ESOL: Estimating Aqueous Solubility Directly from Molecular Structure

John S. Delaney*

Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, United Kingdom

Received October 29, 2003

This paper describes a simple method for estimating the aqueous solubility (ESOL − *E*stimated *SOL*ubility) of a compound directly from its structure. The model was derived from a set of 2874 measured solubilities using linear regression against nine molecular properties. The most significant parameter was calculated $\log P_{octanol}$, followed by molecular weight, proportion of heavy atoms in aromatic systems, and number of rotatable bonds. The model performed consistently well across three validation sets, predicting solubilities within a factor of 5−8 of their measured values, and was competitive with the well-established "General Solubility Equation" for medicinal/agrochemical sized molecules.

INTRODUCTION

Aqueous solubility is one of the key physical properties of interest to a medicinal[1] or agrochemical[2] chemist. Solubility affects the uptake/distribution of biologically active compounds in living material and the environment, thus affecting their potential efficacy and marketability. Accurate equilibrium solubility determination is a time-consuming experiment, and it is useful to be able to assess solubility in the absence of a physical sample.

There have been many methods developed that predict solubility, either solely from molecular structure or using more easily obtained measurements. Group contribution methods[3,4] analogous to CLOGP[5] have been developed as have a variety of methods based on some form of nonlinear regression combined with topological parameters.[6−8] Jorgensen and Duffy have used Monte Carlo simulation with QSPR,[9] while Gasteiger[10] and Wegner[11] used other 3D descriptors. Approaches based on quantum mechanics calculations have also been tried.[12,13] Solvatochromic methods such as LFER[14] form a distinct class of prediction method, as does mobile order theory.[15]

A small group of methods have been founded on the observation that $\log P_{octanol}$ (logP) shows a strong correlation with aqueous solubility. The preeminent such method is probably the "General Solubility Equation" (GSE[16]), which has just two variables—logP and melting point (T_m). These parameters handle the partition between liquid compound and water (logP) and correct for the transition from solid to liquid (T_m). Octanol partition can be calculated with reasonable accuracy from a compound's structure,[17] but estimating melting point is far harder. Where a measured melting point is available, GSE becomes the method of choice, while other methods, based solely on structure, have to be used in situations where T_m is not available.

Two recent papers have discussed structure-only methods based on logP estimates.[18,19] Butina used an AI technique to produce a small number of rules and local models based on a proprietary logP estimate and additional structural terms. Cheng used a genetic algorithm to select variables for a linear model, with AlogP98[20] (a logP estimate similar to CLOGP) being the most significant contributor. The method described in this paper (named ESOL for *E*stimated *SOL*ubility) is in a similar vein, relying on CLOGP version 4.17[5] to provide a reasonably accurate logP estimate, which is then augmented by a small number of additional terms. If ESOL is distinctive, it is in terms of its relative simplicity versus its predictive performance. Only nine molecular descriptors (used in an earlier paper on bioavailability[2]) were initially considered using straightforward linear regression, with the final model having four parameters, just two more than the GSE. I have found that ESOL works particularly well on compounds of agrochemical interest, often out-performing the GSE in terms of average absolute error of prediction and the method is fast enough to be used on large numbers of "virtual" compounds such as putative compound libraries or potential vendor purchases.

METHOD

The molecules used to derive the model were initially described by 9 parameters calculated directly from their 2D connectivity. This initial set of parameters included the following:

1. clogP—calculated using Daylight CLOGP version 4.72[5] (fragment database version 17)
2. Molecular weight (MWT)
3. Rotatable bonds (RB)—calculated using an in-house program from SMILES. Daylight SMARTS[5] substructures define rotatable bonds.[21]
4. Aromatic proportion (AP)—calculated using an in-house program from SMILES. Uses the Daylight SMARTS definition of aromatic ([a]) to count "aromatic atoms". The proportion of heavy atoms in the molecule that are in an aromatic ring.[10]
5. Non-carbon proportion—calculated using an in-house program from SMILES. The proportion of heavy atoms in the molecule that are not carbon ([!#6] in SMARTS).
6. H-bond donor count—calculated using an in-house program from SMILES.[22]
7. H-bond acceptor count—calculated using an in-house program from SMILES.[22]

* Corresponding author e-mail: john.delaney@syngenta.com.

