ADMET & DMPK 8(3) (2020) 180-206; doi: http://dx.doi.org/10.5599/admet.794

ADMET

Open Access : ISSN : 1848-7718

http://www.pub.iapchem.org/ojs/index.php/admet/index

Original scientific paper

Can small drugs predict the intrinsic aqueous solubility of 'beyond Rule of 5' big drugs?

Alex Avdeef [1,*] and Manfred Kansy[2]

[1] *in-ADME Research, 1732 First Avenue #102, New York, NY 10128 USA*
[2] *79111 Freiburg im Breisgau, Germany*

*Corresponding Author: E-mail: alex@in-adme.com; Tel: +1-646-678-5713; ORCID ID: 0000-0002-3139-5442

Received: February 18, 2019; Revised: April 14, 2020; Available online: April 25, 2020

Abstract

The aim of the study was to explore to what extent small molecules (mostly from the Rule of 5 chemical space) can be used to predict the intrinsic aqueous solubility, S_0, of big molecules from beyond the Rule of 5 (bRo5) space. It was demonstrated that the General Solubility Equation (GSE) and the Abraham Solvation Equation (ABSOLV) underpredict solubility in systematic but slightly ways. The Random Forest regression (RFR) method predicts solubility more accurately, albeit in the manner of a 'black box.' It was discovered that the GSE improves considerably in the case of big molecules when the coefficient of the log P term (octanol-water partition coefficient) in the equation is set to -0.4 instead of the traditional -1 value. The traditional GSE underpredicts solubility for molecules with experimental S_0 < 50 µM. In contrast, the ABSOLV equation (trained with small molecules) underpredicts the solubility of big molecules in all cases tested. It was found that the errors in the ABSOLV-predicted solubilities of big molecules correlate linearly with the number of rotatable bonds, which suggests that flexibility may be an important factor in differentiating solubility of small from big molecules. Notably, most of the 31 big molecules considered have negative enthalpy of solution: these big molecules become less soluble with increasing temperature, which is compatible with 'molecular chameleon' behavior associated with intramolecular hydrogen bonding. The X-ray structures of many of these molecules reveal void spaces in their crystal lattices large enough to accommodate many water molecules when such solids are in contact with aqueous media. The water sorbed into crystals suspended in aqueous solution may enhance solubility by way of intra-lattice solute-water interactions involving the numerous H-bond acceptors in the big molecules studied. A 'Solubility Enhancement–Big Molecules' index was defined, which embodies many of the above findings.

Keywords

aqueous intrinsic solubility; Rule of 5 (Ro5), beyond Ro5 (bRo5); General Solubility Equation (GSE); Abraham Solvation Equation (ABSOLV); Partial Least Squares (PLS); Random Forest regression (RFR); intramolecular hydrogen bonding (IMHB); Solubility Enhancement–Big Molecules (*SEBM*)

Introduction

The aqueous solubility of compounds is an important physical property to assess in pharmaceutical research and development [1-4]. Solubility of potentially promising compounds not yet synthesized may be estimated computationally. Many methods for predicting solubility have been described [5-7], based on linear quantitative structure-property relationship (QSPR) approaches [8-16]. More recent methods have

evolved using machine learning statistical methods [17-20]. The molecular descriptors needed for these predictions are most often calculated from two-dimensional (2D) structures.

In the early 1990s, attrition due to poor oral bioavailability and pharmacokinetics (PK) was responsible for nearly 40 % of compounds being rejected in clinical studies [21]. Lipinski's Rule of 5 (Ro5) emerged as part of the effort to address critical issues underlying the high attrition [1]. The Ro5 guidelines suggest that compounds are more likely to be orally bioavailable if three or more of these rules are adhered to: molecular weight, $M_W \leq 500$ Da, calculated octanol-water partition coefficient, *clogP* ≤ 5, number of H-bond donors, *NHD* ≤ 5, and number of H-bond acceptors, *NHA* ≤ 10. High-throughput screening strategies of physicochemical properties of research compounds led to improvements. By the new millennium, attrition due to PK was reduced to below 10 % [22].

However, many recently approved drugs are larger, more lipophilic, and possess more H-bond acceptors, compared to drugs in the Ro5 chemical space [22]. Many of the newly-approved therapeutics are used in immunosuppression, treatment of infectious/viral diseases, and in oncology. Since 2014, an increasing number of 'beyond the Rule of 5' (bRo5) commentaries have stressed that the strict adherence to the Ro5 may result in lost opportunities [21-30]. Cell-permeable and orally-bioavailable drugs can be discovered far into bRo5 chemical space. Some of these drugs are derived from natural products, which appear to be better suited for the newer targets which possess large and flat binding sites. Nevertheless, concerns have been raised over the expected higher pharmacokinetic risks from bRo5 compounds: low solubility, poor cell permeability, increased cellular efflux, and extensive metabolism. Medicinal chemists have applied tactics to lessen some of the risks: (a) reducing or shielding polarity by N-methylation, or by bulky side chains, (b) selecting compounds with flexible rings structures allowing for conformational lability, and (c) selecting compounds which can reversibly form multiple *intra*molecular H-bonds (IMHB) to shield polar groups during passage across lipoidal cell barriers, in the manner of 'molecular chameleons' [26-31].

Although most of the bRo5 commentaries have emphasized permeability, absorption, and potency topics, Bergström *et al*. [25] focused on solubility aspects and the promising computational bio-pharmaceutical modeling strategies to help identify 'formulate-ability' during lead optimization and early development stages of bRo5 compounds. Caron *et al.* [29] considered case studies of kinetic solubility (measured in pH 7.4 phosphate buffer containing 1-5 % DMSO) of bRo5 molecules, in terms of the tendency to form IMHBs and their effect on solubility.

In our preceding study [20], three methods of solubility prediction were compared: (a) Yalkowsky's General Solubility Equation (GSE) [8], (b) Abraham Solvation Equation (ABSOLV) [11], and (c) Random Forest regression (RFR) [19] statistical machine learning. The linear ABSOLV and the RFR multiple decision-tree methods were trained with molecules in the *Wiki-pS_0* database. Thirty of the most important descriptors identified in the RFR analysis [20] were subjected to Principle Components Analysis (PCA). The scores plot had the appearance of a 'comet' – with a dense symmetrical core of Ro5 compounds about the origin of the first two principle components and a long sparsely-populated tail of big molecules queuing far into the *lower-right* quadrant. The molecules in the tail have high H-bond acceptor strength (*NHA*), topological polar surface area (*TPSA*), fraction of sp^3 carbons (FractionCSP3), and possess $M_W > 800$ Da – many of the recognized hallmarks of bRo5 chemical space.

The aim of the present study was to explore to what extent small molecules (mostly Ro5) can be used to predict the intrinsic aqueous equilibrium solubility of big molecules (all bRo5 drugs), i.e., can the 'head predict the tail'?

Methods

Computational models

Three computational approaches described below span from the theoretically sound and easy-to-use GSE, the sound and flexible ABSOLV, and the more accurate (but somewhat of a 'black box') RFR.

General Solubility Equation (GSE)

Expanding on the work of Irmann [32] and Hansch *et al.* [33], Yalkowsky and coworkers in 1980 developed and thereafter popularized the General Solubility Equation (GSE), to enable the prediction of the solubility of organic molecules in water [8,9,34-38]. Just two variables, melting point (*mp* in °C) and octanol-water partition coefficient, log *P*, are used in the equation to predict solubility (in log molar units):

$$\log S_0^{GSE} = 0.5 - \log P - 0.01(mp - 25) \quad (1)$$

Below, the derivation of Eq. (1) is briefly reviewed in terms of its underpinning assumptions to determine if any are incompatible with bRo5 molecules. Also, the thermodynamics of solubility are well cast by Eq. (1), which can apply to all models tested here. It is useful to start by dissecting aqueous solubility in terms of the Gibbs free energy based on the thermodynamic fusion cycle. The dissolution of a non-ionized crystalline substance suspended in water can be viewed in terms of two major contributions: (a) *crystal lattice* – energy has to be provided to break down the lattice to form a hypothetical 'supercooled' liquid (sliq), and (b) *solvation* – energy is released when the liquefied solute dissolves in water. The total solubility of the solute in water is the product of the above two contributions (lattice and solvation), which in logarithmic terms can be stated as the sum [36,37]:

$$\log S = -\frac{\Delta S_m}{2.3RT}(T_m - T) + \log S_w^{sliq} \quad (2)$$

Crystal lattice effect

The lattice contribution, first term on the right side of Eq. (2), arises from the application of the van't Hoff equation, where ΔS_m (kJ/mol·K) is the entropy of melting (fusion) and T_m is the melting point (in K units). By the 'Walden's rule' approximation [36,37], ΔS_m = 0.0565 kJ/mol·K for many organic compounds (*particularly for rigid planar molecules, but less so for spherical molecules*). At 25 °C, 2.3 *RT* = 5.706 kJ/mol·K. On substituting these constants, Eq. (2) at 25 °C reduces to:

$$\log S = -0.01(mp - 25) + \log S_w^{sliq} \quad (3)$$

Solvation effect

The solvation contribution, right-most term in Eq. (3), was investigated by Hansch and coworkers [33]. For 156 simple organic liquid solutes, they demonstrated that solubility correlated with octanol-water partition data as described by a Collander-like linear equation: log *S* = c_0 + $c_1$log *P*. Octanol, possessing nearly identical calculated H-bond donor and acceptor strength, was selected as a model organic solvent. For a series of aromatics, alkyl halides, and alkanes liquid solutes, c_0 intercepts were calculated to be +0.34, +0.83, and -0.25, respectively. The derived c_1 slope factors were -1.0 for aromatics, -1.22 for alkyl halides, and -1.24 for alkanes.

For a liquid solute, log *P* relates to the Gibbs free energy for the sum of solute-solute and water-water cohesive interactions minus twice the solute-water adhesive interactions [8]. For a liquid solute with the aqueous solubility of S_w^{liq} and the solubility in octanol as S_{oct}^{liq}, it can be approximated that $P = S_{oct}^{liq}/S_w^{liq}$ (*assuming activity equals concentration and that solute aggregates/micelles don't form* [39]). It follows that

if the slope c_1 = -1, then the intercept c_0 = log S_{oct}^{liq}.

Yalkowsky and coworkers surmised that c_0 = 0.5, by the following reasoning. Entropy of mixing favors complete miscibility of the two liquids; *i.e.*, the mole fraction = 0.5 [8]. (*This is likely to be valid for apolar* [37], *but may not be accurate for large polar molecules like those found in the bRo5 chemical space.*) Since the molar concentration of pure octanol is 6.32 M, the log S_{oct}^{sliq} = log (6.32x0.5) = 0.5 (*assuming the solute liquid density is near that of octanol*). On rearranging the log form of *P* defined as the solubility ratio, log S_w^{sliq} = 0.50 – log *P*. On substitution of the latter term into Eq. (3), the GSE, Eq. (1), is so derived.

The above considerations suggest that Eq. (1) may have possible limitations in bRo5 chemical space: (i) Lattice energy of rigid-planar molecules may be different from those of spherical molecules; (ii) Octanol as the model for the supercooled liquid solute may not be accurate for large polar or conformation-flexible molecules; (iii) Non-ideal activity may arise due to solute self-aggregation (e.g., dimer formation of vancomycin), possible micellization (e.g., ubiquinone, iodoxamic acid), and 'molecular chameleon' IMHB effects [29,31].

Using Calculated log P (clogP)

The original two variables (*mp*, log *P*) were taken to be experimental values. In a pharmaceutical research setting, such experimental values may not be available in early studies. It has become a common practice to use calculated values, *clogP*, in place of measured log *P* in Eq. (1). Apparently, the accuracy of GSE lessens, but only slightly. The use of calculated *mp* is less frequent, since the accuracy of such predicted values is thought to be relatively low. In the present investigation, experimental values were applied when available, and were calculated in a small number of instances [40].

Abraham Solvation Equation (ABSOLV)

Abraham and Le [11] amended the Abraham Solvation Equation [41] to predict intrinsic solubility (log molar):

$$\log S_0^{ABSOLV} = c_0 + c_1 A + c_2 B + c_3 S_\pi + c_4 E + c_5 V + c_6 A{\cdot}B \quad (4)$$

The independent variables are the five solute descriptors accounting for the transfer of solute from one phase to another: *A* is the sum of H-bond acidity (similar to *NHD*), *B* is the sum of H-bond basicity (similar to *NHA*), S_π is the dipolarity/polarizability (subscripted here, so as not to be confused with solubility), *E* is an excess molar refraction in units of $(cm^3{\cdot}mol^{-1})/10$, and *V* is the McGowan characteristic volume in units of $(cm^3{\cdot}mol^{-1})/100$.

In principle, the five solute variables could account for any shortcomings of just using *clogP*, as in Eq. (1). The *A·B* cross-term in Eq. (4) was intended to address *inter*molecular H-bond interactions between acid and base functional groups in the solid or liquid environment. Its inclusion, as an alternative to using the *mp* term in Eq. (1), was intended to improve the prediction accuracy of Eq. (4).

The c_0-c_6 coefficients in Eq. (4) are usually determined by multiple linear regression (MLR), trained on a set of intrinsic solubility values of a diverse collection of molecules. The five Abraham solvation descriptors may be calculated from 2D structure (introduced as a SMILES text or as coordinates in a 'mol' format) using the program ABSOLV [42] (*cf.*, www.acdlabs.com). In the present study, the seven MLR coefficients were re-determined using our own training data (*Wiki-pS_0* database), with log S_0 data weighted in the regression analysis according to estimated measurement errors [20].

