluble$_{tr}$,[c] %	soluble$_{tr}$,[d] %	less soluble$_{te}$,[e] %	soluble$_{te}$,[f] %
less soluble	100	0	86	14
soluble	5	95	29	71

[a] The complete list and the predictions are given in the Supporting Information. The model used for prediction is presented in Figure 4. [b] The classes were based on the position of the relation between ClogP and log S_0 for a general data set; the cutoff value was −3.31 on a log molar scale. [c] Fourteen compounds were included in this group. [d] Six compounds were included in this group. [e] Twenty-nine compounds were included in this group. [f] Twenty-four compounds were included in this group.

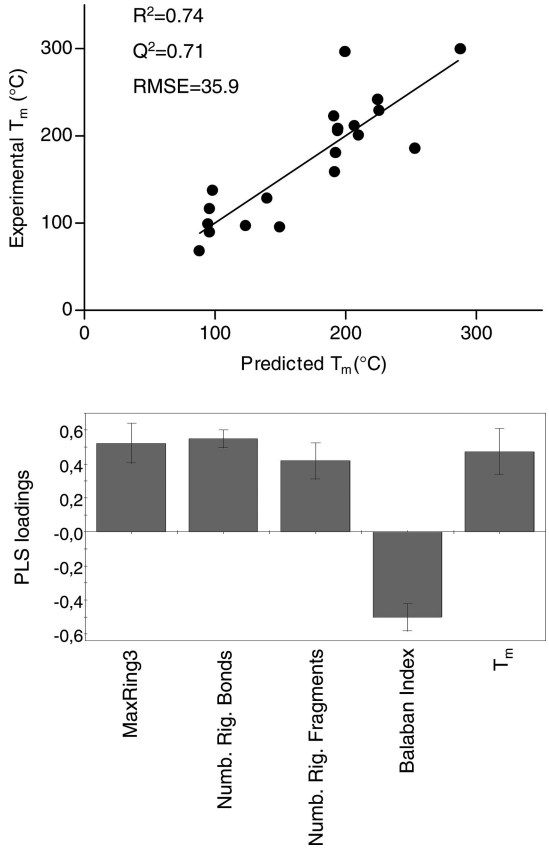

Figure 6. PLS melting point model. (a, top) Observed vs predicted T_m. (b, bottom) Loadings of molecular descriptors: the third largest ring (MaxRing3), the number of rigid bonds (Numb. Rig. Bonds), the number or rigid fragments (Numb. Rig. Fragments), and the Balaban index. Error bars correspond to the limits of the 95% confidence interval.

and 8. Hence, we speculate that the limited range of the number of hydrogen bond donors for this data set results in the intermolecular hydrogen bonds within the crystal not being captured by the models. It is probable that as the lipophilicity decreases, the hydrogen bonding capacity will be a more pronounced determinant for the compounds displaying a solid-state limited solubility. However, such compounds seldom do become truly poorly soluble compounds as indicated in Figure 3, where only a few compounds with a log $P < 0$ have a lower solubility than 100 μM.

The accurate computational classification of the compounds in this work highlights the opportunities to use rapidly calculated molecular descriptors as identifiers of compounds exhibiting solid-state limited solubility. Since most reports on predictions of solubility have lacked a specific component dealing with the solid state properties, we believe an improved computational filter for identifying target compounds behaving like "brick dust" can be developed on the basis of the findings in this work.

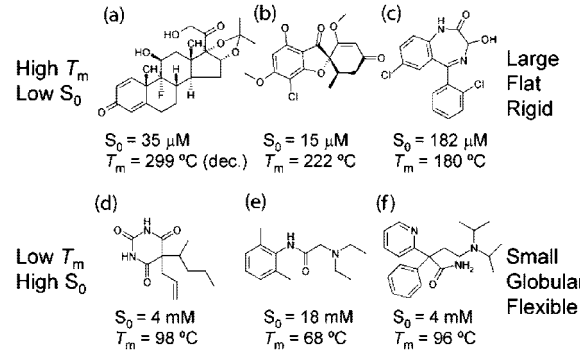

Figure 7. Structural features that are indicative of low solubility owing to the stability of the crystal structure (top panel) and of high solubility caused by weak crystal structure (bottom panel) for the data set studied. The compounds are (a) triamcinolone acetonide, (b) griseofulvin, (c) lorazepam, (d) secobarbital, (e) lidocaine, and (f) disopyramide.

However, the investigation was based on marketed drugs and a future challenge is to forecast lead compounds of "brick dust" type. The application of the current model to this chemical space is currently under investigation in our laboratories.

