molecular pharmaceutics

articles

Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle

David S. Palmer,[†] Antonio Llinàs,[†] Iñaki Morao,[‡] Graeme M. Day,[§] Jonathan M. Goodman,[†] Robert C. Glen,[†] and John B. O. Mitchell*,[†]

The Pfizer Institute for Pharmaceutical Materials Science and Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom, and Pfizer Global Research and Development, Sandwich Laboratories, Sandwich, Kent CT13 9NJ, United Kingdom

Received July 3, 2007; Revised Manuscript Received November 14, 2007; Accepted December 1, 2007

Abstract: We report methods to predict the intrinsic aqueous solubility of crystalline organic molecules from two different thermodynamic cycles. We find that direct computation of solubility, via *ab initio* calculation of thermodynamic quantities at an affordable level of theory, cannot deliver the required accuracy. Therefore, we have turned to a mixture of direct computation and informatics, using the calculated thermodynamic properties, along with a few other key descriptors, in regression models. The prediction of log intrinsic solubility (referred to mol/L) by a three-variable linear regression equation gave $r^2 = 0.77$ and RMSE = 0.71 for an external test set comprising drug molecules. The model includes a calculated crystal lattice energy which provides a computational method to account for the interactions in the solid state. We suggest that it is not necessary to know the polymorphic form prior to prediction. Furthermore, the method developed here may be applicable to other solid-state systems such as salts or cocrystals.

Keywords: ADME; QSPR; crystal; lattice energy; solvation; pharmacokinetics

Introduction

Intrinsic solubility for an ionizable molecule is defined as the concentration of the unionized molecule in saturated aqueous solution at thermodynamic equilibrium at a given temperature.[1] It is related to both pH-dependent solubility and dissolution rate by models such as the Henderson–Hasselbalch equation and Noyes−Whitney equation, respectively.[2–4] For the pharmaceutical industry, the intrinsic solubility of a potential drug candidate is a useful indicator of potential bioavailability and, therefore, there has been much interest in prediction of solubility from molecular structure as an *in silico* method to guide drug discovery.[5,6]

Ab initio methods for the prediction of the solubility of drugs do not yet exist due to the difficulties of modeling both crystalline and solution phases. Instead, the most commonly used methods to predict solubility are empirical quantitative structure–property relationships (QSPR). Since 1990, more than 90 different QSPR models have been published for the prediction of the solubility of organic

* Corresponding author. Mailing address: Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, U.K. Phone: +44-1223-762983. Fax: +44-1223-763076. E-mail: jbom1@cam.ac.uk.

[†] PIPMS and UCMSI, Cambridge.

[‡] Pfizer, Sandwich.

[§] PIPMS, Cambridge.

(1) Horter, D.; Dressman, J. B. Influence of Physicochemical Properties on Dissolution of Drugs in the Gastrointestinal Tract. *Adv. Drug Delivery Rev.* **2001**, *46*, 75–87.

(2) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; Oxford University Press: New York, 1999; p 67.

(3) Noyes, A.; Whitney, W. R. The Rate of Solution of Solid Substances in their own Solutions. *J. Am. Chem. Soc.* **1897**, *19*, 930–934.

(4) Hamlin, W. E.; Northam, J. I.; Wagner, J. G. Relationship Between In Vitro Dissolution Rates and Solubilities of Numerous Compounds Representative of Various Chemical Species. *J. Pharm. Sci.* **1965**, *54*, 1651–1653.

(5) Nigsch, F.; Klaffke, W.; Miret, S. In Vitro Models for Intestinal Absorption. *Exp. Opin. Drug. Metab. Toxicol.* **2007**, *3*, 545–556.

(6) Dokoumetzidis, A.; Macheras, P. A Century of Dissolution Research: From Noyes and Whitney to the Biopharmaceutics Classification System. *Int. J. Pharm.* **2006**, *321*, 1–11.

10.1021/mp7000878 CCC: $37.00
Published on Web 02/22/2008

molecules in water.[7–9] Although many are able to predict the solubility of simple organic molecules reasonably accurately, few are able to predict the solubility of drug molecules with an accuracy much better than approximately 0.7–1 log solubility (referred to mol/L). Three reasons are often given to explain this observation: (1) models do not account for the influence of the solid state;[10] (2) the quality of experimental data is poor;[11,12] and (3) the models are often trained on nondruglike molecules and, hence, the predictions for drug molecules may be poor.[13] In this work, we develop a method for including the influence of the solid state in models to predict solubility and we report some new solubility data for drug molecules, measured using Sirius's Chasing Equilibrium method (CheqSol).[14]

QSPR models attempt to establish a mathematical relationship between the physical property of interest (e.g., solubility) and molecular descriptors calculable from a simple computational representation of the molecule. QSPRs have been used widely for the prediction of many different physical properties and bioactivities.[15] The benefits of these models are that they are computationally inexpensive and may offer reasonably accurate predictions for molecules similar to those in the training set. The problems are that QSPR models may be less accurate for molecules dissimilar to those in the training set and they may also be black boxes, which do not provide information about the underlying physical chemistry. There are also some specific problems with using these methods for predicting solubility. The solubility of a crystalline molecule depends upon the free energy required to remove molecules from the crystal lattice as well as the free energy change for solvation. The molecular descriptors employed in QSPRs normally quantify the properties of single molecules and, thus, they may not be adequate to model the solid state, which depends upon the arrangement and interactions of molecules in the crystal lattice. However, some methods for including solid-state effects into QSPR models have been suggested by, for example, Abraham et al.,[16] Klamt et al.[17] and Johnson et al.[18] The linear solvation−free energy relationships proposed by Abraham et al. model hydrogen bonding in the crystal with a "hydrogen bond acidity × hydrogen bond basicity" term in the regression equation (a similar scheme has also been adopted by Cheng et al.[19]). However, the model ignores other nonbonded interactions and is obviously not a complete solution to the problem. The COSMOtherm method of Klamt et al. makes an estimate of the Gibbs free energy of fusion ($\Delta G_{(fus)}$) from a three-variable linear QSPR model. However, the estimation of $\Delta G_{(fus)}$ is approximate and does not offer a clear improvement over other methods for solubility prediction. Recently, Johnson et al.[18] have used molecular dynamics simulations of the crystal lattice, at a series of escalating temperatures, to calculate a measure of crystal stability, which they use as a *post hoc* correction for a traditional QSPR. For prediction of the intrinsic solubility of 26 drug molecules, this method gave a RMSE = 0.86 in log solubility.

Solubility can also be related to one of two thermodynamic cycles into which it is possible to decompose the thermodynamic equilibrium: (i) transfer of the molecules from crystal to supercooled liquid to solution; (ii) transfer of the molecules from crystal to vapor to solution.[20] The benefits of deriving predictive models from these thermodynamic cycles are that they inherently include information about the solid state and the underlying physical chemistry. The problems include the difficulty in calculating the related free energy terms.

The thermodynamic cycle of crystal to supercooled liquid to solution is problematic because the Gibbs free energy change for transfer from crystal to supercooled liquid is not easily accessible by either experiment or computation. Lüder et al. have developed Monte Carlo simulations to predict

(7) Dearden, J. C. In Silico Prediction of Aqueous Solubility. *Expert Opin. Drug. Discov.* **2006**, *1*, 31–52.
(8) Delaney, J. S. Predicting Aqueous Solubility from Structure. *Drug Discov. Today* **2005**, *10*, 289–295.
(9) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241.
(10) Johnson, S. R.; Zheng, W. Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption. *AAPS J.* **2006**, *8*, E27–E40.
(11) Llinas, A.; Burley, J. C.; Box, K. J.; Glen, R. C.; Goodman, J. M. Diclofenac Solubility: Independent Determination of the Intrinsic Solubility of Three Crystal Forms. *J. Med. Chem.* **2007**, *50*, 979–983.
(12) Clark, T. *Does quantum chemistry have a place in cheminformatics? Molecular Informatics Confronting Complexity*; Hicks, G. M., Kettner, C., Eds.; Proceedings of the Beilstein-Institut Workshop, 2002. See: http://www.beilstein-institut.de/bozen2002/proceedings/. Accessed 21 July 2006.
(13) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355–66.
(14) Stuart, M.; Box, K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases. *Anal. Chem.* **2005**, *77*, 983–990.
(15) Debnath, A. K. Quantitative Structure-Activity Relationship (QSAR) Paradigm - Hansch Era to New Millennium. *Mini Rev. Med. Chem.* **2001**, *1*, 187–195.
(16) 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.
(17) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS. *J. Comput. Chem.* **2002**, *23*, 275–281.
(18) Johnson, S. R.; Chen, X.-Q.; Murphy, D.; Gudmundsson, O. A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects. *Mol. Pharmaceutics* **2007**, *4*, 513–523.
(19) Cheng, A.; Merz, K. M., Jr. Prediction of Aqueous Solubility of a Diverse Set of Compounds using Quantitative Structure−Property Relationships. *J. Med. Chem.* **2003**, *46*, 3572–80.
(20) Grant, D. J. W.; Higuchi, T. *Solubility Behaviour of Organic Compounds*; John Wiley and Sons: New York, 1990; pp 373–378.

