Available online at www.sciencedirect.com

ScienceDirect

Advanced Drug Delivery Reviews 59 (2007) 533 – 545

Advanced DRUG DELIVERY Reviews

www.elsevier.com/locate/addr

Computational approaches to determine drug solubility ☆

Bernard Faller *, Peter Ertl

Novartis Institutes for BioMedical Research, Novartis Campus, CH-4056 Basel, Basel, Switzerland

Received 23 March 2007; accepted 10 May 2007
Available online 26 May 2007

Abstract

Water solubility is an important molecular property for successful drug development as it is a key factor governing drug access to biological membranes. There have been a number of review articles addressing computational models to predict water solubility with emphasis on the accuracy of the various prediction methods. This paper briefly reviews the available models and focuses on the value which can be extracted by comparing calculated and measured solubility, discusses the potential and limitations of the main computational approaches, and provides guidelines as to when to trust the computed value.

Keywords: Aqueous solubility; Water solubility; log*P*; Drug transport

Contents

List of Abbreviations: α^H, hydrogen-bond acidity; β^H, hydrogen-bond basicity; π^H, polarizability term; CLog*P*, calculated log*P*; ELog*P*, experimentally measured log*P*; log*D*, distribution coefficient; MP, melting point; PSA, polar surface area; R_2, excess molar refraction; relVol, relative volume calculated as a ratio between molecule volume and number of nonhydrogen atoms in molecule; S_0, intrinsic aqueous solubility; V_x, McGowan molecular volume.

☆ This review is part of the *Advanced Drug Delivery Reviews* theme issue on “Drug solubility: How to measure it, how to improve it”.

* Corresponding author.

E-mail addresses: bernard.faller@novartis.com (B. Faller), peter.ertl@novartis.com (P. Ertl).

0169-409X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.addr.2007.05.005

1. Introduction

Solubility plays an essential role in drug disposition, since the maximum rate of passive drug transport across a biological membrane, the main pathway for drug absorption, is the product of permeability and solubility. Poor solubility has been identified as the cause of numerous drug development failures. It is one of the components of the Biopharmaceutical Classification Scheme (BCS) and is particularly important for immediate release BCS class II drugs, for which absorption is limited by solubility (thermodynamic barrier) or dissolution rate (kinetic barrier). Further, if the solubility is incorrectly estimated, this can lead to erroneous interpretation of results in a number of in-vitro assays and weaken SAR.

Poor aqueous solubility is caused by two main factors: i) high lipophilicity and ii) strong intermolecular interactions which make the solubilization of the solid energetically costly. What is meant by good and poorly soluble depends partly on the expected therapeutic dose and potency. As a rule of thumb, a compound with an average potency of 1mg/kg should have a solubility of at least 0.1g/L to be adequately soluble. If a compound with the same potency has a solubility of less than 0.01g/L it can be considered poorly soluble.

Poor aqueous solubility can often be overcome by appropriate formulation work. However, this approach is expensive and without guarantee of success. It is much better to improve solubility by chemistry means through adequate changes in the molecule itself. To this end, it is desirable to determine the aqueous solubility of candidates as early as possible in the discovery process. Even though higher throughput assays have recently become available, the generation of high quality solubility data remains a relatively expensive and time-consuming activity. Therefore, the development of models to predict the aqueous solubility of drug candidates from their chemical structure has attracted considerable attention. Predictive models based on molecular descriptors also help understanding what feature(s) limits solubility and can thus provide useful information to medicinal chemists.

2. Computational strategies to predict solubility

2.1. Issues with experimental determinations

Despite the availability of a number of methods to measure solubility [1–5] it remains a challenge to collect homogenous, high quality datasets in an appropriate time-frame. The lack of high quality solubility datasets in turn, has a large impact on our ability to create predictive models for solubility. Common issues encountered can be divided into two categories:

2.1.1. Data scattering due to assay limitations

There are a number of issues which can affect the quality of the data. If the traditional shake-flask method is used, adsorption to the vial or to the filter, incomplete phase separation, compound instability and slow dissolution can affect the result. When the potentiometric method is used, inaccurate pK_a determination, compound degradation during the titration, slow dissolution or incorrect data analysis can affect the data quality.

2.1.2. Imprecise definition of the term “solubility”

It is very important to define the experimental conditions well. Analogous to log*P* and log*D*, one needs to distinguish the intrinsic solubility, S_0, from the solubility measured at a given pH value in a defined medium. Intrinsic solubility refers to the solubility of the unionized species. Artursson et al. [6] has shown that this parameter is relatively independent of the nature of the medium used. In contrast, solubility measured at a fixed pH value may be highly dependent on the nature and concentration of the counter ions present in the medium. This is especially critical for poorly soluble compounds which are strongly ionized at the pH of the measurement. Finally, it is important to note that single pH measurements (using the shake-flask method, for example) cannot distinguish between soluble monomers and soluble aggregates of the drug molecules (which may range from dimers to micelles), unless more sophisticated experiments are performed [7].

2.2. Predicting intrinsic solubility

The number of original research papers in the area of water solubility prediction is really overwhelming. Just in a single journal – *Journal of Chemical Information and Modeling* – 43 papers focusing on the development of various methods for solubility prediction have been published since 1995. On the one hand this clearly documents the importance of this topic, but on the other hand shows that no really satisfactory approach to solubility prediction is available yet. Recently, several good reviews on this topic have appeared (for example [8–11], and in [11] the performance of 34 literature models is compared) and we recommend interested readers to check these sources. The purpose of this section is not to list all these papers and to provide detailed description of all the methods published, but rather to provide a general overview of different conceptual approaches to predict solubility. Thus, only citations to selected publications that are particularly interesting from the methodological point of view are included. The advantages and disadvantages of the various approaches, particularly with respect to the drug discovery and design processes, are then discussed.

2.2.1. Fragment-based models

Fragment-based models try to predict solubility as a sum of substructure contributions — such as contributions of atoms, bonds or larger substructures. This approach is based on a general assumption that molecule properties are determined completely by molecule structure, and may be approximated by contributions of fragments in the molecule. Fragment-based methods work very well for purely additional molecule properties (such as log$P$ or molar refractivity) where substructures have rather constant contributions to the studied property. This is, however, not the case for solubility, where effects like electron donating/accepting contributions of substituents, and intramolecular hydrogen bonding can play an important role. Such complex effects cannot be properly described solely by fragment contributions. On the other hand, the fragment contributions approach offers the possibility of describing, at least partially, the effect of crystal packing [12], How? which would otherwise be accessible only via expensive computations [13]. When applying fragment contribution methods for prediction of solubility, one needs to use quite a large number of fragments to get reasonable performance (usually over 100) and this sets also high requirements on the number of data points needed to develop the model (the rule of thumb is to use minimally 5 to 6 data points per parameter/fragment). This precludes the development of "local" fragment-based models for smaller data sets. Numerous methods for solubility prediction based on fragment contributions exist, some of the more popular approaches include models of Huuskonen [14], Klopman [15] or Tetko et al. [16]. The introduction of fragments into correlation can also improve models based on physicochemical properties, as discussed in Section 2.2.4.

2.2.2. Models based on logP

The inverse relation between solubility and lipophilicity has been recognized for a long time and empirical relationships between logS_0 and log$P$ have been reported.

$$\log \quad S_0 = 0.978 - 1.339 \qquad \log \quad P \text{ with} \\ n = 1.56, r^2 = 0.874 \text{ for liquid solutes} \tag{1}$$

[17]

$$\log \quad S_0 = 1.17 - 1.38 \qquad \log \quad P \text{ with} \\ n = 300, r^2 = 0.931 \text{ for crystalline solutes} \tag{2}$$

[18]

However, when more complex, drug-like molecules were added to the set, the relation deteriorated and it became obvious that additional parameters are required. For example, for the 60 drugs listed in Table 1, Eq. (2) gives $r^2=0.542$ with CLog$P$ and 0.569 with ELog$P$, values considerably lower than those obtained with simple compounds.

The octanol–water partition coefficient (log$P$), characterizing molecule hydrophobicity, is probably the single most important parameter influencing solubility. Hansch, who introduced log$P$ in QSAR studies, formulated the first correlation between log$P$ and solubility [17]. In this publication the experimentally determined log$P$ was used. However correlation with calculated log$P$ provided equally valid results. Yalkowsky and coworkers proposed a so-called general solubility equation (GSE) [19], in which the correlation with log$P$ is improved by the addition of an experimental melting point (MP). This equation is physically reasonable because it covers the effect of hydrophobicity (log$P$) as well as the effect of crystal packing (approximated by MP). The GSE has the form

$$\log \ S_0 = 0.5 - 0.01(\mathrm{MP} - 25) - \log \ P. \tag{3}$$

The disadvantage of the method is the necessity to know the experimental melting point. When MP is not available, it was suggested to use a median value of 125°C instead [8].

