molecular pharmaceutics

A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects

Stephen R. Johnson,*,† Xue-Qing Chen,† Denette Murphy,‡ and Olafur Gudmundsson†

Bristol-Myers Squibb, Co., Princeton, New Jersey 08543, and Bristol-Myers Squibb, Co., New Brunswick, New Jersey 08903

Received March 20, 2007; Revised Manuscript Received April 16, 2007; Accepted May 8, 2007

Abstract: The optimization of aqueous solubility is an important step along the route to bringing a new therapeutic to market. We describe the development of an empirical computational model to rank the pH-dependent aqueous solubility of drug candidates. The model consists of three core components to describe aqueous solubility. The first is a multivariate QSAR model for the prediction of the intrinsic solubility of the neutral solute. The second facet of the approach is the consideration of ionization using a predicted pK_a and the Henderson–Hasselbalch equation. The third aspect of the model is a novel method for assessing the effects of crystal packing on solubility through a series of short molecular dynamics simulations of an actual or hypothetical small molecule crystal structure at escalating temperatures. The model also includes a Monte Carlo error function that considers the variability of each of the underlying components of the model to estimate the 90% confidence interval of estimation.

Keywords: Aqueous solubility; crystal packing; QSAR; nonparametric confidence interval estimation

Introduction

The importance of aqueous solubility in the discovery and development of new drugs is difficult to overstate. This critical physical property affects many aspects of the discovery and development cycle, from the accuracy of screening assays to the selection of a final dosage form for clinical trials.[1] Nonetheless, limited aqueous solubility is a perpetual obstacle in the successful discovery and development of a new drug as typical medicinal chemistry routes to improved pharmacological potency frequently increase lipophilicity as compounds move from early leads to drugs.[2]

Not surprisingly, a seemingly endless number of papers have appeared in the literature dealing with solubility prediction using computational models. Methods employing linear statistics,[3−6] sophisticated machine learning approaches,[7−17] or continuum solvation methods[18−20] have all been reported

* Author to whom correspondence should be addressed. Mailing address: Pharmaceutical Research Institute, Bristol-Myers Squibb, Co., P.O. Box 4000, Princeton, NJ 08543-4000, 609-252-3003 (voice), 609-252-6030 (fax), stephen.johnson@bms.com.

† Bristol-Myers Squibb, Co., Princeton.

‡ Bristol-Myers Squibb, Co., New Brunswick.

[1] Chen, X.-Q.; Antman Melissa, D.; Gesenberg, C.; Gudmundsson Olafur, S. Discovery pharmaceutics—challenges and opportunities. *AAPS J.* **2006**, *8*, E402−408.

[2] Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

[3] Chen, X.-Q.; Cho, S. J.; Li, Y.; Venkatesh, S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J. Pharm. Sci.* **2002**, *91*, 1838−1852.

[4] 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*, 3572−3580.

[5] Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

[6] Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

10.1021/mp070030+ CCC: $37.00 © 2007 American Chemical Society
Published on Web 06/01/2007

in recent years for the estimation of aqueous solubility. For the most part, these reports have utilized large data sets coupled with powerful statistical approaches to generate predictions. Several reviews of aqueous solubility prediction have appeared recently.[21−24]

These reviews have clearly shown that reliable prediction of aqueous solubility has proven quite difficult. Far from being a simple property, solubility is affected by molecular properties, pH, and crystal packing among other factors. As most drugs are electrolytes, the effect of ionization is critical to useful solubility estimation. Most computational models have completely ignored this aspect of solubility, or have made an implicit assumption that sophisticated learning algorithms can predict solubility at the pH at which the underlying data was measured. More recently, a number of reports have made explicit use of a predicted pK_a to estimate the solubility at a particular pH.[10,25−28]

As mentioned above, solubility is also dependent on the crystal form of the solid material used in the measurement. A recent review gives an excellent overview of the role of the solid state on pharmaceutical developability considerations.[29] Most interesting out of this review is the notion that solubility differences among polymorphs is typically less than 10-fold.[29] Other reports put the typical effect of different polymorphs at less than a 2-fold change in solubility.[30] The difference in solubility between crystalline material and solubility from amorphous material, however, can be quite substantial.[29] This may imply that including information regarding crystal packing is important for accurate prediction, but that highly accurate information regarding the exact crystal form may not be necessary. However, most computational models of solubility have ignored, or treated implicitly, the effect of crystal packing. This is likely because of the difficulty in modeling the condensed phase, and a lack of information on the precise crystal form used to derive the experimental data. There are a few notable exceptions, however, that include factors relating to crystallinity.[28,31−33]

(7) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.

(8) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.

(9) Engkvist, O.; Wrede, P. High-Throughput, In Silico Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247−1249.

(10) Hansen, N. T.; Kouskoumvekaki, I.; Jorgensen, F. S.; Brunak, S.; Jonsdottir, S. O. Prediction of pH-Dependent Aqueous Solubility of Druglike Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 2601−2609.

(11) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G., et al. A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674−679.

(12) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodiversity* **2004**, *1*, 1829−1841.

(13) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.

(14) Yan, A.; Gasteiger, J.; Krug, M.; Anzali, S. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75−87.

(15) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds by topological descriptors. *QSAR Comb. Sci.* **2003**, *22*, 821−829.

(16) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.

(17) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; Ter Laak, A., et al. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. *J. Chem. Inf. Model.* **2007**, *47*, 407−424.

(18) 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.

(19) Ikeda, H.; Chiba, K.; Kanou, A.; Hirayama, N. Prediction of solubility of drugs by conductor-like screening model for real solvents. *Chem. Pharm. Bull.* **2005**, *53*, 253−255.

(20) Oleszek-Kudlak, S.; Grabda, M.; Shibata, E.; Eckert, F.; Nakamura, T. Application of the conductor-like screening model for real solvents for prediction of the aqueous solubility of chlorobenzenes depending on temperature and salinity. *Environ. Toxicol. Chem.* **2005**, *24*, 1368−1375.

(21) Gudmundsson, O. S.; Venkatesh, S. Strategies for in silico and experimental screening of physicochemical properties. *Biotechnol.: Pharm. Aspects* **2004**, *1*, 393−412.

(22) Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discovery Today* **2005**, *10*, 289−295.

(23) Johnson, S. R.; Zheng, W. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **2006**, *8*, E27−E40.

(24) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1*, 31−52.

(25) Dewitte, R. S.; Kolovanov, E. D. Predicting molecular physical properties. *Biotechnol.: Pharm. Aspects* **2004**, *1*, 27−52.