ΔG(liq-water).[21] However, Monte Carlo simulations are computationally expensive and, hence, the method is not applicable to high-throughput drug discovery. Also, the supercooled liquid state of most drugs at room temperature is not accessible and so it is necessary to carry out simulations at elevated temperatures. The most successful method for prediction of solubility from this thermodynamic cycle is the general solubility equation (GSE), which relates log *S* to melting point and the logarithm of the octanol–water partition coefficient (log *P*).[2] It can be derived (if some assumptions are made about the entropy of melting) from the thermodynamic cycle of gas to supercooled liquid to solution. Dannenfelser et al.[22,23] have provided models for the prediction of ΔS_m, and Wassvik et al. [24] have demonstrated that the GSE is more accurate if experimental values of ΔS_m are used. However, the best models for predicting melting point still give 40–50 °C predictive errors,[25,26] and so the GSE is not usually applicable to as yet unsynthesized molecules.

The thermodynamic cycle for transfer from crystal to vapor to solution has been the subject of both experimental and computational studies. Reinwald et al. predicted aqueous solubility of drugs from experimental enthalpies of sublimation and calculated hydration energies.[27] Unfortunately, this method was only accurate for one molecule from a data set of 12, and no computational procedure was suggested for the calculation of $\Delta H_{(sub)}$. Perlovich et al. have published a series of papers that investigate the thermodynamic properties of drugs by experiment and computation, but as yet they have not provided any methods for the prediction of solubility from structure alone.[28,29] The most successful application of a thermodynamic cycle via the vapor has been the prediction of the solubility of liquids (and a small number of low molecular weight solids) from both experimental and calculated vaporisation and hydration energies by Thompson et al.[30] The authors report predictive mean unsigned errors in the range of 0.4–0.6 in log solubility for a data set comprising simple low molecular weight compounds.

In this study, we propose a method for the prediction of solubility that combines QSPR with thermodynamic calculations that explicitly consider the solid state. We test this model on a small data set of drug molecules, for which we measure intrinsic solubilities and characterize the precipitates by powder X-ray diffraction. The computational work was carried out before the characterization of the precipitates, when no information about the crystal form was available.

Methods

The intrinsic solubility of a crystalline molecule is related to the Gibbs free energy difference between crystal and solvated forms of the molecule, which may be obtained from a thermodynamic cycle via the vapor. If the activity coefficient for the solute in solution is assumed to be unity, then the relationship between intrinsic solubility (S_o) and the overall change in Gibbs free energy is

$$\Delta G^*_{(sol)} = \Delta G^*_{(sub)} + \Delta G^*_{(hydr)} = -R\mathrm{T} \ln S_o V_m \qquad (1)$$

where $\Delta G^*_{(sol)}$ is the Gibbs free energy for solution, $\Delta G^*_{(sub)}$ is the Gibbs free energy for sublimation, $\Delta G^*_{(hydr)}$ is the Gibbs free energy for hydration, R is the molar gas constant, T is the temperature (298 K), V_m is the molar volume of the crystal, S_o is the intrinsic solubility in moles per liter, and the superscript * denotes that we are using the Ben−Naim terminology, which refers to the Gibbs free energy for transfer of a molecule between two phases at a fixed center of mass in each phase.[31,32]

If octanol is selected as an intermediate solvent between gaseous and solution states, then the relationship between solubility and the Gibbs free energies may be given as

$$\Delta G^*_{(sol)} = \Delta G^*_{(sub)} + \Delta G^*_{(solv)} + \Delta G^*_{(tr)} = -RT \ln S_o V_m \quad (2)$$

where $\Delta G^*_{(solv)}$ is the Gibbs free energy for solvation in octanol and $\Delta G^*_{(tr)}$ is the Gibbs free energy for transfer of a

(21) Luder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Kjellander, R. In Silico Prediciton of Drug Solubility: 2. Free energy of Solvation in Pure Melts. *J. Phys. Chem. B* **2007**, *111*, 1883–1892.
(22) Dannenfelser, R. M.; Yalkowsky, S. H. Estimation of Entropy of Melting from Molecular Structure - a Non-Group Contribution Method. *Ind. Eng. Chem. Res.* **1996**, *35*, 1483–1486.
(23) Dannenfelser, R. M.; Yalkowsky, S. H. Predicting the Total Entropy of Melting: Application to Pharmaceuticals and Environmentally Relevant Compounds. *J. Pharm. Sci.* **1999**, *88*, 722–724.
(24) Wassvik, C. M.; Holmen, A. G.; Bergstrom, C. A.; Zamora, I.; Artursson, P. Contribution of Solid-State Properties to the Aqueous Solubility of Drugs. *Eur. J. Pharm. Sci.* **2006**, *29*, 294–305.
(25) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422.
(26) 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*, 1177–1185.
(27) Reinwald, G.; Zimmerman, I. A Combined Calorimetric and Semiempirical Quantum Chemical Approach To Describe the Solution Thermodynamics of Drugs. *J. Pharm. Sci.* **1998**, *87*, 745–751.
(28) Perlovich, G. L.; Kurkov, S. V.; Hansen, L. K.; Bauer-Brandl, A. Thermodynamics of Sublimation, Crystal lattice energies and crystal structures of racemates and enantiomers: (+)- and (−)-Ibuprofen. *J. Pharm. Sci.* **2004**, *93*, 654–666.
(29) Perlovich, G. L.; Kurkov, S. V.; Bauer-Brandl, A. Thermodynamics of solutions II. Flurbiprofen and diflunisal as models for studying solvation of drug substances. *Eur. J. Pharm. Sci.* **2003**, *19*, 423–432.
(30) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances. *J. Chem. Phys.* **2003**, *119*, 1661–1670.
(31) Ben-Naim, A. Standard Thermodynamics of Transfer. Uses and Misuses. *J. Phys. Chem.* **1978**, *82*, 792–803.
(32) Ben-Naim, A.; Marcus, Y. Solvation Thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016–2027.

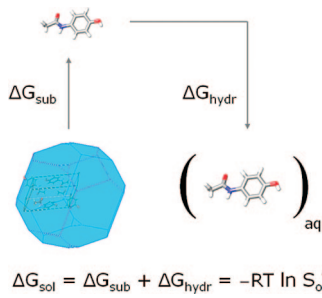

Figure 1. Thermodynamic cycle for transfer from crystal to gas and then to aqueous solution.

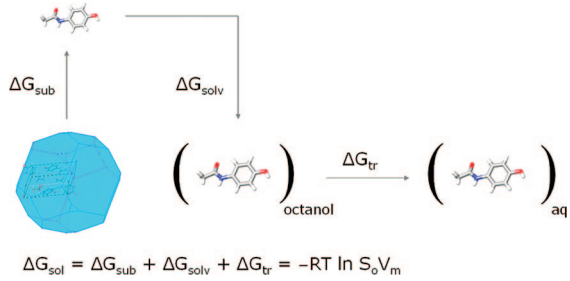

Figure 2. Thermodynamic cycle for transfer from crystal to gas to octanol and then to aqueous solution.

molecule from octanol to water, which can be approximated by

$$\Delta G^*_{(\mathrm{tr})} = 2.303RT \log P \qquad (3)$$

where P is the experimental or calculated octanol–water partition coefficient. Equation 3 is not rigorously correct because the mutual solubility of octanol and water is not zero, but it has been shown to be acceptable.[20] The first and second thermodynamic cycles are illustrated in Figures 1 and 2, respectively.

In principle, each of the Gibbs free energies in eqs 1−3 can be calculated. $\Delta G^*_{(\mathrm{sub})}$ can be calculated by model potential-based lattice energy and lattice dynamics simulations, $\Delta G^*_{(\mathrm{hydr})}$ and $\Delta G^*_{(\mathrm{solv})}$ can be calculated by quantum mechanics with an appropriate model for the solvent, and $\Delta G^*_{(\mathrm{tr})}$ can be estimated from a calculated log P value.

Polymorphism−the existence of crystal structures of chemically identical substances with different regular 3D arrangements of molecules in the crystal lattice−is an additional concern for solubility prediction. Different polymorphs have different apparent solubilities,[33] as exemplified by the problem that Abbott laboratories had with Ritonavir. The average difference in molar solubilities between polymorphs has been estimated to be approximately 2-fold.[34] When this value is compared with the average error in models to predict solubility (6- to 10-fold molar solubility), this suggests that, although knowledge of the correct crystalline form would, of course, be beneficial, an improvement in models to predict solublity may be possible without this information. Similar assumptions have been made in the work of Ouvrard and Mitchell[35] and Chickos et al.[36] on the prediction of lattice energies from structure, and the same idea has been proposed in a review by Johnson et al.[10]

In the next section, we present solubility measurements and characterization for 34 druglike molecules and compare these to values calculated by the methods discussed above. Four thermodynamic parameters are calculated for each molecule $\Delta G^*_{(\mathrm{sub})}$, $\Delta G^*_{(\mathrm{hydr})}$, $\Delta G^*_{(\mathrm{solv})}$ and $\Delta G^*_{(\mathrm{tr})}$. We also demonstrate that an empirical parametrization is necessary in order to provide useful predictions of solubility.

Data Set. A data set of 34 chemically diverse molecules was selected. With the exception of naphthalene, all molecules are drugs or drug precursors. The molecules were selected on the basis that an experimental crystal structure was available in the Cambridge Structural Database and that a reliable value of solubility was available or could be measured. The chemical structures of the molecules in the data set are illustrated in Figure 3.

