Trends in PhysChem Properties of Newly Approved Drugs over the Last Six Years; Predicting Solubility of Drugs Approved in 2021

Alex Avdeef (alex@in-ADME.com)
Manfred Kansy

Research Article

Keywords: Flexible-Acceptor General Solubility Equation, Abraham Solvation Equation, Kier molecular flexibility index, intrinsic solubility, partial least squares, Random Forest Regression

Posted Date: July 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1848437/v1

Abstract

For the 'small-molecule' NMEs (new molecular entities) approved in 2021 by the US FDA, quantitative solubility values were found for 28 drugs, nearly all from published New Drug Applications (NDAs). Comparisons of physicochemical properties over the last six years indicate that the NMEs are slowly continuing to increase in size and decrease in solubility. Since 2016, the intrinsic solubility values (S_0) have decreased on the average by 0.50 log unit, the calculated octanol-water partition coefficients (*clogP*) have increased by 0.34 log unit, and the molecular weights (*MW*) have increased by 22 g·mol^{-1} (to 477, compared to 298 in older drugs). The average number of Hbond acceptors has remained constant, while the average number of H-bond donors and the Kier Φ molecular flexibility indices have decreased slightly. The reported solubility data for the 2021 small-molecule NMEs were processed using the program *p*DISOL-X to obtain S_0 values, normalized to 25 ^{o}C. The S_0 values ranged from 2 ng·mL^{-1} (avacopan) to 43 mg·mL^{-1} (viloxazine). In the new set, *MW* spanned from 233 g·mol^{-1} (dexmethylphenidate) to 1215 g·mol^{-1} (voclosporin). Values of *clogP* ranged from − 0.3 (serdexmethylphenidate, a quaternary ammonium molecule) to 8.1 (avacopan). Five different *in-silico* models were used to predict the aqueous intrinsic log solubility of the 28 novel NMEs: (i) Yalkowsky's General Solubility Equation (GSE(*classic*)), (ii) Abraham's Linear Solvation Equation (ABSOLV), (iii) Avdeef-Kansy 'Flexible-Acceptor' General Solubility Equation ((GSE(Φ,*B*)), (iv) Breiman's Random Forest Regression (RFR), and (v) consensus model based on (ii) and (iii) above. The various models were retrained with an enlarged version of the Wiki-pS_0 database (currently at 7655 log S_0 entries of drug-relevant molecules). The consensus model (r^2 = 0.67, RMSE = 1.08) just slightly outperformed the other four models. The relatively-simple consensus prediction equation can be easily incorporated into spreadsheet calculations. As new drugs are approved, it will be important to continue monitoring the quality of measured solubility. Matching prediction to measurement is valuable when prediction methods are applied to virtual libraries, in order to seek opportunities to minimize pharmacokinetic risks of large, but otherwise promising, candidate molecules.

1 Introduction

The downward trend in the pharma R&D productivity from 1996 to 2011, as indicated by counting the number of new molecular entities (NMEs) approved each year, started to reverse after 2012. Although approvals of biologics have been steadily increasing, most new market introductions are still so-called 'small molecules' [1]. In the last two decades of considerations of Lipinski's 'Rule of 5' (Ro5) drug chemical space [2], many emerging NMEs are larger, less soluble, more lipophilic, and possess more Hbond acceptors, when compared to older drugs [3]. NMEs outside of Ro5 space are dubbed 'beyond the Rule of 5' (bRo5) drugs [3–8].

In 2021, 50 drugs were approved by the FDA. Of these drugs, 72% are considered 'small molecule' NMEs. But these 'small molecules' are trending to larger sizes. Large molecules may be burdened with pharmacokinetic (PK) risks, because of poor solubility or low cell permeability, elevated cellular efflux, and increased metabolism. In drug discovery and early development, several strategies to mitigate some of the risks have been tried [5–9]. Flexible molecules with the capability to form intramolecular H-bonds (IMHBs) have been of particular interest, since these may increase drug solubility in water (*e.g.*, by adopting hydrophilic 'extended' conformations) and enhance permeability across biomembranes (*e.g.*, by adopting hydrophobic 'folded' conformations) [7–9].

Important factors for the productivity gain since 2012 include improved methods, focusing both on biological activity, as well as on absorption-distribution-metabolism-excretion-toxicity (ADMET) characteristics. Given the large number of molecules considered in discovery projects, the *in-silico* prediction of molecular properties is a valuable first step in prioritizing molecules for further (resource-costly) *in-vitro* and *in-vivo* screening.

Current trends in ADMET *in-silico* modeling approaches place increasing emphasis on *calculated* input parameters. Physicochemical properties are still key as inputs (*e.g.*, octanol-water partition coefficients, log *P*, ionization constants, pK_a, etc.), but often these are calculated rather than measured values. The risk may be that overall simulation approaches might have substantial aggregate uncertainty, something that is not usually discussed.

Modeling approaches need reliable input parameters (descriptors) to validate and to further improve the quality and support necessary to shift from more simple simulation of *in-vivo* profiles to challenging real blind prediction of *in-vivo* ADMET and PK/PD (pharmacokinetics/pharmacodynamics) profiles. Thus, early optimization and selection of suitable biologically-active candidate molecules could be improved. Reduced or eliminated animal testing would make pharma R&D even more successful. With a huge number of input descriptors one can get an impressive fitting of *in-vivo* profiles. Questionable trends are evident, that although multifactorial fitting to *in-vivo* profiles appear satisfactory, they often lack knowledge gain and possess limited predictive power. Correct prediction (blind) of human *in-vivo* profiles, based on fewer/no animal experiments, with a minimal number of input descriptors, is a desirable goal. In multifactorial fitting, the use of measured descriptors needs to be clearly differentiated from descriptors which are purely calculated. The uncertainty in the calculated descriptors would be helpful in estimating the aggregation of errors in overall prediction. The increasing use of machine learning methods and artificial intelligence can lead to accurate predictions. However, understanding the basis of such prediction may not readily suggest the steps to take to improve the properties of tested compounds.

Solubility plays a key part in deeper understanding of PK risks [4]. To predict the solubility of novel compounds, a manually curated database (*WikipS*$_0$) of intrinsic solubility values of druglike molecules was assembled in 2011, with entries added steadily since then. A comprehensive mass-action nonlinear regression program, *p*DISOL-X, to analyze solubility-pH data was developed in parallel, with enhancements added periodically. To predict drug solubility, different computational methods were examined in a series of studies [10–13]. Initially [10], Breiman's Random Forest regression (RFR) machine learning method [14] was tested and compared to predictions of Yalkowsky's General Solubility Equation (GSE) [15] and Abraham's Solvation Equation (ABSOLV) [16]. Whether prediction models trained with small drugs could be used to predict the solubility of large bRo5 drugs was then considered [11]. A way to modify the traditional GSE by implementing Kier's molecular flexibility index (Φ) and Abraham's basicity descriptor (B) resulted in the novel 'FlexibleAcceptor' GSE(Φ,B) equation [12]. This equation was directed [13] to predict the intrinsic solubility values of 72 NMEs recently-approved by the FDA (20162020) [17–21].

In the present study, the predictions are extended to the 2021 newly-approved drugs. Also examined are the trends in the physicochemical properties. The upward growth in the number of H-bond acceptor values may have leveled most recently, but *MW* and *clogP* values appear to be still slightly on the increase, underlying concomitant lower solubility. *In-silico* models to predict solubility of such NMEs and of molecules not yet synthesized is expected to be an asset for early risk assessment [4].

2 Computational Methods And Data Sources

2.1 Thermodynamic Basis of the General Solubility Equation (GSE)

Yalkowsky and coworkers [15, 22, 23] developed the General Solubility Equation (GSE), Eq. 1, to predict the solubility of liquid/solid nonelectrolytes (mostly industrial organic chemicals) in water. The method is particularly appealing since it requires no 'training.' Merely the melting point (*mp* in ^{o}C) and the octanol-water partition coefficient, either measured (log *P*) or calculated (*clogP*), are prerequisites for predicting solubility (log molar units):

$$\log S_0^{\,\mathrm{GSE(classic)}} = 0.5 - 1.0 \log P - 0.01(mp - 25) \quad (1)$$

The thermodynamic basis of the equation was reviewed recently [13]. Briefly, the dissolution of a crystalline substance in water consists of two main contributions: (i) crystal lattice effect (XTL), *i.e.*, the energy needed to break down the lattice to form a hypothetical 'supercooled liquid' (SCL), and (ii) solvation effect, *i.e.*, the energy released as the SCL dissolves in water. The total solubility can be expressed as [22, 23]:

$$\log S = \log S_w^{SCL} + \log S_w^{XTL} \quad (2)$$

where $\log S_w^{XTL} = -\Delta S_m (T_m - T) / (2.303\, R T)$; ΔS_m is the standard molar entropy of phase transformation, T is the absolute temperature (K) and T_m is the melting point (K). At 25 ^{o}C:

$$\log S \approx \log S_w^{SCL} - 0.010(mp - 25) \quad (3)$$

Hansch and coworkers [24] showed that log *S* of simple liquid solutes correlated linearly with the octanol-water partition coefficients, $\log P \approx \log (S_{oct}^{liq} / S_w^{liq})$. On re-arrangement,

$$\log S_w^{liq} \approx \log S_{oct}^{liq} - \log P \quad (4)$$

where $\log S_{oct}^{liq}$ is the log solubility of a liquid solute in octanol, ranging from − 0.3 to + 0.9 for small molecules [24]. Yalkowsky and coworkers rationalized log $S_{oct}^{SCL} = 0.5$ in Eq. 1 [15].