10.1021/ci034243x CCC: $27.50
Published on Web 03/13/2004

Table 1. Summary of Regression Statistics

no. of compds	R^2	SE	F-statistic	intercept	clogP	MWT	RB	AP
2874	0.72	0.97	1865	2.8	−53	−32	6.4	−10

8. Polar surface area—calculated using Peter Ertl's program from SMILES.[23]

Multiple linear regression (Microsoft Excel 2000) was performed on a training set of 2874 neutral compounds (12% of original set of compounds predicted to be charged at pH7[24] were excluded) with measured aqueous solubilities (log M/L at 25 °C) against the eight parameters. The significance of each parameter was assessed in terms of its absolute t-statistic. It was found that only the first four parameters (clogP, MWT, RB, and AP) made significant contributions to the model (absolute t-statistic > 2), and the other five parameters were discarded. This result was checked using stepwise multiple regression within the Cerius2[25] molecular modeling package with the same outcome. This is a crude way of selecting parameters for a model and is open to the charge that statistics related to model quality will be overoptimistic (there are 70 ways of selecting 4 parameters from a choice of 8). By way of justification, I would state that logP was always going to be included in any model (i.e. it should be treated as a "given"), and there were reasonable grounds for treating molecular weight the same way (see ref 26 where the GSE was augmented with a molecular weight term). This reduces the number of combinations down to no more than 35, and, combined with the large number of data points, the model seems to be statistically significant. A summary of the final regression statistics and the t-statistics for the intercept and the four parameters is given in Table 1.

All four parameters made significant ($P < 0.01$) contributions to the model. The final equation for solubility (S_w) in M/L was

$$\mathrm{Log}(S_w) = 0.16 - 0.63\ \mathrm{clogP} - 0.0062\ \mathrm{MWT} + 0.066\ \mathrm{RB} - 0.74\ \mathrm{AP}$$

The training set was made up from three sources, each with different properties. The "Small" set mainly consisted of low molecular weight organic compounds compiled from the literature by Abraham's group at University College London (particularly Joelle Le[14]) and provided to the author by Kei Enomoto. The "Medium" set was made up of pesticide products[27] (using the reported measured solubilities in the Pesticide Manual) of moderate molecular weight (200−300). The "Large" set had Syngenta proprietary compounds with equilibrium solubilities measured in-house, their molecular weight tending toward the heavier end of the scale (300−400). The overall property averages, with their standard deviations in brackets, for the three subsets are shown in Table 2.

The method was tested with a blind test set of 528 in-house compounds and two sets of literature compounds.[9,28]

RESULTS

The performance of the final equation was judged by the average absolute error (AAE) of prediction, standard error (SE), and correlation coefficient (R^2). These were compared with GSE results for 1305 compounds from the training set with measured melting points. The AAE results were further broken down by subset (Table 3).

Graphs of the estimates from the two methods for the 1305 compounds are shown (Figures 1 and 2).

The three subsets of the collection showed some variation in average absolute errors. The GSE performed much better on the "Small" set than on the other two sets, while ESOL worked about as well across all three sets. The GSE was originally derived from a set of molecules similar to our "small" set (reference compounds) which may explain why it did well here, while ESOL seemed to work better on pesticide/drug sized compounds, a noted weakness of the GSE.[29]

To validate the model, three sets of compounds were used. I obtained more in-house measured solubility results to create a blind test set with 528 compounds. ESOL produced an R^2 of 0.55, a standard error of 0.96, and an average absolute error of 0.83. This seemed consistent with the regression results obtained for the training set.

The model was also assessed on a small set of literature compounds to compare its performance against other published studies. This set of 21 compounds has been used extensively for solubility prediction method validation since its introduction by Yalkowsky.[28] I have used this set with Klopman's amended experimental solubility values,[4] and the results are summarized in Table 4 (all S_w values in log M/L).