Solubility Measurements. The thermodynamic solubility of the nonionized form of 20 compounds (intrinsic solubility) at 25 °C was determined by the CheqSol method.[14] Compounds used for this study were purchased from Sigma (Poole, Dorset, UK), and because of their high purity (typically more than 99.5% pure according to the Sigma-Aldrich certificate of analysis), they were used without further purification. The solubility of each molecule was measured 10 times, and statistically treated results are presented in Table 1. The standard deviations for the measured intrinsic solubilities are typically between 0.01 and 0.03 (log S_0, where S_0 is referred to units of mol/L). Each experiment was then repeated to allow the final precipitates to be fingerprinted by powder X-ray diffraction. Crystal structures were identified either by comparison to PXRD patterns simulated from entries in the Cambridge Structural Database (CSD)[37] or by solving the PXRD patterns directly using DASH.[38]

For the remaining 14 molecules in the data set, solubilities were obtained from the data sets of Bergstrom[39] and of

(33) Llinas, A.; Box, K. J.; Burley, J. C.; Glen, R. C.; Goodman, J. M. A New Method for the Reproducible Generation of Polymorphs: Two Forms of Sulindac with very different Solubilities. *J. Appl. Crystallogr.* **2007**, *40*, 379–381.
(34) Pudipeddi, M.; Serajuddin, A. T. M. Trends in Solubility of Polymorphs. *J. Pharm. Sci.* **2005**, *94*, 929–939.
(35) Ouvrard, C.; Mitchell, J. B. O. Can we predict lattice energy from molecular structure. *Acta Crystallogr.* **2003**, *B59*, 676–685.
(36) Chickos, J. S.; Annuziata, R.; Ladon, L. H. Estimating Heats of Sublimation of Hydrocarbons. A Semiempirical Approach. *J. Org. Chem.* **1986**, *51*, 4311–4314.
(37) Cambridge Structural Database; Cambridge Crystallographic Data Centre: **2007**. See: http://www.ccdc.cam.ac.uk/. Accessed 19th April 2007.
(38) David, W. I. F.; Shankland, K.; Shankland, N. Routine Determination of Molecular Crystal Structures from Powder Diffraction Data. *Chem. Commun.* **1998**, 931.
(39) Bergstrom, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477–1488.

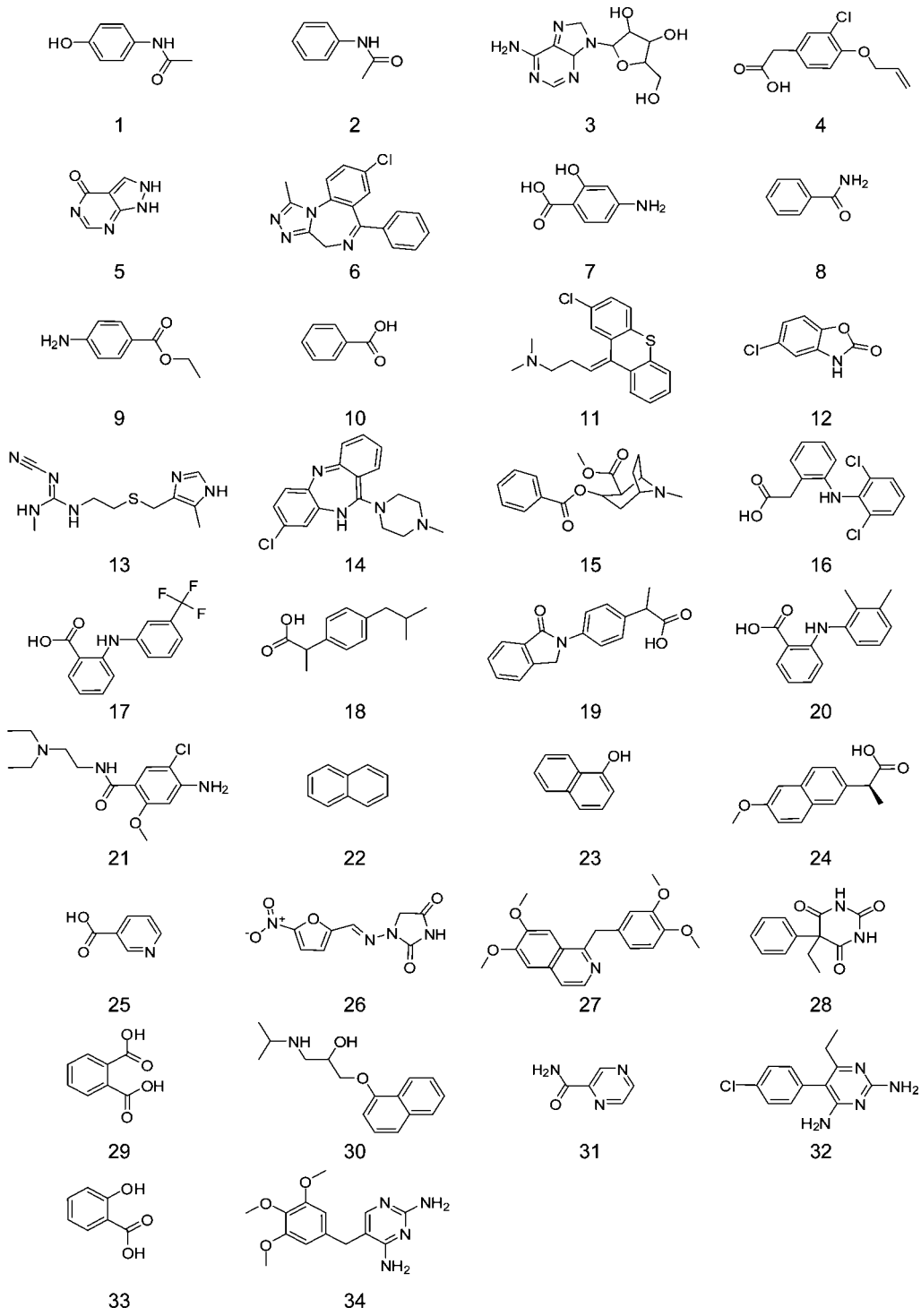

Figure 3. Chemical structures of all 34 molecules in the data set (numbers refer to row number in Table 1).

Rytting.[40] The standard deviation of the solubility values in the complete data set is 1.54 log solubility and the mean is −2.97 log solubility (referred to mol/L). Molecular weights range from 121.14 to 339.39 Da, with a mean of 218.35 Da.

Selection of Crystal Structures. For the lattice energy calculations, a single polymorph for each molecule was selected from the CSD using the following algorithm:

(40) Rytting, E.; Lentz, K. A.; Chen, X. Q.; Qian, F.; Vakatesh, S. Aqueous and Cosolvent Solubility Data for Drug-like Organic Compounds. *AAPS J.* **2005**, *7*, E78–105.

(41) Gavezzotti, A.; Filippini, G. In *Theoretical Aspects and Computer Modeling*; Gavezzotti, A., Ed.; J. Wiley and Sons: Chichester, 1997; pp 61–97.

(42) Gavezzotti, A.; Price, S. Crystal Structure Calculations: 2. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; John Wiley & Sons: Chichester, 1998; Vol. 1, p 641.

Table 1. Crystallographic Data for the 34 Molecules in the Training Data Set

no.	molecule	log *S*(exp) (mol/L)	ref	polymorph observed in experiment (1)	polymorph used for calculations (2)	are (1) and (2) the same?
1	acetaminophen	−1.02		HXACAN01	HXACAN01	yes
2	acetanilide	−1.40	40	*a*	ACANIL01	
3	adenosine	−1.73	40	*a*	ADENOS10	
4	alclofenac	−3.13	39	*a*	FICJAC	
5	allopurinol	−2.26	39	*a*	ALOPUR	
6	alprazolam	−3.60	39	*a*	MENMIB	
7	4-aminosalicylic acid	−1.96	40	*a*	AMSALA01	
8	benzamide	−0.95	40	*a*	BZAMID02	
9	benzocaine	−2.41		QQQAXG01	QQQAXG02	no
10	benzoic acid	−1.58	39	*a*	BENZAC02	
11	chlorprothixene	−6.75		*c*	CMAPTX	
12	chlorzoxazone	−2.59		NEWKOP	NEWKOP	yes
13	cimetidine	−1.69		*b*	CIMETD	No
14	clozapine	−3.24		NDNHCL01	NDNHCL01	yes
15	cocaine	−2.25	39	*a*	COCAIN10	
16	diclofenac	−5.34		SIKLIH01	SIKLIH01	yes
17	flufenamic acid	−5.33		FPAMCA	FPAMCA11	no
18	ibuprofen	−3.62	39	*a*	IBPRAC01	
19	indoprofen	−4.82	39	*a*	LEKMET	
20	mefenamic acid	−6.35		XYANAC	XYANAC	yes
21	metoclopramide	−3.58		METPRA	METPRA	yes
22	naphthalene	−3.61	40	*a*	NAPHTA04	
23	1-naphthol	−1.96		NAPHOL01	NAPHOL01	yes
24	naproxen	−4.15		COYRUD	COYRUD11	yes
25	nicotinic acid	−0.85	40	*a*	NICOAC02	
26	nitrofurantoin	−3.26		*b*	LABJON	no
27	papaverine	−4.34		MVERIQ01	MVERIQ01	yes
28	phenobarbital	−2.28		PHBARB06	PHBARB	no
29	phthalic acid	−1.47		PHTHAC01	PHTHAC01	yes
30	propranolol	−3.49		*c*	IMITON	
31	pyrazinamide	−0.91	40	*a*	PYRZIN	
32	pyrimethamine	−4.11		MUFMAB	MUFMAB	yes
33	salicylic acid	−1.94		SALIAC03	SALIAC03	yes
34	trimethoprim	−2.90		AMXBPM10	AMXBPM10	yes

[a] For 14 molecules, solubility data were taken from the literature and no information about polymorphic form was available. [b] For two molecules (cimetidine and nitrofurantoin), comparison between the PXRD pattern and entries in the CSD revealed that a new polymorph had been formed. [c] For propanolol and chlorprothixene, not enough precipitate could be obtained for characterization by PXRD.