Furthermore, we attempted to improve the accuracy of Eq. (4) when applied to big compounds, by introducing a nonlinear term,

$$\log S_0^{\text{mod-ABSOLV}} = \log S_0^{\text{ABSOLV}} + c_7 B^{+z} \quad (5)$$

Due to potentially high linear correlations among the descriptors, partial least squares (PLS) regression (open source package in R: https://cran.r-project.org/web/packages/pls/) was used (instead of MLR) to determine c_0-c_7 for given values of z. Several z values in the range 0.9 to 2.0 were tested; the best-fit exponent was selected as the minimum point in the fit of PLS root-mean-square error (RMSE) vs. z.

Random Forest regression

Of the new machine-learning statistical approaches, the Random Forest regression (RFR) method is thought to be one of the most accurate in predicting solubility [17-20]. RFR can be employed 'off the shelf,' requiring only minimal learning [19]. The provided default 'tuning' parameters are nearly optimal. However, its 'black box' nature makes the outcome of the analysis challenging to interpret in terms of the descriptors used, even when the most important descriptors are quantitatively ranked in RFR.

The method was introduced in 2001 by Brieman [43], and is implemented in the open-source 'randomForest' library for the R statistical software [44-46]. RFR works by constructing an ensemble of hundreds of decision trees [47]. The tutorial chapter by Walters [19] is highly recommended as a means to get started with RFR.

The first applications of RFR to predict solubility appeared in 2007 [17,48]. Schroeter *et al.* [48] used S_w and S_{pH} data to train a RFR method, using ~4000 measurements mostly taken from secondary sources [12,49,50] and some from in-house (Bayer Schering Pharma) sources. For the Huuskonen data [12] as test set, RMSE = 0.66 log unit (n=1290) was reported. For the solubility data in the domain of applicability (DOA) matching that of research compounds (10^{-3} to 10^{-7} M solubility), the RFR method indicated RMSE ~ 0.85 log.

In the Palmer *et al.* [17] RFR study, aqueous solubility values of 998 structurally diverse druglike solid organic compounds were gathered from similar secondary sources [12,51,52]. The authors used the Molecular Operating Environment (MOE) [53] to generate 126 2D (*clogP*, *MR*, charged-surface properties, atom, group, and H-bond counts, connectivity and topological indices) and 36 3D (total potential energy, electrostatic contributions, molecular shape, and solvent-accessible surface area) descriptors. Randomly splitting the entire data into a training set (70 %) and an internal validation set (30 %) produces a good measure of the model predictivity of compounds not included in the training set: $r^2 = 0.89$, RMSE = 0.69 log, n = 330.

More recently, Walters [19] critically compared the Huuskonen thermodynamic S_w values (n = 1274) [12], the Llinas *et al.* thermodynamic S_0 values (n = 94) [54] and PubChem (n=1000) kinetic high-throughput solubility [55] databases using the RFR framework. Avdeef [20] applied RFR, trained with 6355 log S_0 values, to predict the solubility of four well-publicized small external test sets, occasioning in RMSE from 0.66 to 1.05 log.

Data

Wiki-pS_0 database

The intrinsic aqueous solubility database *Wiki-pS_0* (*in-ADME* Research) [20,56] was used. It now contains 6473 log S_0 (log molar) entries, based on measured aqueous solubility values of 3065 different compounds (excluding agrochemicals) collected from 1415 cited references. The most reliable published data had been determined by the saturation shake-flask (SSF) method, particularly when measured *as a function of pH*. In the majority of the cases, the literature data were further processed, using *p*DISOL-X (*in-ADME* Research) [56-61], to extract intrinsic solubility (S_0) values from reported aqueous free-acid/base or salt solubilities

(S_w), or solubilities at specified pH (S_{pH}), or log *S*-pH profiles. All of the molecules are solids at room temperature (except propofol). There are 1127 log S_0 entries derived from 10167 individual-pH log *S* measurements. About half of the data sources originate from secondary listings and the rest are from primary sources. In the case of secondary sources, the citations to the original work were generally available, and in many cases were consulted for clarifications. Melting points are included in the database. When measured *mp* were not found (19 % of entries), *mp* were calculated by the Lang and Bradley method [40] in the QsarDB open repository of data and prediction programs (http://qsardb.org/repository/-predictor/10967/104?model=rf).

Physicochemical properties of the big molecules

In this study, the compounds in *Wiki-pS$_0$* were divided into two groups: 'big' (M_W ≥ 800 Da, structures in **Appendix**) and 'small' molecules. The demarcation was motivated by the shape of the principle components scores plot (Fig. 10 in [20]). There are 31 molecules (58 log S_0 entries) in the 'big' set. Table 1 lists their characteristic properties. Figure 1 shows the distribution of big-molecule log S_0 values. On the average, the big molecules are less soluble (-4.52) than the small molecules (-3.12).

Table 1. Big-molecule (bRo5) physicochemical properties

Compound	*log S_0*	SD	n	M_W	*mp*	*clogP*	ΔH^0_{sol}	*NHA*	*NHD*	*nROT*	*A*	*B*	S_π	*E*	*V*	Ref
Amphotericin B	-3.52	0.69	2	924	179	0.71	-33	17	12	3	3.55	5.99	5.12	4.47	7.12	[63,64]
Anidulafungin	-4.36	0.45		1140	250	-0.93	0	17	14	14	3.67	8.22	10.4	7.8	8.4	[63]
Bryamycin	-4.14	0.29	3	1665	210	0.77	-16	31	17	12	4.47	11.56	14.55	10.51	11.65	[64-66]
Cyclosporine A	-5.03	0.16	6	1203	151	3.27	-40	12	5	15	1.25	7.61	10.16	4.23	10.02	[63,67-69]
Dactinomycin	-3.22	0.16	2	1255	242	0.73	-10	18	5	8	1.36	8.49	11.45	6.12	9.49	[70]
Docetaxel	-5.14	0.05	2	808	232	3.26	-12	14	5	8	1.03	4.01	4.18	3.47	5.92	[63,71]
Everolimus	-5.02	0.58		958	138	6.20	-36	14	3	9	0.63	4.73	4.73	3.29	7.68	[63,72]
Gramicidin A	-4.16	0.41		1882	229	4.37	-31	17	21	53	5.57	11.3	17.8	10.26	14.81	[73]
Gramicidin S	-3.89	0.35		1141	169	1.23	-14	12	10	16	2.46	7.42	10.69	5.38	9.12	[64]
Iodipamide	-5.47	0.67		1140	307	6.85	45	4	4	9	2.25	1.87	4.84	5.86	4.38	[50]
Iodoxamic Acid	-5.49	0.36		1288	224	6.13	35	8	4	19	2.25	2.73	5.48	6.01	5.46	[74]
Ivermectin	-5.56	0.39	5	875	140	5.60	-33	14	3	8	0.68	4.23	3.21	3.24	6.72	[72,75-78]
Leuprolide	-3.15	0.20		1209	153	-1.44	-11	14	15	32	4.25	8.66	11.75	7.23	9.21	[63]
Nafarelin	-5.61	0.52		1323	189	-1.62	2	15	16	33	4.74	9.32	13.46	8.93	9.88	[79]
Nystatin	-4.1	0.39		926	170	0.94	-35	17	12	3	3.55	5.93	5.02	4.32	7.16	[80]
Oxytocin	-1.2	0.17		1007	164	-3.61	2	15	12	17	3.96	7.67	11.34	5.9	7.47	[81]
Paclitaxel	-6.53	0.14	2	854	216	3.74	-3	14	4	10	0.9	4.13	5.22	4.05	6.2	[82,83]
Paclitaxel analog12	-5.48	0.67		808	187	3.20	-14	14	5	8	1.03	4.01	4.18	3.47	5.92	[71]
Paclitaxel analog17	-4.52	0.48		802	179	3.32	-24	14	5	9	1.03	3.97	3.67	2.87	6.02	[71]
Paclitaxel analog23	-5.78	0.73		807	187	3.23	-13	13	6	8	1.33	4.15	4.31	3.63	5.96	[71]
Rapamycin	-5.55	0.69		914	184	6.18	-30	13	3	6	0.63	4.51	4.57	3.26	7.34	[84]
Rifabutin	-4.09	0.66	3	847	176	4.62	-9	14	5	4	1.31	4.39	4.43	4.24	6.47	[63,85,86]
Rifampicin	-2.96	0.27	9	823	164	4.34	-6	15	6	4	2.55	4.66	4.67	4.73	6.21	[87-91]
Roxithromycin	-3.98	0.37		837	120	2.21	-47	17	5	13	1.05	5.12	2.9	2.58	6.55	[63]
Solithromycin	-6.23	0.14		845	189	4.60	-12	15	2	11	0.35	4.46	4.7	3.67	6.44	[92]
Stevioside	-2.83	0.17		805	198	-2.94	-21	18	11	9	2.74	5.49	4.29	4.25	5.67	[93]
Tacrolimus	-5.42	0.54	2	804	126	4.64	-27	12	3	7	0.71	3.98	3.98	2.82	6.38	[63,94]
Telithromycin	-3.02	0.17		812	188	4.93	-13	14	1	11	0.12	4.40	4.53	3.49	6.32	[63]
Temsirolimus	-5.06	0.59		1030	134	5.72	-39	16	4	10	1.02	5.07	5.25	3.46	8.17	[63]
Ubiquinone	-7.56	1.65	2	863	48	17.85	-53	4	0	31	0.00	2.20	1.16	2.15	7.95	[95,96]
Vancomycin	-2.13	0.14		1449	175	0.11	-22	25	19	13	5.81	10.56	12.32	9.73	9.88	[97]

[a] log S_0 averaged for n > 1 sources (references in last column). SD is the estimated standard deviation in the measured value. ΔH^0_{sol} (kJ/mol) are calculated enthalpies of solution (see text). *nROT* is the number of rotatable bonds in the molecule. For the other terms, *cf. Abbreviations and definitions*.

Figure 2 shows the trend between measured log S_0 and *clogP* (calculated in RDKit [62]: Wildman-Crippen sum of atomic contributions – cf., *Abbreviations and definitions*) for the two groups of molecules. The

scatter is substantial. Nevertheless, the small molecules (green circles) show the expected -1 slope, whereas the big molecules (red squares) show an apparent slope of -0.239. This is an important characteristic differentiating the two groups.

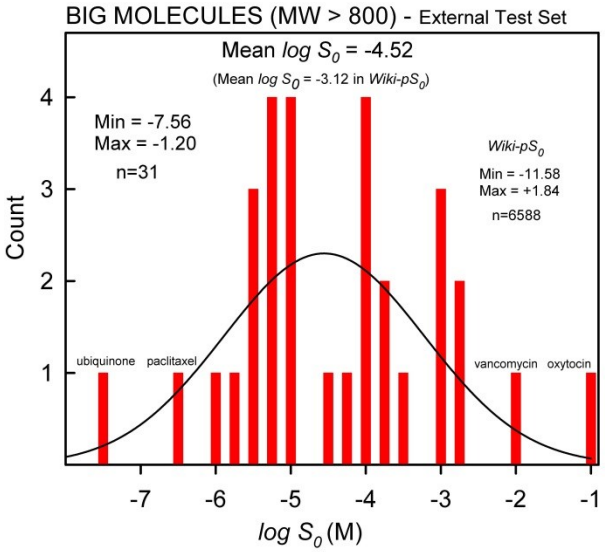


Figure 1. Distribution of the big-molecule intrinsic aqueous solubility values in *Wiki-pS_0*

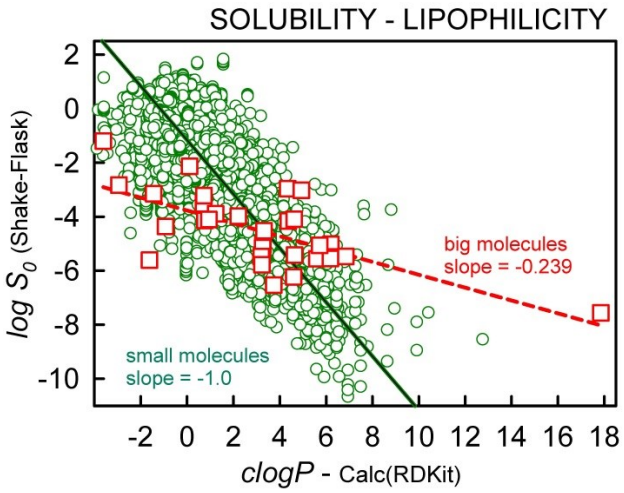


Figure 2. Plot of log S_0 versus octanol-water partition coefficient, *clogP*, calculated using the RDKit software [62]. Squares refer to big molecules; circles refer to small molecules

Characteristics of the big molecules

Figure 3 shows property distributions as possible indicators of bRo5 'big-drug-likeness.' Frame (a) shows the *clogP* distribution: on the average, *clogP* of the 'big' set (3.17) is greater than that of the 'small' set (1.89) [20]. Frame (b) shows the distribution of molecular weights about the mean value 1034 Da (compared to 280 Da in the entire set [20]). Frame (c) considers H-bonding characteristics. The red bars (tallest near 5) refer to H-bond donor counts (*NHD*). The black bars (tallest near 15) refer to H-bond acceptors (*NHA*). In the small-molecule set, the *NHA* and *NHD* groups overlap considerably, as illustrated elsewhere [20]. But, in the big-molecule set (Fig. 3c, Table 1) the *NHA* and *NHD* distributions are wider apart: the acceptor count increases, but not so much the donor count. This is an important characteristic differentiating the big-small molecule groups.