Liu, Huuskonen, Wegner, Gasteiger, and Tetko all used some form of neural network as their preferred method for model production, while ESOL, Kuhne, and GSE used multiple linear regression. The way the molecules were described also varied with Wegner and Gasteiger favoring 3D descriptors, Liu, Huuskonen, Kuhne, and Tetko utilizing 2D/connectivity-based descriptors, and ESOL and GSE using whole molecule descriptors. Despite their relative simplicity, ESOL and GSE perform remarkably well, achieving comparable results to most of the other methods.

Finally, I ran ESOL against a set of 150 literature compounds,[9] which had also been studied using the GSE.[30] These molecules were split by the authors into a set of reference organic molecules (65 compounds, mean MWT = 114), a set of drug/drug-like molecules (64 compounds, mean MWT = 269), and a set of heterocycles/pesticides (21 compounds, mean MWT = 171). Taking the whole set, ESOL produced an AAE of 0.71 against GSE's AAE of 0.45 (Figures 3 and 4). However, if the drug/drug-like subset is

Table 2. Molecular Property Averages/Standard Deviations for Whole Training Set and "Small", "Medium", and "Large" Subsets

	measured solubility	clogP	molecular weight	rotatable bonds	aromatic proportion	non-carbon proportion	donor count	acceptor count	PSA	no. of compds
"Small"	−3.1 (2.1)	2.7 (2.0)	205 (103)	1.2 (2.2)	0.37 (0.34)	0.26 (0.18)	0.7 (1.1)	1.0 (1.3)	35 (36)	1144
"Medium"	−4.0 (2.0)	3.3 (1.8)	294 (89)	1.7 (1.9)	0.36 (0.23)	0.34 (0.13)	0.5 (0.8)	1.7 (1.1)	51 (27)	485
"Large"	−4.3 (1.3)	3.5 (1.4)	341 (71)	2.0 (1.7)	0.45 (0.16)	0.32 (0.08)	0.6 (0.8)	2.2 (1.1)	59 (23)	1245
all	−3.7 (1.8)	3.0 (1.8)	279 (108)	1.6 (2.0)	0.40 (0.26)	0.30 (0.13)	0.7 (0.9)	1.7 (1.3)	40 (31)	2874

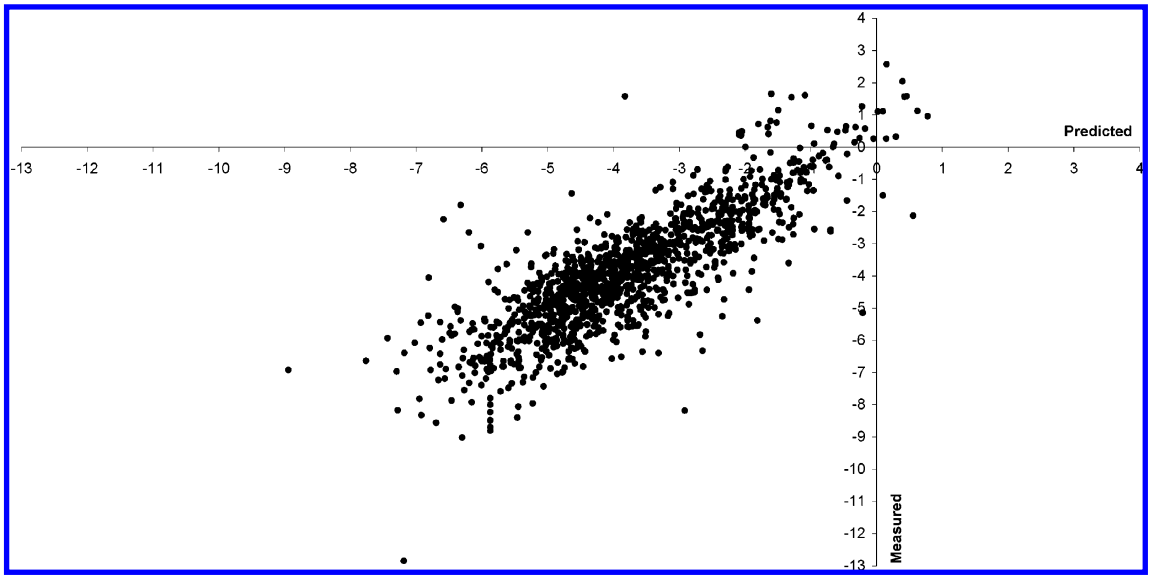

Figure 1. ESOL predicted solubilities for 1305 training compounds with measured melting points.