Hansch's studies suggest that the constant coefficients in Eq. 1 might need to be modified for compounds from novel classes of chemical space. If the 'supercooled liquid' form of a large polar solute is not fully miscible with octanol, then the log S_{oct}^{SCL} contribution could a negative number. A large molecule with a decreased S_{oct}^{SCL} (due to decreased miscibility with octanol) is expected to have an increased S_w^{SCL}. This would lessen the contribution of lipophilicity to the predicted solubility.

2.2 'Flexible-Acceptor' General Solubility Equation, GSE(Φ,B)

It was found [12] that the sum of Kier's molecular flexibility (Φ) [25] and Abraham's [16, 26] basicity descriptor, B, could be incorporated into a *nonlinear* variant of the GSE to produce a trainable model suitable to predict solubility of various classes of drugs, including large NMEs (*MW* > 800 g/mol). The resultant GSE(Φ,*B*) has the form:

$$\log S_0^{\,GSE(\Phi,B)} = c_0 + c_1 \cdot clogP + c_2 \cdot (mp - 25)/100 \quad (5)$$

with the variable coefficients modeled here as:

$$c_0 = b_0 + b_1 \exp(-b_2 \cdot (\Phi + B)) \quad (5a)$$

$$c_1 = b_3 + b_4 [\, 1 - \exp(-b_5 \cdot (\Phi + B)) \,] \quad (5b)$$

$$c_2 = b_6 + b_7 \cdot (\Phi + B) \quad (5c)$$

The c-coefficients as functions of $\Phi + B$ were determined by partial least squares (PLS open-source package from https://cran.r-project.org/web/packages/pls) analysis of solubility data sorted on values of $\Phi + B$ and uniformly binned into 18 groups of 123–775 points, to ensure nearly constant $\Phi + B$ increments, as described previously [12, 13]. Since our last study [13], the database has accumulated nearly 1000 new entries. So, a new set of bconstants was determined in the current investigation, using drug-relevant molecules as the training set, but excluding new drugs from the training. Values of Φ were calculated from the two kappa and the heavy atom count descriptors provided by the Landrum's RDKit open-source chemoinformatics library [27]. Table 1 lists these Φ and *B* values.

***** Table 1 goes here *****

2.3 Abraham Descriptors and the ABSOLV Linear Model for Predicting Solubility

Abraham introduced five solvation descriptors: *A, B,* S_π, *E*, and *V* [16, 26]. Two of these constitute H-bond potentials: *A* is the H-bond acidity (donor strength) and *B* is the H-bond basicity (acceptor strength) of the solute. S_π is the dipolarity/polarizability, *E* is an excess molar refraction in units of (cm^3/mol)/10, and *V* is the McGowan characteristic molar volume in units of (cm^3/mol)/100. Values of the descriptors were calculated from 2D structures using the ABSOLV algorithm [26] (*cf.*, www.acdlabs.com) and are listed in Table 1 for the new drugs.

Abraham and Le [16] amended the ABSOLV model to predict intrinsic solubility (log molar):

$$\log S_0^{ABSOLV} = d_0 + d_1 A + d_2 B + d_3 S_\pi + d_4 E + d_5 V + d_6 A \bullet B \quad (6)$$

The independent variables are the five solute descriptors, plus the cross product of the H-bond terms. The seven d-coefficients were determined by PLS regression, using the training set database, exclusive of the new drugs set. Quaternary ammonium drugs and drugs with *MW* > 800 Da were each treated separately. The rest of the molecules were divided into four acid-base classes – with reference to predominant charge state at pH 7.4: acids(-), bases(+), neutrals(0), and zwitterions(±), as was done previously [10]. For each class, separate sets of dcoefficients were determined by PLS regression.

2.4 Statistical Machine Learning Random Forest Regression (RFR) Model

The RFR open-source 'randomForest' library for the R statistical software was downloaded from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. The method works by constructing an ensemble of hundreds of decision trees employing about 200 RDKit-generated molecular descriptors [27]. The method was re-trained with the presently enlarged database, excluding the newly-approved drugs.

2.5 Sources of Solubility Data for the Test (New Drugs) and Training (Wiki-pS_0 Database) Sets

The 2021 mini-review of FDA drug approvals by Mullard [1] was a convenient starting point to identify the new drugs and to begin the search for their solubility values. Since the drugs are new, there are hardly any journal publications reporting properties of the compounds. Almost all the data were found in FDA filing documents. As part of the New Drug Application (NDA) process, the FDA Center for Drug Evaluation and Research (CDER, www.accessdata.fda.gov) publishes reports listing some physicochemical properties of compounds under consideration.

There was virtually no experimental detail about the measurements in the published regulatory reports. Many of the reported solubility values are of drugs in water (S_w), with saturation pH not reported. When the temperature was not stated or was reported as 'room' or 'ambient', it was assumed to be 23 °C for the purpose of calculations here. In the dearth of experimental detail, it is a challenge to assess the quality of the reported measurements in most of the FDA reports. Nevertheless, there are high quality data in some of the documents, where solubility measurements were published as a function of pH. Examples of some of these are presented below.

Of the 36 small-molecule NMEs approved in 2021, 39 independent *quantitative* solubility measurements were found only for 28 NMEs [28–62], given that some solubility data are redacted or presented as qualitative values (*e.g.*, 'insoluble', 'poorly soluble', 'very soluble') in FDA reports. The reported values were transformed into the intrinsic solubility scale, S_0, using known (or predicted when unavailable) pK_a values, and adjusted to 25 °C [63] using the program *p*DISOLX (*inADME* Research) [64–70]. Table 1 lists the solubility data (normalized as intrinsic values), along with the pK_a values used in the data analysis at the temperatures of measurement.

The *Wiki-*pS_0 intrinsic aqueous solubility database of mostly druglike molecules (currently with 7655 deeply-curated entries) was used to train the ABSOLV, GSE(*Φ,B*) and RFR models. Several hundred values from the database have already been published [10–13, 63–74], and the entire database is currently being prepared for publication as a book. The newly-approved drugs were used as external test sets and were excluded from the training process.

The structures of the 28 new drugs considered here are shown in Fig. 1. In dual-API drug products, each API was treated as a separate 'drug' in the data analysis.

*** Fig. 1 **goes here** ***

2.6 Sources of Octanol-Water Partition Coefficients (*clogP*) and Melting Points (*mp*)

Values of *clogP* were used in Eqs. 1 and 5 in place of experimental log *P* values. These were calculated by the Wildman-Crippen sum of atomic contributions method in the open-source RDKit chemoinformatics library [27]. Experimental *mp* values were employed where available or were calculated otherwise [75].

3 Results And Discussion

3.1 Data Reduction

About two-thirds of the drugs in Table 1 had their solubility measurements performed in two or more pH buffer solutions. This generally leads to more reliable determinations of log S_0, provided pK_a values are confidently known. For the rest of the drugs, S_0 values were determined from reported water solubility (S_w) values. In these cases, the *p*H of the saturated solutions was also calculated, assuming the Henderson-Hasselbalch (HH) equation is valid and the pK_a value is

reliable. When aggregates/complexes form or when supersaturation persists in the suspension, the HH equation does not accurately predict the shape of the log *S*-*p*H curve for ionizable molecules [65–70]. There is no direct way to recognize such anomalies just from a single S_w measurement.

In cases where measured pK_a values could not be found, they were calculated using the ChemAxon MarvinSketch v5.3.7 program (ChemAxon Ltd., https://www.chemaxon.com), as indicated by *italic* values in Table 1. In a few cases, it was possible to determine pK_a values directly in the analysis of the log *S*p*H* profiles (underlined values in Table 1).