2.2.3. Models based on solvation properties

Abraham and Le [20] proposed an elegant method to predict solubility by considering solute–solvent interactions. The equation has the form

$$\log \quad S_0 = 0.52 - 1.00R_2 + 0.77\pi^{\mathrm{H}} + 2.17\Sigma\alpha^{\mathrm{H}} \\ - 4.24\Sigma\beta^{\mathrm{H}} - 3.36\Sigma\alpha^{\mathrm{H}}\Sigma\beta^{\mathrm{H}} - 3.99V_{\mathrm{x}} \tag{4}$$

where R_2 is excess molar refraction, π^{H} is dipolarity/polarizability, α^{H} and β^{H} are hydrogen-bond acidity and basicity, respectively, and finally V_{x} is the McGowan's molecular volume (which also characterizes the hydrophobicity of the solute). The inclusion of additional hydrogen-bonding cross-correlation terms (which describe intramolecular hydrogen bonding and may account for solid state effects) improves the correlation with solubility. The advantage of the Abraham equations is the fact that each parameter used has a clear physical meaning and therefore the equation is easy to interpret. Most of the coefficients used in the Abraham's equation were derived from experimental measurements on relatively simple small molecules. To obtain reliable values of Abraham parameters for complex multifunctional drug-like structures is not an easy task, limiting to some extent the applicability of this elegant equation. An alternative is to use the ABSOLV module of the ADME Boxes software [21] which calculates these parameters from fragment contributions.

2.2.4. Hybrid models

Numerous other approaches to calculated water solubility have been proposed. The list of molecular descriptors used in this endeavor is nearly unlimited. One of the most useful descriptors for correlating with solubility is the polar surface area (PSA), which characterizes molecule polarity and hydrogen bonding features. PSA, defined as a sum of surfaces of polar atoms [22], is conceptually easy to understand and seems to encode in an optimal way a combination of hydrogen bonding features and molecular polarity. Descriptors calculated by quantum chemical methods (COSMO-RS approach) [23] have also been shown to provide good correlation with experimental solubility values of drugs and pesticides. The method can be

Table 1
Measured and calculated molecular properties of 60 generic drugs

cpd #	INN	log1/S_0 (M)	Elog$P$	CLog$P$	$\Sigma\alpha^H$	$\Sigma\beta^H$	π^H	R_2	V_x	MW	MI_{volume}	relVol	PSA	# Rot. bonds	Solubility ref.	log$P$ ref.
1	Acyclovir	2.24	−1.6	−0.72	0.92	2.04	2.33	2	1.52	225.2	187.6	11.73	119	4	35	32
2	Amiloride	3.36	−0.2	0.11	1.09	2.1	2.44	2.11	1.51	229.6	176.8	11.79	157	2	35	33
3	Amiodarone	8.1	7.8	8.90	0	1.36	2.48	3.39	3.75	645.3	437.0	14.10	43	11	30 p114	30 p60
4	Amitriptyline	5.19	4.62	4.90	0	1.05	1.33	1.92	2.40	277.4	285.1	13.58	3	3	35	30 p60
5	Amoxicilin	2.17	−1.7	−1.80	1.32	2.57	3.23	2.79	2.54	365.4	306.9	12.28	158	4	35	41
6	Ampicillin	1.69	−2.17	−1.20	0.94	2.33	3.02	2.62	2.48	349.4	298.9	12.45	138	4	30 p114	30 p60
7	Atenolol	1.3	0.22	−0.11	0.66	1.87	1.8	1.32	2.18	266.3	260.9	13.73	85	8	4	38
8	Atropine	1.61	1.89	1.30	0.31	1.48	1.72	1.54	2.28	289.4	279.0	13.29	50	5	35	30 p60
9	Benzoic acid	1.59	1.96	1.90	0.66	0.38	0.83	0.86	0.93	122.1	111.0	12.34	37	1	36	39
10	Carbamazepine	3.4	2.2	2.40	0.33	1.07	2.05	2.09	1.81	236.3	215.1	11.95	46	0	Novartis internal	43
11	Chlorpromazine	5.27	5.4	5.30	0	1.11	1.91	2.33	2.41	318.9	285.4	13.59	32	4	35	30 p60
12	Cimetidine	1.43	0.48	0.40	0.84	1.65	1.97	1.58	1.96	252.3	231.5	13.61	114	8	36	30 p61
13	Ciprofloxacin	3.73	−1.08	−0.70	0.62	1.84	2.28	2.22	2.30	331.3	285.5	11.89	73	3	35	30 p61
14	Clozapine	3.7	4.1	3.70	0.19	1.85	1.81	2.55	2.43	326.8	292.3	12.71	31	1	34	34
15	Corticosterone	3.2	1.9	2.30	0.43	1.73	3.09	1.79	2.74	346.5	335.4	13.41	75	2	42	45
16	Cortisone	3	1.9	1.30	0.42	2.08	3.48	1.89	2.75	360.5	337.2	12.97	92	2	42	45
17	Desipramine	3.81	3.8	4.50	0.14	0.98	1.58	1.81	2.26	266.4	270.4	13.52	15	4	35	30 p61
18	Diclofenac	5.59	4.5	4.70	0.71	0.88	2.05	1.99	2.03	296.1	238.7	12.56	49	4	4	30 p61
19	Diltiazem	2.95	2.9	3.60	0	1.96	2.87	2.53	3.14	414.5	377.7	13.03	84	7	36	30 p61
20	Dipyridamole	5	3.5	1.50	1.3	3.98	3.68	3.28	3.87	504.6	475.4	13.20	145	12	46	46
21	Doxycycline	2.35	0.42	−0.50	1.86	3.19	4.28	3.53	3.10	444.4	377.8	11.81	182	2	35	30 p61
22	Erythromycin	3.14	2.54	1.60	1.47	5.13	3.7	2.83	5.77	733.9	709.3	13.91	194	7	35	30 p62
23	Famotidine	2.48	−0.81	−0.60	1.15	2.3	2.97	2.66	2.26	337.4	262.2	13.11	238	7	4	30 p62
24	Flurbiprofen	4.36	4	3.70	0.59	0.63	1.42	1.44	1.84	244.3	220.8	12.27	37	3	4	30 p62
25	Furosemide	4.75	2.56	1.90	1.03	1.65	3.08	2.22	2.10	330.7	249.5	11.88	131	5	4	30 p62
26	Glyburide	5.9	4.4	4.20	0.89	2.05	3.8	2.73	3.56	494.0	424.7	12.87	114	9	46	46
27	Hydrochlorothiazide	2.63	0	−0.40	1	1.54	2.38	2.31	1.73	297.7	202.5	11.91	135	1	4	30 p62
28	Ibuprofen	3.62	4.13	3.70	0.58	0.62	0.9	0.87	1.78	206.3	211.2	14.08	37	4	4	30 p62
29	Indomethacin	5.2	3.5	4.20	0.46	1.18	2.6	2.46	2.53	357.8	303.2	12.13	69	4	35	30 p62
30	Ketoprofen	3.33	3.16	2.80	0.5	0.9	1.88	1.58	1.98	254.3	234.8	12.36	54	4	4	30 p62
31	Labetalol	3.45	1.33	2.50	0.79	2.01	2.33	2.13	2.64	328.4	314.8	13.12	96	8	4	30 p63
32	Lasinavir	4	3.3	4.10	0.74	3.29	3.36	2.26	5.26	659.8	636.3	13.54	154	20	34	34
33	Mefenamic acid	6.6	5.1	5.30	0.71	0.93	1.83	1.74	1.92	241.3	228.0	12.67	49	3	46	46
34	Methotrexate	4.29	0.54	−0.50	1.59	3.45	4.42	3.55	3.22	454.4	387.4	11.74	211	9	35	30 p62
35	Metolazone	4.1	4.1	2.00	0.55	1.76	3.12	2.72	2.50	365.8	295.0	12.29	101	2	34	34
36	Metoprolol	1.2	1.95	1.50	0.23	1.61	1.39	1	2.26	267.4	273.0	14.37	51	9	36	38
37	Nadolol	1.57	0.85	0.40	0.86	2.15	1.87	1.64	2.49	309.4	302.5	13.75	82	6	36	38
38	Naproxen	4.21	3.24	2.80	0.56	0.8	1.5	1.63	1.78	230.3	214.0	12.59	47	3	4	30 p63
39	Norfloxacin	2.78	1.49	−0.80	0.63	1.85	2.22	1.99	2.27	319.3	279.2	12.14	73	3	30 p114	40
40	Nortryptiline	4.18	4.39	4.30	0.23	0.54	1.3	1.89	2.26	263.4	268.2	13.41	12	3	30 p114	30 p64
41	Phenazopyridine	4.24	3.31	2.10	0.47	1.17	1.81	2.03	1.64	213.2	193.0	12.06	90	2	35	30 p64
42	Phenytoin	4.13	2.24	2.10	0.51	1.19	1.88	2.15	1.87	252.3	223.9	11.78	58	2	4	30 p64
43	Pindolol	3.7	1.83	1.70	0.3	1.54	1.62	1.64	2.01	248.3	242.8	13.49	57	6	30 p114	38
44	Piroxicam	5.48	1.98	1.90	0.68	1.85	2.35	2.67	2.25	331.3	268.1	11.65	108	2	30 p114	30 p60
45	Primaquine	2.77	3	2.60	0.45	1.58	1.79	1.87	2.15	259.3	256.9	13.52	60	6	35	30 p64
46	Probenecid	5.68	3.7	3.40	0.5	1.45	2.35	1.38	2.16	285.4	255.6	13.45	83	7	37	30 p64
47	Progesterone	4.4	3.9	3.70	0.01	1.09	2.55	1.44	2.62	314.5	319.1	13.87	34	1	42	45
48	Promethazine	4.39	4.05	4.40	0	1.18	1.78	2.19	2.28	284.4	271.6	13.58	32	3	35	30 p64
49	Propoxyphene	5.01	4.37	5.20	0	1.33	1.6	1.58	2.91	339.5	346.2	13.85	30	9	36	30 p64
50	Propranolol	3.62	3.48	2.80	0.25	1.3	1.53	1.73	2.15	259.3	257.8	13.57	41	6	4	38
51	Quinine	2.82	3.5	2.80	0.27	1.74	1.74	2.36	2.49	324.4	310.8	12.95	46	4	36	30 p60
52	Rufinamide	3.5	0.9	0.50	0.34	1.2	2.04	1.45	1.54	238.2	189.4	11.14	74	3	34	34
53	Tamoxifen	7.55	5.26	6.80	0	1.23	1.74	2.22	3.17	371.5	376.1	13.43	12	8	35	30 p65
54	Terfenadine	6.69	5.52	6.10	0.48	2.29	2.21	2.72	4.01	471.7	478.8	13.68	44	9	36	30 p65
55	Testosterone	4.06	3.3	3.20	0.32	1.05	2.28	1.54	2.38	288.4	291.5	13.88	37	0	46	43
56	Theophylline	1.38	0	0.00	0.31	1.76	2.05	1.3	1.22	180.2	150.7	11.59	69	0	35	30 p60
57	Trovafloxacin	4.53	0.15	−0.20	0.63	2.06	3.07	2.75	2.62	416.4	327.9	10.93	100	3	36	30 p65
58	Valsartan	4.2	3.9	4.90	0.74	2.25	2.99	2.4	3.41	435.5	408.7	12.77	112	10	34	34
59	Verapamil	4.67	4.33	4.50	0	1.89	2.23	1.75	3.79	454.6	454.3	13.77	64	14	35	30 p66
60	Warfarin	4.74	3.54	2.90	0.36	1.11	2.05	2.09	2.31	308.3	277.2	12.05	64	4	35	30 p66