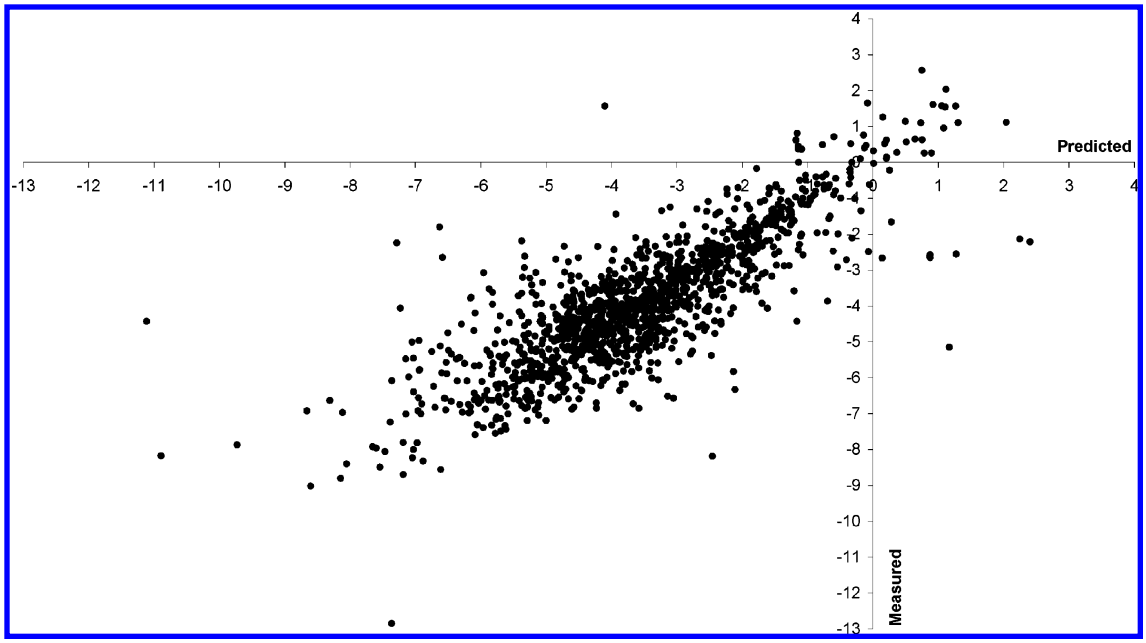

Figure 2. GSE predicted solubilities for 1305 training compounds with measured melting points.

Table 3. Comparison of ESOL and GSE Results Across Training Subsets

	R^2 (all)	SE (all)	AAE (all)	AAE ("Small")	AAE ("Medium")	AAE ("Large")
ESOL	0.69	1.01	0.75	0.75	0.81	0.71
GSE	0.67	1.05	0.81	0.47	0.90	0.93

considered, the gap between the AAEs is substantially reduced (ESOL AAE = 0.60, GSE AAE = 0.50). Once again, it seemed that ESOL became more competitive with GSE for larger molecules.

DISCUSSION

The aim of this work was to produce a robust alternative to solubility estimation by GSE where the melting point of the compound was unknown. GSE sets a very high standard for small, organic reference molecules, and ESOL struggles to match it in terms of average absolute error. A point worth noting with the "Small" molecule set is that a substantial proportion of the compounds has melting points (T_m) below 25 °C, while the "Medium" and "Large" sets have very few compounds such as this. If these low melting point compounds are excluded (i.e. compounds where T_m makes no contribution to the GSE) the GSE's average absolute error jumps from 0.51 to 0.65. The same effect can be observed in the 150-member literature set where exclusion of compounds with low melting points raises the GSE's average absolute error from 0.45 to 0.54. Excluding the same compounds from ESOL has little effect on its average absolute error. The results for larger compounds suggest that ESOL and GSE tend to converge in terms of their absolute errors, with ESOL maintaining a pretty consistent value across a wide range of molecular weights (150−500)—predicted solubilities within a factor of 5−8 of their measured values.