Examples of experimental log *S*-*p*H profiles reported for some of the new drugs are shown in Fig. 2. The circle symbols represent the measured pH-dependent log *S* values at 25 ^{o}C. The solid curves represent best-fit regression curves. The dashed curves were calculated by the Henderson-Hasselbalch equation, using the best-fit log S_0 and the supplied/refined pK_a. Frame c is that of an ampholyte (sotorasib). The rest of the frames are of bases (daridorexant, finerenone, ponesimod, vericiguat, and avacopan). The data from the first five drugs appear to be well defined by the HH equation. It was possible not only to determine the best-fit log S_0, but also the values of pK_a (frames a, b, d, e) and the pK_{sp} (frames c, e).

*** Fig. 2 **goes here** ***

When profiles deviate from expected HH shapes, it may be possible to assess (and to correct for) the degree to which the measurements may be supersaturated or if aggregates/complexes are forming [65–70]. Figure 2f (avacopan) shows such an example of 'anomaly,' where for *p*H > 4, the reported solubility points are higher than that expected for a solution saturated in the free base (dashed curve). The solubility values in the *p*H 1–3 interval, which lie on the diagonal portion of the HH curve, define the intrinsic value, based on the reported pK_a. The suspension above *p*H 4 may have been: (a) supersaturated with respect to the free base during the measurement, or (b) due to solid being amorphous in the pH > 4 region, or (c) due to the formation of aggregates of the *neutral* molecule, or (d) complex formation between the free base and the buffers in solution [67, 68]. Solid-state characterization or LC/MS analysis of saturated solutions may be able to rule out some of the possibilities. The unfilled circle symbols above pH 6 were assigned zero weights in the regression analysis. Had only a S_w measurement been reported in water for avacopan and the pK_a was not known, the intrinsic solubility might have been determined at an order of magnitude too high. There would have been no hint of any 'anomaly.'

3.2 Trends in the PhysChem Properties of the Newly-Approved Drugs since 2016

Figure 3 shows the trends in distribution of several physicochemical properties of the FDA-approved NMEs covering the period of 2016–2021. Each bar is the average of a particular property for the drugs reported in a particular year. The dashed horizontal lines represent the average values of the property in the training set. Evidently, the properties of the test compounds (new drugs) exceed those of the training set molecules. The dotted lines represent trends in the property of the test set drugs, based on weight linear regression analysis covering the six-year period. Since 2016, on the average, the S_0 values have decreased by 0.50 log unit, the *clogP* values have increased by 0.34 log unit, and the *MW* values have increased by 22 g·mol^{-1} (to 477, compared to 298 in the training set). The average number of Hbond acceptors (*NHA*) has remained constant, while the average number of H-bond donors (*NHD*) and the Kier Φ molecular flexibility indices decreased slightly.

*** Fig. 3 **goes here** ***

3.3 Training the Prediction Models

3.3.1 Determination of the Three GSE Coefficients from Training Set iso-($\Phi + B$) Bins

The training set solubility data were sorted by $\Phi + B$ into 18 bins of increasing values. (Fewer bins were used in our previous study [13].) For a narrow range of $\Phi + B$ values in each bin, the three GSE c-coefficients in Eq. 5 were determined by linear PLS regression, in a similar way that Hansch *et al.* [24] had trained the GSE for different chemical classes of compounds. The c-constants are depicted by the points on the three curves in Fig. 4. The best-fit equations (*cf.*, Eqs. 5a-c) as functions of $\Phi + B$ are listed in the figure. The c_0 and c_1 functions follow the previously reported trends [13]. Evidently, solubility dependence on flexibility and H-bond acceptor strength are mediated by solution-phase interactions [76]. The crystal lattice contribution depicted by the c_2 function appears to show an upward trend with increasing $\Phi + B$, which was not evident in the earlier study [13] based on a larger test set of newly approved drugs and a smaller training set of drug-relevant compounds. The solubility of the most flexible molecules appears not to depend on crystal lattice contributions, where $c_2 \sim 0$ for $\Phi + B \sim 25$.

*** Fig. 4 **goes here** ***

From the thermodynamics considerations, the c_0 coefficient may be viewed as a measure of the solubility of the 'supercooled' liquid solute in octanol ($c_0 \approx \log S_{oct}^{SCL}$). Increasingly flexible molecules with strong H-bond acceptor character appear to be less miscible with octanol, as suggested by the decreasing c_0 coefficients with increasing $\Phi + B$ (Fig. 4). Between bins 1 and 18, S_{oct}^{SCL} decreases by five orders of magnitude. The SGE(*classic*) model assumes a constant 0.5 intercept in Eq. 1, which appears to be more consistent with rigid molecules ($\Phi + B \sim 2$). Given that the c_1 coefficient also changes with $\Phi + B$, the precise thermodynamic interpretation of the c_0 coefficient is less clear than in the classical derivation [15, 22, 23] where c_1 is constant at -1.

3.3.2 ABSOLV Training

As was done previously [10], the training set molecules were considered separately in each of four acid-base classes – with reference to predominant charge state at pH 7.4: acids(-), bases(+), neutrals(0), and zwitterions(±) (*cf.*, Fig. 7 in Ref [10]). In addition, the quaternary ammonium drugs, and drugs with *MW* > 800 Da were treated as separate classes. The d-coefficients in Eq. 6 for each of the six classes were determined by PLS regression using the log S_0 values from the database, excluding those of the 2021 NMEs. Table 2 summarizes the dcoefficients by classes.

*** Table 2 goes here ***

The d-coefficients in Table 2 are close to those reported in Table 1 of Ref [10]. Although there are about 1000 additional entries in the present database compared to that used previously, the statistics in the older study are slightly better (*e.g.*, the RMSE values for acids, bases, neutrals, and zwitterions were 0.98, 0.87, 1.01, and 0.77, resp., in Ref [10]). The residual plots in Fig. 7 in Ref [10] are visually indistinguishable from those presently calculated (data not shown).

3.3.3 Random Forest Model Training

As done previously [10–13], the Random Forest Regression (RFR) internal validation was applied to randomly-selected 30% of the database, based on training using the other 70% of the database (exclusive of new drugs). For molecules like those of the current database, it is expected that their log S_0 could be predicted with $r^2 = 0.89$, RMSE = 0.66, with 73% of the molecules 'correctly' predicted. The actual prediction statistics of the test compounds did not reach the expectations of the training set.

3.5 Solubility Prediction Results for the Newly-Approved Drugs

3.5.1 Model Testing

Figure 5 shows the results of the predictions of the solubility of the newly-approved drugs (external test sets) by the four models, as measured log S_0 vs. calculated log S_0 correlation plots. Table 3 summarizes the results. The solid diagonals are identity lines. The dashed diagonals are ± 0.5 log unit displaced from the identity lines. The measure of prediction performance (MPP) is indicated by the pie-charts as the percentage of predicted values that are within ± 0.5 log unit of the observed values [77]. Briefly, the four results (Figs. 5a-d) look similar. All the r^2 were in the range of 0.57 to 0.60, and RMSE values were between 1.18 and 1.22, as MPP values range from 25–46%. The GSE(*classic*) had the highest MPP and appeared to have a symmetrical distribution of residuals about the identity line. The other three models tended to overpredict solubility of drugs with log $S_0 < 7$, which may hint that those molecules possessed structural features not common to the training database. All models showed a systematic negative bias, ranging from − 0.22 to -0.37 log.

*** Fig. 5 goes here ***

*** Table 3 goes here ***

The consensus based on the average of the ABSOLV and GSE(*Φ,B*) models produced the best statistics, as indicated in Fig. 5e, with $r^2 = 0.67$ and RMSE = 1.08. The discrimination between the four models was higher in our previous study [13], covering the NMEs from 2016–2020.

3.5.2 More is Needed than Just Increasing the Size of the Training Set

Although the database has steadily increased in size over the last ten years, it has been our observation that this alone has not proportionately improved its ability to predict the solubility of drugs. Generally, the GSE(*classic*) underperformed when compared to the other models. The GSE(*Φ+ B*) matched the performance of the RFR model. Metrics such as those in Fig. 5 are comparable to those previously reported [10–13], although for the 2021 NMEs, the statistics are somewhat worse than those for the 2016–2020 NME set. Solubility prediction depends on multi-dimensional factors: *e.g.*, quality of measurements (training and test sets), distribution of training set molecules in chemical space in relation to the tested drugs, sensitivity of descriptors used in prediction models. Simply increasing the size of the solubility training set may not lead to improved predictions. Compiling a large database aimed at maximizing chemical diversity may be an inefficient strategy for predicting the solubility of novel molecules, given the enormous size of the chemical space, and since drugs appear to exist there as small tight clusters, as pointed out by Lipinski [78]. It would be helpful if the quality of future measurements were to improve. This could be better assessed in peer-reviewed publications than in regulatory filings. New descriptors which can better differentiate the factors affecting solubility also can be important for narrowing the gap between the accuracy of the prediction models and that of the experimental data.

