Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules

James F Blake

A few major advances have occurred in the area of physicochemical modeling of organic compounds during the past several years, spurred on by changes in the pharmaceutical industry. Recent advances include the ability to categorize and screen the overall physicochemical properties of potential drug candidates based entirely on their molecular structures and the ability to model the components that contribute to the oral absorption characteristics of potential drug candidates.

Addresses
Pfizer Inc., Eastern Point Road, Groton, CT 06340, USA;
e-mail: blakejf@pfizer.com

Current Opinion in Biotechnology 2000, **11**:104–107

0958-1669/00/$ – see front matter

Abbreviations

NPSA	non-polar surface area
***P*c**	permeability coefficient
PSA	polar surface area

Introduction

It has been estimated that roughly 10% of the compounds that enter development eventually become marketed drugs and 40% of compounds fail due to poor pharmacokinetic properties [1]. The ability to predict so called ADME (absorption, distribution, metabolism, and excretion) properties from molecular structure would have a tremendous impact on the drug discovery process both in terms of cost and the amount of time required to bring a new compound to market. This review covers computational approaches to the prediction of various physicochemical properties of complex organic molecules from their molecular structures without the need of any experimentally derived parameters. The environment for the application of these computational approaches is the drug discovery and preclinical drug development settings. Specifically, we are interested in the prediction of physicochemical properties that play a critical role in determining the oral absorption characteristics of therapeutic agents, namely, aqueous thermodynamic solubility and permeability across biological membranes. Dressman *et al.* [2] have related this dependence to the absorption potential (AP) of a compound via:

$$\mathrm{AP} = \log\left(P \cdot F_{\mathrm{non}} \cdot S_{\mathrm{o}} \cdot V_{\mathrm{L}} \cdot X_{\mathrm{o}}^{-1}\right) \quad (1)$$

Where P is the octanol–water partition coefficient, F_{non} is the fraction in nonionized form at pH 6.5, S_{o} is the intrinsic solubility, V_{L} is the volume of the lumenal contents, and X_{o} is the dose administered. Johnson *et al.* [3] formulated the maximum absorbable dose (MAD) of a compound as follows:

$$\mathrm{MAD} = K_{\mathrm{a}} \cdot S \cdot SIWV \cdot SITT \quad (2)$$

Where K_{a} is related to rat intestinal perfusion rate, S is the aqueous solubility at pH 6.5, and *SIWV* and *SITT* are the small intestine (SI) water volume and residency time, respectively. Both equations (1) and (2) describe the quantity of drug that could be absorbed under ideal conditions. The fundamental importance of log*P* (as a measure of lipophilicity/hydrophobicity) in the distribution of compounds in various biological systems and a number of computational methods for the estimation of log*P* from the molecular structure have been surveyed by Buchwald and Bodor [4] and will not be discussed herein. From equation (2), it is clear that maximum absorption occurs for highly soluble compounds with high permeability characteristics. It is also clear that in order to provide guidance to medicinal chemists, one must address the relative contributions of both solubility and permeability in absorption modeling. Additionally, one must be able to make reasonably accurate predictions for complex molecules before they are synthesized in order to reduce discovery time, which excludes the use of any experimentally derived parameters. Curatolo [5•] has recently reviewed the relationship between the physical chemical parameters of drug candidates and how they affect oral delivery. Readers interested in the complex relationships among the various ADME parameters and how they influence the overall human pharmacokinetic characteristics are referred to a review by Obach *et al.* [6]. The ultimate goal of modeling studies is the ability to predict the oral absorption properties of a compound based entirely on its molecular structure. This is a very challenging problem and the ability to predict the individual components of equation (2) would also be of great value.