The literature methods mentioned earlier tend to fall into a number of distinct types depending on the combination of molecular descriptors used (2D/connectivity,[3,6−8] 3D,[10,11] or

Table 4. Experimental against Assorted Predicted Solubilities for a Set of 21 Compounds[a]

common name	CAS no.	experimental values[4,28]	ESOL	Liu[7]	Huuskonen[6]	Kuhne[3]	Wegner[11]	Gasteiger 10	Tetko[8]	GSE
antipyrine	60−80−0	−0.56	−1.79	−1.41	−1.29	−1.9	−1.74	−1.31	−0.89	−0.56
aspirin	50−78−2	−1.72	−1.98	−2.1	−1.69	−1.93	−1.81	−1.87	−1.81	−1.62
atrazine	1912−24−9	−3.85	−2.95	−1.51	−3.51	−3.95	−2.82	−3.83	−3.7	−3.5
benzocaine	94−09−7	−2.32	−2.38	−1.45	−1.79		−2.05	−2.19	−1.63	−2.06
chlordane	57−74−9	−6.86	−5.88	−7.32	−7.29	−6.51	−6.47	−7.66	−7.23	−5.3
chlorpyrifos	2921−88−2	−5.49	−4.89	−4.5	−5.61	−3.75	−6.41	−4.79	−5.31	−4.17
DDT	50−29−3	−8.08	−6.47	−7.93	−7.67	−7.75	−6.85	−7.86	−7.59	−7.1
diazepam	439−14−5	−3.76	−4.03	−4.08	−4.05	−4.51	−4.14	−4.81	−4.37	−3.66
diazinon	33−41−5	−3.64	−3.95	−3.56	−4.01	−4.98	−4.18	−2.66	−3.43	
diuron	330−54−1	−3.8	−3.30	−2.85	−2.86	−3.38	−3.98	−5.04	−4.91	−3.52
lindane	58−89−9	−4.64	−3.97	−4.91	−4.71	−5.08	−3.98	−5.04	−4.91	−4.13
malathion	121−75−5	−3.37	−3.40	−2.52	−3.24	−3.48	−2.96	−2.79	−3.73	−2.2
nitrofurantoin	67−20−9	−3.47	−1.36	−2.89	−3.42	−2.62	−2.82	−2.52	−3.09	
parathion	56−38−2	−4.66	−3.93	−3.64	−4.13	−4.59	−4.06	−3.66	−4.31	−2.97
2,2′,4,5,5′-PCB	37680−73−2	−7.89	−6.61	−7.55	−7.21	−7.47	−7.66	−7.85	−7.57	−6.99
phenobarbital	50−06−6	−2.34	−2.33	−2.5	−2.97	−2.41	−2.36	−2.8	−2.89	−2.38
phenolphthalein	77−09−8	−2.9	−3.91	−4.16	−3.99	−4.61	−4.64	−4.62	−4.31	−4.52
phenytoin	57−41−0	−3.99	−3.09	−3.09	−3.4	−5.25	−2.9	−3.18	−3.52	−4.28
prostaglandin E2	363−24−6	−2.47	−2.87	−3.8	−3.29		−3.98	−3.07	−3.52	−1.93
testosterone	58−22−0	−4.09	−3.64	−4.49	−3.98	−4.62	−4.27	−4.52	−4.13	−4.02
theophylline	58−55−9	−1.39	−1.52	−0.73	−1.71	0.54	−1.21	−1.27	−0.69	−1.91
		R^2	0.85	0.81	0.93	0.78	0.82	0.85	0.91	0.86
		SE	0.78	0.87	0.55	0.97	0.84	0.77	0.61	0.79
		AAE	0.69	0.72	0.44	0.74	0.64	0.63	0.48	0.65

[a] Blank denotes missing predicted value.