Conclusions

We have investigated which molecular characteristics will enhance the risk of producing a solid-state limited solubility. Descriptors such as aromaticity and rigidity were found to govern solid-state limited solubility, flagging large, flat, and rigid molecules with extended ring structures and conjugated π-electron systems as less soluble. When these descriptors were used to predict an external test set, 86% of all less soluble compounds were identified. We suggest that such calculated molecular descriptors can be used for rapid identification of synthetic target compounds with a high risk of having solid-state limited solubility and thereby reduce the risk of synthesizing "brick dust" molecules.

Experimental Section

Data Set Selection and Compounds Studied. A data set was selected to incorporate as wide a range of solubility as possible, simultaneously maintaining a narrow variation in ClogP. On the basis of the GSE (Figure 1),[5] we decided to focus on compounds with a ClogP of around 2. Druglike compounds with such a value were compiled from Bergström et al.,[13] the AquaSol database (University of Arizona, AZ), and SciFinder Scholar (American Chemical Society). In total, this gave 73 compounds; of these, 20 were commercially available, not too expensive in their free form, stable in solubility studies and differential scanning calorimetry (DSC) experiments and maintained the requirements for structural diversity (Figure 2 and Table S1, Supporting Information). Chlorpropamide was purchased from MP Biomedicals LLC, OH; phenytoin from Lancaster Synthesis Ltd., Heysham, Lancashire, U.K.; desalkylflurazepam, lorazepam, and oxazepam from Larodan Fine Chemicals AB, Malmö, Sweden. All other drugs were from Sigma-Aldrich Chemie GmbH, Germany. The purity of the drugs used was greater than 98%, with the exception of griseofulvin (a natural product), which had a purity of 96%.

Differential Scanning Calorimetry (DSC). Thermograms were recorded using a Seiko instrument consisting of a DSC220C analysis module with automatic cooling controller (Seiko Instruments Inc., Japan). Triplicate samples of 1–3 mg were weighed in aluminum pans, which were then sealed and pierced (TA Instruments, DE). Samples of each compound were heated from room temperature to approximately 50 °C above the melting temperature at a rate of 10 °C min^{-1} and purged with nitrogen gas at a flow rate of 80 mL min^{-1}. If any anomalies such as the existence of an asymmetric peak shape, multiple melting endotherms, or recrystallization exotherms were detected, samples were run at a heating rate of 2 °C min^{-1} to allow

further investigations to be made. Omeprazole and triamcinolone acetonide decomposed directly after melting, even though increased heating rates were used. Chlorpropamide, disopyramide, and triazolam exhibited behavior indicative of polymorphism and were therefore converted to the polymorph with the highest melting point.

Solubility Determinations. The S_0 of the crystalline compounds was determined in quadruplicate using the shake-flask method, as described in detail previously.[14] The method was modified for the determination of two of the compounds, corticosterone and omeprazole. Corticosterone exhibited a large standard deviation for the initial measurements after 24 h, so a time study over 11 days (264 h) was undertaken This was undertaken because long equilibration times have often been observed for steroids.[14,15] Our results revealed that corticosterone needed at least 144 h to equilibrate. The data for corticosterone that was used in the analysis were collected at the end of the period, after 264 h. Omeprazole was not chemically stable for 24 h in water, so the stability at different pHs (7, 7.5, 10, and 12) was tested (MilliQ water adjusted with 0.01 M NaOH). At pH 10 and above, water solutions of omeprazole were stable over 24 h, and pH 10 was chosen for the determination. Every solubility value of the quadruplicate was then calculated individually using the Henderson–Hasselbalch equation. For the calculation, experimental values obtained from Wan et al.[16] were used, giving pK_a of 8.9 (acid) and 4.1 (base).

Molecular Descriptors. CLOGP, version 4.9, from Daylight Chemical Information Systems, Inc. (Aliso Viejo, CA), was used to calculate the log *P* values. Further, a total of 93 2D descriptors related to the molecular size, polarity, flexibility, charge distribution, and connectivity were calculated with the AstraZeneca in-house program Selma. They included well-known and commonly used 2D descriptors from different commercial sources, such as the molecular refractivity (CMR), atom counts, rings counts, number of rotatable and rigid bonds, all obtained from Daylight; BCUT (Burden−Chemical Abstracts−University of Texas) parameters;[17] topological indices (Wiener, [18] Balaban, [8] Motoc and Randić[19] indexes); shape and connectivity indices from the Kier and Hall suits of descriptors;[20] element counts; Gasteiger charges;[21] and HYBOT hydrogen bond parameters (TimeTec Inc., Newark, DE).