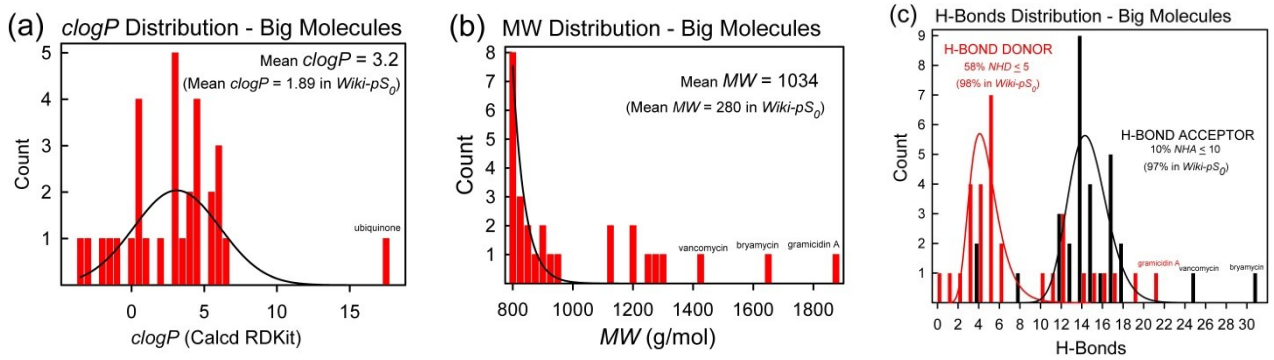


Figure 3. Big-molecule property distributions: **(a)** *clogP*, **(b)** molecular weight (M_W), and **(c)** number of H-bond donors (*NHD*) and acceptors (*NHA*). The separation between the groups is greater than that found in small molecules [20].

Results and discussion

GSE applied to big compounds

Hydrophilicity (*clogP*) effect in big lipophilic molecules

The linear dependence of log S_0 on *clogP* (Fig. 2) was further analyzed in the context of Eq. (1). MLR of the SD-weighted log S_0 data confirmed the large difference in *clogP* contributions in the two groups of molecules:

$$\log S_0^{SMALL} = -0.28(\pm 0.04) - 0.83(\pm 0.01)\ clogP - 0.0062(\pm 0.0002)\cdot(mp - 25) \quad (6a)$$

$$r^2 = 0.60,\ RMSE = 1.16\ log,\ MPP = 37\ \%,\ n = 6392$$

$$\log S_0^{BIG} = -1.77(\pm 0.93) - 0.40(\pm 0.07)\ clogP - 0.010(\pm 0.005)\cdot(mp - 25) \quad (6b)$$

$$r^2 = 0.61,\ RMSE = 0.89\ log\ unit,\ MPP = 36\ \%,\ n = 31$$

The extent of 'correct' predictions is defined here by MPP (measure of prediction performance: percentage of the absolute residuals ≤ 0.5 log unit).

Apparently, *crystal lattice* contributions are not appreciably different in the two groups of molecules; the refined temperature coefficients in the two sets are close to the GSE value (-0.01) in Eq. (1). Hence, solution-phase interactions appear to dominate solubility [98].

The intercept constants suggest that big-molecule 'supercooled' liquid solutes are less miscible in octanol by 1-2 orders of magnitude than suggested by the original Yalkowsky analysis [8,37]. The intercepts in Eqs. (6a) and (6b) are nearer to those of alkane solutes found in the Hansch *et al.* [33] study, compared to the constant in Eq. (1). For the big molecules, the highly negative intercept (*i.e.*, *decreased* solubility of the supercooled liquid in the octanol phase) depresses the solute water solubility by a constant amount.

Countering that, the -0.4 slope factor lessens the contribution of lipophilicity to the calculated solubility of big molecules. The net result is that the traditional GSE overpredicts S_0 for big molecules with experimental solubility above ~50 µM *(e.g., oxytocin, nafarelin)*, and underpredicts S_0 below the crossover point *(e.g., everolimus, telithromycin).*

General Solubility Equation (GSE)

Figure 4a shows the relationship between the measured solubility of small molecules and that calculated by the classic ('untrained') GSE. (Permanently-charged quaternary amines and big molecules are excluded in the training.) The r^2, RMSE, MPP statistics are nearly identical to those associated with Eq. (6a), suggesting that 'training' does not improve the GSE predictivity for small molecules.

However, the performance of the 'untrained' GSE degrades when the equation is applied to the big molecules, as shown in Figure 4b, with $r^2 = 0.0$, RMSE = 3.0 log (2.3 without ubiquinone), and MPP = 16 %.

Figure 4c plots the big-molecule 'trained' GSE result (*cf.*, Eq. 6b). *Note that this is not the equivalent of Ro5 molecules predicting the solubility of bRo5 molecules.* Rather, it highlights the hydrophilicity solvation effect of big molecules discussed in the preceding section. The traditional GSE requires adjustments when it comes to predicting the solubility of big molecules (*cf.*, Fig. 4b).

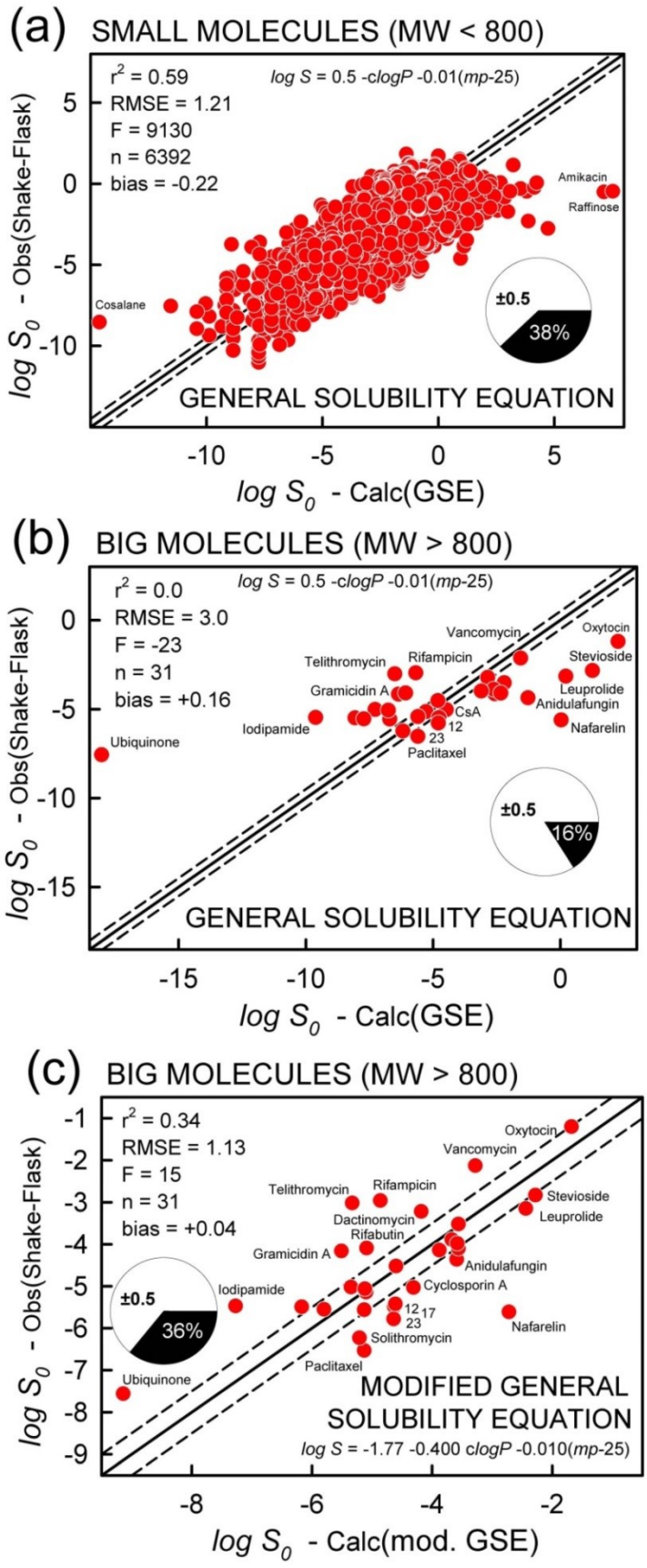


Figure 4. The log *P* in the figure refers to calculated octanol-water partition coefficients, *clogP*. The solid diagonals are the identity lines, and the dashed lines refer to ±0.5 log deviations. The MPP pie charts refer to percentage of 'correct' prediction, based on absolute residuals being ≤ 0.5 log. The prediction of log S_0 values of **(a)** small molecules and **(b)** big molecules using the classical General Solubility Equation (numeric compound labels are of paclitaxel analogs). **(c)** When using just the big-molecule data, the three constants in the GSE (Eq. 1) subjected to MLR analysis (*cf.*, Eq. 6b) produce the modified GSE, which is valid only for molecules with *MW* > 800 Da. There are not enough big molecules in the *Wiki-pS_0* database to test the predictiveness of Eq. (6b).

Abraham solvation model (ABSOLV) – weighted MLR to predict solubility of big compounds

Abraham linear equation for solubility prediction

The ABSOLV MLR analysis of the small-molecule data, weighted according to the estimated errors in the measured log S_0 values produced the equation

$$\log S_0^{ABSOLV} = -0.15(\pm0.05) + 0.20(\pm0.06)A + 1.78(\pm0.04)B - 0.11(\pm0.04)S_\pi - 1.17(\pm0.03)E - 1.49(\pm0.03)V + 0.01(\pm0.03)A{\cdot}B \quad (7)$$

$r^2 = 0.67$, RMSE = 1.08 log unit, MPP = 38%, n = 6392

The plot of measured log S_0 as a function of the calculated values according to Eq. (7) is shown in Figure 5a. This trained ABSOLV model only slightly outperforms the small-molecule untrained/trained GSE model (Fig. 5a/Eq. 7 compared to Fig. 4a/Eq. 6a). The $A{\cdot}B$ cross term contribution appears to be negligible.

The application of Eq. (7) to the big-molecule set produced unidirectionally-skewed plot, as shown in Figure 5b. According to the ABSOLV model trained on small molecules*, the solubility of all big molecules is underpredicted.* For example, the gramicidin A measured log S_0 = -4.16 ±0.41 is underestimated by 10 orders of magnitude. Vancomycin is underestimated by nearly 5 orders of magnitude.

An effort was made to improve the fit. A distinguishing characteristic of big compounds is that they contain a high level of H-bond *basicity* (*B*) character (Table 1). We tested several nonlinear contributions of the *B* descriptor, with the aim of amplifying its uniquely high positive impact on solubility (Eq. 7). In order to avoid difficulties due to descriptor correlations, PLS regression was used in place of MLR. The modified model, depicted in Figure 5c, is the best improvement that was found. The modified solvation model consisted of an additional nonlinear term, B^{+z}, with z > 1. The best-fit value of z was determined to be 1.11. This new descriptor was expected to amplify the positive H-bond acceptor contribution in Eq. (7) in the case of big molecules. Other modifications were explored, but only the latter descriptor appeared to improve ABSOLV to a level slightly better than that of the classic GSE (Fig. 4b).

On inspection, the systematic errors in Figure 5b were found to correlate with the number of rotatable bonds (*nROT*): $\log S_0^{Obs} - \log S_0^{ABSOLV} = 0.75 + 0.13\ nROT$, with $r^2 = 0.44$ and RMSE = 1.62. Adding 0.75 + 0.13 *nROT* to Eq. (7) reduced the RMSE from 3.4 to 1.6 and the bias from 2.6 to 0.13 (r^2 remained unchanged). However, this did not result in a significantly improved training-set model when *nROT* was added to the list of ABSOLV descriptors in a repeated PLS analysis. Flexibility appears to be important, but *nROT* is not significantly predictive in the *training* process. Caron *et al.* [30] demonstrated that *nROT* may have limitations because it neglects the contribution to flexibility from cyclic fragments in big molecules.

Random forest regression using RDKit combined with Abraham descriptors and melting points

Descriptors

For the RFR model building, the 190 RDKit [62] descriptors (excluding those which were zero for all compounds) were combined with the *mp* and the ABSOLV descriptors. The *Abbreviations and definitions* section below identifies and defines the most important descriptors used in the RFR algorithm.