4 Conclusion

If good practices in solubility measurement were adhered to, as detailed in the recent data-quality 'white paper' by experts from six countries [69], and the experimental details were more transparent, newly-reported measurements could be expected to achieve precision approaching that of the curated database used as the training set (average interlaboratory SD < 0.2 log unit). Presently, the data quality in the database is not the limiting factor in prediction, given that the best prediction root-mean-square error achieved in this study is above a log unit. The benchmark statistical machine learning approaches are probably up to the task in narrowing the gap between prediction and measurement. The Flexible-Acceptor GSE(*Φ,B*) performed nearly as well as the benchmark Random Forest regression method in predicting the aqueous intrinsic solubility of the newly-approved drugs since 2016. The consensus model based on the average predictions of the ABSOLV and GSE(*Φ,B*) methods was found to reduce the prediction biases in the separate methods, but perhaps even more significant, it slightly *outperformed* the Random Forest regression method overall. This is an advantage since the relatively-simple consensus model can be readily incorporated into spreadsheet calculations.

As new drugs are approved, it will be important to continue monitoring the quality of measured solubility. Matching prediction to measurement can be of immense practical value when prediction methods are applied to virtual libraries, in order to seek opportunities to minimize pharmacokinetic risks of large, but otherwise promising, candidate molecules.

Abbreviations

S_0	aqueous intrinsic solubility (i.e., the solubility of the *uncharged* form of the API)
S_w	solubility of the pure API (active pharmaceutical ingredient) in pure water
n	number of measurements of log S_0 in the training/test set
MPP	The measure of prediction performance [77] refers to the percent of 'correct' predictions, as defined by the count of absolute residuals $\|\log S_0^{obs} - \log S_0^{calc}\| \leq 0.5$ divided by n. MPP is represented as a pie chart in the correlation plots (Fig. 5).
RMSE	root-mean-square error, accounting for bias in the prediction of external test set solubility values: $RMSE = [\ 1/(n-1)\ \Sigma_i\ (y_i^{obs} - bias - y_i^{calc})^2\ ]^{1/2}$, where $y = \log S_0$.
r^2	coefficient of determination, accounting for bias in prediction of external test set solubility values [79]: $r^2 = 1 - \Sigma_i\ (y_i^{obs} - bias - y_i^{calc})^2\ /\Sigma_i\ (y_i^{obs} - \langle y \rangle)^2$, where $y = \log S_0$, and $\langle y \rangle$ is the mean value of observed log S_0.
bias	intercept in the regression fit: $y^{obs} = a + b\ y^{calc}$, where the slope factor is fixed at unity.
SD	standard deviation: $SD = [\ 1/n\ \Sigma_i\ (y_i^{obs} - \langle y \rangle)^2\ ]^{1/2}$, where $\langle y \rangle$ = mean value of log S_0.

Declarations

Funding Declaration The authors did not receive support from any organization for the submitted work.

Conflict of Interest The authors declare that they have no known competing financial interests that could have appeared to influence the work reported in this paper.

Acknowledgements This study is dedicated to the memory of Professor Michael Abraham, whose pioneering work in the critical role of hydrogen bonding in solvation has influenced the authors deeply. The complete *Wiki-pS_0* database is planned to be released in book form: A. Avdeef. *Intrinsic Aqueous Solubility - Curated Data for Pharmaceutical Research* (under discussion with publisher).

References

1. Mullard, A.: 2021 FDA drug approvals. The FDA approved 50 novel drugs in 2021, including the first KRAS inhibitor for cancer and the first anti-amyloid antibody for Alzheimer's disease. Nat. Rev. Drug Discov. **21,** 83-88 (2022); https://doi.org/10.1038/d41573-022-00001-9**.**
2. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. **23**, 3-25 (1997).
3. Leeson, P.D.: Molecular inflation, attrition & the rule of five. Adv. Drug Deliv. Rev. **101**, 22-33 (2016).
4. Bergström, C.A.S.; Charman, W.N.; Porter, C.J.H.: Computational prediction of formulation strategies for beyond-rule-of-5 compounds. Adv. Drug Deliv. Rev. **101**, 6-21 (2016).
5. Krämer, S.D.; Aschmann, H.E.; Hatibovic, M.; Hermann, K.F.; Neuhaus, C.S.; Brunner, C.; Belli, S.: When barriers ignore the rule-of-five. Adv. Drug Del. Rev. **101**, 62-74 (2016).
6. Ermondi, G.; Vallaro, M.; Goetz, G.; Shalaeva, M.; Caron, G.: Updating the portfolio of physicochemical descriptors related to permeability in the beyond the rule of 5 chemical space. Eur. J. Pharm. Sci. **146**, 105274 (2020); https://doi.org/10.1016/j.ejps.2020.105274.
7. Caron, G.; Kihlberg, J.; Ermondi, G.: Intramolecular hydrogen bonding: An opportunity for improved design in medicinal chemistry. Med. Res. Rev. **39**, 1707-1729 (2019); https://doi: 10.1002/med.21562.
8. Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G.: Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. Drug Discov. Today **25**, 621-627 (2020); https://doi.org/10.1016/j.drudis.2020.01.012.
9. Carrupt, P.A.; Testa, B.; Bechalany, A.; el Tayar, N.; Descas, P.; Perrissoud, D.: Morphine 6-glucuronide and morphine 3-glucuronide as molecular chameleons with unexpected lipophilicity. J. Med. Chem. **34**, 1272-1275 (1991).
10. Avdeef, A.: Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with *Wiki-pS_0* database. ADMET & DMPK **8**, 29-77 (2020); http://dx.doi.org/10.5599/admet.766.
11. Avdeef, A.; Kansy, M.: Can small drugs predict the intrinsic aqueous solubility of 'beyond Rule of 5' big drugs? ADMET & DMPK **8** (2020). http://dx.doi.org/10.5599/admet.794.
12. Avdeef, A.; Kansy M.: Flexible-Acceptor General Solubility Equation for beyond Rule of 5. Drugs. Mol. Pharm. **17**, 3930-3940 (2020); https://doi.org/10.1021/acs.molpharmaceut.0c00689.
13. Avdeef, A.; Kansy, M.: Predicting solubility of newly-approved drugs (2016-2020) with a simple ABSOLV and GSE(Flexible-Acceptor) consensus model outperforming random forest regression. J. Solution Chem. (2022) Feb 7;1-36; https://doi: 10.1007/s10953-022-01141-7.
14. Breiman, L.: Random forests. Mach. Learn. **45**, 5-32 (2001).
15. Yalkowsky, S.H; Valvani, S.C.: Solubility and partitioning I: Solubility of nonelectrolytes in water. J. Pharm. Sci. **69**, 912-922 (1980).
16. Abraham, M.H.; Le, J.: The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. J. Pharm. Sci. **88**, 868-880 (1999).
17. Mullard, A.: 2016 FDA drug approvals. FDA approval count fell last year, despite a steady regulatory filing rate. Nat. Rev. Drug Discov. **16**, 73-76 (2017).