used to predict solubility in almost any arbitrary solvent. A drawback lies in the demands on computational time (about 2h of computational time per molecule), although this has been considerably improved by the COSMOfrag approach [24]. Clark [25] used a set of quantum chemically calculated parameters to obtain correlations with solubility. Unlike simple 2D descriptors, descriptors which require 3D molecule structure are not so commonly used for solubility prediction. Although this description of molecules is closer to physical reality, the necessity to handle conformational problems and transform descriptors to an alignment-free form adds complexity to the calculations. Moreover, reported results of 3D approaches do not provide any considerable improvement in comparison with 2D approaches.

Table 2
Measured and calculated molecular properties of 53 Novartis in-house compounds

Cpd#	Exp log1/S_0(M)	ELog*P*	CLog*P*	$\Sigma\alpha^H$	$\Sigma\beta^H$	π^H	R_2	V_x	MW	MI_{volume}	relVol	PSA	# Rot. bonds
61	2.8	3.4	2.747	0	1.44	1.82	1.71	2.46	296.4	296.9	13.49	24	3
62	5.2	4.6	4.912	0.13	1.62	1.01	2.32	3.09	392.6	373.5	14.37	40	6
63	5.4	4.7	5.471	0.13	1.62	1.01	2.32	3.23	406.7	390.3	14.46	40	6
64	4.9	4.6	5.265	0.48	2.13	3.33	3.09	3.72	510.6	453.0	12.24	69	10
65	5.5	4.8	4.772	0.48	2.14	3.4	3.31	3.75	522.6	459.3	12.09	69	7
66	4.4	5.3	6.919	0.6	2.31	2.79	1.84	4.24	487.7	508.5	14.53	82	6
67	5.7	4.9	4.053	0.48	1.86	3.15	3.02	3.39	477.5	412.2	11.78	66	7
68	5.4	4.4	4.353	0.13	1.61	1.01	2.32	2.95	378.6	356.7	14.27	40	6
69	3.9	3.6	5.373	0.72	3.16	3.92	3.96	4.63	599.7	559.9	12.73	124	11
70	3.6	2.8	3.943	0.41	2.24	3.18	2.47	3.80	468.7	460.7	13.55	78	10
71	3.8	3.5	3.914	0.86	2.38	3.65	2.9	3.78	489.6	455.2	12.65	90	11
72	5.1	5	5.783	1.14	2.85	3.89	3.58	4.25	594.7	519.6	12.08	109	9
73	3.5	3.1	4.245	0.51	2.37	3.23	2.58	3.90	483.7	473.1	13.52	90	11
74	5.2	5.1	6.407	0.44	2.23	3.51	3.29	4.13	575.5	498.4	12.78	101	12
75	3.1	2.8	3.349	0.21	1.11	1.17	0.87	2.13	286.3	263.3	13.17	29	5
76	5.0	2.4	3.159	1.2	2.18	3.86	3.42	2.92	429.5	351.4	11.71	149	5
77	5.9	4.7	5.247	0.89	2.78	3.78	3.46	4.06	578.6	497.5	11.84	100	8
78	4.9	4.4	5.098	0.7	2.54	3.44	3.43	3.76	536.6	460.8	11.82	97	7
79	5.3	2.8	1.463	0.19	2.11	3.19	2.53	2.95	397.4	356.0	12.28	112	7
80	6.0	5.3	6.024	0.51	1.93	3	2.71	3.81	573.4	463.8	12.21	92	10
81	5.6	4.8	6.085	0.89	2.81	3.77	3.47	4.34	606.7	530.8	12.06	100	9
82	5.2	4.3	4.904	1.34	2.95	3.82	3.46	4.29	593.7	523.6	12.18	109	10
83	5.8	5.2	4.52	1.19	2.44	3.6	3.82	4.29	580.1	522.6	12.75	111	8
84	3.3	2.2	2.379	0	2.79	3.94	3.09	3.64	491.6	440.6	12.59	84	5
85	4.1	4.4	4.917	0.21	1.32	2.14	1.56	2.81	392.4	345.9	12.35	56	6
86	5.6	3.1	3.191	0.37	2.61	2.99	3.43	3.39	462.5	413.6	12.16	93	6
87	3.3	2.5	1.984	0.7	1.84	2.2	1.6	2.60	409.4	320.4	11.87	113	6
88	6.1	3.5	2.118	0.58	1.27	2.73	1.98	2.08	308.3	251.5	11.98	109	6
89	5.1	4.9	6.019	0.41	2.15	3.38	3.32	3.97	544.5	477.3	12.90	63	8
90	5.6	5	4.994	0.48	2.36	2.8	2.41	3.72	513.6	455.7	12.32	71	9
91	5.1	5	5.605	0.48	2.3	3.28	3.26	3.89	536.6	476.1	12.21	69	8
92	6.0	4.3	5.431	0.89	3.03	4.17	4.14	4.38	621.7	538.4	11.96	126	11
93	6.4	5	6.685	0.41	2.26	3.87	4.1	3.96	554.5	476.9	12.55	74	7
94	4.7	4.8	5.391	0.13	1.65	0.95	2.31	3.23	406.7	389.6	14.43	40	6
95	3.6	3.5	4.11	0.14	2.57	3.06	3.53	3.83	542.0	473.2	12.45	90	8
96	6.1	5.1	6.414	0.51	2.17	3.47	3.17	4.14	593.5	503.3	12.58	101	12
97	4.9	2.8	3.671	0.41	2.22	3.64	3.64	3.41	440.6	411.4	12.47	69	6
98	4.1	3.1	3.671	0.41	2.17	3.6	3.62	3.41	440.6	411.4	12.47	69	6
99	4.7	3.1	4.271	0.41	2.41	3.51	3.6	3.70	468.6	444.8	12.71	69	7
100	4.4	3.4	4.271	0.41	2.41	3.51	3.6	3.70	468.6	444.8	12.71	69	7
101	3.7	3.3	4.289	0.96	1.64	3.15	3.14	2.79	386.5	334.5	12.39	112	6
102	6.3	4.7	5.892	0.73	2.53	3.27	3.2	4.60	571.8	555.2	13.22	92	8
103	5.5	3.5	1.854	0.78	1.78	2.37	2.58	2.51	379.4	306.3	11.78	104	6
104	5.2	5.4	4.29	1.43	2.2	3.95	3.45	3.67	517.0	433.9	12.40	133	8
105	3.8	1.75	1.151	0.57	1.5	2.94	2.26	2.57	358.4	305.8	12.23	92	5
106	3.7	1.5	1.424	0.7	1.93	2.8	2.51	2.40	403.3	297.2	11.01	124	4
107	4.4	1.6	1.636	0.63	2.08	2.96	2.87	2.63	377.4	315.9	12.15	124	4
108	3.5	1.8	0.984	0.7	2.1	2.63	2.01	2.59	422.4	323.0	11.54	116	4
109	3.4	0.9	0.45	1.01	2.01	2.26	1.78	2.18	367.3	269.5	11.23	124	4
110	3.2	0.4	0.903	0.93	2.27	2.59	2	2.42	396.4	299.0	11.50	127	6
111	3.4	−0.5	−1.763	1.47	3.31	3.19	3.46	3.42	463.0	405.1	12.66	214	13
112	3.3	1.9	0.695	0.31	2.13	3.3	2.3	2.70	360.4	329.8	12.69	100	7
113	4.4	1.8	1.121	0.7	1.79	2.3	1.81	2.35	393.3	293.4	11.29	113	4