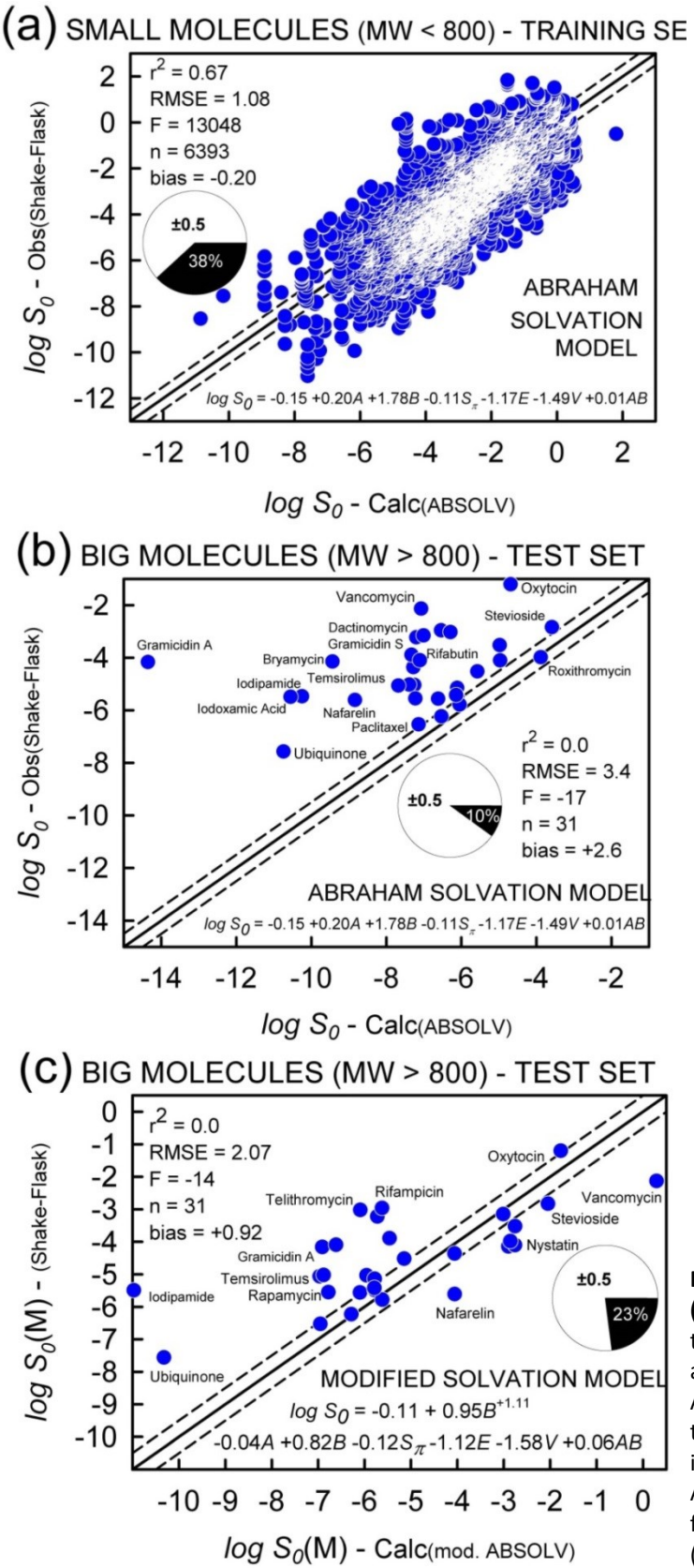


Figure 5. The prediction of log S_0 values of **(a)** small molecules and **(b)** big molecules using the Abraham Solvation Equation (ABSOLV). **(c)** An additional nonlinear descriptor was added to the ABSOLV equation (cf., Eq. 5), which was then trained with the small-molecule set. This improved the prediction accuracy of the modified ABSOLV equation. The pie chart denotes MPP, the fraction of 'correctly' predicted molecules (absolute residuals ≤ 0.5 log unit).

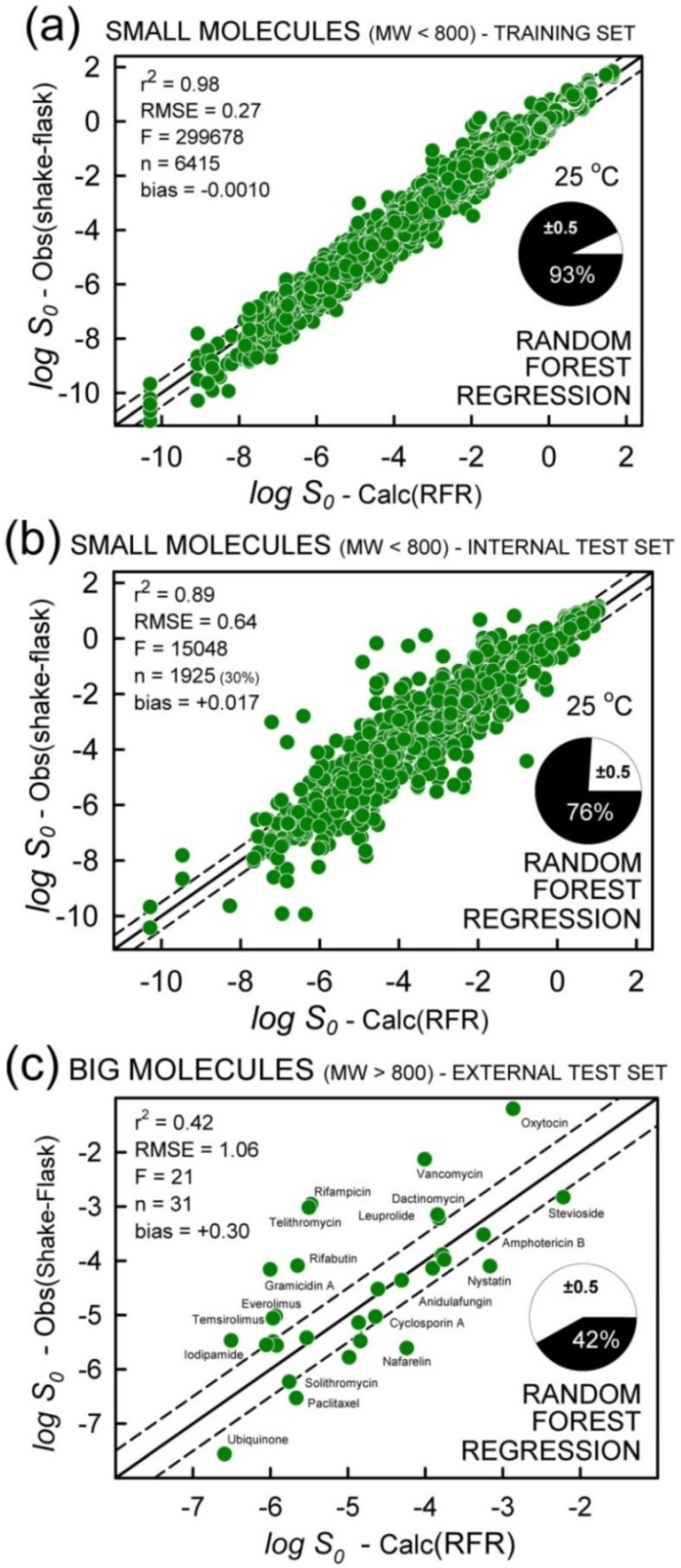


Figure 6. Random Forest regression analysis. (**a**) Training set using the small molecules. (**b**) Internal validation test set, based on 30% of the small molecules randomly selected. **(c)** External test set prediction of big molecules, not used in the method training.

Training set and internal validation

Figure 6a shows the small-molecule training-set RFR analysis, resulting in the metrics: r^2 = 0.98, RMSE = 0.27 log, bias = -0.001. This quality of fit indicates how well the model can incorporate the information presented by the descriptors and relate it to solubility in the training set [18]. The internal validation set of 1925 small-molecule solubility values (30 %), randomly selected by the method, better indicates the ability of RFR to predict external test compounds which are unknown to the training process.

Figure 6b shows the *internal* validation test set prediction results: r^2 = 0.89, RMSE = 0.64 log, bias = 0.017. This performance is to be expected for *external* test molecules, *provided they are adequately represented in the chemical space of the training set.*

Big-molecule external test set prediction

Figure 6c illustrates the degree to which the RFR method, trained by small molecules, can predict the solubility of big molecules. The relative accuracy of the prediction (r^2 = 0.42, RMSE = 1.06 log, MPP = 42 %) evidently exceeds that of the GSE and ABSOLV methods (predictive r^2 = 0 and RMSE > 2 log). To wit, small molecules in the training set provide enough subtle 'clues' for the method to extract a sensibly accurate prediction of big-molecule solubility.

The most important descriptors (RDKit terminology – *cf. Abbreviations and definitions*) were found to be *MolLogP >> MolMR > Ipc >> LabuteASA > BertzCT > HeavyAtomMolWt > MolWt > Chi1 > SMR_VSA7 > mp > Chi0v > SMR_VSA10 > PEOE_VSA7 > fr_benzene > Chi1v > E > Chi4v > B*. Some of these are highly intercorrelated. There were additional ~100 descriptors that played lesser and somewhat hidden roles.

In RFR, relationships between descriptors and the model are difficult to extract, and the influence of each compound property on calculated solubility cannot be readily deduced [98]. A major disadvantage to a medicinal chemist is that the RFR result does not directly suggest how compounds could be altered to increase/decrease their solubility. Unlike the intuitive and appealing descriptors in GSE and ABSOLV, many of the RDKit descriptors used are more abstract and not easy to interpret regardless of the modeling method [99].

Solubility Enhancement–Big Molecules (SEBM)

Table 2 lists the calculated log S_0 values of the big molecules. The last column lists the 'Solubility Enhancement–Big Molecules' – the ratio of the observed S_0 to that calculated by the ABSOLV approach (*cf.*, Eq. 7). The scale quantifies the big-molecule solubility enhancement not predicted by small molecules. A similar ratio using the classic GSE indicates two zones: (a) 'enhancement' for compounds to the left of the identity diagonal line in Figure 4b, and (b) 'attenuation' for compounds to the right of the line. The GSE zoning is directly linked to the partition coefficient (cf., Fig. 4c). The ABSOLV-based *SEBM* assigns a unified enhancement to all molecules, and separately addresses the role of H-bonding and molecular size (as well as the other Abraham solvation descriptors), whereas the GSE confines the relationship mainly to one descriptor – *clogP*, whose value may not be accurately calculated or measured for large molecules (*e.g.*, ubiquinone).

Figure 7 is a plot of log *SEBM* as a function of *nROT*. Although noisy, a trend is evident. The unfilled circles in the figure refer to two external test compounds, big molecules recently approved as drugs: givosiran [100] and tenapanor [101], with M_W 1711 and 1145 Da, respectively.

Factors that may shed light on the unusual intrinsic aqueous solubility of big molecules

Lipophilicity behavior of big vs. small molecules differs

From the GSE analysis, the notable characteristic distinguishing small from big molecules is the dependence on lipophilicity (Fig. 2, Eq. 6b). Big *lipophilic* molecules (ubiquinone, iodipamide, everolimus, telithromycin) are more soluble and big *hydrophilic* molecules (oxytocin, stevioside, nafarelin) are less soluble than predicted by the traditional GSE (*cf.*, Fig. 4b). The empirical Eq. (6b) compensates for this tilted relationship with the less negative (-0.4) *clogP* factor and the more negative intercept factor (-1.77) than those in the GSE (-1 and 0.5, resp.), as illustrated in Figure 4c. The solubility-partition correlation using

octanol works well for small molecules, but octanol does not appear to match the big-molecule solubility-partitioning behavior in the same way, either because the big molecules are uncharacteristically more soluble in water (extra strong solute-water adhesive interactions) and/or less soluble in the octanol phase (extra strong solute-solute cohesive interactions).

Table 2. Calculated log S_0 and 'Solubility Enhancement-Big Molecules'

Compound	*Obs*	GSE [a]	ABSOLV [b]	RFR [c]	SEBM [d]
Amphotericin B	-3.52	-1.75	-4.86	-3.25	22
Anidulafungin	-4.36	-0.82	-7.16	-4.30	628
Bryamycin	-4.14	-2.12	-9.33	-3.90	156099
Cyclosporine A	-5.03	-4.03	-7.12	-4.65	123
Dactinomycin	-3.22	-2.40	-7.07	-3.82	7080
Docetaxel	-5.14	-4.83	-5.97	-4.86	7
Everolimus	-5.02	-6.83	-7.25	-5.93	172
Gramicidin A	-4.16	-5.91	-14.26	-6.00	12667592811
Gramicidin S	-3.89	-2.17	-7.20	-3.78	2056
Iodipamide	-5.47	-9.17	-10.11	-6.51	43888
Iodoxamic Acid	-5.49	-7.62	-10.42	-5.96	85035
Ivermectin	-5.56	-6.25	-6.48	-5.92	8
Leuprolide	-3.15	0.67	-6.89	-3.84	5514
Nafarelin	-5.61	0.48	-8.73	-4.24	1314
Nystatin	-4.10	-1.89	-4.84	-3.16	5
Oxytocin	-1.20	2.72	-4.57	-2.87	2350
Paclitaxel	-6.53	-5.15	-7.00	-5.67	3
Paclitaxel analog12	-5.48	-4.32	-5.97	-4.84	3
Paclitaxel analog17	-4.52	-4.36	-5.44	-4.61	8
Paclitaxel analog23	-5.78	-4.35	-5.91	-4.98	1
Rapamycin	-5.55	-7.27	-7.09	-6.05	35
Rifabutin	-4.09	-5.63	-6.97	-5.65	765
Rifampicin	-2.96	-5.23	-6.40	-5.47	2780
Roxithromycin	-3.98	-2.66	-3.74	-3.76	1
Solithromycin	-6.23	-5.74	-6.39	-5.76	1
Stevioside	-2.83	1.71	-3.45	-2.22	4
Tacrolimus	-5.42	-5.15	-6.01	-5.53	4
Telithromycin	-3.02	-6.06	-6.15	-5.50	1340
Temsirolimus	-5.06	-6.31	-7.54	-5.97	301
Ubiquinone	-7.56	-17.58	-10.60	-6.59	1102
Vancomycin	-2.13	-1.11	-6.97	-4.01	69548

[a] Calculated log S_0 in Fig. 4b. [b] Calculated log S_0 in Fig. 5b. [c] Calculated log S_0 in Fig. 6c.
[d] Observed S_0 divided by the value calculated in ABSOLV analysis: SEBM = S_0^{Obs}/S_0^{ABSOLV}.