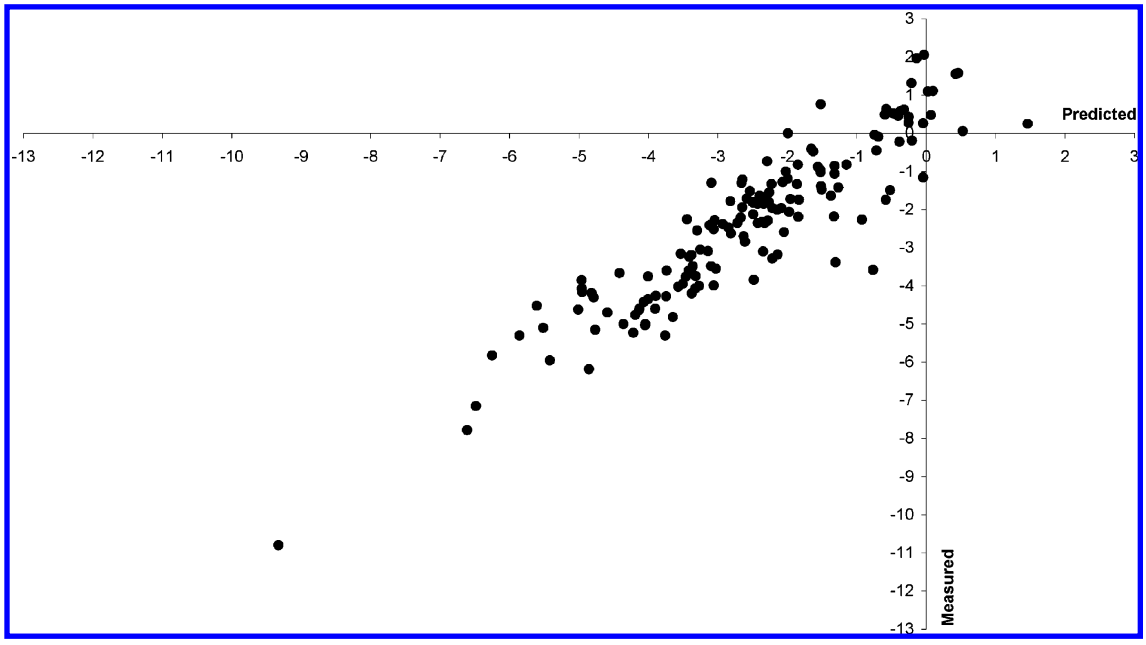

Figure 3. ESOL predicted solubilities for 150 compounds from Jorgensen and Duffy.[9]

whole molecule[16]) and the method used to derive the model (broadly linear regression or some form of nonlinear, AI technique particularly neural networks). There does not seem to be a clearly superior type of molecular descriptor, as the three types seem to produce broadly comparable results. Speed of computation and immunity from conformational uncertainty might favor 2D and whole molecule descriptors for use with large libraries, although the GSE's need for a melting point value rules it out for this type of application. The model building techniques offer a starker choice. It seems clear from Table 4 that nonlinear methods produce better models when judged by summary statistics. This must be weighed against increased model complexity and the rather gnomic nature of neural nets.

In comparison to many of the methods in the literature for predicting solubility directly from structure, ESOL is remarkably simple. It was not possible to directly compare its performance against the Cheng model,[19] but the reported statistics (AAE = 0.77 on a validation set of drugs) indicated that they were similar. Butina[18] does include estimates for 10 (out of 21) compounds in one of the literature data sets,[28] and a like-for-like comparison with ESOL results shows ESOL performing slightly better. Starting from a relatively accurate (CLOGP version 4.17[5]) logP estimate clearly helps enormously, and the molecular weight factor has been noted before.[26] The aromatic proportion (AP) parameter seems to be key to the method's success, and while it has been mentioned in relevant literature,[10] it assumes a much higher