Prediction of oral absorption

We begin our review with computational methods that address the overall oral absorption characteristics of potential drug candidates. The goal is to develop models and rules that will allow medicinal chemists to design compounds, or combinatorial libraries, that are heavily biased in favor of good oral absorption characteristics. In order to accomplish this, one must first define, or categorize, compounds that possess the desired properties. To this end, Lipinski *et al.* [7] were the first to conduct a systematic comparison of the physicochemical properties of compounds that are orally absorbed versus those compounds not expected to be well absorbed. The latter set were chosen to increase the size of the data set for analysis (i.e. there are not very many data points on poorly absorbed compounds). From a survey of compounds that entered Phase II clinical studies, Lipinski *et al.* found that poor absorption or permeability was probable for a compound when the molecular weight was above 500 amu (atomic mass unit), the calculated log*P* (ClogP; [8]) was

above five, and there are more than five hydrogen bond donors (sum of OH and NH groups), or ten hydrogen bond acceptors (sum of N and O atoms) present. If any two of these properties or features are present, the compound is likely to experience poor oral absorption characteristics. This is the so-called 'rule of 5'. The terms it employs are straightforward and are easily implemented by the practicing medicinal chemist. This work is important because it demonstrates that simple rules can be employed to categorize the properties of more favored compounds, and has proved very useful in the design and selection of compounds from virtual libraries. Of the top 100 best selling drugs in 1998 (74 are administered orally), only four fell outside the 'rule of 5' (CA Lipinski, personal communication). It should be mentioned that the 'rule of 5' addresses passive absorption and does not consider whether a compound might be involved in active transport mechanisms [9]. Clarithromycin is an example of a compound that has been shown to be actively transported but also violates the 'rule of 5'.

The 'rule of 5' provides an alert for potential absorption or permeation difficulties, but does not provide a quantitative estimate of a compound's absorption profile. The first true quantitative structure-property relationship (QSPR) study on a large (86 compounds) and diverse set of drug and 'drug-like' compounds was conducted by Wessel and coworkers [10••]. In this study, the measured human intestinal absorption (HIA) was correlated with six descriptors computed from the three-dimensional molecular structure (three charged partial surface area [CPSA] descriptors [11] and three topological and geometric descriptors). The CPSA descriptors are related to the physicochemical properties of molecules and have been successfully used in a variety of property prediction applications [10••,11–15]. In their study, a nonlinear model was used to fit %HIA with root-mean-square errors of 9.4% for the training set and 16.0% for an external prediction set. Given the diversity of the molecules under study and the complex nature of oral absorption, these results are very encouraging. As mentioned above, absorption depends on both the permeability and solubility of a compound; focused studies on each of these properties are described in the sections that follow.

Prediction of permeability

As it is difficult to measure intestinal permeability *in vivo* in humans, a number of models have been developed to provide a measure of the local absorption rate. The effectiveness of using Caco-2 monolayers [16–18], which are derived from a human cell line, in the modeling of human intestinal permeability has been reviewed recently by Lennernäs [19]. From an experimental standpoint, Pade and Stavchansky [20] have demonstrated that permeability coefficients derived from Caco-2 cells coupled with solubility measurements are reasonable predictors of the fraction of drug absorbed. Accordingly, there have been a number of recent efforts aimed at predicting the apparent permeability coefficient (P_c) from Caco-2 membrane permeability studies. While log*P* has been employed in a number of studies involving biological membranes (as a measure of lipophilicity/hydrophobicity of a compound) [21], another popular quantity known as the polar surface area (PSA) of a molecule has been used in a variety of applications [22,23••,24–26]. The PSA is generally defined as the sum of the van der Waals (or solvent-accessible) surface areas of oxygen and nitrogen atoms including attached hydrogens [22,23••]. The PSA of a compound can be related to its hydrogen bond accepting and donating ability.