18. Mullard, A.: 2017 FDA drug approvals. The FDA approved 46 new drugs last year, the highest total in more than two decades. Nat. Rev. Drug Discov. **17**, 81-85 (2018).
19. Mullard, A.: 2018 FDA drug approvals. The FDA approved a record 59 drugs last year, but the commercial potential of these drugs is lackluster. Nat. Rev. Drug Discov. **18**, 85-89 (2019).
20. Mullard, A.: 2019 FDA drug approvals. The FDA approved 48 new drugs last year, keeping up the momentum of recent years. Nat. Rev. Drug Discov. **19**, 79-84 (2020).
21. Mullard, A.: 2020 FDA drug approvals. The FDA approved 53 novel drugs in 2020, the second highest count in over 20 years. Nat. Rev. Drug Discov. **20**, 85-90 (2021).
22. Jain, N.; Yalkowsky, S.H.: Estimation of the aqueous solubility I: Application to organic nonelectrolytes. J. Pharm. Sci. **90**, 234-252 (2001).
23. Ran, Y.; Jain, N.; Yalkowsky, S.H.: Prediction of aqueous solubility of organic compounds by the General Solubility Equation (GSE). J. Chem. Inf. Comput. Sci. **41**, 1208-1217 (2001).
24. Hansch, C.; Quinlan, J.E.; Lawrence, G.L.: Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. J. Org. Chem. **33**, 347-350 (1968).
25. Kier, L.B.: An index of molecular flexibility from kappa shape attributes. Quant. Struct.-Act. Relat. **8**, 221-224 (1989).
26. Platts, J.A.; Butina, D.; Abraham, M.H.; Hersey A.: Estimation of molecular linear free energy relation descriptors using a group contribution approach. J. Chem. Inf. Comput. Sci. **39**, 835-845 (1999).
27. Landrum, G.; Lewis, R.; Palmer, A.; Stiefl, N.; Vulpetti, A.: Making sure there's a give associated with the take: Producing and using open-source software in big pharma. J. Cheminformatics **3**, 1-1 (2011); http://www.rdkit.org/.
28. Schoepfer, J.; Jahnke, W.; Berellini, G.; Buonamici, S.; Cotesta, S.; Cowan-Jacob, S.W.; Dodd, S.; Drueckes, P.; Fabbro, D.; Gabriel, T.; Groell, J.-M.; Grotzfeld, R.M.; Hassan, A.Q.; Henry, C.; Iyer, V.; Jones, D.; Lombardo, F.; Loo, A.; Manley, P.W.; Pellé, X.; Rummel, G.; Salem, B.; Warmuth, M.; Wylie, A.A.; Zoller, T.; Marzinzik, A.L.; Furet, P.: Discovery of asciminib (ABL001), an allosteric inhibitor of the tyrosine kinase activity of BCR-ABL1. J. Med. Chem. **61**, 8120-8135 (2018); https://doi: 10.1021/acs.jmedchem.8b01040.
29. Food and Drug Administration (USA): Asciminib (Scemblix). Novartis. NDA 215358. Multi-Discipline Review. 24 June 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215358Orig1s000,Orig2s000MultidisciplineR.pdf. Accessed 28 Jan 2022.
30. Food and Drug Administration (USA): Asciminib (Scemblix). Novartis. NDA 215358. Product Quality Review(s). 24 June 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215358Orig1s000,Orig2s000ChemR.pdf. Accessed 28 Jan 2022.
31. Food and Drug Administration (USA): Avacopan (Tavneos). ChemoCentrix. NDA 214487.Multi-Discipline Review. 7 Jul 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214487Orig1s000MultidisciplineR.pdf. Accessed 30 Jan 2022.
32. Food and Drug Administration (USA): Avacopan (Tavneos). ChemoCentrix. NDA 214487. Product Quality Review(s). 19Mar 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214487Orig1s000ChemR.pdf. Accessed 28 Jan 2022.
33. Food and Drug Administration (USA): Belumosudil.Mesylate (Resurock). Kadmon. NDA 214783. Product Quality Review(s). 2 June 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214783Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
34. Food and Drug Administration (USA): Belzutifan (WELIREG). Merck. NDA 215383. Product Quality Review(s). 15 Jan 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215383Orig1s000ChemR.pdf. Accessed 28 Jan 2022.
35. Food and Drug Administration (USA): Cabotegravir(Cabenuva Kit), ViiV. NDA 212887Orig1s000, 212888Orig2s000. Product Quality Review(s). 30 Nov 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/212887Orig1s000,212888Orig1s000ChemR.pdf. Accessed 26 Jan 2022.
36. Food and Drug Administration (USA): Daridorexant (Quviviq). Idorsia Pharmaceuticals Ltd. NDA 214985. Product Quality Review(s).13 Aug 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2022/214985Orig1s000ChemR.pdf. Accessed 19 Feb 2022.
37. Food and Drug Administration (USA): Serdexmethylphenidate chloride & Dexmethylphenidate hydrochloride (Azstarys). Commave Therapeutics. NDA 212994. Multi-Discipline Review. 2 Mar 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/212994Orig1s000MultidisciplineR.pdf. Accessed 6 Feb 2022.
38. Food and Drug Administration (USA): Drospirenone+Estetrol (Nextstellis). Mayne Pharma. NDA 214154. Product Quality Review(s). 9 April 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214154Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
39. European Medicines Agency: Fexinidazole (Fexinidazole Winthrop), CHMP assessment report, Procedure No. EMEA/H/W/002320/0000. 15 Nov 2018; https://www.ema.europa.eu/en/documents/outside-eu-assessment-report/fexinidazole-winthrop-assessment-report_en.pdf. Accessed 1 Feb 2022.
40. Food and Drug Administration (USA): Finerenone (Kerendia). Bayer. NDA 215341. Product Quality Review(s). 31 Mar 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215341Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
41. Food and Drug Administration (USA): Infigratinib (Truseltiq). QED Therapeutics. NDA 214622. Product Quality Review(s). 20 Sep 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214622Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
42. Food and Drug Administration (USA): Infigratinib (Truseltiq). QED Therapeutics. NDA 214622. Multi-Discipline Review. 29 Sep 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214622Orig1s000MultidisciplineR.pdf. Accessed 27 Jan 2022.
43. Food and Drug Administration (USA): Maralixibat Chloride (Livmarli). Mirum. NDA 214662. Product Quality Review(s). 22Sep 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214662Orig1s000ChemR.pdf. Accessed 1 Feb 2022.
44. Food and Drug Administration (USA): Maribavir (Livtencity). Takeda. NDA 215596. Product Quality Review(s). 22 Sep 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215596Orig1s000ChemR.pdf. Accessed 28 Jan 2022.