1. Extract all entries for the single molecule that have 3D coordinates (no cocrystals, solvates, salts, etc.).

2. Exclude disordered structures.

3. Exclude entries with reported errors.

4. Select the entry with the lowest *R* factor.

The selected polymorphic form for each molecule is listed in Table 1.

Calculation of $\Delta G^{*}_{(sub)}$. The Gibbs free energy for sublimation was first calculated assuming a 1 atm standard state in the gas (denoted by the superscript °) and was then converted to the related Gibbs free energy given in the Ben−Naim terminology. $\Delta G^{\circ}_{(sub)}$ was calculated from the Gibbs–Helmholtz equation, where $\Delta H^{\circ}_{(sub)}$ was computed from a calculated lattice energy and $\Delta S^{\circ}_{(sub)}$ was considered to be the difference between the entropy of an ideal gas and the entropy of the crystal at 298 K (where the latter was estimated from the calculated phonon modes of the crystal).

The enthalpy of sublimation, $\Delta H^{\circ}_{(sub)}$, can be approximated from the crystal lattice energy, U_{latt}, by[41]

$$\Delta H^{\circ}_{(sub)} = -U_{lattice} - 2RT \quad (4)$$

The $-2RT$ term arises because the lattice energy does not include lattice vibrational energies (which can be approximated by $6RT$ for crystals of rigid molecules oscillating in a harmonic potential) the energy of the vapor is $3RT$ and a $PV = RT$ correction is necessary to change energies into enthalpies, thus yielding $-6RT + 3RT + RT = -2RT$.[42]

Crystal lattice energies were calculated with DMAREL 3.11[43] from the energy-minimized crystal structures. The repulsion-dispersion contributions to the intermolecular potential were evaluated as

(43) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. The Relaxation of Molecular Crystal Structures using a Distributed Multipole Electrostatic Model. *J. Comput. Chem.* **1995**, *16*, 628–647.

$$U_{\mathrm{rep-disp}} = \sum_{M,N}^{N_{\mathrm{mol}}} \left(\sum_{i \in M < k \in N} U_{ik} \right) = \sum_{M,N}^{N_{\mathrm{mol}}} \left(\sum_{i \in M < k \in N} \left(A_{\iota\kappa} e^{-B_{\iota\kappa} R_{ik}} - \frac{C_{\iota\kappa}}{R_{ik}^6} \right) \right) \quad (5)$$

where atoms i and k in molecules M and N are of types ι and κ, respectively, and the parameters $A_{\iota\kappa}$, $B_{\iota\kappa}$, and $C_{\iota\kappa}$ are characteristic of the atom types. The atom−atom potential parameters were taken from Williams and Houpt (C−C, H_C−H_C, N−N, O−O, F−F),[44] Coombes et al. (H_P−H_P),[45] Hsu and Williams (Cl−Cl),[46] and Filippini and Gavezzotti (S−S);[47] here, H_C are hydrogen atoms bonded to carbon and H_P are polar hydrogen atoms (bonded to either oxygen or nitrogen). Potential parameters for interactions between different atoms were constructed as geometric averages for parameters A and C and arithmetic averages for parameter B. Repulsion−dispersion interactions were evaluated up to a 15 Å cutoff.

Electrostatic contributions to the intermolecular potential were calculated from a distributed multipole representation of the electron distribution, which was evaluated by single point calculation, using the B3LYP hybrid functional and 6-31G* basis set in CADPAC,[48] including multipoles up to the hexadecapole. Ewald summation was used for charge–charge, charge−dipole, and dipole–dipole interactions, while all higher order electrostatic terms (up to R^{-5}) were summed to a 15 Å cutoff between molecular centers of mass. The basis set and functional were selected from HF/6-31G*, HF/6-31G**, B3LYP/6-31G*, or B3LYP/6-31G** by comparison to the lattice energies calculated from distributed multipoles evaluated at the MP2/6-31G** level, for a representative subset of the complete data set. The agreement between B3LYP and MP2 results was better than between HF and MP2, and no improvement was observed when the 6-31G** basis set was used as compared to the 6-31G* basis set. Thus, the B3LYP hybrid functional and 6-31G* basis set were selected.

Calculation of $\Delta S^\circ_{(\mathrm{sub})}$. The molar entropy change for sublimation was calculated as $\Delta S^\circ_{(\mathrm{sub})} = S_{(\mathrm{rot,gas})} + S_{(\mathrm{trans,gas})} - S_{(\mathrm{ext,cryst})}$, where $S_{(\mathrm{rot,gas})}$ and $S_{(\mathrm{trans,gas})}$ are the rotational and translational contributions to the entropy of the gas at 298 K, respectively, and $S_{(\mathrm{ext,cryst})}$ is the intermolecular vibrational contribution to the entropy of the crystal at 298 K. The gain in conformational entropy for flexible molecules was ignored but corrected for later by the inclusion of the fraction of rotatable bonds as a parameter. The change in electronic entropy was assumed to be zero. The intra- and intermolecular contributions to the entropy of the crystal were considered to be decoupled, such that the change in intramolecular vibrational entropy for transfer from crystal to gas was taken to be zero.

***S*(gas).** The rotational and translational entropies of the gas ($S_{(\mathrm{rot,gas})}$ and $S_{(\mathrm{trans,gas})}$) were calculated from statistical thermodynamics, assuming an ideal gas at 298 K.[49]

***S*(crystal).** From the Third Law of Thermodynamics, the entropy of all perfect crystalline substances is zero at $T = 0$ K. At 298 K, it is necessary to consider intermolecular and intramolecular vibrations. Translational and rotational entropies are assumed to be negligible, and the crystal lattice is considered to be infinite and perfect. The vibrational terms arise from the intramolecular vibrations and from the phonon modes of the crystal. The latter were calculated using the rigid molecule lattice dynamics implemented in DMAREL 3.11,[50] with the same model potential used for the lattice energy minimizations. Only the $6N - 3$ (where N is the number of molecules in the unit cell) optical zone-center ($k = 0$) phonons were calculated; the remaining three acoustic modes have zero frequency at $k = 0$. The density of states was calculated using a hybrid Debye−Einstein approximation for $k \neq 0$, where the frequencies of the optical phonons were assumed to be independent of k and the acoustic contribution was modeled by the Debye approximation, with the Debye cutoff frequency estimated by extrapolating the acoustic modes to the zone boundary, using sound velocities calculated from the elastic stiffness tensor. The resulting free energy expression is given in ref 72. In these calculations, it is assumed that vibrations are harmonic and coupling between inter and intramolecular vibrations is ignored.

There is only a small amount of accurate data in the literature for $\Delta H^\circ_{(\mathrm{sub})}$, $\Delta S^\circ_{(\mathrm{sub})}$, and $\Delta G^\circ_{(\mathrm{sub})}$ for druglike molecules. However, for those molecules for which data are available there is some agreement between calculated and experimental values (Table 2).[51–53]

Conversion of $\Delta G^\circ_{(\mathrm{sub})}$ to $\Delta G^*_{(\mathrm{sub})}$. In order to evaluate eqs 1 and 2, we convert $\Delta G^\circ_{(\mathrm{sub})}$ to $\Delta G^*_{(\mathrm{sub})}$ using the following

(44) Williams, D. E.; Houpt, D. J. Fluorine Nonbonded Potential Parameters Derived from Crystalline Perfluorocarbons. *Acta Crystallogr.* **1986**, *B42*, 286–295.

(45) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(46) Hsu, L. Y.; Williams, D. E. Intermolecular Potential−Function Models for Crystalline Perchlorohydrocarbons. *Acta Crystallogr.* **1980**, *A36*, 277–281.

(47) Filippini, G.; Gavezzotti, A. Empirical Intermolecular Potentials For Organic Crystals: the '6-exp' Approximation Revisited. *Acta Crystallogr.* **1993**, *B49*, 868–880.

(48) *CADPAC: Cambridge Analytical Derivatives Package* A suite of chemistry programs developed by Amos, R. D. with contributions from Alberts, I. L., Andrews, J. S., Colwell, S. M., Handy, N. C., Jayatilaka, D., Knowles, P. J., Kobayashi, R., Laidig, K. E., Laming, G. 1995; Issue 6. See: http://www-theor.ch.cam.ac.uk/software/cadpac.html. Accessed 19th April 2007.

(49) McQuarrie, D. A. *Statistical Mechanics*; Harper and Row: New York, 1976; pp 30−140.

(50) Day, G. M.; Price, S. L.; Leslie, M. Atomistic Calculations of Phonon Frequencies and Thermodynamic Quantities for Crystals of Rigid Organic Molecules. *J. Phys. Chem. B* **2003**, *107*, 10919–10933.

(51) Perlovich, G.; Kurkov, S. V.; Kinchin, A. N.; Bauer-Brandl, A. Thermodynamics of Solutions III: Comparison of the Solvation of (+)-Naproxen with other NSAIDS. *Eur. J. Pharm. Biopharm.* **2004**, *57*, 411–420.