(26) Bergstrom, C. A. S.; Luthman, K.; Artursson, P. Accuracy of calculated pH-dependent aqueous drug solubility. *Eur. J. Pharm. Sci.* **2004**, *22*, 387−398.

(27) Lobell, M.; Sivarajah, V. In silico prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pKa and AlogP98 values. *Mol. Diversity* **2003**, *7*, 69−87.

(28) Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H. Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR Comb. Sci.* **2003**, *22*, 258−262.

(29) Huang, L.-F.; Tong, W.-Q. Impact of solid state properties on developability assessment of drug candidates. *Adv. Drug Delivery Rev.* **2004**, *56*, 321−334.

(30) Pudipeddi, M.; Serajuddin, A. T. M. Trends in solubility of polymorphs. *J. Pharm. Sci.* **2005**, *94*, 929−939.

(31) Raevsky, O. A.; Raevskaja, O. E.; Schaper, K.-J. Analysis of water solubility data on the basis of HYBOT descriptors. Part 3. Solubility of solid neutral chemicals and drugs. *QSAR Comb. Sci.* **2004**, *23*, 327−343.

(32) Ran, Y.; He, Y.; Yang, G.; Johnson, J. L. H.; Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* **2002**, *48*, 487−509.

This report will detail our efforts at developing a computational model of aqueous solubility. Our approach is to divide solubility into three primary components: the intrinsic solubility of the neutral form free of its crystal lattice, the effect of crystal packing, and the effect of ionization. The intrinsic solubility model uses a linear QSPR model trained using data from the literature. Consistent with previous literature, the Henderson−Hasselbalch equation is used with an estimated pK_a to account for ionization. Crystal packing is assessed by tracking the unit cell integrity over a series of short molecular dynamics simulations at progressively higher temperatures. Finally, a Monte Carlo routine is utilized to generate error bars for the prediction based on the estimated variability of each underlying model component. Together, these represent an interpretable, modular approach to modeling solubility. Such a computational framework could serve as a reasonable basis for the iterative improvement of solubility estimation over time.

Methods

Materials. Commercially available compounds tested and all other chemicals used were obtained from Sigma Chemical Company (St. Louis, MO).

Solubility Measurements. An excess of each compound was equilibrated with either a 25 mM phosphate buffer (pH 6.5) or in water. Samples were mixed in vials for at least 1 day at room temperature (22 ± 2 °C) until equilibrium was reached using a shaker. After visual inspection more drug substance was added, if needed. At the end of the equilibration the samples were filtered using a syringe and Gelman Acrodisc 13 CR PTFE 0.45 μm membrane. The pH of the filtrate was measured, diluted as appropriate, and subsequently analyzed by RP-HPLC.

Computational Methods. For the purposes of developing a computational model of aqueous solubility, we postulated an *ansatz* describing the functional form for the estimation of solubility at an arbitrary pH:

$$\mathrm{Log}(S_{\mathrm{pH}}) = \mathrm{Log}(S_{\mathrm{o}}) + \min\left[\mathrm{Log}\left[10^{\sum_i^{\mathrm{Nacids}}(\mathrm{pH}-\mathrm{p}K_{\mathrm{a}\,i})+\sum_j^{\mathrm{Nbases}}(\mathrm{p}K_{\mathrm{a}\,j}-\mathrm{pH})+1}\right],4.25\right] - \chi_{\mathrm{pack}}\exp(-F_{\mathrm{I}}) \quad (1)$$

where the first term is a predicted intrinsic solubility, the second term describes the impact of ionization using the Henderson−Hasselbalch equation, and the third term captures the influence of crystal packing forces that are mitigated by the degree of ionization (F_I) of the solute. χ_{pack} is approximated by X_{pack} with an assumed functional form:

$$\chi_{\mathrm{pack}} \approx X_{\mathrm{pack}} \equiv \frac{0.1}{{D_{300}}^2+1} + \frac{0.2}{{D_{400}}^2+1} + \frac{0.3}{{D_{500}}^2+1} + \frac{0.6}{{D_{600}}^2+1} + \frac{0.8}{{D_{700}}^2+1} \quad (2)$$

where D_T is the slope of the mean square atomic displacement versus time curve from a series of short molecular dynamics simulations at temperature T (in kelvins). X_{pack} will be discussed in more detail below.

Intrinsic Solubility Prediction. Compounds with measured solubility were taken from the literature.[34] Only compounds with a molecular weight below 500 daltons and no ionizable centers were selected. Ionizable compounds were eliminated by ionizing compounds using LigPrep,[35] and then removing any compounds containing a formal charge. Removing these compounds was necessary, as we could not verify at what pH the literature solubility measurements were determined. The compounds meeting these criteria were then viewed visually, and only those considered to be "reasonable" structures were included. There was no formal definition for what functionality passed this inspection. The goal was simply to avoid compounds that contained strange functionality. Compounds were divided into a low MW set (MW ≤ 250) and a high MW set (250 < MW ≤ 500). Splitting the data by molecular weight is consistent with observations of solvent entropy scaling differently with larger solutes compared to smaller solutes.[36,37] This resulted in 219 compounds in the high MW set and 362 in the low MW set. The high MW set was randomly divided into a training set of 199 compounds and a test set of 20 compounds. The low MW set was randomly divided into a 326 compound training set and a 36 compound test set. The list of compounds is given in the Supporting Information.

Compounds were modeled in the neutral form. A single conformation was generated using Omega[38] and minimized using Batchmin[39] with OPLS2003 and implicit water. Using these conformations, many molecular descriptors were calculated for the training and validation data used in the intrinsic solubility model development. Included among these were the VolSurf descriptors[40] using both the Dry and water probes, clogP, polar surface area, and counts of H-bond acceptors and H-bond donors. A total of 86 descriptors were calculated for use in predicting the intrinsic solubility.

A supervised feature selection algorithm using simulated annealing with a leave-10%-out PRESS cross-validation function was employed to select features for the intrinsic solubility model. This selection algorithm is similar to others

(33) 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.

(34) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(35) *LigPrep*, 20113 ed.; Schrodinger, LLC: New York.

(36) Huang, D. M.; Chandler, D. The Hydrophobic Effect and the Influence of Solute-Solvent Attractions. *J. Phys. Chem. B* **2002**, *106*, 2047−2053.