Ermondi *et al.* [27,28] estimated lipophilicity of nine bRo5 drugs using the well-tested small-molecule ElogP and the new 'block relevance' BRlogP chromatographic methods, to investigate the role played by molecular flexibility. They also subjected the molecules to conformational analysis, in order to calculate lipophilicity of various conformers. ElogP chromatographic method appeared to provide an environment in which flexible compounds are driven to assume a more 'folded' apolar conformation (as expected in octanol), whereas the BRlogP method favored an 'extended' polar conformation for such molecules (as expected in water). Lipophilicity of bRo5 compounds strongly depends on their chameleonic properties: closed form preferred in apolar environments and open form in aqueous media. It is suggested that a non-

traditional lipophilicity scale is needed for many bRo5 compounds, which takes into account the solute conformational flexibility and the polarity of the dissolution media [27,28].

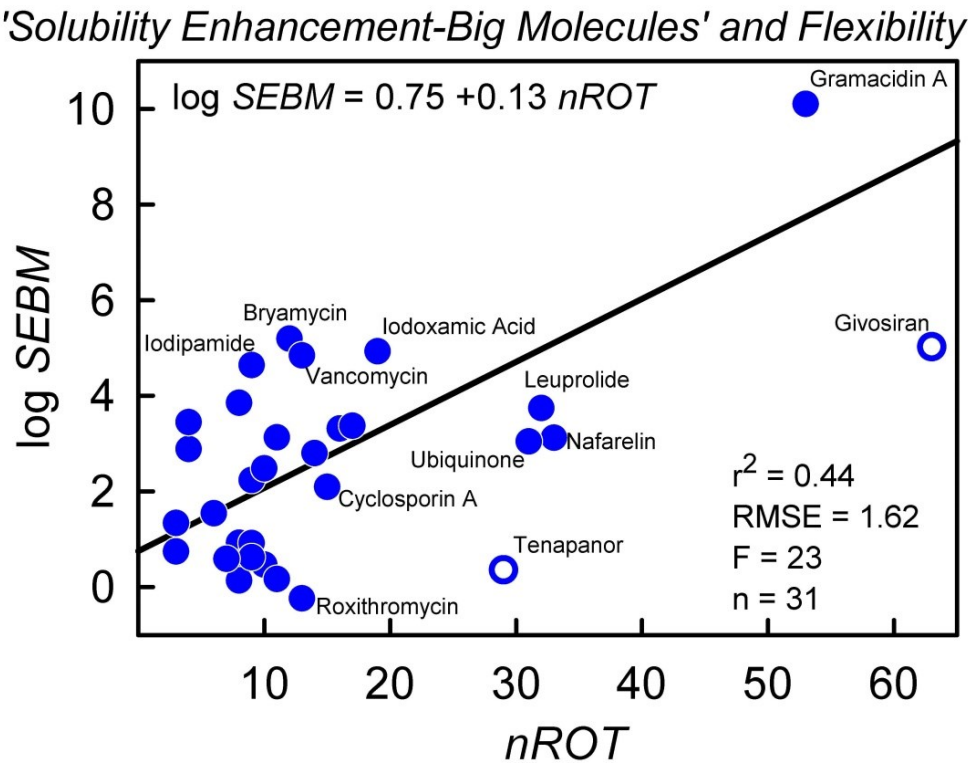


Figure 7. Logarithm of the Solubility Enhancement–Big Molecules as a function of the number of rotatable bonds: log *SEBM* = log S_0^{Obs} – log S_0^{ABSOLV}.

Crystal structures of the big molecules and their 'hydration' in the solid state

The Appendix shows the 2D structures of the big molecules selected for the study. Many of these are derived from natural products, possessing flexible cyclic and polycyclic components in there structures. The crystal structures of only about half of these molecules have been deposited in the Cambridge Crystallographic Data Centre (CCDC). In many of the compounds the crystal lattices contain *internal* void space that may be filled non-stoichiometrically with water, either fixed at certain positions by H-bonds, or mobile in channels. There are numerous sites with which water could interact by donating H-bonds, possibly competing with acceptor groups in IMHB networks, to form stoichiometric hydrates.

Since most of the crystals chosen for structure determination were grown in semi- or non-aqueous media, the reported X-ray structure of the molecules may not precisely reflect the conformational state found in crystals under conditions where they are equilibrated in aqueous solution, or of dissolved molecules in their unhindered states of hydration. In an exceptional study, the aqueous environment was well mimicked in the synchrotron X-ray determination of the structure of the glycopeptide antibiotic vancomycin [102]. Vancomycin crystals grown by the 'hanging drop' method were transferred into a pH 4.6 acetate buffer solution containing 2.2 M NaCl and a cryoprotectant solvent. The suspension was then flash frozen for the low-temperature data collection. The crystal lattice was found to contain an H-bonded dimer of vancomycin, 2 chlorides, 1 acetate, and *105 solvent water molecules* in the asymmetric unit. The organization of the lattice water was not described in the publication.

Zografi and coworkers have conducted pioneering research [103-106] on the influence of adsorbed and absorbed water on the solid state properties of crystalline/amorphous solids, including multicomponent forms such as drug salts and cocrystals. The presence of stoichiometric/nonstoichiometric water in

crystalline solids is expected to impact the thermodynamic activity of the solid and thus could affect equilibrium solubility.

Enthalpy of solution of big molecules is negative

A computational procedure to normalize solubility data determined at various temperatures to values at a 'reference' temperature (*e.g.*, 25 °C) was recently described [107]. The enthalpies of solution, ΔH_{sol}, were predicted from 2D structure, from which the temperature dependence of log S_0 is calculated as:

$$\log S_0^{\text{ref}} = \log S_0^T - 0.175\ \Delta H_{\text{sol}} \cdot \left(1 - \frac{298.15}{T}\right) \quad (8)$$

Small molecules, especially weak acids, generally have *positive* enthalpies of solution. For example, naproxen has the calculated ΔH_{sol} = +29 kJ/mol. Its solubility at T = 310.15 K (37 °C) is log S_0 = -4.03. The value decreases by 0.2 log (Eq. 8) to -4.23 at 25 °C.

Particularly interesting in light of the current study is that just about all the big molecules studied here have *negative* enthalpies of solution (Table 1). The prediction equations [107], based on Abraham descriptors,

$$\Delta H_{sol}{}^{\text{ACIDS}} = 17.1 + 0.024\ mp + 5.8\ A - 1.9\ B + 3.0\ S_\pi + 2.7\ R - 0.4\ V - 4.3\ AB \quad (9a)$$

$$\Delta H_{sol}{}^{\text{NON-ACIDS}} = 11.2\ I_B + 7.3\ I_N + 8.9\ I_Z + 0.039\ mp + 1.9\ A - 9.1\ B + 6.5\ S_\pi + 10.2\ \text{R} - 8.7\ V - 0.9\ AB \quad (9b)$$

indicate that high H-bond basicity (*B*) and large molar volumes (*V*) correlate with negative (exothermic) enthalpies of solution. Acids (*e.g.*, iodipamide and iodoxamic acid) are less inclined to be exothermic, compared to non-acids. (In Eq. (9b), the indicator indices default to zero, except that for a basic molecule, I_B =1; for a neutral molecule, I_N = 1; for an ampholyte, I_Z =1.)

For big basic molecules this means that as temperature rises, the solubility *decreases*. If water is sorbed into the void/channel spaces of crystals containing big molecules, then the negative enthalpy could be rationalized in terms of H-bonding effects. Since water H-bonds weaken with rising temperature, the proportion of the 'extended' (water soluble) conformer may shift in favor of the 'folded' conformer, which is expected to be less soluble in water. With weakened water binding, the intramolecular H-bond interactions may stabilize the structure in a folded form. In this way, negative enthalpy is consistent with the conformational flexibility of 'molecular chameleons' [26-31], and highlights the possible role of sorbed water influences on solubility.

Conclusion

We have shown that traditional approaches to the prediction of solubility of big molecules (bRO5) do not work very well, unless modified. On the other hand, the RFR method works reasonably well, but it is not easy to understand what specific contributions the various molecular descriptors provide to the overall prediction.

We attempted to link the Solubility Enhancement–Big Molecules (*SEBM*) to other physicochemical properties. A trend was evident in the log *SEBM* vs. *nROT* plot, suggesting that flexibility appears to enhance the solubility of big molecules. In the SGE model, a different lipophilicity scale might improve the performance of the approach, as empirically suggested in Figure 4c and as suggested by the chromatographic studies of Ermondi *et al.* [27,28]. The introduction of a nonlinear H-bond basicity term in the case of the ABSOLV approach is empirical, and it is not clear how to relate it to first-principle thermodynamic treatment.

Most of the big molecules have negative enthalpy of solution. That is, their solubility *decreases* with increasing temperature. This hints of an important H-bonding role for water sorbed into the solid state of the large molecules. Such molecules appear to have void spaces in their crystal lattices, sufficient to accommodate many water molecules under equilibrium conditions *with crystals wet by aqueous media*.

The observation that the RFR method appears to work encourages us to further search for 3D-based descriptors arising from 'conformational lipophilicity' analysis akin to that developed by Caron and coworkers [27-30]. The accurate prediction of the solubility of newly approved molecules originating from the bRo5 chemical space would help in selecting/prioritizing candidates in early drug discovery, particularly if the bRo5 molecular basis of solubility were better understood.

Abbreviations and definitions

S_0	"intrinsic" solubility (i.e., the solubility of the *uncharged* form of the compound)
RMSE	root-mean-square error: RMSE = $[\ 1/n\ \Sigma_i\ (y_i^{obs} - y_i^{calc})^2\ ]^{1/2}$, where y^{obs}/ y^{calc} = observed/calculated value of log S_0 according to model, n = number of measurements of log S_0
r^2	squared linear correlation coefficient, $r^2 = 1 - \Sigma_i\ (y_i^{obs} - y_i^{calc})^2\ /\ \Sigma_i\ (y_i^{obs} - \langle y \rangle)^2$, where y = log S_0, and <y> is the mean value of log S_0
SD	standard deviation: $SD = [\ 1/n\ \Sigma_i\ (y_i^{obs} - \langle y \rangle)^2\ ]^{1/2}$, where n = number of measurements, <y> = mean value of log S_0
F	F-statistic: $F = (n-p-1)/p \cdot \Sigma_i\ (y_i^{obs} - \langle y \rangle)^2\ /\ \Sigma_i\ (y_i^{obs} - y_i^{calc})^2$, where p = number of regression parameters
MPP	Measure of prediction performance [108]. It refers to the percent of 'correct' predictions, as defined by the count of absolute residuals $\lvert \log S_0^{obs} - \log S_0^{calc} \rvert \le 0.5$ divided by the number of measurements. MPP is represented as a pie chart in the correlation plots.

Abraham solvation descriptors

A	H-bond total acidity
B	H-bond total basicity
S_π	dipolarity/polarizability due to solute-solvent interactions between bond dipoles and induced dipoles
E	excess molar refraction ($dm^3\ mol^{-1}$ / 10); which models dispersion force interaction arising from π- and n-electrons of the solute
V	McGowan molar volume ($dm^3\ mol^{-1}$ / 100)
A·B	acid-base H-bonding product descriptor used in ABSOLV solubility prediction

Most important RDKit descriptors in RFR analysis

Subdivided Surface Area Molecular Descriptors [109]

LabuteVSA	sum of atomic contributions [110] to the accessible van der Waals surface area
MolLogP	sum of atomic contributions to octanol/water partition coefficient, log *P*
MolMR	sum of atomic contributions to molar refractivity, *MR*
SMR_VSAk	sum of accessible van der Waals surface area for those atoms with atomic contribution to molar refractivity; k refers to a small domain of atomic-contribution to MR; intended to capture *molecular size & polarizability*
PEOE_VSAk	intended to capture *direct electrostatic interactions* in a particular range; based on iterative equalization of atomic *orbital electronegativities* [111].

Complexity descriptors

BertzCT complexity index, based on size, symmetry, branching, rings, multiple bonds, and heteroatoms characteristic of solute [112].

Ipc content information of topological graph [113] - entropy of atomic distribution in solute

Topological and electrotopological connectivity indices

Chi0, Chi0n, Chi0v, Chi1, Chi1n, Chi4n, Chi4v, α – Kier-Hall topological connectivity and shape indices [114,115]; numerical representations of topology of solute calculated from graphical depiction of the molecule

Atomic and subroup counts, *HeavyAtomCount, NumberAromaticCarbocycles, NumberAromaticRings, RingCount, fr_benzene*

Availability of the *Wiki-pS$_0$* Database

The entire *Wiki-pS$_0$* database is planned to be released in book form: A. Avdeef. *Intrinsic Aqueous Solubility Data for Pharmaceutical Research.* Wiley-Interscience, Hoboken, NJ (under discussion with publisher). A sampling is presented in Table A5 in [20].

Acknowledgements

We are grateful for helpful suggestions made by Prof. George Zografi (University of Wisconsin-Madison) regarding the role of sorbed water in crystalline solids.

Conflict of interest*:* The authors declare no conflict of interest.

References

[1] 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 Delivery Rev.* **23** (1997) 3-25.