(52) Sabbah, R.; Le, T. H. D. Etude Thermodynamique des Trois Isomeres de l'acide hydroxybenzoïque. *Can. J. Chem.* **1993**, *71*, 1378–1383.

Table 2. Thermodynamic Properties of Sublimation from Experiment and As Calculated in This Study[a]

	experimental			calculated		
molecule	$\Delta H^{\circ}_{(sub)}$ (kJ mol^{-1})	$\Delta S^{\circ}_{(sub)}$ (J mol^{-1} K^{-1})	$\Delta G^{\circ}_{(sub)}$ (kJ mol^{-1})	$\Delta H^{\circ}_{(sub)}$ (kJ mol^{-1})	$\Delta S^{\circ}_{(sub)}$ (J mol^{-1} K^{-1})	$\Delta G^{\circ}_{(sub)}$ (kJ mol^{-1})
1-naphthol	91.20	187.50	35.30	93.13	192.80	35.60
salicylic acid	96.27			87.36	192.65	29.92
acetaminophen	117.90	190.00	60.00	115.71	200.10	56.06
naproxen	128.30	234.23	58.50	130.20	208.75	67.97

[a] All sublimation values in this table are reported relative to a 1 atm standard state in the gas phase.

equation, which is derived considering the work for isothermal reversible expansion of an ideal gas

$$\Delta G^{*}_{(sub)} = \Delta G^{\circ}_{(sub)} - RT \ln\left(\frac{V_m p_o}{RT}\right) \quad (6)$$

where V_m is the molar volume of the crystal and p_o is equal to 1.0135×10^5 Pa. Calculated values of $\Delta G^{*}_{(sub)}$ are presented in Table 3.

Calculation of $\Delta G^{*}_{(hydr)}$ and $\Delta G^{*}_{(solv)}$. $\Delta G^{*}_{(hydr)}$ and $\Delta G^{*}_{(solv)}$ were calculated using two different implicit models for solvent and quantum mechanics: the self-consistent reaction field (SCRF) model[54,55] using B3LYP/6-31G* as implemented in Jaguar[56] and the SM5.4 model using the AM1 and PM3 semiempirical Hamiltonians, as implemented in Spartan.[57] The SCRF method is able to model both aqueous and nonaqueous solvents, which permits the calculation of both gas to water and gas to octanol solvation energies. The semiempirical method in Spartan was used in order to provide a less computationally expensive alternative for the gas to water hydration energies.

For the SCRF calculations, the following procedure was defined for both gas–water and gas−octanol calculations. For each molecule, a low-mode conformational search was carried out in gas and then solvent (both water and octanol) using the MMFFs forcefield with a Generalized-Born surface area model for solvent, in Macromodel v.9.1.[58] The global minimum energy conformers in the gas and solution phase were then geometry optimized at the B3LYP/6-31G* level in Jaguar. Solvation energies were calculated as the difference in energy between the global minimum energy conformer in the gas phase and the global minimum energy conformer in the solution phase. For the aqueous solution-phase simulations, the dielectric constant for water was set to 80.37, the molecular weight to 18.02 g/mol, and the density to 0.99823 g/mL (from which the probe radius was calculated to be 1.40 Å). For the octanol solution phase, the dielectric constant for octanol was set to 10.30, the molecular weight to 130.23 g/mol, and the density to 0.82620 g/mL (giving a probe radius of 3.15 Å).[59]

For the semiempirical calculations, the lowest energy conformer of each molecule previously calculated by molecular mechanics was reoptimized at AM1 and PM3 semiempirical levels of theory. The aqueous solvation energies were calculated from these molecular geometries using the SM5.4 model of Cramer and Truhlar[60] as implemented in Spartan.

Statistical Analysis. Statistical analyses were carried out in the R Statistical Computing Environment.[61] lm(), anova(), and step() commands were used for multilinear and stepwise regression and additional analyses were implemented using purpose-written R scripts.

Molecular Descriptors. In addition to the thermodynamic parameters discussed above, we also calculated the fraction of rotatable bonds (b_rotR) for each molecule using the Molecular Operating Environment software. Results are provided in Table 3.[62–64]

(53) Perlovich, G. L.; Volkova, T. V.; Bauer-Brandl, A. Towards an Understanding of the Molecular Mechanism of Solvation of Drug Molecules: A Thermodynamic Approach by Crystal Lattice Energy, Sublimation, and Solubility exemplified by Paracetamol, Acetanilide, and Phenacetin. *J. Pharm. Sci.* **2006**, *95*, 2158–2169.

(54) Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M.; Goddard, W. A.; Honig, B. Accurate First Principles Calculation of Molecular Charge Distributions and Solvation Energies from Ab Initio Quantum Mechanics and Continuum Dielectric Theory. *J. Am. Chem. Soc.* **1994**, *116*, 11875–11882.

(55) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. New Model for Calculation of Solvation Free Energies: Correction of Self-Consistent Reaction Field Continuum Dielectric Theory for Short-Range Hydrogen-Bonding Effects. *J. Phys. Chem.* **1996**, *100*, 11775–11788.

(56) Jaguar 7.1; Schrodinger: Portland, OR, 2006. See: http://www.schrodinger.com/. Accessed 19 April 2007.

(57) Spartan' 02; Wavefunction, Inc.: Irvine, CA, 2002. See: http://www.wavefun.com/. Accessed 19 April 2007.

(58) Macromodel v.9.1; Schrodinger: Portland, OR, 2006. See: http://www.schrodinger.com/. Accessed 19 April 2007.

(59) Merck Index; Merck & Co., Inc.: Whitehouse Station, NJ, 2006.

(60) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Model for Aqueous Solvation Based on Class IV Atomic Charges and First Solvation Shell Effects. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(61) R Development Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2005. See: http://www.r-project.org/. Accessed 19th April 2007.

(62) Hester, J. B., Jr. Von Voigtlander, P., 6-Aryl-4H-s-triazolo[4,3-a][1,4]benzodiazepines. Influence of 1-Substitution on Pharmacological Activity. *J. Med. Chem.* **1979**, *22*, 1390–1398.

(63) Abiraj, K.; Srinivasa, G. R.; Gowda, D. C. Simple and Efficient Reduction of Aromatic Nitro Compounds Using Recyclable Polymer-Supported Formate and Magnesium. *Aust. J. Chem.* **2005**, *58*, 149–151.

Table 3. Experimental Values and Calculated Thermodynamic Parameters for Molecules in the Training Data Set

								B3LYP/6-31G* SCRF		$\Delta G^*_{(tr)}$					
row	molecule	log S_{exp} (mol/L)	ref (log S_{exp})	mp (°C)	ref (mp)	V_m (dm^3/mol)	$\Delta G^*_{(sub)}$ (kJ/mol)	$\Delta G^*_{(hydr)}$ (kJ/mol)	$\Delta G^*_{(solv)}$ (kJ/mol)	from log P(exp) (kJ/mol)	from ClogP (kJ/mol)	log P (exp)	ClogP	b_rotR	U_{latt} (kJ/mol)
1	acetaminophen	−1.02		170	68	0.117	52.18	−68.27	−64.98	2.63	2.80	0.46	0.49	0.182	−120.67
2	acetanilide	−1.40	40	114.3	68	0.107	31.24	−41.69	−44.63	6.62	6.62	1.16	1.16	0.200	−96.76
3	adenosine	−1.73	40	235.5	68	0.175	120.3	−89.28	−77.21	−5.99	−17.58	−1.05	−3.08	0.095	−196.72
4	alclofenac	−3.13	39	92.5	68	0.168	66.43	−44.85	−59.30	14.16	15.58	2.48	2.73	0.333	−137.71
5	allopurinol	−2.26	39	350	68	0.084	60.74	−73.23	−73.74	−3.14	−4.74	−0.55	−0.83	0.000	−126.45
6	alprazolam	−3.60	39	230	61	0.221	20.41	−58.85	−74.74	12.10	14.56	2.12	2.55	0.040	−91.06
7	4-aminosalicylic acid	−1.96	40	153	62	0.100	47.38	−50.14	−61.50	5.08	6.05	0.89	1.06	0.091	−113.20
8	benzamide	−0.95	40	129.1	68	0.089	33.59	−43.71	−50.47	3.65	3.71	0.64	0.65	0.111	−97.65
9	benzocaine	−2.41		92	68	0.131	37.66	−38.20	−55.20	10.62	10.96	1.86	1.92	0.250	−105.19
10	benzoic acid	−1.58	39	122.4	68	0.092	29.42	−29.52	−40.22	10.67	10.73	1.87	1.88	0.111	−93.80
11	chlorprothixene	−6.75		97.5	68	0.245	70.18	−21.92	−51.07	29.57	31.28	5.18	5.48	0.130	−143.37
12	chlorzoxazone	−2.59		191.5	68	0.103	45.45	−34.69	−49.94	9.47	10.67	1.66	1.87	0.000	−110.68
13	cimetidine	−1.69		142	68	0.192	98.06	−70.71	−83.94	2.28	2.17	0.40	0.38	0.471	−171.01
14	clozapine	−3.24		184	68	0.248	81.23	−53.21	−63.69	18.44	14.38	3.23	2.52	0.039	−155.70
15	cocaine	−2.25	39	98	68	0.243	68	−37.67	−55.13	13.13	14.67	2.30	2.57	0.208	−142.19
16	diclofenac	−5.34		157	40	0.202	70.79	−35.65	−52.17	25.74	27.00	4.51	4.73	0.200	−143.27
17	flufenamic acid	−5.33		133.5	68	0.190	62.12	−38.65	−58.94	29.97	31.57	5.25	5.53	0.191	−134.87
18	ibuprofen	−3.62	39	76	68	0.175	48.11	−23.05	−48.26	22.66	21.01	3.97	3.68	0.267	−119.72
19	indoprofen	−4.82	39	214	68	0.212	96.69	−57.20	−72.35	15.81	15.64	2.77	2.74	0.130	−170.72
20	mefenamic acid	−6.35		231	68	0.190	66.24	−36.43	−56.49	29.23	30.20	5.12	5.29	0.158	−139.47
21	metoclopramide	−3.58		147.25	68	0.232	90.9	−48.86	−66.64	14.96	12.73	2.62	2.23	0.400	−164.34
22	naphthalene	−3.61	40	80.2	68	0.104	18.39	−7.50	−23.94	18.84	18.95	3.30	3.32	0.000	−79.22
23	1-naphthol	−1.96		95	68	0.112	31.86	−31.62	−40.30	16.27	15.13	2.85	2.65	0.000	−98.09
24	naproxen	−4.15		153	68	0.184	62.97	−25.78	−34.47	18.15	16.10	3.18	2.82	0.167	−135.17
25	nicotinic acid	−0.85	40	236.6	68	0.084	32.95	−39.52	−46.41	2.05	4.57	0.36	0.80	0.111	−98.13
26	nitrofurantoin	−3.26		263	68	0.144	86.98	−36.52	−37.15	−2.68	−2.68	−0.47	−0.47	0.167	−156.97
27	papaverine	−4.34		147.5	68	0.263	94.95	−51.60	−69.07	16.84	21.58	2.95	3.78	0.222	−169.86
28	phenobarbital	−2.28		174	68	0.171	56.81	−65.59	−63.07	8.39	7.82	1.47	1.37	0.111	−127.70
29	phthalic acid	−1.47		230	68	0.104	67.09	−60.14	−66.46	4.17	4.17	0.73	0.73	0.167	−139.33
30	propanolol	−3.49		129	63	0.223	81.12	−32.64	−50.07	20.46	15.70	3.59	2.75	0.300	−155.68
31	pyrazinamide	−0.91	40	192	68	0.085	32.42	−43.60	−49.84	−3.42	−3.88	−0.60	−0.68	0.111	−93.98
32	pyrimethamine	−4.11		233.5	68	0.189	68.89	−54.87	−65.32	15.35	17.12	2.69	3.00	0.111	−142.13
33	salicylic acid	−1.94		158	68	0.096	26.54	−25.94	−44.31	12.90	12.50	2.26	2.19	0.100	−92.32
34	trimethoprim	−2.90		203	68	0.225	99.42	−57.70	−70.92	5.19	5.59	0.91	0.98	0.227	−176.10