In addition to the various descriptors used, a broad range of statistical and data-mining techniques have been applied in the field of solubility prediction. Besides classical linear regression and PLS approaches also neural networks [16], support vector machines [26], Monte Carlo simulations [27], genetic algorithms [28] or cellular automata [29]. Comparison of results of various methods may be found in the review by Taskinen et al. [11].

2.3. Solubility of ionizable compounds

The aqueous solubility of ionizable compounds obviously depends on pH. For compounds with one ionizable group, the solubility pH-profile can be calculated from the intrinsic solubility (S_0) and the pK_a using the Henderson–Hasselbalch type equations shown below:

$$S = S_0\left(10^{-\mathrm{p}K_a+\mathrm{pH}} + 1\right)\text{for acids} \qquad (5)$$

$$S = S_0\left(10^{\mathrm{p}K_a-\mathrm{pH}} + 1\right)\text{for bases.} \qquad (6)$$

Therefore, $S=S_0$ when pH < pK_a for acids and when pH > pK_a for bases. However, this ideal relationship between solubility and pH does not always hold true [6] and a number of additional factors influence solubility of ionized molecules (for further discussion of this topic, refer to the contributions from Avdeef and from Serajuddin in this issue).

The first factor which accounts for deviation from the equations above is the effect of salt precipitation. In the presence of simple, monovalent ions, simple rules can be used as a first estimate. For example, in 0.15M NaCl, a weak acid starts to precipitate at about 4 log units above its intrinsic solubility value and 3 log units for a base. This simple rule has been called the “sdif3–4” effect [30]. The neutral and the salt species can co-precipitate and the equilibrium constant describing the equilibrium between the two solids has been called K_a^{Gibbs} [3]. The chemical equilibrium associated with the species present in solution is:

$$\begin{array}{ccccc} & & K_a & & \\ & A^- + H^+ & \leftrightarrow & HA & \\ K_{SP} & \updownarrow & & \updownarrow & K_{HA}(s) \\ & A^-Na^+(s) + H^+ & \leftrightarrow & HA(s) + Na^+ & \\ & & K_a^{\mathrm{Gibbs}} & & \end{array}$$

The chemical equations above show that K_{SP} and K_a^{Gibbs} are conditional constants which depend on the nature and concentration of the counter ions. The consequence is that the solubility of ionized molecules at a given pH value can change depending on the composition of the medium.

An additional factor which may cause deviation from the Henderson–Hasselbalch equation is the formation of soluble aggregates of the drug (from dimmers to micelles) [31].

3. Model performance

In order to illustrate performance of water solubility prediction in a real world scenario, we present here results obtained for 60 common drugs, as well as 53 in-house molecules. Both sets are diverse and cover a range of chemical functionalities. Experimental log$P$ and log(1 / S_0) data for 60 drugs have been collected from the literature [4,20,30,32–46] and checked for correctness.

A set of molecular properties for all molecules was calculated using the ADME Boxes v3.5 software (http://www.ap-algorithms.com), Molinspiration property calculator package v2007.01 (http://www.molinspiration.com) and CLog$P$ 4.71 from Biobyte (http://www.biobyte.com). All descriptors have been calculated based on the molecular topology (atomic connectivity) only. Compound properties are shown in Table 1 (drugs) and Table 2 (in-house compounds).

The analysis below is restricted to intrinsic solubility (S_0) values. As we have discussed in paragraph 2.3, solubility values at a given pH may largely depend on the nature of the buffer used, and therefore it is more difficult to get homogeneous datasets of adequate quality. By contrast, intrinsic solubility is less affected by the composition of the buffer used. Solubility at a specific pH value can be calculated from intrinsic solubility using Eqs. (5) and (6), provided that the pK_a is known (appropriate equations are given in Avdeef, this issue).

3.1. log(1 / S_0) model for 60 drugs and 53 in-house molecules

In a first step we checked the correlation between experimental and calculated log$P$ values for both compound sets. The correlation for the 60 generic drugs is good. CLog$P$ provides correlation with statistical parameters, $r^2=0.902$ and standard error = 0.651, whereas the correlation for 53 in-house molecules has a somewhat lower coefficient ($r^2=0.833$) but better standard error = 0.583. The dependence of experimental and calculated log$P$ for both sets is shown in Fig. 1.

In a second step we tried to correlate log(1 / S_0) with various molecular descriptors. As expected, log$P$ is the only single parameter which reasonably correlates with log(1 / S_0). For the 60 drugs CLog$P$ provides a slightly better correlation than the

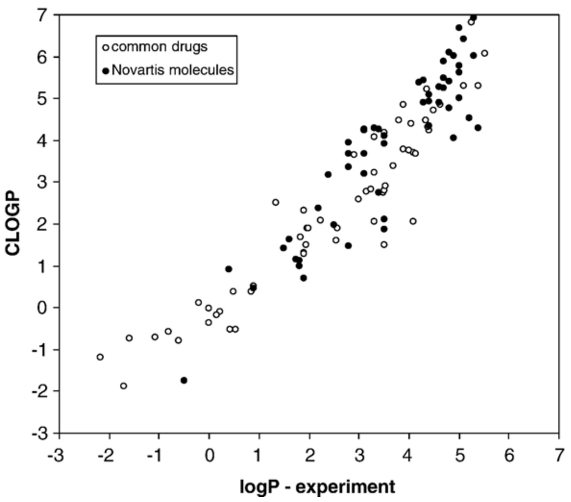

Fig. 1. Correlation between experimental and calculated log$P$ for 60 generic drugs (opencircles) and 53 Novartis molecules (filled circles).

experimental log*P* (r^2=0.569 and 0.542, respectively), while for 53 in-house molecules, experimental log*P* is better (r^2=0.482) than the calculated one (r^2=0.329). We assume that the main reason for inaccurate calculation of log*P* is the use of incorrect tautomeric forms of structures in the calculations.

From our experience this “tautomeric problem” is one of the major issues when calculating log*P* values. Some of drug-like structures may exist in dozen of tautomeric forms, energetically quite close. And in some cases the form present in the database (and therefore used for calculation of properties) does not correspond to the form present in the test tube. The differences between calculated log*P* values among various tautomers may be quite large [47], as is exemplified by Fig. 2, where various tautomeric forms of one of the Novartis structures used in this study (substituents shown only schematically) together with CLog*P* values of these forms are shown. The calculated log*P* values range from −0.03 to 3.43. Out of 53 Novartis compounds used in the study 28 have more than 1 possible tautomeric form (24 out of them more than 2 possible tautomers). This clearly indicates the necessity to pay attention to the distribution of tautomers. The issue is less pronounced in the case of common generic drugs, where the correct tautomeric form is usually well known. With new compounds the tautomeric structure stored in the database and actual structure often do not match. This may also explain the fact, that for 60 common drugs the correlation of solubility with log*P* provides very similar correlations with either experimental or calculated log*P* (actually correlation with CLog*P* is slightly better), while for Novartis structures the correlation is considerably better when using the experimental log*P* values.