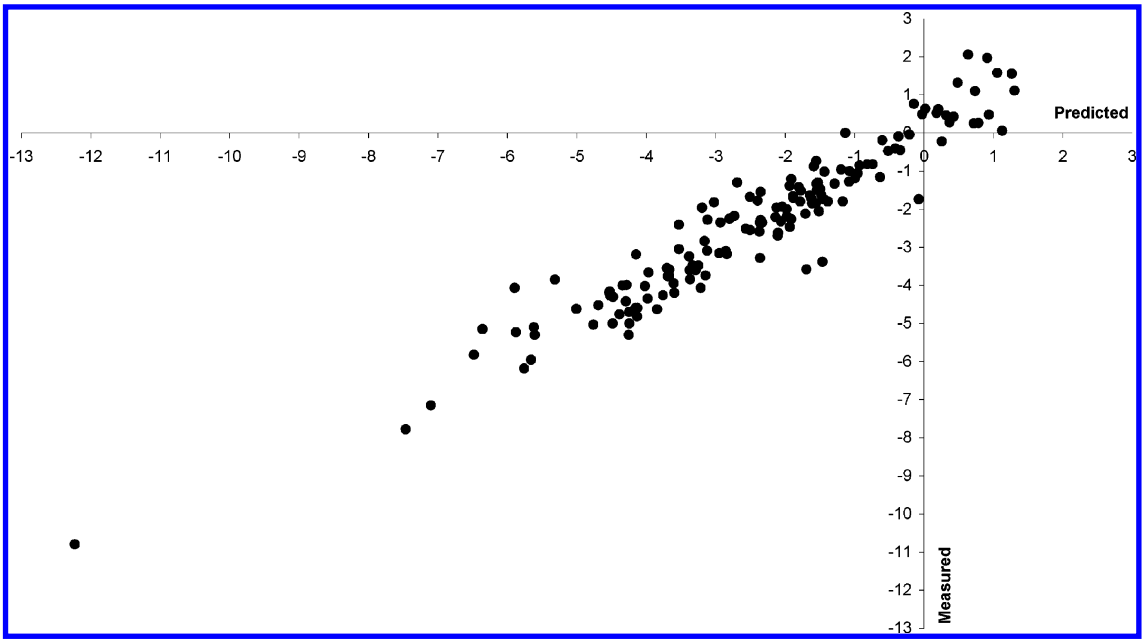

Figure 4. GSE predicted solubilities for 150 compounds from Jorgensen and Duffy.[9]

profile here. The most obvious interpretation of its role seems to be as a measure of flexibility, although it might also have a weak relationship to T_m. The rotatable bond count (RB) seems to have an obvious relationship to the entropic component of solubility. Another interesting point is the complete lack of explicit H-bonding variables in the final model. They did not contribute in a significant way to the regression equation and other attempts to introduce them, such as multiplying donor and acceptor counts, failed to improve matters. The counts seem to be too crude to capture the solute's interactions either within the crystal or with the solvent. An area that might offer scope for improvement could be to include functional group specific correction terms, something that might be addressed in a future study—cursory analysis has not revealed any systematic trends in errors.

ESOL seems to be a viable alternative to GSE for predicting the solubility of pesticide/drug-like molecules. The fact that it works so well is something of a surprise and begs the question why. GSE divides the solubility prediction problem into a liquid−liquid partition term and a solid−liquid state change term. The three non-logP terms in the ESOL equation could be acting in either or both of these terms.

One possibility is that the additional parameters are acting to improve the CLOGP estimate of logP, a plausible explanation given the importance of the logP term in both ESOL and GSE. I have tested this proposition using a set of 489 in-house compounds with measured logPs, solubilities, and melting points. Using the measured logP values improves both models, the effect being slightly more marked for ESOL (AAE goes from 0.68 to 0.50 moving from CLOGP to measured logP) than GSE (AAE goes from 0.87 to 0.73). However, the difference in improvement is not large, indicating that logP estimate enhancement is not the main effect of the additional parameters.

Another possibility is that the combination of MWT, RB, and AP are modeling T_m, which would mean that the ESOL equation essentially reduces to the GSE. Although I have not been able to find a decent linear correlation between these parameters and T_m, there may be some more subtle relationship at work. Flexible compounds often have relatively low melting points due to poor crystal packing while highly aromatic compounds can form dense crystals with high melting points. Larger compounds also tend to have higher melting points. These observations suggest that something about T_m is being captured by these variables, perhaps warranting further investigation.

The standard GSE makes an explicit conjecture that may offer another way forward—entropy of melting (ΔS_m) is assumed to be constant for all molecules (Walden's rule[31]). This means that $T_m\Delta S_m$ (which models solid−liquid state change in the GSE) reduces to T_m. This assumption is known to be good for small, rigid molecules, less so for larger, more flexible molecules. Dannenfelser and Yalkowsky modeled ΔS_m[32] using molecular symmetry and flexibility as parameters. Since most drugs and agrochemicals lack any symmetry, their "molecular rotational symmetry number" would equal 1 for vast majority of compounds in this study. In this case the model reduces to a measure of molecular flexibility. This seems consistent with the importance of AP and RB in the ESOL equation, as both have a bearing on flexibility.