Results

We have defined two different thermodynamic cycles: (a) transfer from crystal to gas to water and (b) transfer from crystal to gas to octanol to water. In the Introduction, the suggestion was made that it might be possible to predict the solubility directly from the calculated Gibbs free energy change, thus eliminating the problems encountered when using a training set. However, when this *ab initio* method was evaluated for both thermodynamic cycles (without an empirical correction), there was little or no correlation between predicted and experimental solubility values. For the thermodynamic cycle of crystal to gas to water, the solubilities of three molecules were predicted with an absolute error of less than 1 log solubility unit (alclofenac, ibuprofen, diclofenac). For the thermodynamic cycle of crystal to gas to octanol to water, the solubilities of five molecules were predicted with absolute error less than 1 log solubility unit (alclofenac, ibuprofen, mefenamic acid, flufenamic acid, cimetidine). There was one large outlier in the predictions from both thermodynamic cycles (alprazolam). The source of this error could not be identified because separate experimental sublimation and solvation-free energy data were not available for this molecule. Analysis of the error in prediction for all 34 molecules revealed a correlation between the error and the molecular size and flexibility as encapsulated by properties such as the molecular weight and number of rotatable bonds. However, even when these relationships were used to correct the predicted solubility

(64) Bergstrom, C. A.; 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*, 1177–85.

Table 4. Regression Coefficients, *F* Values, and *t* Test Statistics for Eq 7[a]

descriptor	coefficient	standard error	*F* value	Pr(>*F*)	*t* test	Pr(>|*t*|)	standard deviation	mean
$\Delta G^*_{(sub)}$	−0.52	0.08	18.77	0.00	−6.17	0.00	26.13	60.52
$\Delta G^*_{(tr)}$	−0.87	0.08	124.71	0.00	−11.40	0.00	10.67	11.55
b_rotR	0.24	0.08	7.83	0.01	2.82	0.01	0.11	0.16

[a] The data were autoscaled using the standard deviations and means given in the final columns.

values, the predictive error improved only to the standard deviation of the solubility values in the data set. The failure to predict intrinsic solubility directly from the calculated Gibbs free energy is less surprising if we consider that using this model, a 5.7 kJ mol^{-1} error in calculated $\Delta G^*_{(sol)}$ will change the predicted solubility by 1 log unit. Standard errors for the calculation of the separate free energy terms are not easy to estimate due to the sparsity of experimental sublimation and solvation data for drug molecules, but it is reasonable to expect that they are smallest for rigid low molecular weight molecules and largest for conformationally flexible high molecular weight molecules. Even though errors exist in these calculations, the values still contain useful information about the properties of the molecules, which can be used to create successful models for solubility prediction. We investigated this by regression analysis, using both leave-one-out (LOO) cross-validation and prediction of molecules in an external test set in order to verify the predictivity of the models. The regression analysis also permits an analysis of the probable importance of each term in the thermodynamic cycle to the prediction of solubility.

Regression Analyses. Regression analyses were carried out between log intrinsic solubility (referred to mol/L) and the calculated thermodynamic parameters (including the term for the fraction of rotatable bonds in the molecule, b_rotR). In order to select the best model, variable selection was carried out by stepwise linear regression using combined forward and backward sampling. The regression equation for the best selected model is given (for unscaled data) in equation 7:

$$\log S = -0.0308(\pm 0.0050)\Delta G^*_{(sub)} - 0.126(\pm 0.011)\Delta G^*_{(tr)} + 3.31(\pm 1.18)\text{b\_rotR} \quad (7)$$

The three regression variables have satisfactory *t* test results (at the 0.01 level), but the intercept does not pass the *t* test and has been removed. The results for fit to the training data were $r^2 = 0.84$, RMSE = 0.62, and bias = 0.00. Table 4 presents regression coefficients, standard errors, and *F* and *t* test results for scaled data. The importance of each regression variable can be inferred: $\Delta G^*_{(tr)}$ is related to the octanol–water partition coefficient and is a measure of solute–solvent interactions, $\Delta G^*_{(sub)}$ captures the importance of solute–solute interactions, and b_rotR acts as a correction factor for the entropy of solution. We note that the signs of the regression coefficients are as expected; predicted solubility will decrease with increasing $\Delta G^*_{(sub)}$ and $\Delta G^*_{(tr)}$ and will increase with increasing molecular flexibility (as represented by b_rotR). However, the regression coefficient for $\Delta G^*_{(sub)}$ in eq 7 predicts that a 32.5 kJ mol^{-1} increase in $\Delta G^*_{(sub)}$ is necessary to cause a 1 log solubility unit decrease in solubility, which on first inspection seems too large. The b_rotR term will oppose this decrease in solubility for flexible molecules and as such is probably acting as a correction to the calculated $\Delta G^*_{(sub)}$ term.

It is noticeable that neither of the calculated solvation energies $\Delta G^*_{(solv)}$ and $\Delta G^*_{(hydr)}$ are included in the final model. When these variables were forcibly included in the regression models, both failed the *F* and *t* tests. The same was true if $\Delta G^*_{(hydr)}$ values calculated at the B3LYP/6-31G* level were replaced by the results for AM1 or PM3 using the SM5.4 model for water. The parametric *t* test assumes that the variables approximate a Gaussian distribution, but its use rather than nonparametric tests seemed satisfactory based upon an analysis of the regression variables. The *t* test for a variable is defined as the regression coefficient divided by the standard error. The latter depends on sample size, and in general, the *t* test is more demanding for smaller data sets. The failure of $\Delta G^*_{(hydr)}$ and $\Delta G^*_{(solv)}$ to pass the *t* test may be because the relevant information is already contained within other regression variables. It may also be due to known shortcomings in the implicit solvent model (we postpone discussion of this until the next section).

Equation 7 was then simplified by (i) replacing the $\Delta G^*_{(tr)}$ variable with ClogP and (ii) replacing $\Delta G^*_{(sub)}$ with ΔU_{latt}. The former was possible because the effect of the 2.303*RT* factor is absorbed by the regression coefficient (and therefore this replacement could be made without changing the r^2, RMSE, and bias for fit to the training data). The latter was possible because the correlation between $\Delta G^*_{(sub)}$ and ΔU_{latt} is high (Pearson $R = -0.99$). The new regression equation is given for unscaled data in eq 8.

$$\log S = 0.0266(\pm 0.0044)U_{latt} - 0.696(\pm 0.063)\text{ClogP} + 3.31(\pm 1.20)\text{b\_rotR} + 1.39(\pm 0.55) \quad (8)$$

The results for fit to the training data were $r^2 = 0.83$, RMSE = 0.63, and bias = 0.00 (which are given in Table 6). Regression coefficients, standard errors, and F and *t* test results for scaled data are presented in Table 5. U_{latt} and ClogP pass both *F* and *t* tests at the 0.001 level, while b_rotR passes both tests at the 0.01 level.