45. Sun, K.; Welty, D.: Elucidation of metabolic and disposition pathways for maribavir in nonhuman primates through mass balance and semi–physiologically based modeling approaches. Drug Metab. Dispos. **49**, 1025-1037 (2021); https://doi.org/10.1124/dmd.121.000493.
46. Spira, J.; Lehmann, F.: Lyophilized preparations of cytotoxic dipeptides. Patent: US 2014.0128462A1. May 8, 2014; https://patentimages.storage.googleapis.com/a7/4e/ad/af3e43497ed19c/US20140128462A1.pdf.
47. Food and Drug Administration (USA): Mobocertinib (Exkivity). Takeda. NDA 215310. Highlights of Prescribing Information. Sep 2021; https://www.accessdata.fda.gov/drugsatfda_docs/label/2021/215310s000lbl.pdf. Accessed 2 Feb 2022.
48. Food and Drug Administration (USA): Mobocertinib (Exkivity). Takeda. NDA 215310. Product Quality Review(s). 9 Aug 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215310Orig1s000ChemR.pdf. Accessed 28 Jan 2022.
49. Food and Drug Administration (USA): Odevixibat (Bylvay). Albireo. NDA 215498. Product Quality Review(s). 5 Jul 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/215498Orig1s000ChemR.pdf. Accessed 28 Jan 2022.
50. Benet, L.Z.; Broccatelli, F.; Oprea, T.I.: BDDCS applied to over 900 drugs. AAPS J. **13,** 519-547 (2011).
51. Fornells, E.; Fuguet, E.; Mañéa, M.; Ruiz, R.; Box, K.; Bosch, E.; Ràfols, C.: Effect of vinylpyrrolidone polymers on the solubility and supersaturation of drugs; a study using the CheqSol method. Eur. J. Pharm. Sci. **117**, 227-235 (2018).
52. Marano, S.; Barker, S.A.; Raimi-Abraham, B.T.; Missaghi, S.; Rajabi-Siahboomi, A.; Craig, D.Q.M.: Development of microfibrous solid dispersions of poorly water-soluble drugs in sucrose using temperature-controlled centrifugal spinning. Eur. J. Pharm. Biopharm. **103**, 84-94 (2016).
53. Food and Drug Administration (USA): Ponesimod (Ponvory). J&J. NDA 213498. Product Quality Review(s). 12 Nov 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/213498Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
54. Food and Drug Administration (USA): Rilpivirine.HCl (Edurant), Tibotek. NDA 202022. Clinical Pharmacology and Biopharmaceutics Review. 25 Mar 2011; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/202022Orig1s000ClinPharmR.pdf. Accessed 27 Jan 2022.
55. Kommavarapu, P.; Maruthapillai, A.; Palanisamy, K.; Sunkara, M.: Preparation and characterization of rilpivirine solid dispersions with the application of enhanced solubility and dissolution rate. Beni-Suef Univ. J. Basic Appl. Sci. **4**, 71-79 (2015); http://dx.doi.org/10.1016/j.bjbas.2015.02.010.
56. Food and Drug Administration (USA): Sotorasib (Lumakras). Amgen. NDA 214665. Product Quality Review(s). 7 May 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214665Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
57. Fink, C.; Lecomte, M.; Badolo, L.; Wagner, K.; Mäder, K.; Peters, S.-A.: Identification of solubility-limited absorption of oral anticancer drugs using PBPK modeling based on rat PK and its relevance to human. Eur. J. Pharm. Sci. **152** (2020) 105431; https://doi.org/10.1016/j.ejps.2020.105431.
58. Food and Drug Administration (USA): Tivozanib (Fotivda). Aveo. NDA 212904. Product Quality Review(s). 30 Sep 2020. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/212904Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
59. Food and Drug Administration (USA): Umbralisib (Ukoniq). TG Therapeutics. NDA 213176. Product Quality Review(s). 11 Sep 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/213176Orig1Orig2s000ChemR.pdf. Accessed 26 Jan 2022.
60. Food and Drug Administration (USA): Vericiguat (Verquvo). Merck, Sharp & Dohme. NDA 214377. Product Quality Review(s). 1 Feb 2019; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/214377Orig1s000ChemR.pdf. Accessed 26 Jan 2022.
61. Food and Drug Administration (USA): Viloxazine (Qelbree). Supernus Pharmaceutics. NDA 211964. Product Quality Review(s). 5 Mar 2021; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/211964Orig1s000ChemR.pdf. Accessed 27 Jan 2022.
62. Food and Drug Administration (USA): Voclosporin (Lupkynis). Aurinia Pharmaceutics. NDA 213716. Product Quality Review(s). 8 Oct 2020; https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/213716Orig1s000ChemR.pdf. Accessed 26 Jan 2022.
63. Avdeef, A.: Solubility temperature dependence predicted from 2D structure. ADMET & DMPK **3**, 298-344 (2015).
64. Völgyi, G.; Marosi, A.; Takács-Novák, K.; Avdeef, A.: 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**, 48-62 (2013).
65. Avdeef, A.: Anomalous solubility behavior of several acidic drugs. ADMET & DMPK **2**, 33-42 (2014).
66. Avdeef, A.: Phosphate precipitates and water-soluble aggregates in re-examined solubility-*p*H data of twenty-five basic drugs. ADMET & DMPK **2**, 43-55 (2014).
67. Verbić, T. Z.; Avdeef, A.: Solubility-*p*H profile of desipramine hydrochloride in saline phosphate buffer: enhanced solubility due to drug-buffer aggregates. Eur. J. Pharm. Sci. **133**, 264-274 (2019).
68. Marković, O.S.; Patel, N.G.; Serajuddin, A.T.M.; Avdeef, A.; Verbić, T.Ž.: Nortriptyline hydrochloride solubility-pH profiles in a saline phosphate buffer: Drug-phosphate complexes and multiple pHmax domains with a Gibbs phase rule "soft" constraints. Mol. Pharm. **19**, 710-719 (2022); https://doi: 10.1021/acs.molpharmaceut.1c00919.
69. Avdeef, A.; Fuguet, E.; Llinàs, A.; Ràfols, C.; Bosch, E.; Völgyi, G.; Verbić, T.; Boldyreva, E.; Takács-Novák, K. Equilibrium solubility measurement of ionizable drugs – consensus recommendations for improving data quality. ADMET & DMPK **4**, 117-178 (2016).
70. Bergström, C.A.S.; Avdeef, A. Perspectives in solubility measurement and interpretation. ADMET & DMPK **7**, 88-105 (2019).
71. Avdeef, A.: *Absorption and Drug Development,* Second Edition, Wiley-Interscience, Hoboken NJ, 2012.
72. Avdeef, A.: Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. ADMET & DMPK **7**, 210-219 (2019). http://dx.doi.org/10.5599/admet.698.
73. Llinàs, A.; Avdeef, A.: Solubility challenge revisited after ten years, with multi-lab shake-flask data, using tight (SD ~ 0.17 log) and loose (SD ~ 0.62 log) test sets. J. Chem. Inf. Model. **59**, 3036-3040 (2019); https://doi: 10.1021/acs.jcim.9b00345.

74. Llinàs, A.; Oprisiu, I.; Avdeef, A.: Findings of the second challenge to predict aqueous solubility. J. Chem. Inf. Model. **60**, 4791-4803 (2020); https://doi: 10.1021/acs.jcim.0c00701
75. Lang, A.S.I.D.; Bradley, J.-C.: ONS melting point model 010. QsarDB content. Property mpC. http://qsardb.org/repository/predictor/10967/104?model=rf.
76. Hughes, L.D.; Palmer, D.S.; Nigsch, F.; Mitchell, J.B.O.: 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**, 220-232 (2008).
77. Hopfinger, A.J.; Esposito, E.X.; Llinàs, A.; Glen, R.C.; Goodman, J.M.: Findings of the challenge to predict aqueous solubility. J. Chem. Inf. Model. **49** , 1-5 (2009).
78. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Tox. Meth.* **44**, 235-249 (2000).
79. Avdeef, A.: Do you know your r^2? ADMET& DMPK **9(1)** (2021); https://doi.org/10.5599/admet.888.
80. Avdeef, A.; Sugano, K.: Salt solubility and disproportionation - uses and limitations of equations for pH_{max} and the in-silico prediction of pH_{max}. J. Pharm. Sci. **111**, 225-246 (2022); https://doi: 10.1016/j.xphs.2021.11.017.

Tables

Table 1. Physicochemical properties of newly-approved drugs (2021)

Approved API/Drug	CAS	S_0 (µg/mL) [a]	t (°C)	*MW*	*mp* (°C)	pK_a(acid) [b]	pK_a(base) [b]	*clogP* [c]	*A* [d]	*B* [d]	S_π [d]
Asciminib	1492952-76-7	4	23	449.84	198	—	3.93	3.46	1.16	2.18	3.32
Asciminib	1492952-76-7	39	25	449.84	198	—	3.93	3.46	1.16	2.18	3.32
Asciminib [e]	1492952-76-7	19	25	449.84	198	—	3.93	3.46	1.16	2.18	3.32
Asciminib	1492952-76-7	18	37	449.84	198	—	3.81	3.46	1.16	2.18	3.32
Avacopan	1346623-17-3	0.0016	23	581.64	187	—	4.70	8.05	0.61	1.7	3.23
Belumosudil	911417-87-3	0.49	37	452.51	185	—	1.49, 5.15	4.82	0.79	2.04	3.45
Belzutifan	1672668-24-4	10	25	383.34	190	—	—	3.29	0.39	1.51	2.72
Cabotegravir [e]	1051375-10-0	3.6	25	405.35	197	7.26	0.97	0.96	0.57	2.71	3.71
Cabotegravir	1051375-10-0	3.6	37	405.35	197	7.19	1.13	0.96	0.57	2.71	3.71
Daridorexant	1505484-82-1	0.92	23	450.92	193	—	7.30	4.27	0.35	1.95	3.59
Dexmethylphenidate	40431-64-9	49	25	233.30	203	—	*9.09*	2.09	0.13	0.94	1.29
Dexmethylphenidate	40431-64-9	52	25	233.30	203	—	*9.09*	2.09	0.13	0.94	1.29
Drospirenone	67392-87-4	16	37	366.49	198	—	—	4.31	0.00	1.24	3.29
Estetrol	15183-37-6	289	37	304.38	235	10.0	—	1.55	1.31	1.56	2.80
Fexinidazole	59729-37-2	5.9	23	279.31	123	—	1.8	2.63	0.00	1.00	2.22
Finerenone	1050477-31-0	38	37	378.42	181	—	3.27	2.99	0.47	1.95	3.03
Infigratinib	872511-34-7	0.079	37	560.48	177	*9.76*	*1.99, 7.95*	5.35	0.44	2.36	3.63
Maralixibat(+).Cl	228113-66-4	3098	37	710.41	176	—	*2.18, 5.49*	3.63	0.31	2.24	3.31
Maribavir	176161-24-3	885	23	376.24	192	—	4.15	1.77	0.88	2.01	2.52
Maribavir	176161-24-3	728	23	376.24	192	—	4.56	1.77	0.88	2.01	2.52
Maribavir	176161-24-3	362	37	376.24	192	—	4.15	1.77	0.88	2.01	2.52
Melflufen	380449-51-4	3.9	19	498.42	210	—	7.13	3.27	0.47	2.06	3.17
Mobocertinib	1847461-43-1	1.3	37	585.70	182	—	*1.01, 8.65*	5.08	0.56	2.80	3.88
Odevixibat	501692-44-0	0.0043	37	740.93	134	*3.55, 9.17*	—	5.88	1.96	3.34	5.39
Olanzapine	132539-06-1	10	23	312.44	195	—	5.59, 8.03	3.44	0.13	1.45	1.59
Olanzapine	132539-06-1	18	23	312.44	195	—	5.59, 8.03	3.44	0.13	1.45	1.59
Olanzapine	132539-06-1	5.4	37	312.44	195	—	5.36, 7.74	3.44	0.13	1.45	1.59
Ponesimod	854107-55-4	10	37	460.97	174	—	2.17	4.27	0.50	1.99	2.69