(37) Jensen, M. O.; Mouritsen, O. G.; Peters, G. H. The hydrophobic effect: Molecular dynamics simulations of water confined between extended hydrophobic and hydrophilic surfaces. *J. Chem. Phys.* **2004**, *120*, 9729−9744.

(38) *Omega*, 2.1.0 ed.; OpenEye Scientific Software, Inc.: Sante Fe, NM.

(39) *MacroModel*, 9.1 ed.; Schrödinger, LLC: New York, NY.

(40) *VolSurf*, 4.1.4 ed.; Molecular Discovery, Ltd.

reported in the literature.[41,42] Initially, a random subset of descriptors is chosen. Following this, a descriptor is removed and replaced with another feature from the pool. This replacement is done randomly, with the exception that the new feature must have an $r < 0.9$ ($r^2 < 0.81$) to the other features already present in the model. There is a 1% chance at each iteration that the number of descriptors may be decreased or increased. The PRESS score is then evaluated for this new model. Each descriptor used in a model adds $0.002 \times \text{rms}_{\text{av}}$ to the PRESS score, where rms_{av} is the RMSE if every observation was predicted as the average value of the dependent variable. If the score is an improvement over the previous model, the result is accepted and the procedure repeats. If the model is not an improvement, the model may still be accepted based on a probability derived from the Boltzmann distribution.

The implementation used here was programmed in our laboratories and fixes the initial temperature such that 80% of the detrimental steps are accepted originally.[43] The temperature is then decreased by 25% every 1000 iterations. The selection process halts when no model is accepted for 900 iterations.

The selected feature values for the training data were then normalized to a range of 0 and 1. These normalized values were utilized by a least median squares (LMS) regression algorithm[44] to generate coefficients. LMS regression minimizes the median squared residual rather than the sum of the squared residuals as is performed using ordinary multiple linear regression. By minimizing the median squared residual, the impact of compounds with high statistical leverage to alter the derived linear coefficients is limited to yield the linear relationship that best fits the majority of the available data. A simple way to think of observations with high leverage is that they have an unusual combination of values for the independent variables (i.e., descriptors). This leaves these observations far from the bulk of observations in descriptor space used in the regression, providing these observations tremendous impact on the resulting coefficients. It is important to realize that an observation can have substantial leverage but not be an outlier, or could be an outlier and not have leverage. Leverage points that are outliers will significantly distort the derived linear relationship, while leverage points that are not outliers may significantly affect the resulting standard errors of prediction.[45] The LMS algorithm minimizes the impact of leverage data on the coefficients, possibly alleviating some concerns about the quality of data used to derive the intrinsic solubility model. For an in-depth discussion of the algorithm the reader is referred to ref 46. The approach has been used previously in the chemical literature.[47−49]

Crystal Packing Simulations. Compounds were selected for purchase based on the following criteria: (a) having an available neat (no salts or solvates) X-ray structure in the Cambridge Structural Database (CSD),[50] (b) being reasonably drug-like by manual inspection, and (c) being available from a vendor (Table 1). Using our internal protocols, the solubility at a pH of 6.5 was measured. In addition to this data, 37 discovery compounds were also employed. These compounds had previously measured solubility from crystalline material along with solved crystal structures. Together, this collection of compounds was used to assess the ability of the model to capture the effect of crystal packing on solubility.

The crystal packing parameter X_{pack} from eq 2 is based on the hypothesis that crystal lattice energy should be proportional to the stability of a unit cell when additional energy is added to the system.[51,52] To assess this, the crystallographic unit cell was utilized in a series of short molecular dynamics simulations in which the mean square atomic displacement over time was calculated. The MD simulations were performed with Discover3 (cdiscover)[53] using the CFF91 force field. For complexes that could not be parametrized using CFF91, the CVFF force field was used. The simulations are performed as a constant pressure (NTP) simulation at a pressure of 1 bar. The simulation details are shown in Table 2. Separate simulations are

[41] Sutter, J. M.; Jurs, P. C. Selection of molecular descriptors for quantitative structure-activity relationships. *Data Handling Sci. Technol.* **1995**, *15*, 111−132.

[42] Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

[43] Sutter, J. M. Personal communication; Accelrys, Inc.

[44] Massart, D. L.; Kaufman, L.; Rousseeuw, P. J.; Leroy, A. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Anal. Chim. Acta* **1986**, *187*, 171−179.

[45] Croux, C. Are Good Leverage Points Good or Bad? International Conference on Robust Statistics: Lisbon, Portugal, 2006.

[46] Rousseeuw, P. J. L.; Annick M. *Robust Regression and Outlier Detection*; John Wiley & Sons, Inc.: New York, NY, 1987; p 329.

[47] Cruz Ortiz, M.; Sarabia, L. A.; Herrero, A. Robust regression techniques. *Talanta* **2006**, *70*, 499−512.

[48] Igumenova, T. I.; Lee, A. L.; Wand, A. J. Backbone and Side Chain Dynamics of Mutant Calmodulin-Peptide Complexes. *Biochemistry* **2005**, *44*, 12627−12639.

[49] Johnson, S. R.; Jurs, P. C. Prediction of acute mammalian toxicity from molecular structure for a diverse set of substituted anilines using regression analysis and computational neural networks. Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry, [European Symposium on Quantitative Structure-Activity Relationships], 11th, Lausanne, Sept 1−6, 1996; 1997, pp 31−48.

[50] Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *B58*, 380−388.

[51] Gavezzotti, A. The chemistry of intermolecular bonding: organic crystals, their structures and transformations. *Synlett* **2002**, 201−214.

[52] Gavezzotti, A. A Molecular Dynamics Test of the Different Stability of Crystal Polymorphs under Thermal Strain. *J. Am. Chem. Soc.* **2000**, *122*, 10724−10725.

[53] *Discover3 (cdiscover)*; Accelrys, Inc.: San Diego, CA.