van de Waterbeemd *et al.* [22], Palm *et al.* [23••], and Krarup *et al.* [24] have all incorporated PSA calculations in various forms to correlate the P_c (or logP_c) from Caco-2 studies with the PSA of a molecule, after screening a number of possible descriptors. In each of these studies, good correlations were generally obtained when the PSA was used alone or in conjunction with a limited number of descriptors. For example, Palm *et al.* [23••] demonstrated that for a series of nine β-receptor antagonists, the Caco-2 logP_c were highly correlated (cross-validation correlation coefficient q, $q^2 = 0.98$) to the dynamic PSA (i.e. multiple conformations of the compounds were averaged in the PSA calculations). Krarup *et al.* [24] also used dynamic averaging (via molecular dynamics simulations) of the PSA defined by the solvent-accessible surface to predict Caco-2 permeabilities for a series of six β-receptor antagonists and five ester prodrugs (correlation coefficient r, $r^2 = 0.98$). Both groups of researchers have noted that a balance between the nonpolar surface area (NPSA) and PSA may be important in permeability modeling. The incorporation of contributions from the NPSA was investigated by Stenberg *et al.* [26] for a series of 19 oligopeptide derivatives using Boltzmann weighted contributions from the PSA and NPSA fit to a sigmoidal function. For this system, the equation developed from the surface area terms outperformed similar equations that employed experimentally derived parameters, such as octanol–water (*P*), heptane–water, iso-octane–water, and heptane–ethylene glycol partition coefficients. For all of these studies, increasing PSA diminished membrane permeability. No distinction was made between hydrogen bond donating or accepting PSA contributions, which may warrant further studies.

In all of these cases, however, relatively small (~20 compound) data sets of closely related structural analogs were used. As is the case with all regression models, the predictive capabilities for compounds outside the training sets is not known, and should be investigated further to look for more generally applicable models. Although predicting P_c is important, Palm has suggested that PSA calculations could be used as a screening tool to weed out compounds with poor absorption properties, in a similar way to the 'rule of 5' [27]. From a study of 20 relatively diverse compounds, they concluded that structures with a PSA ≥ 140 Å^2 would be poorly absorbed (< 10% fractional absorption), whereas those compounds with a PSA ≤ 60 Å^2 would be well

absorbed (> 90%). This was recently investigated further by Clark [25], wherein the data set from Wessel *et al.* [13] (74 drug and drug-like compounds) was used to demonstrate that compounds with a PSA ≥ 140 Å^2 were indeed found to be poorly absorbed. The results of this study compare favorably with the predictions from 'rule of 5' alerts, and offers chemists an additional screening tool.

Prediction of aqueous solubility

Though aqueous solubility has been extensively studied, computational methods for the estimation of this highly important property are just beginning to demonstrate predictive capabilities for complex molecules. The difficulty lies in the fact that aqueous solubility is not an intrinsic property of the molecular structure *per se*, as the crystallographic environment of the solid must be considered as well. Aqueous solubility can be greatly affected by crystal polymorphism. For a recent review of the factors involved in crystal packing effects, see Dunitz and Gavezzotti [28]. We will restrict our discussion to those methods that have been applied to the prediction of aqueous solubility of relatively complex molecules that exist as solids at room temperature, without the need to incorporate experimentally derived quantities, such as the melting point [29•]. For smaller data sets with relatively simple molecules, a number of researchers have shown that reasonably accurate predictions of logS_w (log of the aqueous solubility expressed as mol dm^{-3}) are possible [12,15,30,31]. For a series of hydrocarbons and halogenated hydrocarbons, Huibers and Katrizky [31] were able to predict logS_w with r = 0.98, and standard error s = 0.4 from a three term equation based on volume, surface area, and topological descriptors. Similarly, Mitchell and Jurs [12] determined a nine-parameter model based on topological, CPSA, and electronic descriptors with an root-mean-square error of 0.394 log units for a collection of relatively simple organic compounds. More recent publications have looked at larger and more drug-like data sets.