[2] D. Hörter, J.B. Dressman. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. *Adv. Drug Deliv. Rev*. **25** (1997) 3-14.

[3] L. Di, P.V. Fish, T. Mano. Bridging solubility between drug discovery and development. *Drug Disc. Today* **17** (2012) 486-495.

[4] C.A.S. Bergström, R. Holm, S.A. Jørgensen, S.B.E. Andersson, P. Artursson, S. Beato, A. Borde, K. Box, M. Brewster, J. Dressman, K.-I. Feng, G. Halbert, E. Kostewicz, M. McAllister, U. Muenster, J. Thinnes, R. Taylor, A. Mullertz. Early pharmaceutical profiling to predict oral drug absorption: Current status and unmet needs. *Eur. J. Pharm. Sci*. **57** (2014) 173-199.

[5] J.C. Dearden. In silico prediction of aqueous solubility. *Expert Opin. Drug Discov*. **1** (2006) 31–52.

[6] J. Taskinen, U. Norinder. In silico prediction of solubility. In: B. Testa, H. van de Waterbeemd (Eds.). *Comprehensive Medicinal Chemistry II*, Elsevier: Oxford, UK, 2007, pp. 627-648.

[7] J. Wang, T. Hou. Recent advances on aqueous solubility prediction. *Comb. Chem. HighThroughput Screen*. **14** (2011) 328-338.

[8] S.H. Yalkowsky, S.C. Valvani. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **69** (1980) 912-922.

[9] S.H. Yalkowsky, S. Banerjee. *Aqueous Solubility: Methods of Estimation for Organic Compounds*. Marcel Dekker, Inc., New York, 1992, p. 142.

[10] G. Klopman, S. Wang, D.M. Balthasar. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci*. **32** (1992) 474-482.

[11] 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.

[12] 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.

[13] I.V. Tetko, V.Yu. 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.

[14] A. Yan, J. Gasteiger. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci*. **43** (2003) 429-434.

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

[16] H. Sun. *A Practical Guide to Rational Drug Design*. Elsevier, Amsterdam, 2015, pp. 193-223.

[17] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B.O. Mitchell. Random Forest models to predict aqueous solubility. *J. Chem. Inf. Model*. **47** (2007) 150-158.

[18] D.S. Palmer, J.B.O. Mitchell. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? *Mol. Pharmaceutics* **11** (2014) 2962-2972.

[19] W.P. Walters. What are our models really telling us? A practical tutorial on avoiding common mistakes when building predictive models. In: J. Bajorath (Ed.). Chemoinformatics for Drug Discovery. John Wiley & Sons, Hoboken, NJ, 2014, pp. 1-31.

[20] A. Avdeef. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with *Wiki-pS_0* database. *ADMET & DMPK* **8** (2020) 29-77; DOI: http://dx.doi.org/-10.5599/admet.766.

[21] B.C. Doak, B. Over, F. Giordanetto, J. Kihlberg. Oral druggable space beyond the Rule of 5: insights from drugs and clinical candidates. *Chem. & Biol*. **21** (2014) 1115-1142.

[22] P.D. Leeson. Molecular inflation, attrition & the rule of five. *Adv. Drug Deliv. Rev.* **101** (2016) 22-33.

[23] P. Matsson, B.C. Doak, B. Over, J. Kihlberg. Cell permeability beyond the rule of 5. *Adv. Drug Deliv. Rev*. **101** (2016) 42-61.

[24] D.A. DeGoey, H.-J. Chen, P.B. Cox, M.D. Wendt. Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection. *J. Med. Chem*. **61** (2018) 2636-2651.

[25] C.A.S. Bergström, W.N. Charman, C.J.H. Porter. Computational prediction of formulation strategies for beyond-rule-of-5 compounds. *Adv. Drug Deliv. Rev*. **101** (2016) 6-21.

[26] S.D. Krämer, H.E. Aschmann, M. Hatibovic, K.F. Hermann, C.S. Neuhaus, C. Brunner, S. Belli. When barriers ignore the "rule-of-five". *Adv. Drug Del. Rev*. **101** (2016) 62-74.

[27] G. Ermondi, M. Vallaro, G. Goetz, M. Shalaeva, G. Caron. Experimental lipophilicity for beyond Rule of 5 compounds. *Future Drug. Discov*. (2019) Published Online: 14 Mar 2019. DOI: https://doi.org/-10.4155/fdd-2019-0002.

[28] G. Ermondi, M. Vallaro, G. Goetz, M. Shalaeva, G. Caron. Updating the portfolio of physicochemical descriptors related to permeability in the beyond the rule of 5 chemical space. *Eur. J. Pharm. Sci.* **146** (2020) 105274. DOI: https://doi.org/10.1016/j.ejps.2020.105274.

[29] G. Caron, J. Kihlberg, G. Ermondi. Intramolecular hydrogen bonding: An opportunity for improved design in medicinal chemistry. *Med. Res. Rev*. **39** (2019) 1707-1729. DOI: https://dx.doi.org/10.1002/med.21562.

[30] G. Caron, V. Digiesi, S. Solaro, G. Ermondi. Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discov. Today* (2020), in press. DOI: https://doi.org/10.1016/-j.drudis.2020.01.012.

[31] P.A. Carrupt, B. Testa, A. Bechalany, N. el Tayar, P. Descas, D. Perrissoud. Morphine 6-glucuronide and morphine 3-glucuronide as molecular chameleons with unexpected lipophilicity. *J Med Chem*. **34** (1991) 1272-1275.

[32] F. Irmann. A simple correlation between water solubility and the structure of hydrocarbons and halogenated hydrocarbons. *Chem. Ing. Tech*. **37** (1965) 789-798.

[33] 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.

[34] D. Alantari, S. Yalkowsky. Comments on prediction of the aqueous solubility using the general solubility equation (GSE) versus a genetic algorithm and a support vector machine model. *J. Pharm. Dev. Technol.* **23** (2018) 739-740.

[35] Y. Ran, S.H. Yalkowsky. Prediction of drug solubility by the General Solubility Equation. *J. Chem. Inf. Comput. Sci.* **41** (2001) 354-357.

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

[37] Y. Ran, N. Jain, S.H. Yalkowsky. Prediction of aqueous solubility of organic compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **41** (2001) 1208-1217.

[38] N. Jain, G. Yang, S.G. Machatha, S.H. Yalkowsky. Estimation of the aqueous solubility of weak electrolytes. *Int. J. Pharm*. **319** (2006) 169-171.

[39] J.C. Dearden. Partitioning and lipophilicity in quantitative structure-activity relationships. *Environ. Health Persp.* **61** (1985) 203-228.

[40] A.S.I.D. Lang, J.-C. Bradley. ONS Melting Point Model 010. QDB archive, DOI: 10.15152/QDB.104. QsarDB content. Property mpC. Steps:Calculate descriptors. SMILES. Calculate. Scroll down to mpC.

[41] M.H. Abraham. Scales of hydrogen bonding - their construction and application to physicochemical and biochemical processes. *Chem. Soc. Revs*. **22** (1993) 73-83.

[42] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci*. **39** (1999) 835-845.

[43] L. Breiman. Random forests. *Mach. Learn*. **45** (2001) 5-32.

[44] A. Liaw, M. Wiener. Classification and regression by Random Forest. *R News* **2** (2002) 18-22.

[45] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (accessed 5 May 2019).

[46] A. Liaw. Random Forests What, Why, And How. https://www.youtube.com/watch?v=XJnjlpW9w5A. (YouTube lecture). https://nyhackr.blob.core.windows.net/presentations/Random-Forests-What-Why-and-How_Andy_Liaw.pdf.

[47] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Classification and Regression Trees. Chapman & Hall/CRC: Boca Raton, 1984.

[48] T.S. Schroeter, A. Schwaighofer, S. Mika, A.T. Laak, D. Suelzle, U. Ganzer, N. Heinrich, K.-R. Müller. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des*. **21** (2007) 485-498.

[49] P. Howard, W. Meylan. PHYSPROP DATABASE. Syracuse Research Corp., N. Syracuse, NY, Sept. 1999. https://www.srcinc.com/what-we-do/environmental/scientific-databases.html (accessed 3 May 2019).

[50] Beilstein Organischen Chemie, Berlin, Springer-Verlag. 4th Ed. (Online: http://www.elsevier.com/-online-tools/reaxys).

[51] S.H. Yalkowsky, Y. He. *The Handbook of Aqueous Solubility Data*. CRC Press, Boca Raton, 2003.

[52] J.S. Delaney. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci*. **44** (2004) 1000-1005.

[53] MOE. Chemical Computing Group Inc., Montreal, QC H3A 2R7, Canada. http://www.chemcomp.com (accessed 6 May 2019).

[54] A. Llinàs, R.C. Glen, J.M. Goodman. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.* **48** (2008) 1289-1303.

[55] R. Guha, T.S. Dexheimer, A.N. Kestranek, A. Jadhav, A.M. Chervenak, M.G. Ford, A. Simeonov, G.P. Roth, C.J. Thomas. Exploratory analysis of kinetic solubility measurements of a small molecule library. *Bioorg. Med. Chem.* **19** (2011) 4127-4134.

[56] A. Avdeef. Suggested improvements for measurement of equilibrium solubility-*pH* of ionizable drugs. *ADMET & DMPK* **3** (2015) 84-109.

[57] G. Völgyi, A. Marosi, K. Takács-Novák, A. Avdeef. Salt solubility products of diprenorphine hydrochloride, codeine and lidocaine hydrochlorides and phosphates – Novel method of data analysis not dependent on explicit solubility equations. *ADMET & DMPK* **1** (2013) 48-62.

[58] A. Avdeef, E. Fuguet, A. Llinàs, C. Ràfols, E. Bosch, G. Völgyi, T. Verbić, E. Boldyreva, K. Takács-Novák. Equilibrium solubility measurement of ionizable drugs – consensus recommendations for improving data quality. *ADMET & DMPK* **4** (2016) 117-178.

[59] O.S. Marković, M.P. Pešić, A.V. Shah, A.T.M. Serajuddin, T.Z. Verbić, A. Avdeef. Solubility-pH profile of desipramine hydrochloride in saline phosphate buffer: enhanced solubility due to drug-buffer aggregates. *Eur. J. Pharm. Sci*. **133** (2019) 264–274.

[60] C.A.S. Bergström, A. Avdeef. Perspectives in solubility measurement and interpretation. *ADMET & DMPK* **7** (2019) 88-105.

[61] A. Avdeef. *Absorption and Drug Development,* Second Edition, Wiley-Interscience, Hoboken NJ, 2012.

[62] G. Landrum, R. Lewis, A. Palmer, N. Stiefl, A. Vulpetti. Making sure there's a "give" associated with the "take": Producing and using open-source software in big pharma. *J. Cheminformatics*. **3** (2011) 1-1; *cf.*, http://www.rdkit.org/ (accessed 5 May 2019).

[63] L.Z. Benet, F. Broccatelli, T.I. Oprea. BDDCS applied to over 900 drugs. *AAPS J*. **13** (2011) 519-547.

[64] M.L. Andrew, P.J. Weiss. Solubility of antibiotics in twenty-four solvents. II. *Antibiot. Chemother*. **9** (1959) 277-279.

[65] K. Florey. Thiostrepton. *Anal. Prof. Drug Subst.* **7** (1978) 423-444.

[66] J.R. Marsh, P.J. Weiss. Solubility of antibiotics in twenty-six solvents. III, *J. Assoc. Off. Anal. Chem.* **50** (1967) 457-462.

[67] G. Ismailos, C. Peppas, J. Dressman, P. Macheras. Unusual solubility behavior of cyclosporin A in aqueous media*. J. Pharm. Pharmacol*. **43** (1991) 287-289.

[68] T. Loftsson, D. Hreinsdóttir, M. Másson. Evaluation of cyclodextrin solubilization of drugs. *Int. J. Pharm.* **302** (2005) 18-28.

[69] S.D. Mithani, V. Bakatselou, C.N. TenHoor, J.B. Dressman. Estimation of the increase in solubility of drugs as a function of bile salt concentration. *Pharm. Res.* **13** (1996) 163-167.

[70] S.N. Giri, L.R. Kartt. Temperature dependent aqueous solubility of actinomycin D. *Specialia* **31** (1975) 482-483.

[71] M.S. Ali, M.Z. Hoemann MZ, Aubé J, Mitscher LA, Georg GI, McCall R, Jayasinghe LR. Novel cytotoxic 3 -(tert-butyl) 3 -diphenyl analogs of paclitaxel and docetaxel*. J. Med. Chem*. **38** (1995) 3821-3828.

[72] R. Takano, K. Sugano, A. Higashida, Y. Hayashi, M. Machida, Y. Aso, S. Yamashida. Oral absorption of poorly water-soluble drugs: computer simulation of fraction absorbed in humans from a miniscale dissolution test. *Pharm. Res*. **23** (2006) 1144-1156.

[73] G.A. Brewer. Gramicidin. *Anal. Prof. Drug Subst*. **8** (1979) 179-218.

[74] D. Pitré, A. Davies, M. Grandi. Profile of Iodoxamic Acid. *Anal. Prof. Drug Subst. Excip.* **20** (1991) 303-335.

[75] N.A. Kasim, M. Whitehouse, C. Ramachandran, G.L. Amidon. Molecular properties of WHO essential drugs and provisional biopharmaceutical classification. *Mol. Pharmaceutics* **1** (2004) 85-96.