Table 1. Compounds Used To Parametrize Crystal Packing Simulation

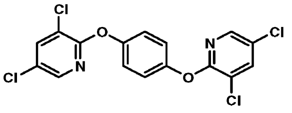
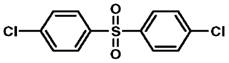
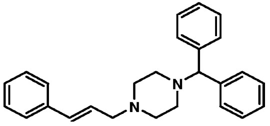
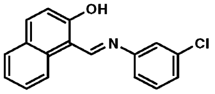
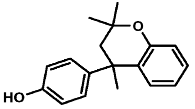
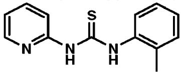
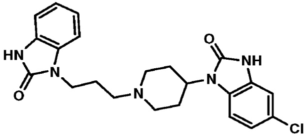
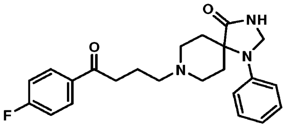
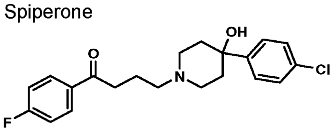
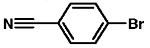
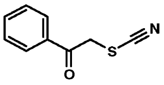
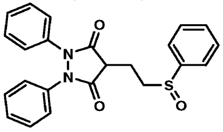
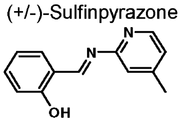
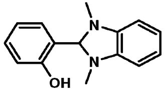
	CSD Entry	Observed Solubility pH6.5,ug/ml	Predicted Solubility pH6.5, ug/mL	Xpack
 1,4-bis(3,5-dichloro-2-pyridyloxy)-benzene	DADDUH	0.016	0.0013	0.55
 Bis 4-Chlorophenyl Sulfone	CLPSUL10	0.05	0.974	0.5
 1-trans-cinnamyl-4-diphenylmethyl-piperazine	CINNAZ	0.24	0.07	0.91
 N-(3-chlorophenyl)-2-hydroxy-1-naphthaldimine	FADBUI	0.81	2.22	0.46
 4-(4-hydroxyphenyl)-2,2,4-trimethylchroman	QQQESP01	0.93	0.597	1.26
 N-(2-pyridyl)-N'-o-tolythiourea	FAYXUY	2.13	54.5	0.32
 Domperidone	BEQJUC	17.1	333	0.62
 Spiperone	FBPAZD01	32	209	0.58
 Haloperidol	HALDOL01	61	131	0.3
 4-Bromobenzonitrile	BRBNIT	157.2	191	0.3
 Phenacyl Thiocyanate	FAGFUO	349.5	509	0.14
 (+/-)-Sulfinpyrazone	COKBOT	2608	871	0.66
 Alpha-(4-Methyl-2 Pyridylimino)-O-Cresol	BUMXOW10	4303.5	12020	0.62
 2-(1,3-Dimethyl-2,3-Dihydro-1H-Benzpimidazol-2-YL)-Phenol	FOKJIY	10732.6	105	1.1

Table 2. Parameters for the Molecular Dynamics Simulations

nonbonded interactions:	
summation method: group based	
cutoff: 45	
spline width: 1.0	
buffer width: 0.5	
minimization:	
iterations: 300	
movement limit: 0.2	
steepest descent convergence: 1000	
conjugate gradient: convergence: 10	method: Polak
Newton convergence: 0.1	method: BFGS
dynamics:	
time: 10 ps	step: 1 fs
ensemble: NPT	temp control: Andersen
integration method: velocity Verlet	
pressure: 1 bar	pressure control: Parrinello

performed over 10 ps at temperatures of 300, 400, 500, 600, and 700 K. The mean square atomic displacement (MSD), relative to the starting unit cell coordinates, is calculated every 10 fs and plotted versus simulation time at each temperature. The value D_T used in eq 2 the slope of the MSD vs time curve at temperature T. Each simulation is repeated 3 times, and the average and standard deviation of X_{pack} from eq 2 is calculated.

The maximum effect of 2 orders of magnitude was selected based on the results presented in Bergstrom et al.[26] In this work, the authors cite the typical increase in solubility upon ionization to be approximately 4.25 orders of magnitude. However, the greatest value they measured was 6.5 orders of magnitude. We hypothesized that this difference was related more to the effect of crystal packing leading to a suppression of the intrinsic solubility than to the heightened impact of ionization for some compounds. When the functional form for eq 2 was postulated, the coefficients were chosen to sum to 2, with larger effects arising as the unit cell maintained organization even at high temperature.

When used for prediction, it is typical for compounds to lack crystallographic data. In this case, virtual crystal structures were simulated using the crystal structure of a related compound. These were generated by overlaying the new compound on to the monomers of the asymmetric unit cell of the known compound. The unit cell was then expanded using the CRYSIN functionality within SYBYL.[54] Any salts or solvates in the observed crystal data were included in the simulated structure.

Confidence Interval Estimate. A Monte Carlo simulation is used to generate an estimate of the 90% confidence interval of prediction. Each of the main components of the model is varied by adding normally distributed noise with a mean of zero and a standard deviation consistent with the variability of the estimated component. While the ACD/pK_a program does provide a specific error term for each estimated pK_a, we have found it to be unrealistically low for most estimates. We have assumed a standard deviation of ±0.5 for the pK_a estimates. The variability of the intrinsic solubility is varied within the 90% confidence interval of prediction from the multiple linear regression. The value of the crystal packing parameter, X_{pack}, is varied within the standard deviation of 3 runs of the MD simulation. Each of these components of the pH-dependent solubility prediction model is varied with the above error bounds over 1000 iterations at each pH. The median calculated solubility value at each pH is used as the point estimate, with the 10th percentile and 90th percentile values used as the bounds for the confidence interval. The method is similar to the percentile method described by Buckland.[55]

Table 3. Descriptors and Coefficients of the Intrinsic Solubility Model

		coeff	
descriptor	label	MW ≥ 250	MW < 250
no. of H-bond acceptors	HA	1.22	0.72
hydrophilic interaction volume at −3.0 kcal/mol	W5	3.98	2.23
hydrophilic capacity			
at −0.2 kcal/mol	CW1	0.85	1.66
at −2.0 kcal/mol	CW4	−6.88	−6.51
hydrophobic interaction volume at −0.6 kcal/mol	D3	−1.94	−1.48
critical packing	CP	−2.42	−1.99
molecular weight	MW	−2.98	−1.46
cLogP	cLogP	−5.69	−6.03
constant		0.1	2.36
R^2 (train/test)		0.88/0.93	0.79/0.85
RMSE (train/test)		0.61/0.50	0.54/0.49

Results and Discussion

The prediction of aqueous solubility at an arbitrary pH begins with an estimate of the intrinsic solubility. The intrinsic solubility of a compound is the aqueous solubility of the nonionized form of the compound. For our purposes, this intrinsic solubility should not include the impact of crystal packing to the extent possible. Second is a correction for the effect of ionization on the pH−solubility curve. Finally, the model includes a procedure for evaluating the impact of crystal packing on the aqueous solubility.