In addition to single-parameter equations we also tried equations with multiple parameters. Several equations with 2 or 3 parameters provide improvements in prediction. However, the addition of fourth parameters further improves the correlation only slightly, with leave-one-out cross-validated correlation coefficients decreasing. The parameters which enhance the correlation of the equation over and above the results with log*P* are: 1) PSA — polar surface area calculated from fragment contributions [22], 2) R_2 — Abraham’s relative molar refractivity [20] and 3) relVol — relative volume calculated as a ratio between molecule volume [48] and number of nonhydrogen atoms in molecule. This parameter describes branching and steric requirements of the molecule.

Table 3
Intercorrelation descriptor matrix for 60 drugs, for comparison we include both experimental and calculated log*P*

r^2	ELog*P*	CLog*P*	PSA	relVol
ELog*P*	1.00	0.90	0.32	0.48
CLog*P*	0.90	1.00	0.37	0.25
PSA	0.32	0.37	1.00	0.13
relVol	0.48	0.25	0.13	1.00

As mentioned, there are several equations with 3 parameters which provide very similar performance. We selected Eq. (7), because it contains parameters which have clear physical meaning, which can be easily interpreted and are not mutually inter-correlated (Table 3):

$$\log\ (1/S_0) = 9.5942 + 0.7555{*}\mathrm{C\ Log}\ P + 0.0088{*}\mathrm{PSA} - 0.6438{*}\mathrm{relVol}. \tag{7}$$

The statistical parameters obtained with Eq. (7) for the data set containing the 60 generic drugs are r^2=0.726, r=0.853, standard error=0.805, n=60.

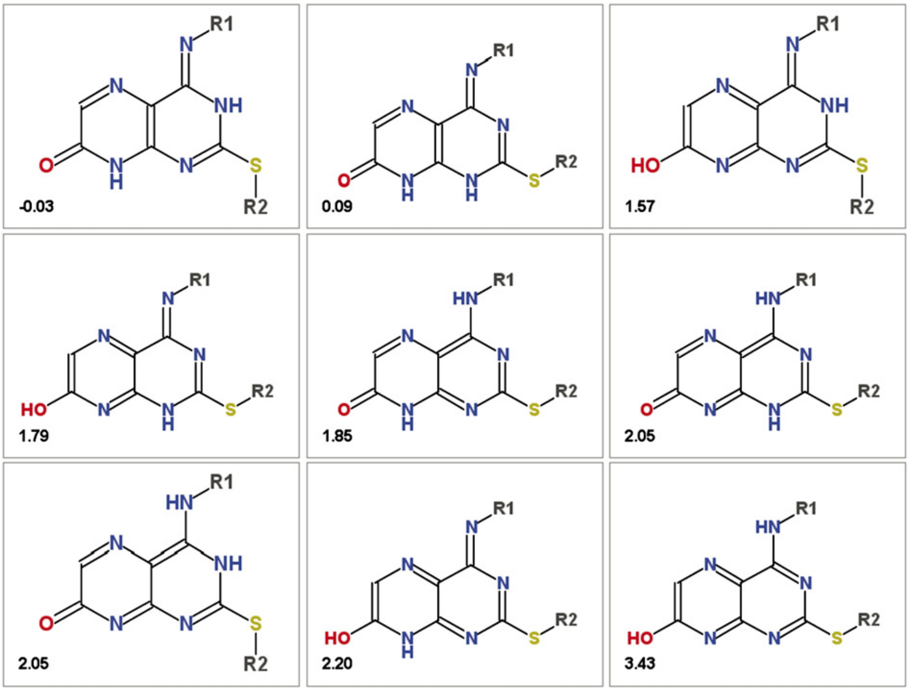

Fig. 2. Calculated log*P* values for a sample structure range from −0.03 to 3.43 based on a particular tautomeric form.

When applying this equation to the 53 in-house compounds, results depend on the type of log$P$ values used. As discussed already, the ELog$P$ provides better prediction than CLog$P$, with statistical parameters r^2=0.524, r=0.724, q^2=0.492, standard error=0.693, n=53. These results are considerably better than the correlation with CLog$P$, r^2=0.399). The correlations for both data sets are shown on Fig. 3.

Fig. 3 shows that the generic drug set covers a relative broad range of solubility values (log(1 / S_0) between 1.2 and 8.1) containing some extremely soluble, as well as insoluble drugs, while for the Novartis molecules the range is narrower (between 2.8 and 6.4). This, of course, affects the comparison of results and provides some "advantage" to the generic set. We therefore also examined Eq. (7) only for the 43 molecules from the drug set with log(1 / S_0) values in the range 2.8–6.4. The resulting correlation coefficient is considerably lower (r^2=0.337, standard error=0.693), providing a more realistic picture of the extent to which it is possible to predict aqueous solubility in the drug discovery environment using log$P$ combined with few additional, simple calculated properties.

3.2. When can one trust the computed value?

The resolution of a high quality solubility assay is usually within 0.6 log units [49]. Therefore, one cannot expect *in silico* models to be more accurate than 0.5 log units i.e. a factor 3–5. There are several reasons which can be the cause of large differences between experimental and calculated values.

- compound is outside the property space of the training set
- molecular descriptors used to characterize the property space are not appropriate to describe the structural diversity of the compounds
- one or several calculated molecular descriptors are erroneous for the compound considered (see the discussion of tautomers above)
- calculated and measured solubility have a different meaning. This may happen for example, when a compound forms soluble aggregates or micelles.
- experimental value underestimates solubility due to adsorption
- experimental value overestimates solubility because the incubation time was insufficient to allow the most stable form to crystallize.

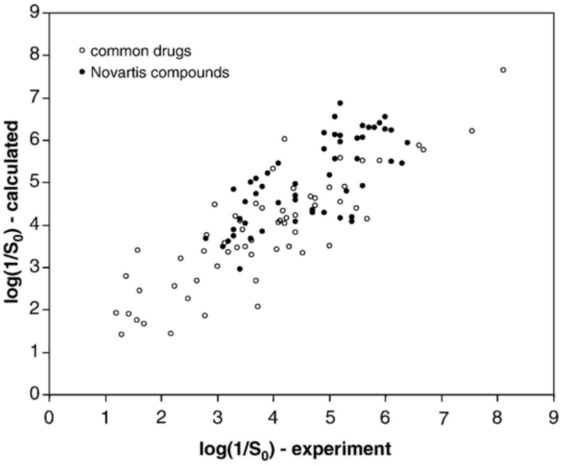

Fig. 3. Experimental vs. predicted (Eq. (7)) log(1 / S_o) for 60 generic drugs (open circles) and 53 in-house molecules (filled circles).

4. Beyond solubility prediction

Fig. 4 shows the relationship between log$P$ and log1 /S_0. One can see that when solubility is expressed in molar terms, almost no compound falls below the unity line for both generic drugs (Fig. 4A) and "new compounds" sets (Fig. 4B). In other words, when the molecular weight is factored in solubility, one can see that lipophilicity dictates the highest solubility limit and the following simple rule can be derived:

$$S_0(M) \leq 10^{-\log P} \quad (8)$$

For the compounds close to the unity line, solubility is mainly lipophilicity driven. However, a number of compounds lie distinctly above the unity line. These compounds are significantly less soluble than predicted by Eq. (8), indicating that other interactions limit their solubility. In order to separate lipophilicity from other contributions to solubility, we introduce the ΔSL parameter which is defined as:

$$\Delta \mathrm{SL} = \log 1/S_0 - \log P. \quad (9)$$

A high ΔSL value means that the contribution of factors other than lipophilicity to solubility is significant and the improvement of the intrinsic solubility by addition of classical formulation ingredients might be modest. Table 4 shows properties of compounds with ΔSL>3. In this context, the solvatochromic equation (Eq. (4)) proposed by Abraham [20] for water solubility appears useful to link this feature to molecular properties.

Eq. (4) shows that the simultaneous presence of hydrogen-bond donors and hydrogen-bond acceptors results in a negative contribution to solubility. Abraham postulated that it enhances the cohesive properties of the solid and makes dissolution energetically more costly. Most of the compounds with ΔSL>3 in Table 4 do have a $\Sigma\alpha^H\Sigma\beta^H$ term higher than 2. Exceptions seem to be zwitterionic compounds, which need to be treated separately. These compounds are indicated with a "z" in the ionization column in Table 4. In contrast, compounds with ΔSL<1 usually have a $\Sigma\alpha^H\Sigma\beta^H$ term lower than 2 (see Table 5) with two exceptions: erythromycin and lasinavir. Due to its chemical structure erythromycin is outside the usual property space of medicinal compounds and lasinavir appears to be a highly flexible molecule, with 20 rotatable bonds.