My conclusion is that the non-logP terms are probably providing an enhanced estimate of $T_m\Delta S_m$ rather than accurately modeling the individual terms. I lacked the data to model ΔS_m, and, from the limited work in this study, it appears that modeling T_m in isolation is difficult.

In summary, ESOL provides a fast and robust method for estimating the solubility of drugs and agrochemicals without recourse to physical measurements.

ACKNOWLEDGMENT

I thank Eric Clarke for inspiring and encouraging this project, Kei Enomoto for supplying solubility data, and Carine Delaney for helpful comments on the style and content of the paper.

Supporting Information Available: A comma-separated value file containing a list of the members of the "Small"

training data set, their measured and ESOL-predicted solubilities, and their structures as SMILES.[5] This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Clarke, E. D.; Delaney, J. S. Physical and Molecular Properties of Agrochemicals: An Analysis of Screen Inputs, Hits, Leads and Products. *Chimia* **2003**, *57*, 731−734.

(3) Kuhne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(4) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Compounds by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.

(5) Daylight Chemical Information Systems, Santa Fe, New Mexico, U.S.A. (www.daylight.com).

(6) 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.

(7) Liu, R.; So, S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.

(8) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(9) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

(10) 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.

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

(12) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arena, A.; Giralt, F. A Fuzzy ARTMAP Based on Quantitative Structure−Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177−1207.

(13) Klampt, 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.

(14) 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.

(15) Ruelle, P.; Kesselring, U. W. The Hydrophobic Effect. 2. Relative Importance of the Hydrophobic Effect on the Solubility of Hydrophobes and Pharmaceuticals in H-bonded Solvents. *J. Pharm. Sci.* **1998**, *87*, 998−1014.

(16) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility 1: Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*(2), 234−252.

(17) Leo, A. J. Calculating log Poct from Structures. *Chem. Rev.* **1993**, 1281−1306.

(18) Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837−841.

(19) Cheng, A.; Merz, K. M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure−Property Relationships. *J. Med. Chem.* **2003**, *46*, 3572−3580.

(20) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(21) The following SMARTS patterns are defined as rotatable: [!X1]-,=-[$([C.;X4])]-&!@[$([C.;X4])]-[!X1], [!X1]:c-&!@[$([C.;X4])]-[!X1], [!X1]-,=C-&!@[$([N.;X4])]-[!X1], [!X1]-[$([C.;X4])]-&!@[$([N.;X3])]-[!X1], [!X1]-[$([C.;X4])]-&!@[$([O.;X2])]-[!X1].

(22) The following SMARTS patterns are defined as H-bond donors: [NH],[NH2],[NH3],[OH],[nH]. The following SMARTS patterns are defined as H-bond acceptors: [O;X1], [n;X2].

(23) Ertl, P.; Rohde, B.; Selzer, P.; Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(24) Kenny, P. Handling Heterocyclic Tautomerism; Daylight User Group Meeting, Santa Fe, New Mexico, February 1999; talk.

(25) Cerius2, Accelrys Inc., San Diego, USA. http://www.accelrys.com.

(26) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Ocatanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.

(27) *Pesticide Manual*, 12th ed.; Tomlin, C., Ed.; British Crop Protection Council: 2000; ISBN 1901396126.

(28) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility, Methods of Estimation for Organic Compounds*; Marcel Dekker: 1992; Chapter 4, pp 128−148.

(29) Morris, J. J.; Bruneau, P. P. Prediction of Physicochemical Properties. Virtual Screening for Bioactive Molecules. In *Virtual Screening for Bioactive Molecules*; Bohm, Schneider, Eds.; Wiley VCH: Chapter 3, pp 33−5, ISBN 3527301534.

(30) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(31) Walden, P. *Elektrochem.* **1908**, *14*, 713.

(32) Dannenfelser, R. M.; Yalkowsky, S. H. Estimation of Entropies of Fusion of Organic Compounds. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 108−111.

CI034243X