The features selected for the estimation of the intrinsic solubility are shown in Table 3 along with their coefficients. Figure 1A shows the calculated versus observed plot for the higher MW nonionizable compounds used in the training and validation of this model. Clearly, the model performs well over the span of the 11 orders of magnitude of the intrinsic solubility data. Figure 1B shows the same data, but focused in on the 0.01 μM to 100 μM range of data that is most relevant to discovery in the pharmaceutical industry. A similar plot of the low MW data is shown in the Supporting Information.

[54] *SYBYL*; Tripos, Inc.: St. Louis, MO.

[55] Buckland, S. T. Algorithm AS 214: Calculation of Monte Carlo Confidence Intervals. *Appl. Stat.* **1985**, *34*, 296−301.

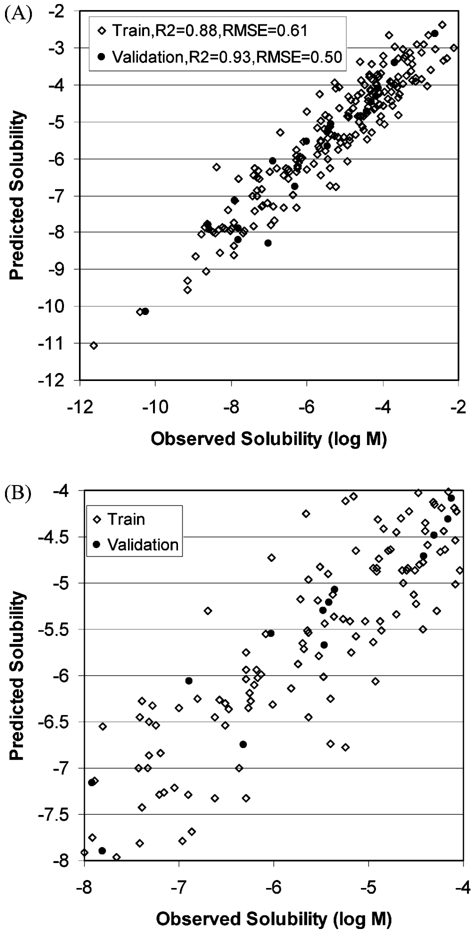


Figure 1. (A) Calculated versus observed plot for intrinsic solubility for nonionizable training and test data. (B) Focused on the pharmaceutically relevant solubility range.

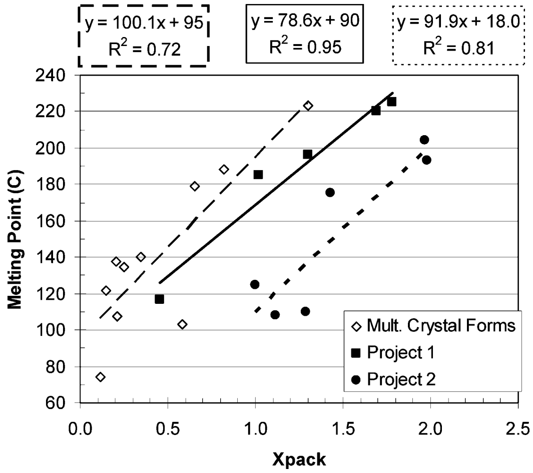


Figure 2. Relationship between X_{pack} and melting point for three groups of compounds. The first group (open diamonds) are 10 different crystal forms of the same compound. Projects 1 and 2 are compounds from two discovery projects.

Solubility is the balance of forces between cavity formation, solute−solute interactions, and solute−solvent interactions. The MW and the critical packing feature relate to the molecular size and shape. They are likely related to the energy required to form a cavity in the solvent large enough to hold the solute molecule. The remaining features most likely relate to intermolecular interactions between the solute and solvent molecules or between multiple monomers of the solute. In general, the coefficients are in line with the basic intuition regarding the role of hydrophilicity and solubility.

The lone exception is the sign of the coefficient on the hydrophilic capacity factor at −2.0 kcal/mol (CW4) descriptor. The coefficient on this descriptor leads to a prediction of lower solubility for compounds with higher hydrophilic potential. At first glance, this is contradictory to the general perception regarding solubility and polarity. However, solute−solute interactions are also very important in aqueous solubility. Raevsky and coauthors[31] uncovered a similar trend in experimental data in which compounds with strong H-bond acceptors and donors have lower than expected solubility. It is possible that the CW4 feature is describing the potential for stronger solute−solute interactions for certain types of polar groups. It is also worth noting that while the coefficient is fairly large for the CW4 feature, most compounds do not have large values for this descriptor. Still, the interplay between the CW2 and CW4 features is particularly important in the model.

Crystal packing has long been a vexing problem with respect to aqueous solubility prediction. Yalkowsky's general solubility equation[28] was an early attempt at incorporating information regarding the crystal lattice into a prediction of aqueous solubility. More recently, several attempts have been made at predicting melting points, a frequent surrogate for crystal packing in solubility models.[28,32,56−60] We hypothesized that the amount of atomic motion observed in a unit cell in response to adding increasing amounts of energy would be related to the effect of crystal packing on solubility. Figure 2 shows the relationship between X_{pack} and melting point for three different groups of crystal structures. One set uses several crystal forms of a single compound, each of which varies by salt or solvate or is a polymorph of another included structure. The other two groups of compounds represent compounds from two different discovery projects. The relationship is roughly linear within each group of

(56) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; et al. Perspective on the Relationship between Melting Points and Chemical Structure. *Cryst. Growth Des.* **2001**, *1*, 261−265.

(57) 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*, 581−590.

(58) Johnson, J. L. H.; Yalkowsky, S. H. Two New Parameters for Predicting the Entropy of Melting: Eccentricity (e) and Spirality (m). *Ind. Eng. Chem. Res.* **2005**, *44*, 7559−7566.

(59) Jain, A.; Yang, G.; Yalkowsky, S. H. Estimation of Melting Points of Organic Compounds. *Ind. Eng. Chem. Res.* **2004**, *43*, 7618−7621.

(60) Bergstroem, 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.