5. Discussion

5.1. Predictive power of current models

As demonstrated in other studies, the inclusion of additional terms to octanol/water log$P$ improves the correlation between

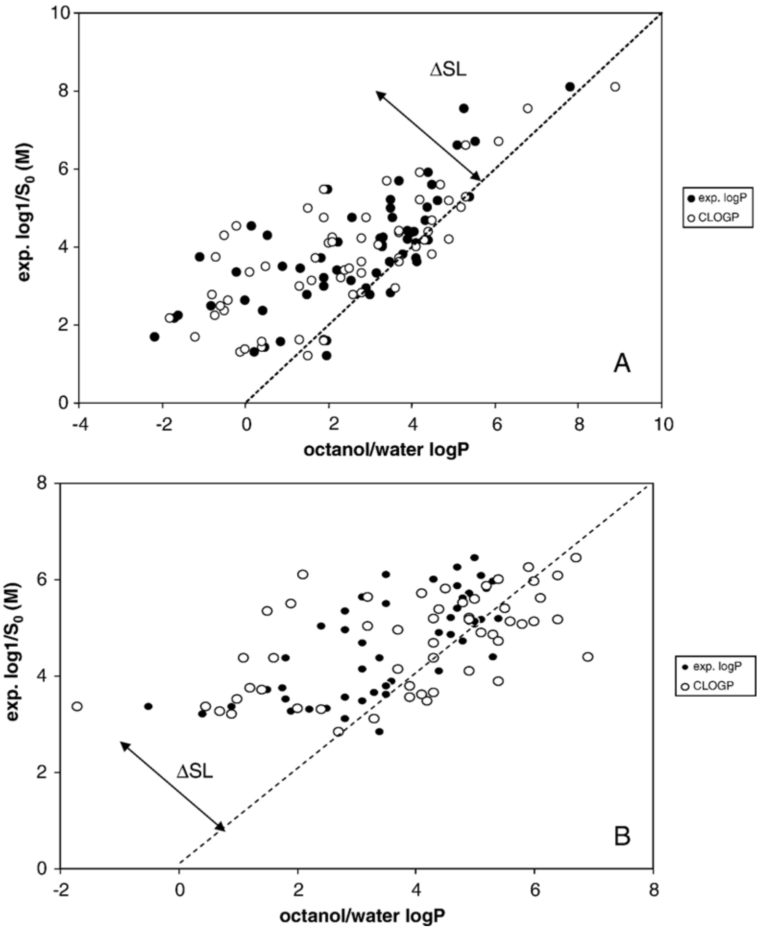

Fig. 4. A: Solubility-lipophilicity relationship for generic drugs. B: Solubility-lipophilicity relationship for research compounds.

calculated and experimental solubility values. The addition of hydrogen-bond terms [20], molecular volume [50], partial atomic charges terms [51] were all shown to be useful descriptors to predict water solubility. The model presented in this study combines log*P*, Polar Surface Area and Relative Volume. The data in Table 6 show that the 3-parameter model correlates better with experimental measurements than the simplest model, which is based solely on log*P*. However, the correlation coefficient remains lower for new compounds compared to generic drugs (0.399 vs. 0.726). One major contributor to this difference lies in the fact that log*P* prediction performs better with generic drugs as opposed to new chemical entities. The substitution of calculated log*P* values with experimental values increases the correlation coefficient from 0.399 to 0.524 for new compounds while it does not improve the predictive power of the model with generic drugs (0.645 vs. 0.726). Improved prediction of solubility by replacing CLog*P* with ELog*P* has already been reported by Glomme and Dressman [46]. Tetko, in a comprehensive review of *in silico* approaches to predict solubility, also pointed out that, in contrast to a common perception, the accuracy of CLog*P* for new compounds needs to be improved [10]. A correlation coefficient between measured and predicted intrinsic water solubility of 0.524 seems a modest

Table 4
Properties of compounds with ΔSL>3

Cpd #	INN	Log1/So (M)	ELogP	CLogP	$^{\Delta}$SL	$\Sigma\alpha^H$	$\Sigma\beta^H$	$\Sigma\alpha^H\Sigma\beta^H$	π^H	R_2	V_x	MW	PSA	# Rot. bonds	Ionization
1	Acyclovir	2.24	−1.6	−0.72	3.84	0.92	2.04	1.9	2.33	2	1.52	225.20	119	4	z
2	Amiloride	3.36	−0.2	0.11	3.56	1.09	2.1	2.3	2.44	2.11	1.51	229.63	157	2	
5	Amoxicilin	2.17	−1.7	−1.8	3.87	1.32	2.57	3.4	3.23	2.79	2.54	365.40	158	4	
6	Ampicillin	1.69	−2.17	−1.2	3.86	0.94	2.33	2.2	3.02	2.62	2.48	349.41	138	4	
13	Ciprofloxacin	3.73	−1.08	−0.7	4.81	0.62	1.84	1.1	2.28	2.22	2.30	331.34	73	3	z
23	Famotidine	2.48	−0.81	−0.6	3.29	1.15	2.3	2.6	2.97	2.66	2.26	337.45	238	7	
34	Methotrexate	4.29	0.54	−0.5	3.75	1.59	3.45	5.5	4.42	3.55	3.22	454.44	211	9	
44	Piroxicam	5.48	1.98	1.9	3.5	0.68	1.85	1.3	2.35	2.67	2.25	331.35	108	2	z
57	Trovafloxacin	4.53	0.15	−0.2	4.38	0.63	2.06	1.3	3.07	2.75	2.62	416.35	100	3	z

Table 5
Properties of compounds with $\Delta SL < 1$

cpd #	INN	log1/S_0 (M)	ELog*P*	CLog*P*	$^{\Delta}$SL	$\Sigma\alpha^H$	$\Sigma\beta^H$	$\Sigma\alpha^H$ $\Sigma\beta^H$	π^H	R_2	V_x	MW	PSA	# Rot. bonds	Ionization
3	Amiodarone	8.1	7.8	8.9	0.3	0	1.36	0.0	2.48	3.39	3.75	645.3	43	11	
4	Amitriptyline	5.19	4.62	4.9	0.57	0	1.05	0.0	1.33	1.92	2.40	277.4	3	3	
8	Atropine	1.61	1.89	1.3	−0.28	0.31	1.48	0.5	1.72	1.54	2.28	289.4	50	5	
9	Benzoic acid	1.59	1.96	1.9	−0.37	0.66	0.38	0.3	0.83	0.86	0.93	122.1	37	1	
11	Chlorpromazine	5.27	5.4	5.3	−0.13	0	1.11	0.0	1.91	2.33	2.41	318.9	32	4	
12	Cimetidine	1.43	0.48	0.4	0.95	0.84	1.65	1.4	1.97	1.58	1.96	252.3	114	8	
14	Clozapine	3.7	4.1	3.7	−0.4	0.19	1.85	0.4	1.81	2.55	2.43	326.8	31	1	
17	Desipramine	3.81	3.8	4.5	0.01	0.14	0.98	0.1	1.58	1.81	2.26	266.4	15	4	
19	Diltiazem	2.95	2.9	3.6	0.05	0	1.96	0.0	2.87	2.53	3.14	414.5	84	7	
22	Erythromycin	3.14	2.54	1.6	0.6	1.47	5.13	7.5	3.7	2.83	5.77	733.9	194	7	
24	Flurbiprofen	4.36	4	3.7	0.36	0.59	0.63	0.4	1.42	1.44	1.84	244.3	37	3	
28	Ibuprofen	3.62	4.13	3.7	−0.51	0.58	0.62	0.4	0.9	0.87	1.78	206.3	37	4	
30	Ketoprofen	3.33	3.16	2.8	0.17	0.5	0.9	0.5	1.88	1.58	1.98	254.3	54	4	
32	Lasinavir	4	3.3	4.1	0.7	0.74	3.29	2.4	3.36	2.26	5.26	659.8	154	20	
35	Metolazone	4.1	4.1	2	0	0.55	1.76	1.0	3.12	2.72	2.50	365.8	101	2	
36	Metoprolol	1.2	1.95	1.5	−0.75	0.23	1.61	0.4	1.39	1	2.26	267.4	51	9	
37	Nadolol	1.57	0.85	0.4	0.72	0.86	2.15	1.8	1.87	1.64	2.49	309.4	82	6	
38	Naproxen	4.21	3.24	2.8	0.97	0.56	0.8	0.4	1.5	1.63	1.78	230.3	47	3	
40	Nortryptiline	4.18	4.39	4.3	−0.21	0.23	0.54	0.1	1.3	1.89	2.26	263.4	12	3	
41	Phenazopyridine	4.24	3.31	2.1	0.93	0.47	1.17	0.5	1.81	2.03	1.64	213.2	90	2	
45	Primaquine	2.77	3	2.6	−0.23	0.45	1.58	0.7	1.79	1.87	2.15	259.3	60	6	
47	Progesterone	4.4	3.9	3.7	0.5	0.01	1.09	0.01	2.55	1.44	2.62	314.5	34	1	
48	Promethazine	4.39	4.05	4.4	0.34	0	1.18	0.0	1.78	2.19	2.28	284.4	32	3	
49	Propoxyphene	5.01	4.37	5.2	0.64	0	1.33	0.0	1.6	1.58	2.91	339.5	30	9	
50	Propranolol	3.62	3.48	2.8	0.14	0.25	1.3	0.3	1.53	1.73	2.15	259.3	41	6	
51	Quinine	2.82	3.5	2.8	−0.68	0.27	1.74	0.5	1.74	2.36	2.49	324.4	46	4	
55	Testosterone	4.06	3.3	3.2	0.76	0.32	1.05	0.34	2.28	1.54	2.38	288.4	37	0	
58	Valsartan	4.2	3.9	4.9	0.3	0.74	2.25	1.7	2.99	2.4	3.41	435.5	112	10	
59	Verapamil	4.67	4.33	4.5	0.34	0	1.89	0.0	2.23	1.75	3.79	454.6	64	14	