Huuskonen *et al.* [32] assembled a data set of 211 drugs and related analogs for their solubility prediction experiments. By screening 101 descriptors, they were able to build an artificial neural network (ANN) model based on electrotopological state indices (E-state) [33]. Using a training set of 160 compounds and a test set of 51 compounds, they found 14 input parameters gave r^2 = 0.86 with s = 0.53 for logS_w of the test set, and r^2 = 0.9 with s = 0.46 for the training set. Although this result is encouraging, care must be taken when using any topological indices in predictive studies because compounds or functional groups that are not well represented in the training set can lead to poorly predictive models for compounds outside the training set. A more general approach would be based on properties computed from the structures, such as the solvatochromic parameters of Abraham *et al.* In their most recent study, Abraham and Le [34••] determined a seven parameter equation with r^2 = 0.92 and s = 0.56 for a 659 compound training set. From the equation, Abraham finds that increasing hydrogen bonding ability always leads to an increase in solubility, as does the polarizability, whereas an increase in volume leads to decreasing solubility. Interestingly, Abraham and Le concludes that for simple systems, standard errors as low as 0.3 log units are attainable, whereas more complicated drug-like data sets will have errors of ~0.5 log units.

Conclusions

Over the past several years, there has been a tremendous shift toward emphasis on optimization of ADME properties early in the life of drug discovery programs. This shift has been fueled to a large extent by an increase in the number of compounds with poor solubility and absorption properties entering development [7]. Research to date has provided a good conceptual framework for the prediction of properties of interest, namely, solubility and permeability, in connection with the overall goal of predicting human absorption [7,10••,23••,34••]. Future research will probably focus on developing models with data sets that are larger and built around more diverse collections of compounds with a wide range of chemical functionalitites. Two strategies are likely to emerge in the area of physicochemical property prediction: those seeking to develop general rules in order to screen large numbers of compounds, and those attempting to provide increasing levels of accuracy for more diverse compounds.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Prentis RA, Lis Y, Walker SR: **Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985).** *Br J Clin Pharmacol* 1988, **25**:387-396.

2. Dressman JB, Amidon GF, Fleisher D: **Absorption potential: estimating the fraction absorbed for orally administered compounds.** *J Pharm Sci* 1985, **74**:588-589.

3. Johnson K, Swindell A: **Guidance in the setting of drug particle size specifications to minimize variability in absorption.** *Pharm Res* 1996, **13**:1795-1798.

4. Buchwald P, Bodor N: **Octanol–water partition: searching for predictive models.** *Curr Med Chem* 1998, **5**:353-380.

5. • Curatolo W: **Physical chemical properties of oral drug candidates in the discovery and exploratory development settings.** *Pharm Sci Tech Today* 1998, **1**:387-393.

A review of the link between parameters such as solubility, permeability, and the dose. Many of the challenges that are faced in developing oral formulations are discussed, including guidelines for acceptable solubility and permeability values for drug candidates.

6. Obach RS, Baxter JG, Liston TE, Silber MB, Jones BC, MacIntyre F, Rance DJ: **The prediction of human pharmacokinetic parameters from preclinical and *in vitro* metabolism data.** *J Pharmacol Exp Ther* 1997, **283**:46-58.

7. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 1997, **23**:3-25.

8. Anon: *ClogP. Daylight Chemical Information Software.* Mission Viejo, CA: Daylight Chemical Information Inc.

9. Doppenschmitt S, Langguth-Spahn H, Regärdh CG, Langguth P: **Role of p-glycoprotein-mediated secretion in absorptive drug**

permeability: an approach using passive membrane permeability and affinity to p-glycoprotein. *J Pharm Sci* 1999, **88**:1067-1072.

10. •• Wessel MD, Jurs PC, Tolan JW, Muskal SM: **Prediction of human intestinal absorption of drug compounds from molecular structure.** *J Chem Inf Comput Sci* 1998, **38**:726-735.
A highly cited paper on the use of CPSA and other descriptors to predict the fractional absorption of compounds based on their molecular structures.

11. Stanton DT, Jurs PC: **Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies.** *Anal Chem* 1990, **62**:2323-2329.

12. Mitchell BE, Jurs PC: **Prediction of aqueous solubility of organic compounds from molecular structure.** *J Chem Inf Comput Sci* 1998, **38**:489-496.