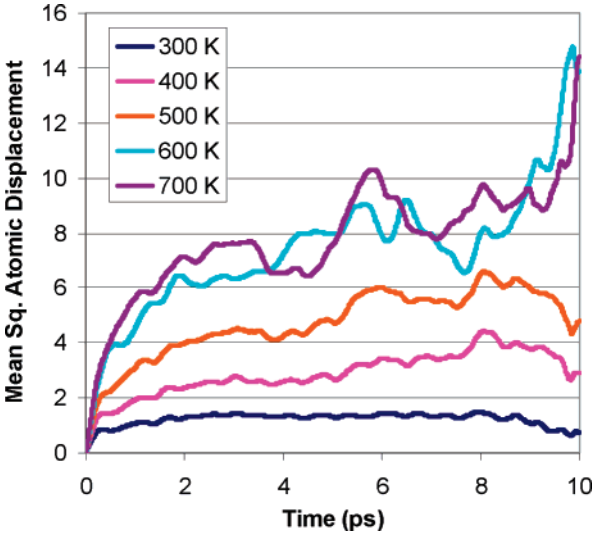


Figure 3. Results from the crystal packing simulation of griseofulvin. The GRISFL03 structure was utilized from the CSD. The slopes of the mean square displacement curves are as follows: 300 K, 0.01; 400 K, 0.24; 500 K, 0.37; 600 K, 0.67; 700 K, 0.57. Equation 2 gives the result $X_{\text{pack}} = 1.57$.

compounds, but shifts to higher values of X_{pack} as the average number of H-bonds in the unit cells decreases. This suggests avenues for further investigation going forward.

Figure 3 shows the plot of the mean squared atomic displacement versus time for griseofulvin, a marketed antifungal known to have solubility limited absorption. This simulation, performed using the GRISFL03 structure from the CSD, gives the value of $X_{\text{pack}} = 1.57$.

Wassvik et al.[61] measured the intrinsic solubility of griseofulvin to be $\log(S_o,M) = -4.83$ (14.8 μM). The intrinsic solubility prediction generated by the model described here is $\log(S_o,M) = -3.33$ (141 μM). Including the effect of crystal packing above, the estimated solubility is $\log(S_o - X_{\text{pack}},M) = -4.90$ (12.5 μM). Predictions of the intrinsic solubility of 25 compounds given in Wassvik et al.[61] were generated using the intrinsic solubility model discussed above. Figure 4A shows the prediction results using only the intrinsic model. The model predicts the intrinsic solubility quite well with an $R^2 = 0.70$ and an RMSE = 0.85 log unit. Figure 4B shows the predictions after correction for crystal packing.

The crystal packing corrections were generated using crystal structures obtained from the CSD. By including a contribution from crystal packing, the R^2 improves to 0.75 while the RMSE is increased to 0.86 log unit. While the overall RMSE does not improve with the inclusion of crystal packing effects, the median absolute error improves dramatically from 0.63 to 0.46. This difference is the result of particularly poor predictions for mifepristone and diazepam. For these two compounds, the inclusion of the crystal packing parameter gives predictions that are noticeably worse than those of the intrinsic solubility model alone. The R^2 improves to 0.85 and the RMSE to 0.62 by excluding these two compounds. Perphenazine is also very poorly predicted, although the prediction does not get any worse by including the crystal packing contribution.

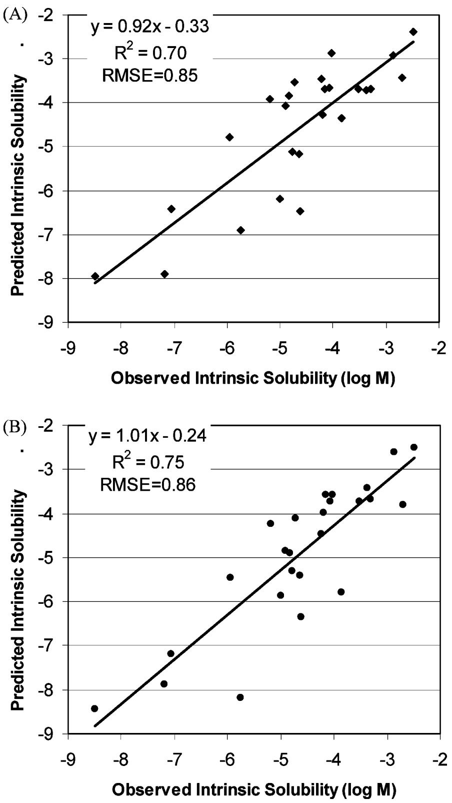


Figure 4. Predicted versus observed results for intrinsic solubility of compounds from Wassvik et al. (A) Results from the intrinsic solubility model only. (B) Predictions including the estimate of the impact of crystal packing.

Of course, most potential drug compounds contain some sort of basic or acidic ionizable functionality. The literature contains several reports of using the Henderson−Hasselbalch equation to correct for ionization. We have followed the same basic approach here. Bergstrom et al.[26] highlighted some of the weaknesses of this approach including the variability in the slope of ionization. We made several unsuccessful attempts (not shown) to correct for the effects of aggregation on the slope of ionization. Included among these were linear and nonlinear QSAR approaches to predict aggregation that incorporated conformational and energetic changes upon ionization. While some of these approaches appeared interesting in training, none proved reliable in improving solubility estimation upon external validation.

For our purposes, we have utilized a predicted pK_a using the ACD/pK_a program.[62] In addition, we cap the contribution of ionization to solubility at 4.25 orders of magnitude for a

[61] Wassvik, C. M.; Holmen, A. G.; Bergstroem, C. A. S.; Zamora, I.; Artursson, P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur. J. Pharm. Sci.* **2006**, *29*, 294−305.

[62] *ACD/pKa*, 4.76 ed.; Advanced Chemistry Development, Inc.: Toronto, ON, Canada.

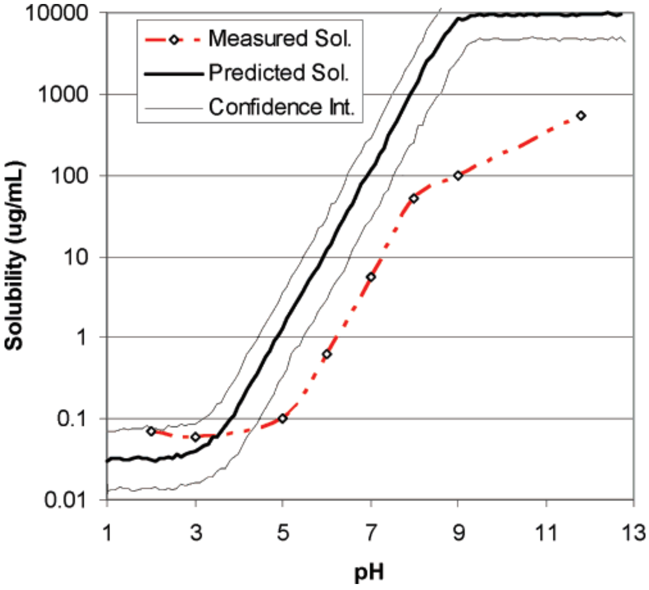


Figure 5. Predicted pH−solubility curve for glyburide. Experimental data from ref 63.

single ionizable center and a total of 5 orders of magnitude for more than a single ionizable center. While the cap of 4.25 is drawn from the literature[26] and may be reasonable in the absence of any other information, the cutoff of 5 orders of magnitude for polyelectrolytic systems is completely arbitrary. It is worth pointing out that, while the predicted impact of ionization is capped at 4.25 log units, it can appear as great as 6.25 log units as a result of the depression of the intrinsic solubility based on the contribution of crystal effects discussed below.