performance compared to the 0.726 obtained with known compounds. However, the direct comparison of these two values might be misleading because the solubility range for generic drugs is wider than what is usually encountered with today's medicinal chemistry compounds and if one restricts the analysis to the micro-molar to milli-molar range, the good correlation obtained with generic drugs dramatically decreases (see Table 6). In this context, great care must be taken when looking at the claims of software vendors. Often, the use of these generic models is a source of great disappointment when applied in a narrower solubility range.

Table 6
Correlation coefficients (r^2) obtained for different models and compound sets

Model	C	log*P* coefficient	PSA coefficient	relVol coefficient	Generic drugs (SE)	New compounds (SE)
CLog*P**	2.578	0.517	–	–	0.569 (1.01)	0.329 (0.82)
ELog*P**	2.471	0.543	–	–	0.542 (1.04)	0.482 (0.72)
3-P model/ CLog*P*	9.594	0.755	0.0088	−0.644	0.726 (0.82)	0.399 (0.78)
3-P model/ ELog*P*	8.981	0.734	0.0058	−0.579	0.654 (0.92)	0.548 (0.68)
3-P model/ CLog*P* [2.8–6.4] $n=43$	9.594	0.755	0.0088	−0.644	0.338 (0.69)	–

*CLog*P*: based on calculated log*P*, ELog*P*: based on experimental log*P*, 3-P model-1/CLog*P*: 3-parameter model with CLog*P*, 3-model/ELog*P*: 3-parameter mode-11 with ELog*P*, 3-P model-2/ELog*P*: 3-parameter model based on ELog*P*, 3-P model-1[2.8–6.4] restricted to the solubility range of new compounds. In all cases, the coefficients are determined using the generic drugs as the training set.

5.2. Can one do better than CLogP?

The predictive power of the models seems modest if judged by their correlation coefficients. However, a more optimistic view is obtained if one looks at the fraction of compounds for which the calculated solubility is within a log unit of the experimental value (Table 6). With generic drugs, 65% are predicted within one log unit using a simple CLog*P* model and up to 80% using the 3-parameter approach. With new compounds 60–80% are predicted within 1 log unit irrespective of the model used while 50–60% are predicted within 5-fold which is close to the variability of experimental determinations. Looking at the fraction of compounds correctly predicted within 5-fold, the advantage of the multi-linear regression approaches appear modest compared to the simple CLog*P* correlation proposed almost 40 years ago by Hansch et al. [17].

5.3. What is the value of solubility models in medicinal chemistry?

There are several ways solubility models can add value in medicinal chemistry. The best known is the prediction of solubility *per se*. In this context, it is important to know when to trust the result (see Section 3.2 for details). However, solubility models also bring value by providing links between predicted solubility and molecular descriptors. For example, the equation proposed by Abraham and Le allows to distinguish between the various H-bond contributions and molecular volume. Additionally, the ΔSL parameter introduced in Section 4 is a way to separate lipophilicity from other contributions.

5.4. What are the differences between generic drugs and new compounds?

The calculated molecular properties shown in Tables 1 and 2 were used to perform a PCA analysis and map the property space covered by generic drugs vs. new compounds (Fig. 5). The score plot shown in Fig. 5 shows that the 60 generic drugs used as training set only partially covers the property space of the new compounds. Compared to generic drugs, the new compounds in the present study have a higher molecular weight, a higher lipophilicity, a higher number of rotatable bonds and a higher hydrogen bond basicity. Before using a prediction model, one should first check whether the test compound fits in the property space covered by the training set.

5.5. What is missing, perspectives

The first missing element is the difficulty to access large datasets of adequate quality and diversity. Oftentimes, the highest data quality is only available with late stage development compounds, which reduces the dataset coverage. In early discovery, the numbers are available but compound are less well characterized. Despite the development of a number of high-throughput assays to measure solubility, the result is often not the intrinsic solubility but rather the solubility at a given pH value in a defined medium — with all its associated limitations (see Section 2). In addition, one does not know whether the equilibrium has been truly reached and if the solubility measured includes soluble aggregates.

Second, inaccurate log*P* values appear to be a major cause of discrepancy between calculated and measured solubility values. Although it is probably not the only reason, tautomers play a major role here and a good estimate of the most likely tautomer in solution could quite significantly improve the prediction power of the models.

Third, appropriate descriptors to account for intermolecular forces need to be improved.

The last element lies in the proper use of *in silico* models. Great disappointment usually occurs when applying a global model to a small set of congeneric compounds because the standard error is usually higher than the solubility differences within the small set. Along the same lines, one need to check if the new compounds fits in the property space covered by the training set. Finally, one cannot expect the accuracy of an *in silico* model to be greater than experimental determinations, which is usually within a factor 5.

How can one move forward? As stated at the beginning of this section, the quality and richness of experimental data are key success factors. The problem is that analytical scientists need to cope with a high number of determinations which leaves little time for more in-depth characterization of the compounds. If the currently available *in silico* models were to be used in appropriate situations, one could focus resources on a fewer number of compounds which are outside the property space of the current training sets, freeing up resources for more in-depth investigations. In parallel, improvements in the computation of molecular descriptors (octanol/water log*P*, hydrogen-bond terms, polarizability) are necessary.

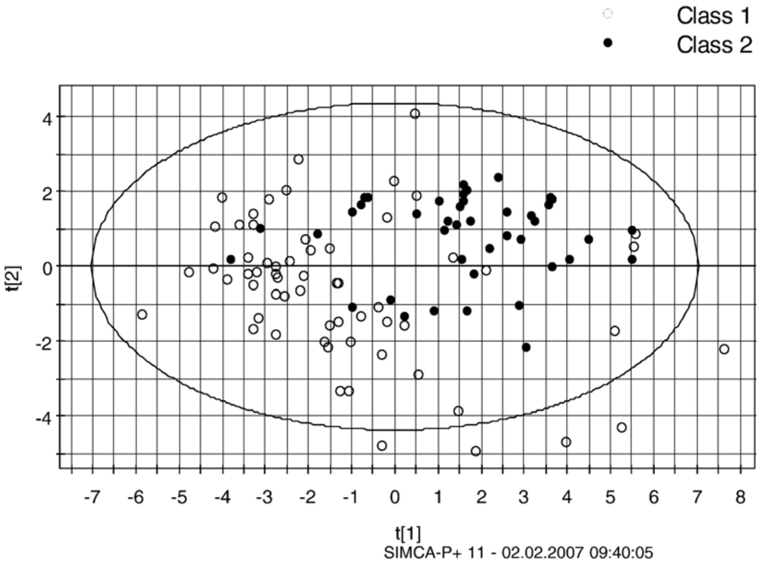

Fig. 5. PCA plot of generic drugs (open circles) and new compounds (filled circles). Parameters used for the PCA analysis are listed in Tables 1 and 2. The PCA plot was generated with SIMCA-P+v11 using a 2 component model.

References

[1] L. Pan, Q. Ho, K. Tsutsui, L. Takahashi, Comparison of chromatographic and spectroscopic methods used to rank compounds for aqueous solubility, J. Pharm. Sci. 90 (2001) 521–529.

[2] D.J.W. Grant, T. Higushi, Solubility Behavior of Organic Compounds, John Wiley, New York, 1990.

[3] A. Avdeef, pH-metric solubility. 1. Solubility-pH profiles from Bjerrum plots. Gibbs buffer and p*K*a in the solid state, Pharm. Pharmacol. Commun. 4 (1998) 165–178.

[4] A. Avdeef, C.M. Berger, C. Brownell, pH-metric solubility. 2: correlation between the acid-base titration and the saturation shake-flask solubility-pH methods, Pharm. Res. 17 (2000) 85–89.