Rilpivirine	500287-72-9	0.011	23	366.42	248	—	5.16	4.99	0.16	1.26	3.13
Rilpivirine [e]	500287-72-9	0.022	37	366.42	248	—	4.99	4.99	0.16	1.26	3.13
Serdexmethylphenidate(+).Cl [e]	1996626-29-9	4957	25	499.51	167	*2.56*	—	-0.28	0.50	2.42	3.81
Serdexmethylphenidate(+).Cl [e]	1996626-29-9	4712	25	499.51	167	*2.56*	—	-0.28	0.50	2.42	3.81
Sotorasib	2296729-00-3	23	37	560.59	189	8.05	4.68	4.48	0.69	2.89	3.72
Tepotinib	1100598-32-0	0.022	37	492.57	178	—	2.72, 9.24	4.01	0.90	2.39	3.56
Tivozanib	475108-18-0	0.045	25	454.86	187	—	5.90	5.64	0.59	1.90	3.61
Umbralisib	1532533-67-7	1.8	23	571.55	141	—	2.71	6.66	0.23	2.17	3.91
Vericiguat	1350653-20-1	1.0	37	426.38	205	—	4.35	2.56	0.84	1.97	3.55
Viloxazine	46817-91-8	42969	23	237.29	89	—	*8.19*	1.45	0.16	1.24	1.74
Voclosporin	515814-01-4	7.6	20	1214.62	144	—	—	3.44	1.25	7.66	10.22

[a] Intrinsic solubility, determined here by refinement of published solubility-pH data at the indicated temperature. 'Room'/unreported temperature assumed to b

[b] Ionization constants, either published values, or, if underlined, which refer to values determined here based on solubility-pH data, or if in *italics*, which refer t calculated by ChemAxon MarvinSketch v5.3.7 program (ChemAxon Ltd., https://www.chemaxon.com), since published values for these were located.

[c] Crippen-type octanol-water partition coefficient, calculated in RDKit[27].

[d] A, B, S_π, E, V = Abraham solvation descriptors; NHA, NHD, Nrot = number of H-bond acceptors, donors, and rotatable groups, resp.; Φ = Kier molecular flexibili

[e] Intrinsic solubility determined from δ-type (disproportionated) salt [80].

Table 2. Determination of ABSOLV coefficients by PLS regression analysis of training set [a]

Abraham Solvation Equation	const	*A*	*B*	S_π	*E*	*V*	*A·B*	r^2	RMSE	n
acids	-0.28	0.25	1.08	0.03	-0.91	-1.79	0.43	0.65	1.17	1579
bases	-0.41	-0.50	1.97	0.28	-1.67	-1.40	0.06	0.64	1.11	932
neutrals	-0.43	-0.15	1.71	0.12	-1.52	-1.35	0.26	0.63	1.12	4205
zwitterions	1.54	-1.45	0.85	-0.24	-0.99	-1.07	0.39	0.70	0.88	642
quaternaries	-0.03	-1.43	0.96	-0.02	-0.59	-0.37	0.10	0.63	0.94	35
big molecules (MW>800 Da)	-3.56	0.51	0.13	-0.04	-0.08	-0.32	0.01	0.29	1.21	78

[a] PLS calculates Pearson's r^2. See text.

Table 3. Predicted intrinsic solubility (log S_0) of newly approved drugs (2021)

Approved Drug	log S_0 [a] (mol·L^{-1})	*SD* [b]	n [c]	GSE(*classic*) [d]	ABSOLV [e]	GSE(*Φ,B*) [f]	RFR [g]	Consensus [h]	Residual [i]
Asciminib	-4.50	0.40	4	-4.69	-4.51	-4.67	-5.04	-4.59	0.1
Avacopan	-8.53		1	-9.17	-6.67	-7.16	-6.40	-6.92	-1.6
Belumosudil	-6.18		1	-5.92	-7.04	-5.48	-5.62	-6.26	0.1
Belzutifan	-4.58		1	-4.44	-3.42	-4.46	-4.64	-3.94	-0.6
Cabotegravir	-5.11	0.08	2	-2.18	-2.44	-2.93	-3.84	-2.68	-2.4
Daridorexant	-5.66		1	-5.45	-6.36	-5.20	-5.51	-5.78	0.1
Dexmethylphenidate	-3.66	0.02	2	-3.37	-2.62	-3.55	-2.75	-3.09	-0.6
Drospirenone	-4.51		1	-5.54	-4.88	-5.46	-5.26	-5.17	0.7
Estetrol	-3.18		1	-3.15	-3.66	-3.29	-2.82	-3.48	0.3
Fexinidazole	-4.65		1	-3.11	-3.98	-3.39	-3.87	-3.68	-1.0
Finerenone	-4.11		1	-4.05	-4.20	-4.22	-4.16	-4.21	0.1
Infigratinib	-6.99		1	-6.37	-6.34	-5.57	-5.50	-5.95	-1.0
Maralixibat(+).Cl	-2.37		1	-4.64	-2.07	-4.74	-5.15	-3.40	1.0
Maribavir	-2.81	0.29	3	-2.94	-3.98	-3.42	-3.70	-3.70	0.9
Melflufen	-5.07		1	-4.62	-4.34	-4.70	-5.20	-4.52	-0.5
Mobocertinib	-5.78		1	-6.15	-7.09	-5.42	-5.44	-6.26	0.5
Odevixibat	-8.38		1	-6.47	-6.85	-5.50	-5.75	-6.18	-2.2
Olanzapine	-4.52	0.34	3	-4.64	-4.33	-4.62	-4.16	-4.47	0.0
Ponesimod	-4.76		1	-5.26	-5.55	-5.01	-5.25	-5.28	0.5
Rilpivirine	-7.45	0.03	2	-6.72	-6.09	-6.10	-7.16	-6.09	-1.4
Serdexmethylphenidate(+).Cl	-2.01	0.02	2	-0.64	-1.15	-2.71	-3.01	-1.93	-0.1
Sotorasib	-4.48		1	-5.62	-5.78	-5.21	-5.08	-5.50	1.0
Tepotinib	-7.50		1	-5.04	-6.21	-4.89	-5.16	-5.55	-2.0
Tivozanib	-7.00		1	-6.76	-5.68	-6.03	-6.22	-5.86	-1.1
Umbralisib	-5.46		1	-7.32	-7.87	-6.29	-6.73	-7.08	1.6
Vericiguat	-5.88		1	-3.86	-5.89	-4.07	-4.64	-4.98	-0.9
Viloxazine	-0.73		1	-1.59	-1.94	-2.30	-1.99	-2.12	1.4
Voclosporin	-5.32	-	1	-4.13	-4.04	-4.53	-4.96	-4.29	-1.0

[a] Logarithm, base 10, of intrinsic solubility, S_0, averaged over n values, normalized to 25 °C [63] ('observed' value).

[b] Estimated standard deviation in the n-determined log S_0 value(s), with the overall average of 0.17 log.

[c] Number of independently reported solubility data sources for the determinations of log S_0.

[d] Yalkowsky's General Solubility Equation[15]: log S = 0.5 - *clogP* - 0.01 (*mp*-25); *clogP* calculated in RDKit [27].

[e] Abraham Solvation Equation (Eq. 6). See text.

[f] 'Flexible-Acceptor' enhanced GSE calculation [12,13], with the three traditional constant coefficients expanded as functions of Kier flexibility index, Φ, and Abraham's H-bond acceptor parameter, B (Eq. 5). See text.

[g] Breiman's Random Forest Regression model [14], trained on *Wiki-p*S_0 values, excluding those of 2021 NMEs.

[h] Consensus model = average of ABSOLV and GSE(*Φ,B*) predictions.

[i] Residual = observed log S_0 minus consensus value.