Statistical Analysis. The principal component analysis (PCA) and the partial least-squares projection to latent structures (PLS) were performed in version 11 of the Simca-P software (Umetrics AB, Umeå, Sweden). The data were mean-centered and scaled to unit variance. A variable selection was applied to decrease the complexity of the models and facilitate interpretation. First, the bottom 50% of the variables exhibiting the lowest level of importance was excluded. Second, variables duplicating the information contained within other variables (residing in the same area of the PLS loading plot) were excluded to leave just a few (3–7) variables representing the key descriptors that encoded the majority of the information related to the response variable. The aim of the variable selection was to maintain predictivity and increase the robustness of the model by removing information that was not directly related to the response variable (i.e., noise). The accuracy of the PLS models was judged by the R^2 and RMSE. The models were validated by cross-validated R^2 (Q^2) and permutation tests (100 iterations) in which the values for the response variable were randomized and the multivariate data analysis was repeated to detect whether chance correlations had occurred. Finally the solubility model was challenged by a test set. The test set applied was extracted from the literature, and since it is well-known that reported solubility values differ largely, we performed this as a qualitative assessment. The experimental solubility data were clustered into two groups based on the cutoff value found by the log S_0−ClogP trend line in Wassvik et al.[6] The mean ClogP value for the 20 compounds investigated in this study ($ClogP_{mean} = 2.28$) was inserted in eq 3 obtained in Wassvik et al., resulting in a cutoff value of log S_0 of −3.31. Compounds having a higher solubility than this were classified as "soluble", whereas compounds with lower solubility were sorted as "less soluble". Thereafter, the remaining compounds from the primary data set selection ($n = 53$) were predicted with the PLS model and the quantitative values transformed into these two qualitative measures.

Acknowledgment. This work was supported by AstraZeneca R&D Mölndal, Sweden; the Swedish Research Council grant no. 9478; the Swedish Fund for Research without Animal Experiments and the Knut and Alice Wallenberg Foundation.

Supporting Information Available: Smiles of the training and test sets used for the solubility model and the tabulated results of the solubility predictions (training and test sets). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44* (1), 235–249.
(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(3) Pouton, C. W. Formulation of poorly water-soluble drugs for oral administration: physicochemical and physiological issues and the lipid formulation classification system. *Eur. J. Pharm. Sci.* **2006**, *29* (3–4), 278–287.
(4) Bergström, C. A. S.; Wassvik, C. M.; Johansson, K.; Hubatsch, I. Poorly soluble marketed drugs display solvation limited solubility. *J. Med. Chem.* **2007**, *50* (23), 5858–5862.
(5) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90* (2), 234–252.
(6) Wassvik, C. M.; Holmen, A. G.; Bergström, C. A. S.; Zamora, I.; Artursson, P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur. J. Pharm. Sci.* **2006**, *29* (3–4), 294–305.
(7) Van den Mooter, G.; Ven den Brande, J.; Augustijns, P.; Kinget, R. Glass forming properties of benzodiazepines and co-evaporate systems with poly(hydroxyethyl methacrylate). *J. Therm. Anal. Calorim.* **1999**, *57* (2), 493–507.
(8) Balaban, A. T. Topological indexes based on toplogical distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55* (2), 199–206.
(9) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1177–1185.
(10) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45* (3), 581–590.
(11) Ghosh, S.; Adsmond, D. A.; Huotari, J.; Grant, D. J. Hydrogen-bond patterns of dialkylpyridone iron chelators and their 1:1 formic acid solvates: description, prediction, and role in crystal packing. *J. Pharm. Sci.* **1993**, *82* (9), 901–911.
(12) Adsmond, D. A.; Grant, D. J. Hydrogen bonding in sulfonamides. *J. Pharm. Sci.* **2001**, *90* (12), 2058–2077.
(13) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and local computational models for prediction of aqueous solubility of drug-like molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1477–1488.
(14) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharmacol. Res.* **2002**, *19* (2), 182–188.
(15) Higuchi, T.; Shih, F. M.; Kimura, T.; Rytting, J. H. Solubility determination of barely aqueous-soluble organic solids. *J. Pharm. Sci.* **1979**, *68* (10), 1267–1272.
(16) Wan, H.; Holmen, A. G.; Wang, Y.; Lindberg, W.; Englund, M.; Nagard, M. B.; Thompson, R. A. High-throughput screening of pKa values of pharmaceuticals by pressure-assisted capillary electrophoresis and mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (23), 2639–2648.
(17) Burden, A. T. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
(18) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
(19) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97* (23), 6609–6615.
(20) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986; p 280.
(21) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36* (22), 3219–3228.

JM701587D