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

[6] C.A.S. Bergstrom, K. Luthman, P. Artursson, Accuracy of calculated pH-dependent aqueous drug solubility, Eur. J. Pharm. Sci. 22 (2004) 387–398.

[7] A. Avdeef, D. Voloboy, A. Foreman, Dissolution and Solubility, in: J.B. Taylor, D.J. Triggle (Eds.), In Comprehensive Medicinal Chemistry II, vol 5, Elsevier, ISBN: 978-0-08-044513-7, 2007, pp. 399–423.

[8] J.S. Delaney, Predicting aqueous solubility from structure, Drug Discov. Today 10 (2005) 289–295.

[9] C.A.S. Bergström, Computational models to predict aqueous drug solubility, permeability and intestinal absorption, Expert Opin. Drug Metab. Tox. 1 (2005) 613–627.

[10] K.V. Balakin, N.P. Savchuk, I.V. Tetko, In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions, Curr. Med. Chem. 13 (2006) 223–241.

[11] J. Taskinen, U. Norinder, In silico predictions of solubility, in: J.B. Taylor, D.J. Triggle (Eds.), in Comprehensive Medicinal Chemistry, Elsevier, 2007, pp. 627–648.

[12] J.S. Chickos, C.M. Braton, D.G. Hesse, J.F. Liebman, Estimating entropies and enthalpies of fusion of organic compounds, J. Org. Chem. 56 (1991) 927–938.

[13] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26.

[14] J. Huuskonen, Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology, J. Chem. Inf. Comput. Sci. 40 (2000) 773–777.

[15] G. Klopman, H. Zhu, Estimation of aqueous solubility of organic molecules by the group contribution approach, J. Chem. Inf. Comput. Sci. 41 (2001) 439–445.

[16] I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva, A.E.P. Villa, Estimation of aqueous solubility of chemical compounds using E-State indices, J. Chem. Inf. Comput. Sci. 41 (2001) 1488–1493.

[17] C. Hansch, J.E. Quinlan, G.L. Lawrence, Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids, J. Org. Chem. 33 (1968) 347–350.

[18] P. Isnard, S. Lambert, Aqueous solubility and n-octanol/water partition coefficient correlations, Chemosphere 18 (1989) 1837–1853.

[19] Y. Ran, N. Jain, S.H. Yalkowski, Prediction of aqueous solubility of organic compounds by the General Solubility Equation (GSE), J. Chem. Inf. Model. 41 (2001) 1208–1217.

[20] M.H. Abraham, J. Le, The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship, J. Pharm. Sci. 88 (1999) 868–880.

[21] http://www.ap-algorithms.com/absolv.htm.

[22] P. Ertl, B. Rohde, P. Selzer, 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. 43 (2000) 3714–3717.

[23] A. Klamt, F. Eckert, M. Hornig, M.E. Beck, T. Bürge, Prediction of aqueous solubility of drugs and pesticides with COSMO-RS, J. Comput. Chem. 23 (2001) 275–281.

[24] M. Hornig, A. KLamt, COSMOfrag: a novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry, J. Chem. Inf. Comput. Sci. 45 (2005) 1169–1177.

[25] T. Clark, Quantum cheminformatics: an oxymoron? in: M.G. Hicks (Ed.), Chemical Data Analysis in the Large. The Challenge of the Automation Age, 2000.

[26] P. Lind, T. Maltseva, Support vector machines for the estimation of aqueous solubility, J. Chem. Inf. Comput. Sci. 43 (2003) 1855–1859.

[27] W.L. Jorgensen, E.M. Duffy, Prediction of drug solubility from Monte Carlo simulations, Bioorg. Med. Chem. Lett. 10 (2000) 1155–1158.

[28] K. Wegner, A. Zell, Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method, J. Chem. Inf. Comput. Sci. 43 (2003) 1077–1084.

[29] L.B. Kier, C.-K. Cheng, A cellular automata model of an aqueous solution, J. Chem. Inf. Comput. Sci. 34 (1994) 1334–1337.

[30] A. Avdeef, Absorption and Drug Development, Wiley-Intersience, New York, 2003.

[31] A. Avdeef, S. Bendels, O. Tsinman, K. Tsinman, M. Kansy, Solubility-excipient classification maps, Pharm. Res. 24 (2007) 536–545.

[32] A. Kristl, S. Pecar, Hydrolipophilic anomalies of some guanine derivatives, Eur. J. Med. Chem. 32 (1997) 3–8.

[33] B. Faller, H-P. Grimm, F. Loeuillet-Ritzler, S. Arnold, X. Briand, High-throughput lipophilicity measurement with immobilized artificial membranes, J. Med. Chem. 48 (2005) 2571–2576.

[34] Physicochemical parameters as tools in drug discovery and lead optimization. B. Faller, F. Wohnsland. Editor(s): B. Testa. Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies, [LogP2000, Lipophilicity Symposium], 2nd, Lausanne, Switzerland, Mar. 5–9, 2000 (2001), Meeting Date 2000, 257–273. Publisher: Verlag Helvetica Chimica Acta, Zurich, Switzerland.

[35] C.A.S. Bergstroem, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, P. Artursson, Absorption classification of oral drugs based on molecular surface properties, J. Med. Chem. 46 (2003) 558–570.

[36] A. Avdeef, C.M. Berger, pH-metric solubility. 3. Dissolution titration template method for solubility determination, Eur. J. Pharm. Sci. 14 (2001) 281–291.

[37] A. Avdeef, Physicochemical profiling (solubility, permeability and charge state), Current Topics in Medicinal Chemistry 1 (2001) 277–351 (Hilversum, Netherlands).

[38] G. Caron, G. Steyaert, A. Pagliara, F. Reymond, P. Crivori, P. Gaillard, P.-A. Carrupt, A. Avdeef, J. Comer, K.J. Box, H.H. Girault, B. Testa, Structure-lipophilicity relationships of neutral and protonated b-blockers. Part 1. Intra- and intermolecular effects in isotropic solvent systems, Helv. Chim. Acta 82 (1999) 1211–1222.

[39] B. Slater, A. McCormack, A. Avdeef, J.E.A. Comer, pH-Metric log P. 4. Comparison of partition coefficients determined by HPLC and potentiometric methods to literature values, J. Pharm. Sci. 83 (1994) 1280–1283.

[40] High-throughput artificial membrane permeability studies in early lead discovery and development. M. Kansy, H. Fischer, K. Kratzat, F. Senner, B. Wagner, I. Parrilla. Editor(s): B. Testa. Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies, [LogP2000, Lipophilicity Symposium], 2nd, Lausanne, Switzerland, Mar. 5–9, 2000 (2001), Meeting Date 2000, 447–464. Publisher: Verlag Helvetica Chimica Acta, Zurich, Switzerland.

[41] S. Winiwarter, N.M. Bonham, F. Ax, A. Hallberg, H. Lennernaes, A. Karlen, Correlation of human jejunal permeability (in Vivo) of drugs with experimentally and theoretically derived parameters, a multivariate data analysis approach, J. Med. Chem. 41 (1998) 4939–4949.

[42] N. Jain, S.H. Yalkowsky, Estimation of the aqueous solubility I: application to organic nonelectrolytes, J. Pharm. Sci. 90 (2001) 234–252.

[43] F. Lombardo, M.Y. Shalaeva, K.A. Tupper, F. Gao, M.H. Abraham, ElogPoct: a tool for lipophilicity determination in drug discovery, J. Med. Chem. 43 (2000) 2922–2928.

[44] M. Ponec, J. Kempenaar, B. Shroot, J.-C. Caron, Glucocorticoids: binding affinity and lipophilicity, J. Pharm. Sci. 75 (1986) 973–975.

[45] T.S. Wiedmann, W. Liang, L. Kamel, Solubilization of drugs by physiological mixtures of bile salts, Pharm. Res. 19 (2002) 1203–1208.

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

[47] P. Pospisil, P. Ballmer, L. Scapozza, G. Folkers, Tautomerism in computer-aided drug design, J. Recept. Signal Transduct. Res. 23 (2003) 361–371.

[48] Molecule volume was calculated as a sum of fragment contributions, http://www.molinspiration.com/services/volume.html.

[49] A.R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson, QSPR studies on vapor pressure, aqueous solubility, and the prediction of water–air partition coefficients, J. Chem. Inf. Comput. Sci. 38 (1998) 720–725.

[50] C.A.S. Bergstrom, U. Norinder, K. Luthman, P. Artursson, Experimental and computational screening models for prediction of aqueous drug solubility, Pharm. Res. 19 (2002) 182–188.

[51] J.W. McFarland, A. Avdeef, C.M. Berger, O.A. Raevsky, Estimating the water solubilities of crystalline compounds from their chemical structures alone, J. Chem. Inf. Comput. Sci. 41 (2001) 1355–1359.