Figure 5 shows the measured[63] and predicted pH−solubility curves for glyburide. The predicted intrinsic solubility, excluding crystal packing, was −5.904 ± 0.351 (log molarity). The crystal packing simulation yields $X_{\text{pack}} = 1.3 \pm 0.25$, which together with the predicted Log S_o gives −7.2 compared to a measured[63] Log S_o = −7.05. The predicted pK_a was 4.86 using ACD/pK_a, with an assumed variability of ±0.5. The measured pK_a is 5.3. Viewed as a whole, the predicted pH−solubility curve is a quite reasonable estimate of the measured curve. However, estimates at a specific pH can have large errors when in the range of pH's affected by ionization as evidenced by the predicted solubility at pH = 6 being 7.4 μg/mL (with a 90% range from 2.4 to 24.1 μg/mL) compared to a measured value of 0.62 μg/mL.

Figure 6A shows the measured[64] and predicted pH−solubility curve for haloperidol as the HCl salt. The crystal packing simulation was performed based on HALDOL01 from the CSD, which is a small molecular crystal structure of the HCl salt of haloperidol. Figure 6B shows the measured[64] and predicted pH−solubility curve for haloperidol as the mesylate salt. The crystal packing simulation was performed based on the CSD entry YANMUW, where the mesylate ion was modeled manually into the crystal structure in place of the saccharinate ion. The replacement was guided by simple pharmacophore and steric considerations. The crystal packing simulation used the CFF91 forcefield, which may have significant shortcomings when modeling charged ions. In both cases, the predictions are quite reasonable compared to the measured values. It is noteworthy that the maximum solubility of the HCl salt is limited by the common ion effect, which is not accounted for in the current model. The model also uses a slope of 1.0 for the ionization curve. The actual slope of the observed pH−solubility plot for the mesylate salt is slightly over 2.

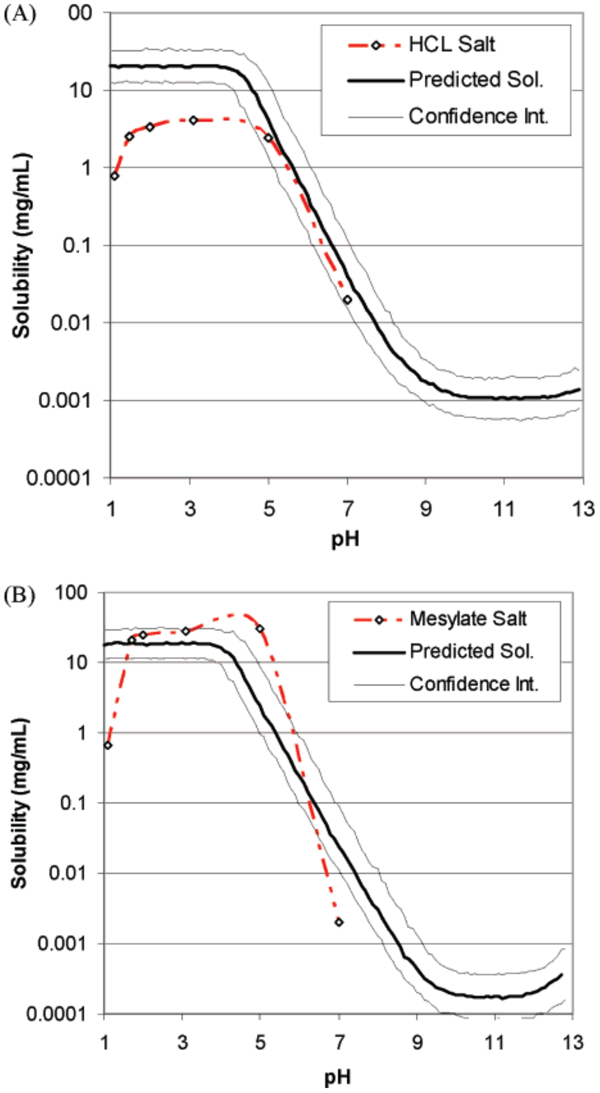


Figure 6. Predicted and observed pH−solubility curves for (A) haloperidol HCl salt and (B) mesylate salt. Data from ref 64.

[63] Glomme, A.; Maerz, J.; Dressman, J. B. Comparison of a miniaturized shake-flask solubility method with automated potentiometric acid/base titrations and calculated solubilities. *J. Pharm. Sci.* **2005**, *94*, 1−16.

[64] Li, S.; Wong, S.; Sethia, S.; Almoazen, H.; Joshi, Y. M., et al. Investigation of Solubility and Dissolution of a Free Base and Two Different Salt Forms as a Function of pH. *Pharm. Res.* **2005**, *22*, 628−635.

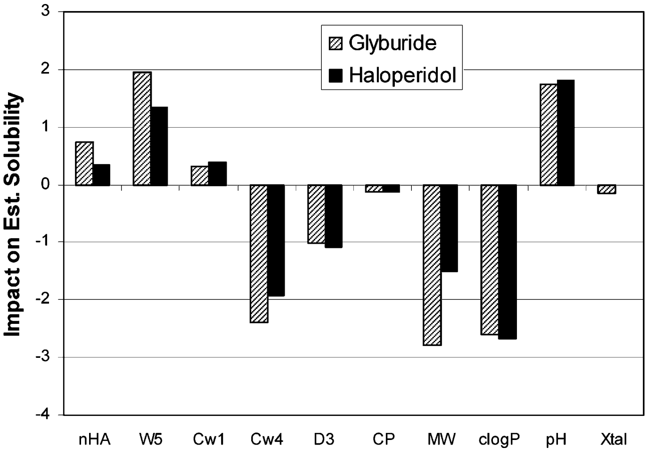


Figure 7. Descriptor impact on solubility predictions for glyburide and haloperidol (HCl salt) at pH = 6.5. pH denotes the impact from ionization at pH = 6.5, Xtal denotes the impact from crystal lattice at pH = 6.5.