[76] B.I. Escher, C. Berger, N. Bramaz, J.-H. Kwon, M. Richter, O. Tsinman, A. Avdeef. Membrane-water partitioning, membrane permeability and non-target modes of action in aquatic organisms of the parasiticides ivermectin, albendazole and morantel. *Envir. Tox. Chem*. **27** (2008) 909-918.

[77] R. Takano, K. Sugano, Y. Hayashi, M. Machida, Y. Aso, S. Yamashita. Annual meeting of the academy of pharmaceutical science and technology, Japan. Tokyo (2005).

[78] D. Singh, K. Pathak. Hydrogen bond replacement - unearthing a novel molecular mechanism of surface solid dispersion for enhanced solubility of a drug for veterinary uses. *Int. J. Pharm*. **441** (2013) 99-110.

[79] S.T. Anik, D.M. Johnson. Preformulation and formulation considerations of peptide drugs: case history of an LHRH analog. In: Lee VHL(ed.). Peptides and Drug Delivery. Marcel Dekker, Inc., New York, 1991, pp 769-784.

[80] G.W. Michel. Nystatin. *Anal. Prof. Drug Subst*. **6** (1977) 341-422.

[81] F. Nachtmann, K. Krummen, F. Maxl, E. Riemer. Oxytocin. *Anal. Prof. Drug Subst*. **10** (1981) 563-600.

[82] C.A.S. Bergström, C.M. Wassvik, U. Norinder, K. Luthman, P. Artursson. Global and local computational models for aqueous solubility prediction of drug-like molecules. *J. Chem. Inf. Comput. Sci.* **44** (2004) 1477-1488.

[83] J. Lee, S.C. Lee, G. Acharya, C.J. Chang, K. Park. Hydrotropic solubilization of paclitaxel: Analysis of chemical structures for hydrotropic property. *Pharm. Res.* **20** (2003) 1022-1030.

[84] P. Simamora, J.M. Alvarez, S.H. Yalkowsky. Solubilization of rapamycin. *Int. J. Pharm*. **213** (2001) 25-29.

[85] B.N. Singh. A quantitative approach to probe the dependence and correlation of food-effect with aqueous solubility, dose/solubility ratio, and partition coefficient (log P) for orally active drugs administered as immediate-release formulations, *Drug Dev. Res*. **65** (2005) 55-75.

[86] G.F. Plöger, M.A. Hofsäss, J.B. Dressman. Solubility determination of active pharmaceutical ingredients which have been recently added to the list of Essential Medicines in the context of the Biopharmaceutics Classification System - biowaiver. *J. Pharm.Sci.* **107** (2018) 1478-1488.

[87] G.G. Gallo, P. Radaelli. Rifampacin. *Anal. Prof. Drug Subst*. **5** (1976) 467-514.

[88] G. Boman, P. Lundgren, G. Stjernstrom. Mechanism of the inhibitory effects of PAS granules on the absorption of rifampicin: adsorption of rifampicin by an excipient, bentonite. *Eur. J. Clin. Pharmacol*. **8** (1975) 293-299.

[89] S.Q. Henwood, M.M. de Villiers, W. Liebenberg, A.P. Lotter. Solubility and dissolution properties of generic rifampicin raw materials. *Drug Dev. Ind. Pharm*. **26** (2000) 403-408.

[90] T.T. Mariappan, S. Singh. Regional gastrointestinal permeability of rifampicin and isoiazid (alone and their combination) in the rat. *Int. J. Tuberc. Lung Dis*. **7** (2003) 797-803.

[91] S. Agrawal, R. Panchagnula. Dissolution test as a surrogate for quality evaluation of rifampicin containing fixed dose combination formulations. *Int. J. Pharm*. **287** (2004) 97-112.

[92] D. Evans, S. Yalkowsky, S. Wu, D. Pereira, P. Fernandes. Overcoming the challenges of low drug solubility in the intravenous formulation of solithromycin. *J. Pharm. Sci.* **107** (2018) 412-418.

[93] M. J. O'Neil, P.E. Heckelman, P.H. Dobbelaar, K.J. Roman (Eds.). *The Merck Index: an Encyclopedia of Chemicals, Drugs, and Biologicals*, The Royal Society of Chemistry, 15th Ed, 2013.

[94] H. Arima, K. Yunomae, K. Miyake, K. Uekama. Comparative studies of the enhancing effects of cyclodextrins on the solubility and oral bioavailability of tacrolimus in rats. *J. Pharm. Sci.* **90** (2001) 690-701.

[95] L.-C. Dong, K.A.U. Co, S. Li, C. Pollock-Dove. Stabilized solubility-enhanced formulations for oral delivery. PCT Application WO 2010/111397 A1 (2010).

[96] E.V. Persson, A.-S. Gustafsson, A.S. Carlsson, R.G. Nilsson, L. Knutson, P. Forsell, G. Hanisch, H. Lennernäs, B. Abrahamsson. The effects of food on the dissolution of poorly soluble drugs in human and in model small intestinal fluids. *Pharm. Res*. **22** (2005) 2141-2151.

[97] P.J. Faustino. Report to Office of Generic Drugs. Vancomycin solubility study. Division of Product Quality Research Office of Testing and Research Center for Drug Evaluation and Research. Food and Drug Administration. 5 Feb 2008. https://wayback.archive-it.org/7993/20170403212949/https:/-www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-en/documents/document/ucm082291.pdf

[98] L.D. Hughes, D.S. Palmer, F. Nigsch, J.B.O. Mitchell. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P. *J. Chem. Inf. Model* **48** (2008) 220-232.

[99] J.C. Dearden. The use of topological indices in QSAR and QSPR modeling. In: K. Roy (Ed.) *Advances in QSAR Modeling. Challenges and Advances in Computational Chemistry and Physics*, vol 24. Springer, Cambridge, 2017, pp. 57-88.

[100] Food and Drug Administration (USA): Givosiran(Givlaari), Alnylam Pharmaceutical Inc. NDA 212194Orig1s000; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212194Orig1s000-ChemR.pdf.

[101] Food and Drug Administration (USA): Tenapanor(Ibsrela), Ardelyx, Inc. NDA 211801Orig1s000; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211801Orig1s000ChemR.pdf.

[102] M. Schäfer, T.R. Schneider, G.M. Sheldrick. Crystal structure of vancomycin. *Structure* **4** (1996) 1509-1515.

[103] C. Ahlneck, G. Zografi. The molecular basis of moisture effects on the physical and chemical stability of drugs in the solid state. *Int. J. Pharm*. **62** (1990) 87-95.

[104] Hancock BC, Zografi G. The relationship between the glass transition temperature and the water content of amorphous pharmaceutical solids. *Pharm. Res.* **11** (1994) 471-477.

[105] A.W. Newman, S.M. Reutzel-Edens, G. Zografi. Characterization of the "hygroscopic" properties of active pharmaceutical ingredients. *J. Pharm. Sci.* **97** (2007) 1047-1059.

[106] A. Newman, G. Zografi. Commentary: an examination of water vapor sorption by multicomponent crystalline and amorphous solids and its effects on their solid-state properties. *J. Pharm. Sci.* **108** (2019) 1061-1080.

[107] A. Avdeef. Solubility temperature dependence predicted from 2D structure. *ADMET & DMPK* **3** (2015) 298-344.

[108] A.J. Hopfinger, E.X. Esposito, A. Llinàs, R.C. Glen, J.M. Goodman. Findings of the challenge to predict aqueous solubility. *J. Chem. Inf. Model*. **49** (2009) 1-5.

[109] P. Labute. A widely applicable set of descriptors. *J. Molec. Graph. Model*. **18** (2000) 464-477.

[110] S.A. Wildman, G.M. Crippen. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci*. **39** (1999) 868-873.

[111] J. Gasteiger, M. Marsali. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **36** (1980) 3219-3228.

[112] S.H. Bertz. The first general index of molecular complexity. *J. Am. Chem. Soc*. **103** (1981) 3599-3601.

[113] D. Bonchev, N. Trinajstić. Information theory, distance matrix, and molecular branching. *J. Chem. Phys*. **67** (1977) 4517-4533.

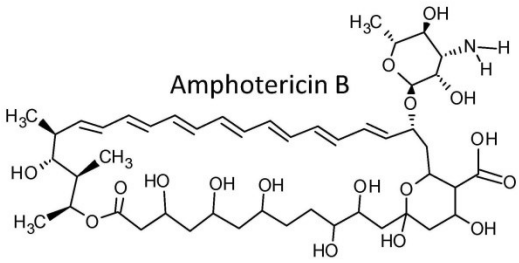
[114] L.H. Hall, L.B. Kier. *Reviews of Computational Chemistry*. In: D. Boyd, K. Lipkowitz (Eds.), VCH Publishers, **2** (1991) 367-422.

[115] L.H. Hall, L.B. Kier. The nature of structure–activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem. - Chimica Therapeutica*. **4** (1997) 307-312.

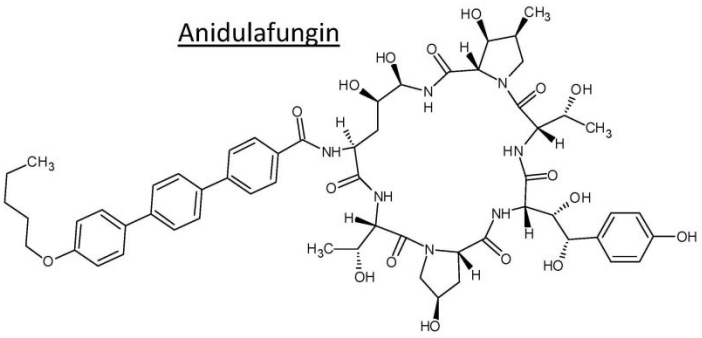
Appendix

Underlined names denote compounds whose crystal structures have not been deposited in the Cambridge Crystallographic Data Centre.