13. Wessel MD, Sutter JM, Jurs PC: **Prediction of reduced ion mobility constants of organic compounds from molecular structure.** *Anal Chem* 1996, **63**:4237-4243.

14. Stanton DT, Jurs PC: **Computer-assisted prediction of gas chromatographic retention indices of pyrazines.** *Anal Chem* 1989, **61**:1328-1332.

15. Sutter JM, Jurs PC: **Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure-property relationship.** *J Chem Inf Comput Sci* 1996, **36**:100-107.

16. Yee S: ***In vitro* permeability across Caco-2 cells (Colonic) can predict *in vivo* (small intestinal) absorption in man – fact or myth.** *Pharm Res* 1997, **34**:1242-1250.

17. Artursson P, Palm K, Luthman K: **Caco-2 monolayers in experimental and theoretical predictions of drug transport.** *Adv Drug Deliv Rev* 1996, **22**:67-84.

18. Artursson P, Borchardt R: **Intestinal drug absorption and metabolism in cell cultures.** *Pharm Res* 1997, **14**:1655-1658.

19. Lennernäs H: **Human intestinal permeability.** *J Pharm Sci* 1998, **87**:403-410.

20. Pade V, Stavchansky S: **Link between absorption solubility and permeability measurements in Caco-2 cells.** *J Pharm Sci* 1998, **87**:1604-1607.

21. Clark DE: **Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration.** *J Pharm Sci* 1999, **88**:815-821.

22. van de Waterbeemd H, Camenisc G, Folkers G, Raevsky OA: **Estimation of Caco-2 cell permeability using calculated molecular descriptors.** *Quantitative Struct Ativ Rel* 1996, **15**:480-490.

23. •• Palm K, Luthman K, Ungell A-L, Standlund G, Beigei F, Lundahl P, Artursson P: **Evaluation of dynamic polar surface area as a predictor of drug absorption: comparison with other computational and experimental predictors.** *J Med Chem* 1998, **41**:5382-5392.
An excellent paper on the use of polar surface area in the modeling of Caco-2 permeability.

24. Krarup LH, Christensen IT, Hovgaard L, Frokjaer S: **Predicting drug absorption from molecular surface properties based on molecular dynamics simulations.** *Pharm Res* 1998, **15**:972-978.

25. Clark DE: **Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption.** *J Pharm Sci* 1999, **88**:807-814.

26. Stenberg P, Luthman K, Artursson P: **Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties.** *Pharm Res* 1999, **16**:205-121.

27. Palm K, Stenberg P, Luthman K, Artursson P: **Molecular surface properties predict the intestinal absorption of drugs in humans.** *Pharm Res* 1997, **14**:568-571.

28. Dunitz JD, Gavezzotti A: **Attractions and repulsions in molecular crystals: what can be learned from the crystal structures of condensed ring aromatic hydrocarbons.** *Acc Chem Res* 1999, **32**:677-684.

29. • Ruelle P, Kesselring UW: **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.
The authors describe the use and development of mobile order theory in solubility modeling. For the alcohol data set considered, excellent results were obtained. Contains a good discussion of the theory employed.

30. Katritzky AR, Want Y, Sild S, Tamm T, Karelson M: **QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients.** *J Chem Inf Comput Sci* 1998, **38**:720-725.

31. Huibers PDT, Katritzky AR: **Correlation of the aqueous solubility of hydrocarbons with molecular structure.** *J Chem Inf Comput Sci* 1998, **38**:283-292.

32. 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.

33. Hall LH, Kier LB: **Electrotopological state indices for atom types: a novel combination of electronic, topological and valence state information.** *J Chem Inf Comput Sci* 1995, **35**:1039-1045.

34. •• Abraham M, Le J: **The correlation and prediction of the solubility of compounds in water using and amended solvation energy relationship.** *J Pharm Sci* 1999, **88**:868-880.
One of the latest best efforts at modeling solubility based on molecular structures. Provides an excellent framework for researchers interested in solubility modeling.