Equation 8 was validated both internally (using cross-validation on the training data) and by prediction of an external test set. The results for LOO cross-validation for eq 8 were $q^2 = 0.78$, RMSE = 0.71, and bias = −0.02 (Table 6). The importance of U_{latt} can be seen by comparison to the LOO cross-validation results for regression against ClogP only, which are $q^2 = 0.53$, RMSE = 1.04, and bias = −0.03. The inclusion of U_{latt} and b_rotR reduces the error in predicted molar solubility from approximately 10- to 6-fold. The statistical significance of eq 8 was further verified using a series of *y*-scrambling tests. The values of log *S* were

Table 5. Regression Coefficients, *F* Values, and *t* Test Statistics for Eq 8[a]

descriptor	coefficient	standard error	*F* value	Pr(>*F*)	*t* test	Pr(>|*t*|)	standard deviation	mean
U_{latt}	0.52	0.09	22.44	0.00	6.02	0.00	29.87	−130.57
ClogP	−0.84	0.08	116.46	0.00	−11.08	0.00	1.87	2.02
b_rotR	0.24	0.09	7.58	0.01	2.75	0.01	0.11	0.16

[a] The data were autoscaled using the standard deviations and means given in the final columns.

Table 6. Regression Statistics for Fit to the Training Data and LOO Cross-Validation for Eqs 7 and 8

	variables	r^2(tr)	RMSE(tr)	bias(tr)	q^2	RMSE(LOO)	bias(LOO)	r^2(te)	RMSE(te)	bias(te)
eq *7*	$\Delta G^*_{(sub)}$, $\Delta G^*_{(tr)}$, b_rotR	0.84	0.62	0.00	0.79	0.70	−0.02			
eq *8*	U_{latt}, ClogP, b_rotR	0.83	0.63	0.00	0.78	0.70	−0.02	0.77	0.71	−0.06

Table 7. Experimental and Calculated Values for 26 Molecules in the External Test Set

molecule	REFCODE	mp (°C)	ref (mp)	log *P*	ref (log *P*)	U_{latt} (kJ/mol)	ClogP	b_rotR	log *S*(exp)	ref (log *S*)	log *S*(pred)	error
4-aminobenzoic acid	ambnac04	188.5	68	0.83	68	−115.05	0.98	0.100	−1.37	40	−2.03	0.66
corticosterone	cortic	181	68	1.94	68	−170.36	2.51	0.071	−3.24	39	−4.66	1.42
danthron	dhanqu01	193	68			−123.08	3.74	0.000	−5.19	40	−4.49	−0.70
dapsone	dapsuo03	175.5	68	0.97	68	−171.69	0.89	0.111	−3.09	40	−3.43	0.34
diazepam	dizpam10	132	68	2.82	68	−133.67	2.96	0.046	−3.75	39	−4.08	0.33
diphenylhydantoin	phydan01	286	68	2.47	68	−138.29	2.08	0.095	−3.86		−3.42	−0.44
estrone	estron10	260.2	68	3.13	68	−147.03	3.38	0.000	−3.95	40	−4.88	0.93
fluconazole	ivuqof	140	65	2.17	66	−160.98	−0.44	0.208	−1.80	39	−1.90	0.10
hydrochlorothiazide	hcsbtz	274	68	−0.07	68	−193.12	−0.36	0.056	−2.68		−3.32	0.64
hydroflumethiazide	ewuhaf01	270.5	68	0.36	68	−201.73	−0.21	0.095	−2.98		−3.52	0.54
isoproturon	jodtur01	158	68	2.87	68	−133.21	2.40	0.267	−3.47	40	−2.94	−0.53
oxytetracycline	oxytet	184.5	68	−0.90	68	−203.26	−1.27	0.056	−3.10		−2.95	−0.15
piroxicam	biyseh	200	68	3.06	68	−153.60	1.89	0.120	−4.80		−3.62	−1.18
primidone	ephpmo	281.5	68	0.91	68	−136.64	0.88	0.118	−2.64	40	−2.47	−0.17
pteridine	pterid11	139.5	68	−0.58	68	−85.70	−0.79	0.000	0.02		−0.34	0.36
pyrene	pyrene02	151.2	68	4.88	68	−105.42	4.95	0.000	−6.18	40	−4.86	−1.32
sparfloxacin	jekmob	260	67	1.70	66	−189.82	−0.60	0.097	−3.37		−2.92	−0.45
sulfadiazine	suldaz01	255.5	68	−0.09	68	−166.57	0.10	0.167	−3.53	40	−2.56	−0.97
sulfamerazine	slfnma01	236	68	0.14	68	−168.52	0.60	0.158	−3.12		−2.99	−0.13
sulfamethazine	slfnmd01	198.5	68	0.89	68	−180.87	1.10	0.150	−2.73		−3.69	0.96
sulfamethoxazole	slfnmb01	167	68	0.89	68	−155.60	0.56	0.167	−2.70	40	−2.59	−0.11
sulfanilamide	sulamd01	165.5	68	−0.62	68	−136.91	−0.57	0.091	−1.36	40	−1.56	0.20
tolbutamide	zzzpus02	128.5	68	2.34	68	−168.22	2.50	0.389	−3.47		−3.54	0.07
1,3,5-trichlorobenzene	tchlbz	63.5	68	4.19	68	−78.00	4.28	0.000	−4.44	40	−3.67	−0.77
triphenylene	triphe11	199	68	5.49	68	−123.89	5.66	0.000	−6.73	40	−5.85	−0.88
uracil	uracil	338	68	−1.07	68	−105.82	−1.06	0.000	−1.49	39	−0.69	−0.80

sorted randomly such that each molecule was associated with an incorrect value of solubility, and leave-one-out cross-validation was repeated. The average results for 10 repetitions were q^2(*y*-scrambling) = −0.23, RMSE(*y*-scrambling) = 1.68, and bias(*y*-scrambling) = 0.00 for leave-one-out cross-validation. As expected, the root-mean-square error for leave-one-out cross-validation of the *y*-scrambled data is approximately equal to the standard deviation of experimental log *S*, which suggests that eq 8 was not arrived at by chance correlation.

In order to validate the selected model, an external data set of 26 molecules was compiled with intrinsic aqueous solubility data measured by the CheqSol method or taken from the literature. For each molecule, a single crystal polymorph was selected from the Cambridge Structural Database using the algorithm defined earlier. Equation 8 was used to predict the solubility of the molecules in the external test set from values of U_{latt}, ClogP, and b_rotR, which were calculated as described in the Methods section. Experimental and predicted values of solubility are given in Table 7, and the correlation diagram is provided in Figure 4.[65–67] The results for prediction of this external test set were r^2(te) =

(65) Milne, G. W. A.; *Drugs: Synonyms & Properties*; Ashgate Publishing Co.: Brookfield, VT, 2000; p1280.

(66) Drugbank. See: http://redpoll.pharmacy.ualberta.ca/drugbank. Accessed 21 July, 2007.

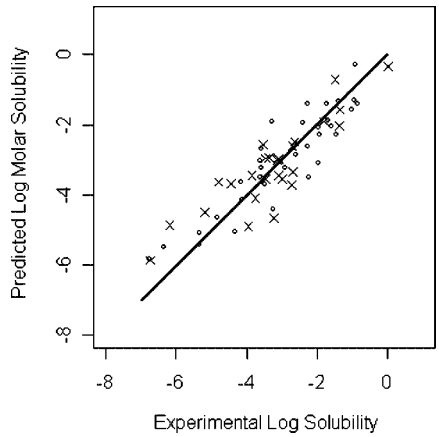

Figure 4. Predicted versus experimental log solubility (referred to mol/L) for eq 8 for both training and external test sets.

0.77, RMSE(te) = 0.71, bias = −0.06, which is satisfactory when compared to the results for leave-one-out cross-validation. Three molecules were predicted with an unsigned error greater than 1 log unit; pyrene (predictive error, –1.3), piroxicam (−1.17), and corticosterone (1.43). The most similar molecule to pyrene in the training data is naphthalene, which during leave-one out cross-validation was predicted with an error of −0.56 in log solubility (referred to mol/L). It is plausible that the lattice energies of these nonfunctional hydrocarbons may be overpredicted by the model potential, but no clear conclusion is possible from our data. Piroxicam and corticosterone are both structurally dissimilar to the molecules in the training data, which may explain the predictive error for these molecules.

Comparison with the General Solubility Equation. To provide a comparison to our results, the GSE was also used to predict solubility for the molecules in the training and external data sets. Experimental melting point and log *P* data for the molecules in both the training and external test sets were extracted from the literature; experimental values and references are provided in Tables 3 and 7.[68] For the training data set, the results for prediction of solubility using experimental melting point and ClogP were $r^2 = 0.68$, RMSE = 0.86, bias = −0.03. For the external test set, the results were $r^2 = 0.37$, RMSE = 1.14, and bias = −0.62. The root-mean-square error for prediction of molecules in the external test set is somewhat higher than normally expected for the General Solubiliy Equation. The reason for the poor prediction is partly due to one outlier, oxytetracycline, for which the solubility was predicted with an error of −3.28 log units. If this outlier is temporarily excluded, the results improve to $r^2 = 0.57$, RMSE = 0.96, and bias = −0.51. However, no explanation could be found for the error in prediction for this molecule, and therefore, there is no justifiable reason to remove this outlier. The comparison between our results and those for the GSE suggest that equation 8 is useful for the prediction of solubility (although we note that the comparison is only approximate, due to the small size of the test set). Results are provided in Table 8.

