EASOL: Development of Data-driven Prediction Model for Estimating Aqueous SOLubility of Molecular Compound

Sanchita Mondal
School of Medical science and Technology
IIT Kharagpur
sanchita.mondal@kgpian.iitkgp.ac.in

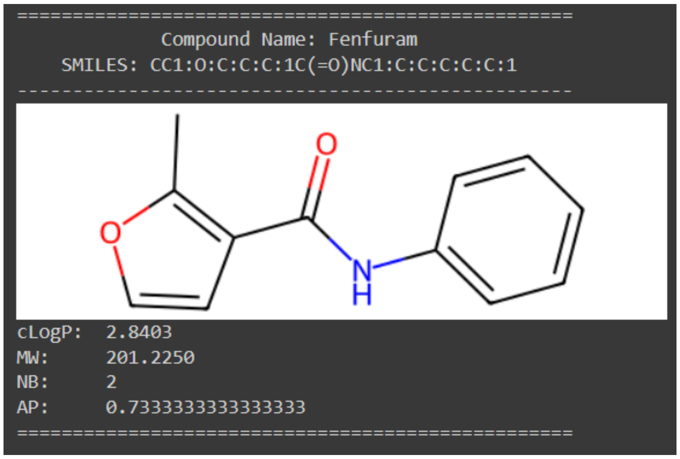
***Abstract*— This study presents a regression technique for determining a compound's aqueous solubility (EASOL-Estimated Aqueous SOLubility) from its structure and feature set. Using linear regression against six molecular characteristics, a set of 1128 observed solubilities was used to create the model. The most important parameter was the measured log solubility in mols per litre, which was followed by the number of rotatable bonds, number of rings, molecular weight, and the polar surface area. The model well performed for medicinal/agrochemical sized compounds, performing consistently well across the datasets and predicting solubilities within a factor of 5-8 of their measured values.**

I. INTRODUCTION

Aqueous solubility is one of the most important physical characteristics for a pharmaceutical or agrochemical chemist. The potential efficacy and marketability of biologically active substances are influenced by their solubility, which also influences how these compounds are absorbed and distributed in living things and the environment. Determining the equilibrium solubility accurately takes time, so being able to determine solubility without a physical sample is helpful.

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 non linear regression combined with topological parameters[6],[7],[8]. A small group of methods have been founded on the observation that $Log_{Poctanol}(LogP)$ shows a strong correlation with aqueous solubility. The preeminent such method is probably the "General Solubility Equation" (GSE[9]), which has just two variabless logP and melting point (T_m). The separameters 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[10],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 insituations where T_m is not available. Two recent papers have discussed structure-only methods based on logP estimates[11],[12]. The method described in this paper (named ESOL for Estimated SOLubility) 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 bio availability[13]) were initially considered using straight forward 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 outperforming 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.



```
==================================================
          Compound Name: Fenfuram
    SMILES: CC1:O:C:C:C:1C(=O)NC1:C:C:C:C:C:1
--------------------------------------------------
```

```
cLogP:   2.8403
MW:      201.2250
NB:      2
AP:      0.7333333333333333
==================================================
```

Fig. 1. Extracted molecular representations and properties using the RDKit library to parse SMILES strings.The rdkit.Chem module handles the parsing of SMILES strings and the calculation of each of the features.

II. METHOD

ESOL is a small dataset consisting of water solubility data for 1128 compounds[1]. The dataset has been used to train models that estimate solubility directly from the extracted features from the chemical structures (as encoded in SMILES strings)[2]. Note that these structures don't include 3D coordinates, since solubility is a property of a molecule and not of its particular conformers.

At first, six characteristics that were derived directly from the 2D connectivity of the molecules were employed to describe them in the model. The following were part of this first set of parameters:

1) Minimum Degree
2) Molecular Weight
3) Number of H-Bond Donors
4) Number of Rings
5) Number of Rotatable Bonds
6) Polar Surface Area
7) measured log solubility in mols per litre

Multiple linear regression was performed on a data set of 1128 neutral compounds with measured aqueous solubilities (log M/L at 25 °C) against the six parameters. The significance of each parameter was assessed in terms of its absolute t-statistic. It was found that only the three parameters (Molecular Weight, Number of Rotatable Bonds, and Polar Surface Area) made significant contributions to the model. This result was checked using stepwise multiple regression. 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 over optimistic (there are 70 ways of selecting parameters from a choice of 6). 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. 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.

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

$$\begin{split} Log(S_w) = -0.499 * MinimumDegree + -0.014* \\ MolecularWeight + 0.073 * Numberof \\ H - BondDonors + -0.413 * Number \\ ofRings + -0.143 * NumberofRotatable \\ Bonds + 0.032 * PolarSurfaceArea + -0.01 \end{split} \tag{1}$$

III. RESULT

The performance of the final equation was judged by correlation coefficient (R2). The score is 68%.

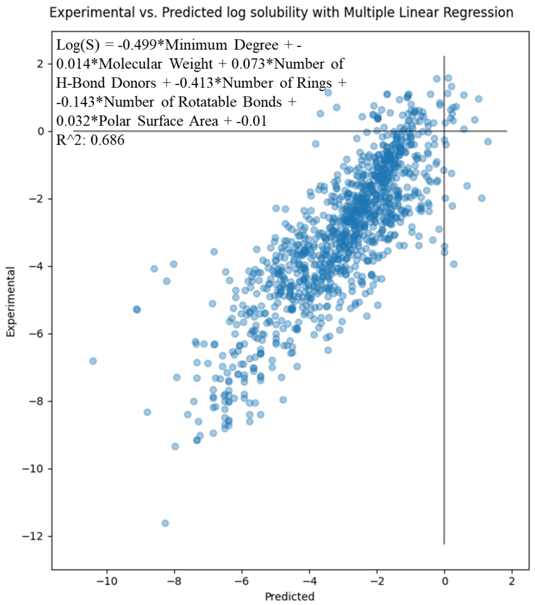


Fig. 2. EASOL predicted solubilities for 1128 training compounds with the six features

References

[1] Delaney, John S. "ESOL: estimating aqueous solubility directly from molecular structure." Journal of chemical information and computer sciences 44.3 (2004): 1000-1005.

[2] Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." Advances in neural information processing systems 28 (2015).

[3] Kuhne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schuurmann, G.Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. Chemosphere1995,30, 2061-2077.(4)

[4] Wang, S., G. Klopman, and D. M. Balthasar. "Estimation of Aqueous Solubility of Organic Compounds by the Group Contribution Approach. Application to the Study of Biodegradation." J. Chem. Inf. Comput. Sci. (1992).

[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 ofOrganic Compounds Based on Molecular Topology.J. Chem. Inf.Comput. Sci.2000,40, 773-777.

[7] Liu, R.; So, S. Development of Quantitative Structure-PropertyRelationship 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. Estimationof Aqueous Solubility of Chemical Compounds Using E-State Indices.J. Chem. Inf. Comput. Sci.2001,41, 1488-1493.

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

[10] Leo, A. J. Calculating log Poct from Structures. Chem. ReV. 1993.

[11] Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. J. Chem. Inf. Comput. Sci. 2003, 43, 837-841.

[12] 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, 3572-35.

[13] 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.

[14]