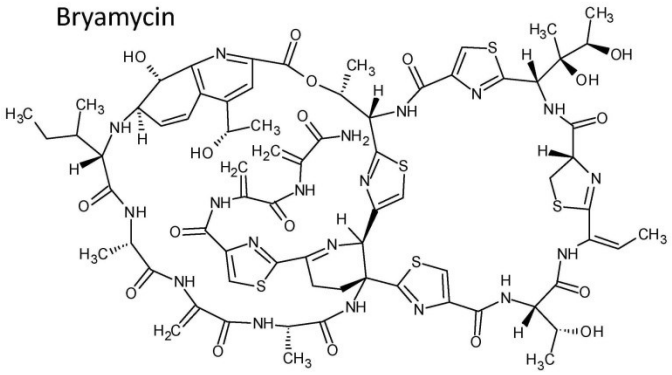
Amphotericin B



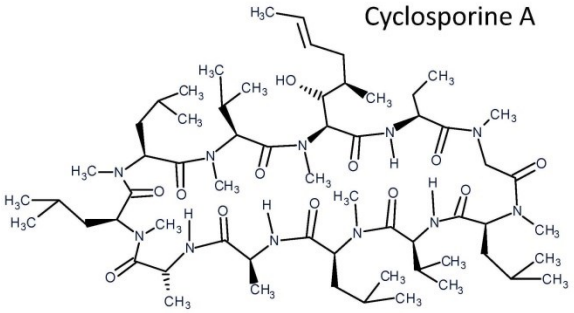
Anidulafungin



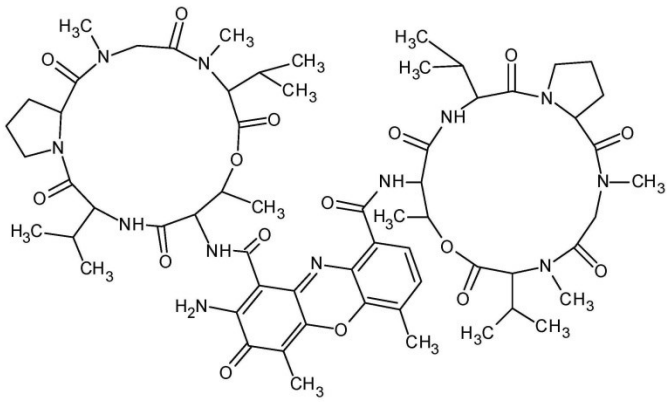
Bryamycin



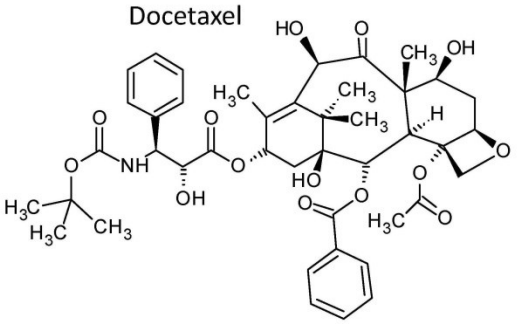
Cyclosporine A



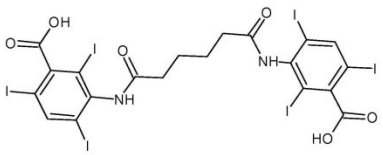
Dactinomycin



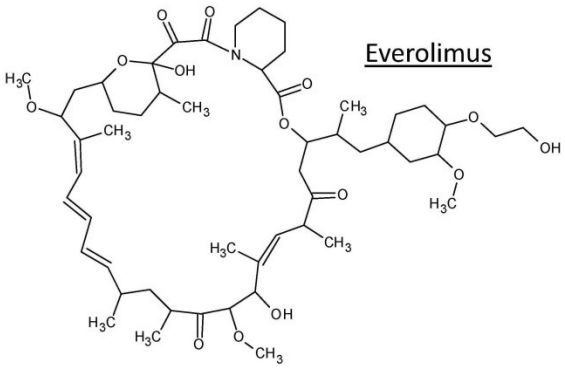
Docetaxel



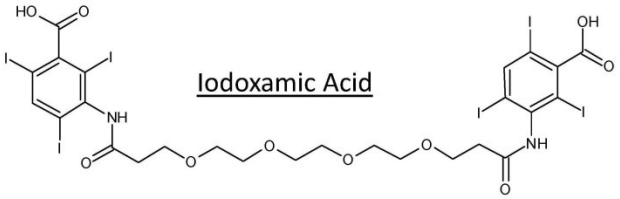
Iodipamide

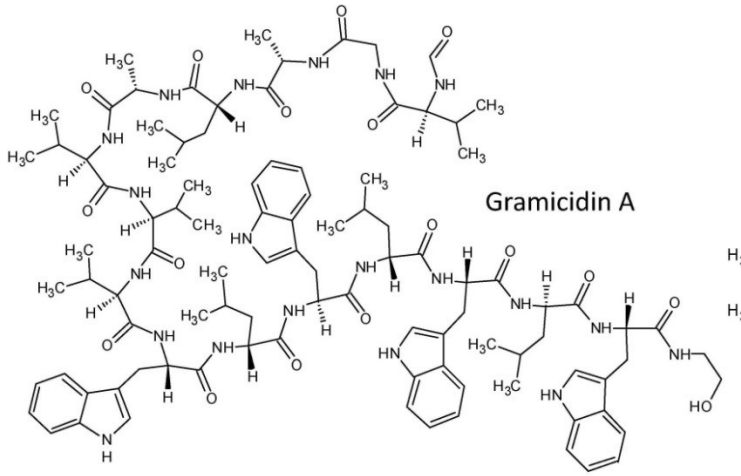


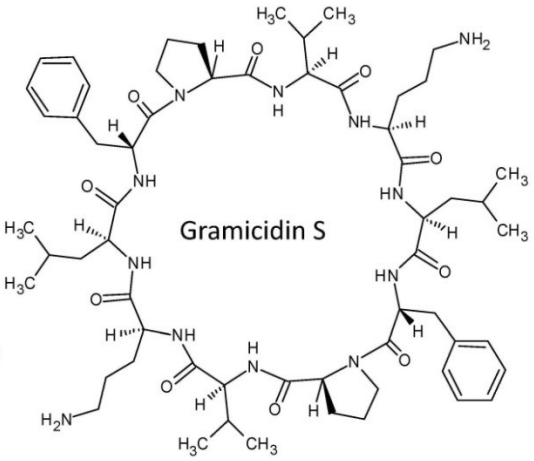
Everolimus

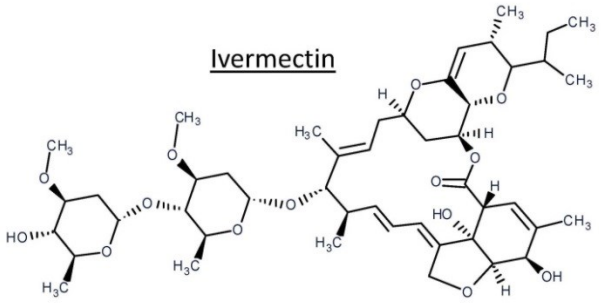


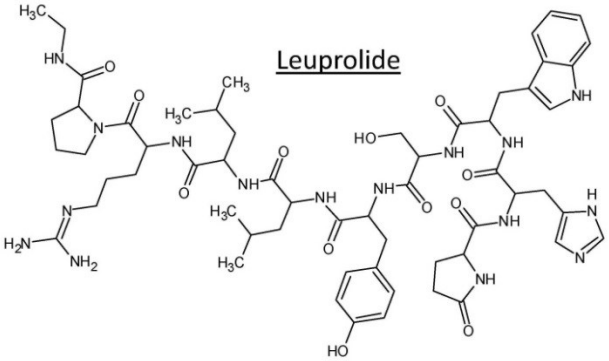
Iodoxamic Acid

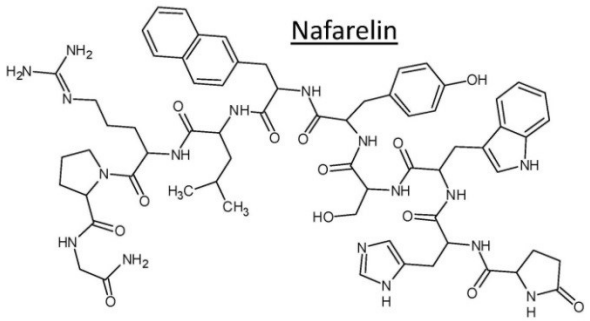


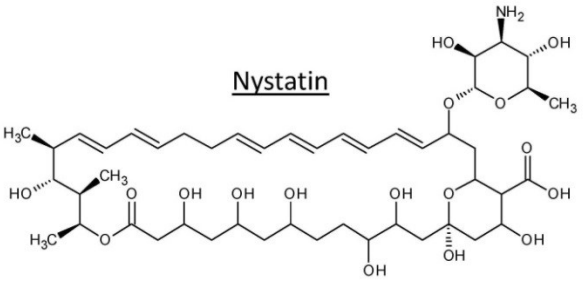

Gramicidin A

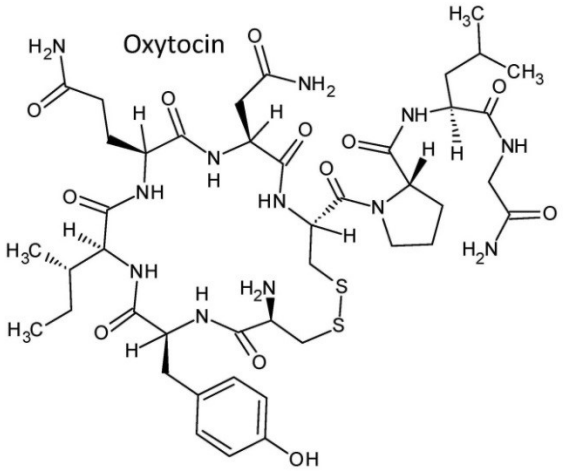

Gramicidin S

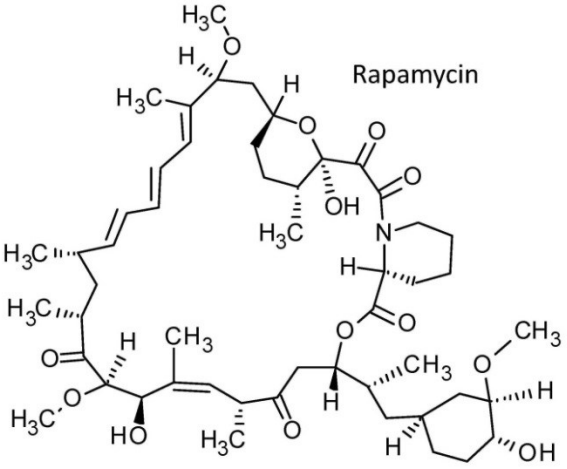

Ivermectin


Leuprolide

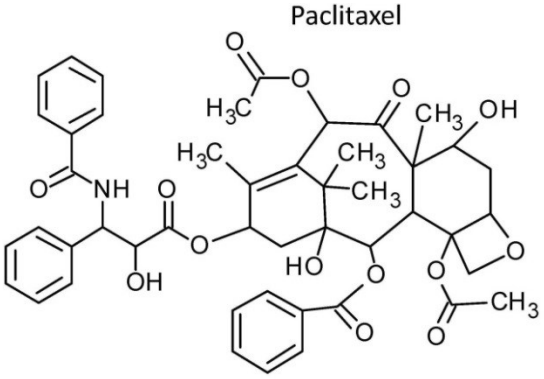

Nafarelin


Nystatin

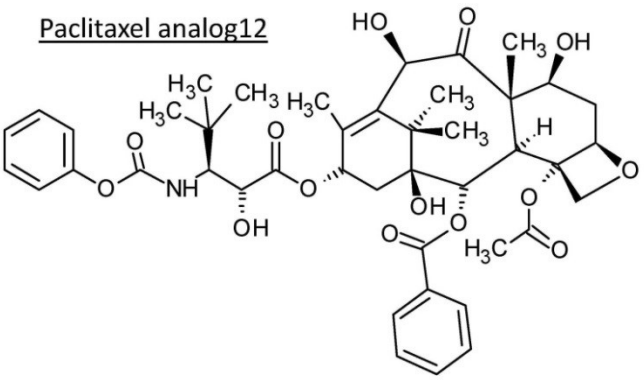

Oxytocin


Rapamycin

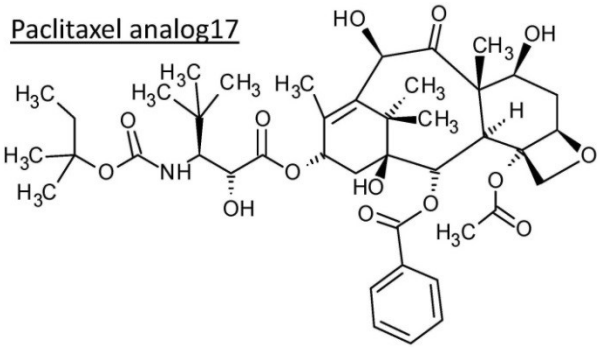
Paclitaxel



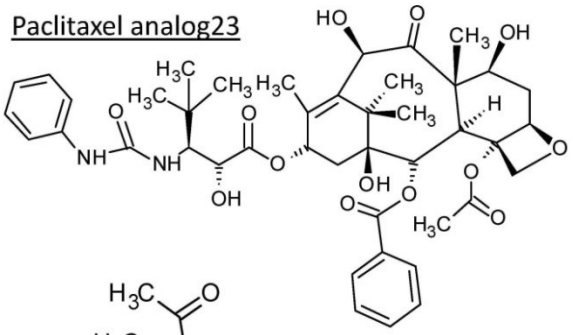
Paclitaxel analog12



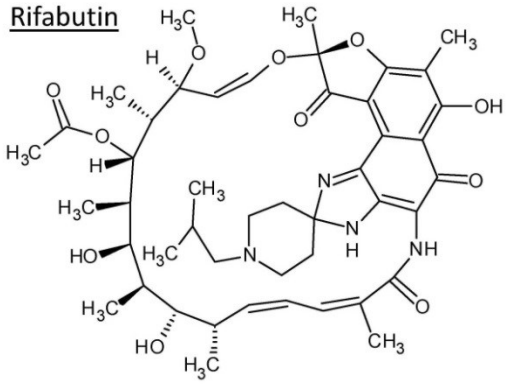
Paclitaxel analog17

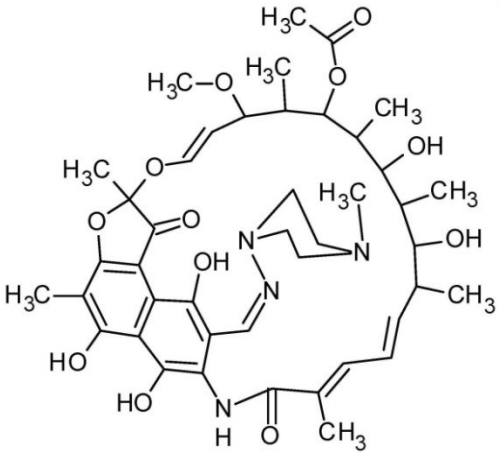


Paclitaxel analog23



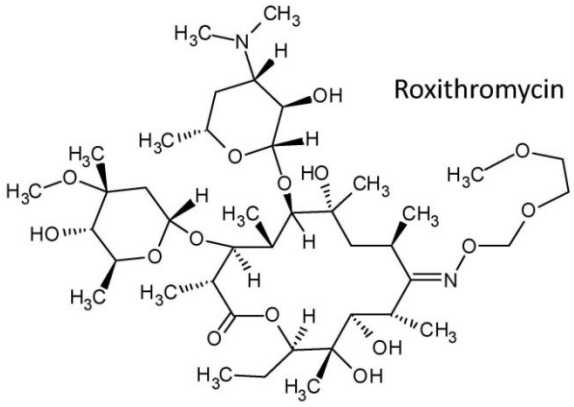
Rifabutin

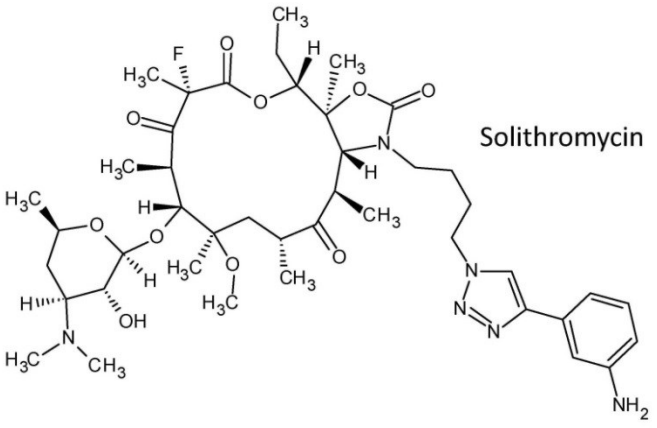




Rifampicin

Roxithromycin





Solithromycin