Discussion

We found that direct *ab initio* calculation of solubilities by computing the component terms of the thermodynamic cycles was of insufficient accuracy to be useful. Thus, it seems that direct calculation of solubilities for druglike molecules, which is required to be affordable and scalable to reasonably high throughput, remains out of reach. Therefore, we have turned to a mixture of direct computation and informatics, using the calculated thermodynamic properties, along with a few other key descriptors, in regression models.

We have derived a regression equation for the prediction of solubility, which contains three variables, whose importance to solubility behavior might have been anticipated. First, the calculated lattice energy (U_{latt}) is a measure of the strength of solute–solute interactions, which will disfavor solvation. The crystal lattice energy is calculated from an experimental crystal structure using a model potential implemented in DMAREL3.11. The reliance on an experimental crystal structure does present an obvious limitation to the method. However, in calculating U_{latt} we have not required the correct polymorphic form of the crystal, only an available experimental structure, which we have selected from the Cambridge Structural Database. This supports the suggestion made in the Methods section that it is not necessary to know the correct polymorphic form in order to include a realistic measure of crystal packing into models to predict solubility (which is in agreement with Johnson et al.[10]). Once a crystal structure is available, the calculation of U_{latt} is only moderately computationally expensive; the calculations took approximately 1−4 CPU hours for the majority of molecules in the data set. In principle, these calculations are automatable, but we found that some manual intervention was often necessary. The final regression variable in eq 8 is the octanol–water partition coefficient calculated by the ClogP program. Log *P* is a measure of molecular hydrophobicity and is the most commonly used variable in QSPR models to predict solubility; log *P* shows a reasonable negative correlation with solubility. The prediction of solubility by a regression against log *P* only for the training and test sets gave r^2(loo) = 0.53, RMSE(loo) = 1.04, bias(loo) = 0.03 and r^2(te) = 0.50, RMSE(te) = 1.07, bias(te) = −0.57. If these results are compared to the results for eq 8, it indicates that the inclusion of terms related to the solute–solute interactions improves the root-mean-square error for prediction by approximately 0.35 log solubility units. This result may be put into context by considering three other studies that have attempted to explicitly include information about the solid state in models to predict solubility. Wassvik et al.[24] found that the inclusion of the experimental enthalpy and entropy of fusion improved

(67) Junichi, M.; Teruyuki, M.; Hiroshi, E.; ShinichiN. Preparation of piperazinylquinolonecarboxylates as bactericides. *Eur. Pat. Appl.* **1987**, 50.

(68) EPI suite; EPA, 2005. See: http://www.syrres.com/esc/epi.htm. Accessed 19 April 2007.

Table 8. Prediction of log *S* by the General Solubility Equation, Using Experimental Melting Points and Experimental Values of Either log *P* or ClogP

	training data set ($n = 34$)			test data set ($n = 26$)		
	r^2	RMSE	bias	r^2	RMSE	bias
GSE (using experimental log *P*)	0.79	0.70	0.01	0.55[a]	0.95[a]	−0.38[a]
GSE (using ClogP)	0.68	0.86	−0.03	0.37	1.14	−0.62

[a] Evaluated for 25 molecules because experimental log *P* data could not be found for danthron.

predictions of solubility, over prediction by log *P* only, by 0.3 log units, which is directly comparable to our results. A computational approach was proposed by Johnson et al.,[18] who used molecular dynamics simulations of the crystal lattice to quantify the crystallinity of the solute. The calculated crystal packing term was used as a *post hoc* correction to a traditional QSPR model. When validated on the data set of Wassvik et al.,[24] the inclusion of the solute–solute term permitted a small improvement in r^2 values from 0.70 to 0.75 and no change in RMSE values. In the General Solubility Equation, the experimental melting point is used as a surrogate for the Gibbs energy of fusion. The equation indicates that a 100 °C change in melting point is necessary to change the predicted solubility by 1 log unit. Consideration of the results of all four of these methods, in particular the agreement between our results and those of Wassvik et al.,[24] indicates that, while solvation interactions have the dominant effect on solubility, solute–solute interactions make a small but non-negligible contribution to solubility behavior.

In the experimental part of this work, the crystal structure of the solute was characterized by repeating the solubility experiment, collecting the precipitate, and characterizing this by powder X-ray diffraction. This approach was used because previous work has shown that the polymorphic form may change during the course of a solubility experiment, in accordance with Ostwald's law.[11] However, despite due care, it was not possible to characterize the precipitate for three of the 20 molecules for which we measured solubility. Therefore, we note that regardless of the polymorphic form chosen for computational methods, it is often not trivial to identify the experimental polymorphic form in equilibrium with the solute in solution.

We are mindful that QSPRs reveal correlations, not causality. Nonetheless, the regression variables in eq 8, U_{latt}, Clog*P*, and b_rotR have well defined chemical meanings, unlike some alternative molecular descriptors. Thus they may be used, with the aforementioned caution of possible partial compensation between U_{latt} and b_rotR, to provide qualitative information about the relative importance of solute–solute interactions and solvation interactions of different molecules. For instance, the experimental solubility of oxytetracycline is –3.10 log units compared to a predicted value of −2.95 log units. The regression variables indicate that oxytetracycline has a low calculated lattice energy (−203.26 kJ/mol), a low calculated log *P* value (−1.27) and only moderate contributions from b_rotR (0.056). Therefore, the regression variables indicate that the solubility behavior of oxytetracycline is a balance between strong solute–solute interactions (which disfavors solvation in water) and relatively high hydrophilicity (which favors solvation in water).

In deriving eq 8, the calculated Gibbs free energies for solvation [$\Delta G^*_{(\mathrm{hydr})}$ or $\Delta G^*_{(\mathrm{solv})}$] were omitted. The reason why these variables were not found to be beneficial to the models is not clear, but may indicate that the information they contain is already accounted for by the ClogP variable or that the calculations are inaccurate. In the SCRF method, $\Delta G^*_{(\mathrm{solv})}$ is calculated as $\Delta G_{(\mathrm{elec})} + \Delta G_{(\mathrm{vdw})} + \Delta G_{(\mathrm{cav})}$.[69] The first term is calculated from quantum mechanics and an implicit continuum model for the solvent; the latter two terms are estimated by an empirical model from the molecular surface area. Problems with the implicit solvent models used to calculate $\Delta G_{(\mathrm{elec})}$ have been discussed elsewhere.[70] The empirical calculation will suffer from inaccuracies due to differences between the molecules in the training set (which are generally low molecular weight and linear or monocyclic) and the drug molecules used in our data set (which are higher molecular weight). The implicit solvent model could be replaced or combined with an explicit solvent model, but this would increase computational expense.

We have derived a method for the prediction of the solubility of drug molecules in their free acid or base form. An interesting possibility is that this method could be extended to the prediction of the solubility of single molecules in other solvents and to the solubility of other solid forms, such as cocrystals, solvates or salts. The latter is of particular interest due to the importance of these crystalline forms in pharmaceutical formulations. The relationship between solid form and solubility is more complicated when there is more than one chemical entity in the lattice. The equilibrium may not exist solely between the multicomponent crystal and the solvated form, but the lattice energy of the multicomponent system will still be a determinant in the solubility. DMAREL 3.11 has already been used for the calculation of the lattice energies of

(69) Cramer, C. J. In *Essentials of Computational Chemistry: Theories and Models*; John Wiley & Sons: Chichester, 2002; p 347.
(70) Koehl, P. Electrostatics Calculations: Latest Methodological Advances. *Curr. Opin. Struct. Biol.* **2006**, *16*, 142–151.
(71) Cruz Cabeza, A. J.; Day, G. M.; Motherwell, W. D. S.; Jones, W. Prediction and Observation of Isostructurality Induced by Solvent Incorporation in Multicomponent Crystals. *J. Am. Chem. Soc.* **2006**, *128*, 14466–14467.
(72) Anghel, A. T.; Day, G. M.; Price, S. L. A Study of the Known and Hypothetical Crystal Structures of Pyridine: Why Are There Four Molecules in the Asymmetric Unit? *CrystEngComm* **2002**, *4*, 348–355.

cocrystals and solvates,[71] which suggests that our method could be applicable to other solid-state systems.

Conclusion

We find that direct computation of solubility, via *ab initio* calculation of thermodynamic quantities at an affordable level of theory, cannot deliver the required accuracy. Therefore, we have turned to a mixture of direct computation and informatics, using the calculated thermodynamic properties, along with one additional descriptor, in regression models. A multilinear regression (eq 8) containing the variables U_{latt}, ClogP, and b_rotR is shown to be able to predict solubility with an error of 0.71 log solubility (referred to mol/L), for an external data set comprising druglike molecules. The model contains a calculated lattice energy (U_{latt}) to account for the interactions in the solid state. We suggest that it is not necessary to know polymorphic form prior to prediction. Furthermore, the method developed here may be applicable to other solid state systems such as salts or cocrystals.

Acknowledgment. We thank Pfizer for sponsoring this work through the Pfizer Institute for Pharmaceutical Materials Science and acknowledge Unilever plc for their financial support of the Unilever Centre for Molecular Science Informatics.

MP7000878