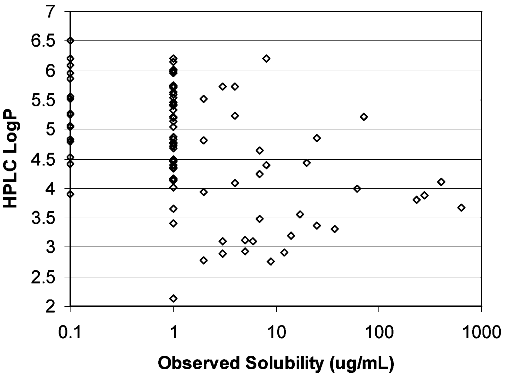


Figure 8. HPLC log *P* (pH = 6.5) compared to the observed solubility at pH = 6.5 for compounds from a single discovery program. Measurements from crystalline material. Note that most measured solubility values of either 0.1 or 1 μg/mL were reported as <0.1 μg/L or <1 μg/mL, respectfully.

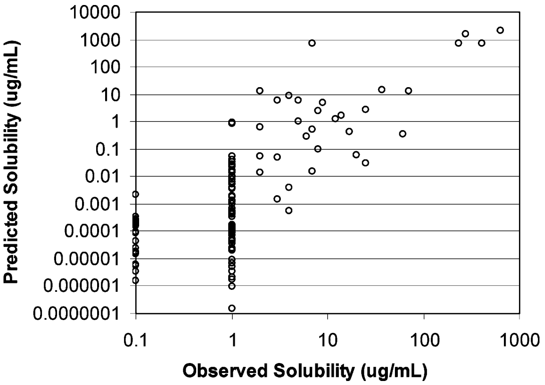


Figure 9. Predicted aqueous solubility at pH = 6.5 compared to the observed solubility at pH = 6.5 for compounds from a single discovery program. Measurements from crystalline material. Note that most measured values of either 0.1 or 1 μg/mL were reported as <0.1 μg/L or <1 μg/mL, respectfully.

The ability to interpret a predictive model is critical to its use in driving discovery projects. This is frequently one of the major criticisms of models for ADMET related properties. Figure 7 shows the descriptor impact on the solubility predictions for haloperidol (HCl salt) and glyburide at a pH of 6.5. Haloperidol and glyburide were predicted to have an aqueous solubility of 130 μg/mL and 24 μg/mL, respectively, at pH of 6.5. The plot is generated by multiplying each normalized feature value by its coefficient in the intrinsic solubility model. Also shown are the contributions from the crystal packing simulation and ionization at pH = 6.5. The plot shows that the difference in the predictions for the two compounds is largely driven by the difference in molecular weights. As discussed above, this feature likely encodes the energy required to create a cavity large enough to accommodate each solute. Note that the effect of crystal packing appears small in Figure 7 because the effect is modulated by the degree of ionization (eq 1). At a pH of 6.5 both glyburide and haloperidol are expected to be ionized.

The optimization of solubility in discovery projects is often driven by a calculated log *P* followed by a measured chromatographic log *P* and measured solubility. Typically this approach is quite reasonable, however occasionally this strategy does not lead to improved solubility. Figure 8 shows the relationship between the log *P* measured by an HPLC assay and the measured aqueous solubility for discovery compounds from a single chemotype. All the measurements were from crystalline material at pH = 6.5. The correlation is quite poor ($r^2 = 0.21$), indicating that crystal packing may substantially affect aqueous solubility. Figure 9 shows the correlation between the predicted solubility at pH = 6.5 and the measured solubility. The crystal packing term, X_{pack}, was calculated using simulated crystal structures based on the structure of a close analogue. The close analogue chosen was one from the same chemical series with a solved X-ray structure and was the closest to being “isographic” with the compound of interest. The correlation between the predicted and measured values is substantially improved ($R^2 = 0.56$) compared to the results for the measured log *P*. While far from ideal, these results imply a much greater utility in evaluating molecular hypotheses with respect to solubility.

Conclusion

We have assembled a computational model for the estimation of aqueous solubility at an arbitrary pH that explicitly accounts for the effects of intrinsic solubility, ionization, and crystal packing. While complicated in functional form, the model is readily interpretable by analyzing the underlying components in detail. In addition, a Monte Carlo error function is employed to provide a confidence interval for the estimate. This confidence interval gives users a better appreciation of the sensitivity of an estimate at a particular pH. This is particularly true when the pH of interest lies within 2 log units of the pK_a of the molecule and the estimation solubility changes rapidly in response to small changes in pH.

Another benefit of the approach employed here is that the method is likely to be more extensible than models that rely on high order machine learning algorithms. We believe this is an important advantage as undoubtedly the method will need further improvement to include the effect of low MW aggregation, the common ion effect, and an expansion of the underlying chemical space captured by the training data. In addition, there is much room for improvement in the current approach to encoding crystal packing. It is quite probable that different computational parameters, possibly including longer simulation times, would improve the quality, consistency, and reliability of the simulation. Our future work will also concentrate on the shift in the relationship between X_{pack} and melting point observed in Figure 2.

While an interesting approach, the crystal packing simulation reported here is computationally intensive relative to most aqueous solubility models, taking approximately 1 CPU hour per temperature on an Opteron Linux workstation. In addition, it requires a real or putative crystal form upon which to act. The use of a simulated crystal structure, while pragmatic, introduces a significant potential source of error into the model. It is difficult to quantify how similar a molecule must be to the chemical structure of an analogue with a solved crystal structure in order for the simulated structure to relevant. Strictly speaking, the structures need to be nearly isographical to have a high likelihood of adopting similar unit cells. One interesting possibility is the use of a simulation similar to the one discussed here to evaluate numerous simulated crystal structures that can be generated from the several crystal structure prediction programs now available. However, as the experimental data for the aqueous solubility of polymorphs leans toward only 2−10-fold differences in solubility, we believe a simulated structure represents an acceptable starting point in the absence of any of other information.

Supporting Information Available: Training and validation data for the high MW and low MW intrinsic solubility models and plot of calculated versus observed instrinsic solubility for the low MW model. This material is available free of charge via the Internet at http://pubs.acs.org.

MP070030+