Figures

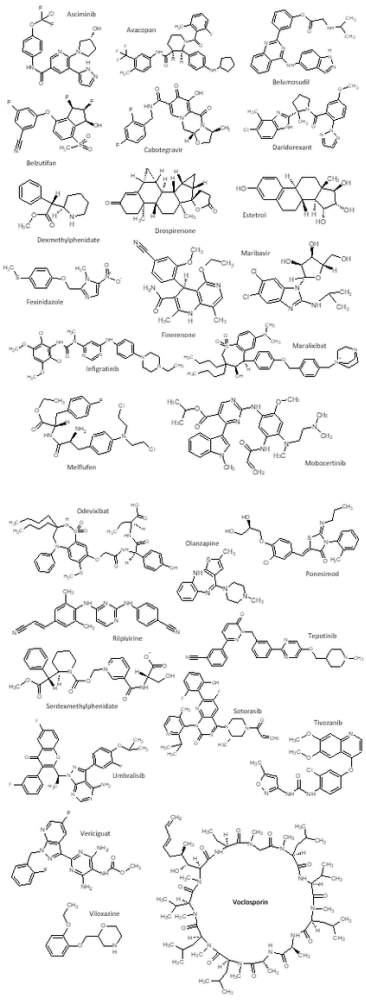

Figure 1

Structures of the 28 NMEs considered in this study

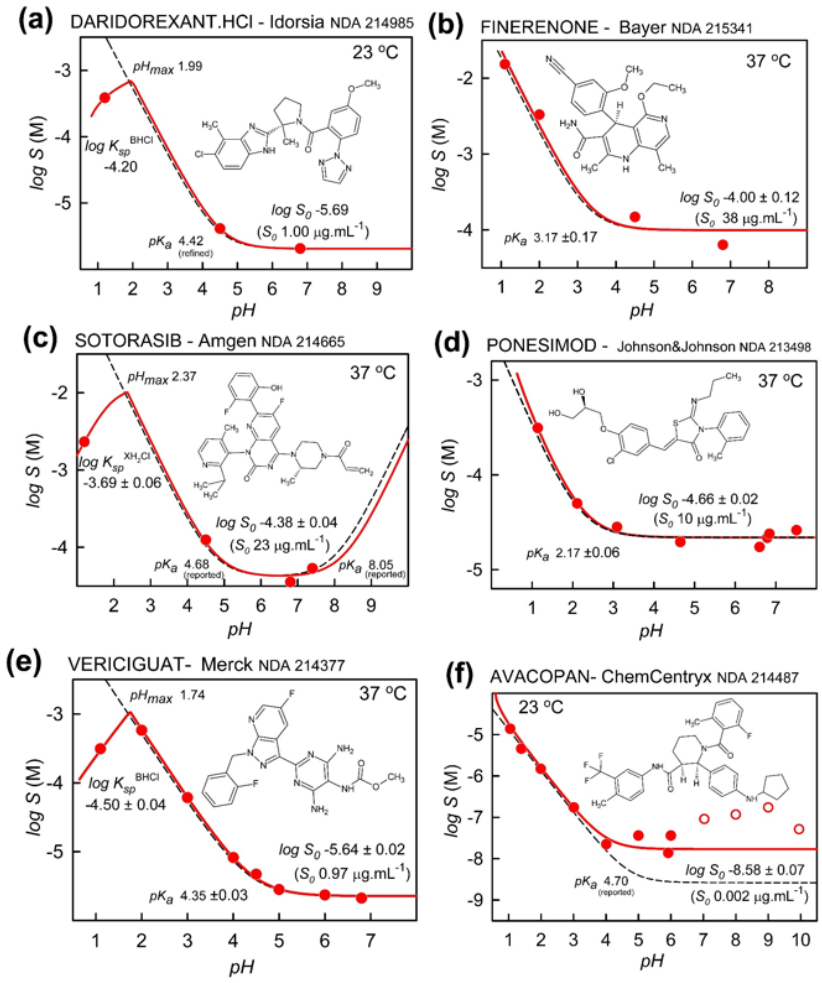

Figure 2

New-drug examples of log *S*-*p*H profiles. The solid red curves are the best fit to the measured data (circles), using the regression analysis program *p*DISOL-X. It was also possible to determine the pK_a values in cases **(a,b,d,e)**. The dashed curves were calculated using the Henderson-Hasselbalch equation, incorporating the pK_a used and the refined log S_0. In cases **(c)** and **(e)**, it was possible to determine the salt solubility products.

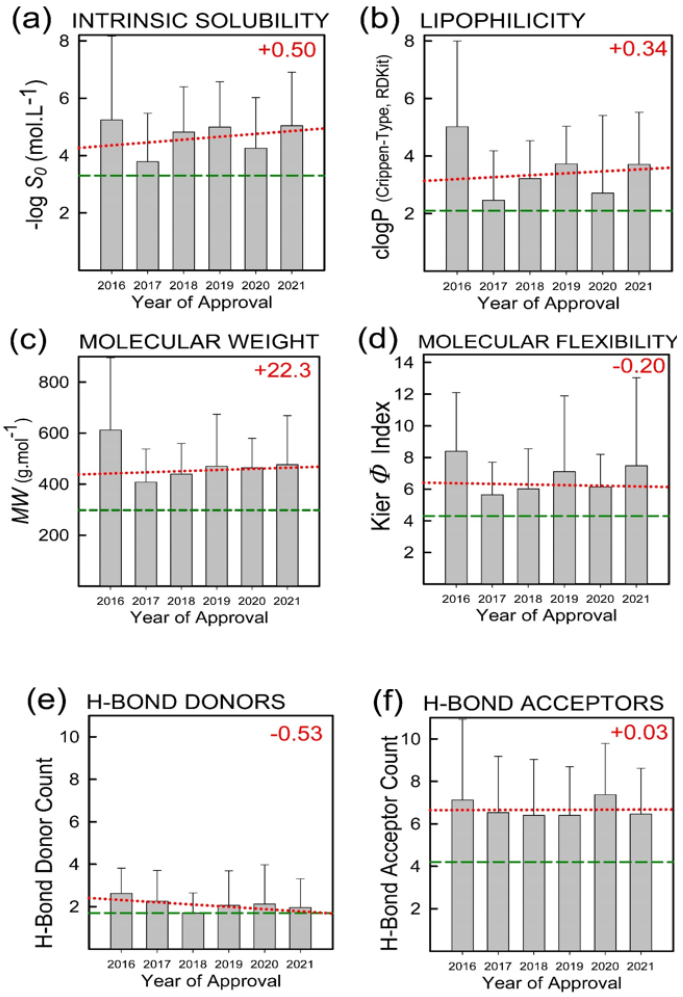

Figure 3

Trends in distribution of several physicochemical properties of the FDA-approved NMEs (test sets) covering the period of 2016-2021. Each bar is the average of a particular property for a given year. The dashed horizontal lines = average values of the property in the training set. The dotted lines = weighted trends in the property of NMEs for the five-year period: log S_0 decreased by 0.50 log, the *clogP* increased by 0.34 log unit, and the *MW* by 22 g·mol^{-1}. The average number of H-bond donors (*NHD*) and the Kier Φ molecular flexibility indices decreased slightly, but the average number of H-bond acceptors (*NHA*) has remained essentially constant.

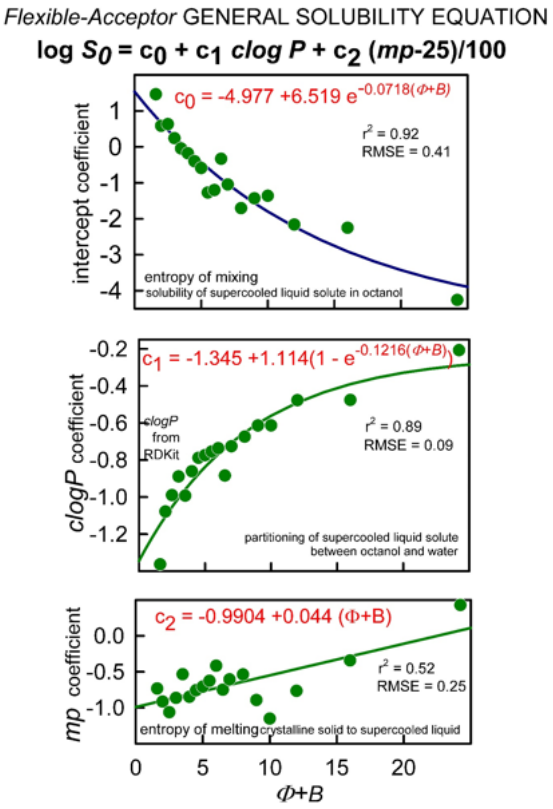

Figure 4

Re-training the Flexible-Acceptor GSE(*Φ,B*) model. The solubility data in the training set were sorted on *Φ+B* and then divided into 18 practically constant-value (*Φ+B*) bins. On the average, each bin contained about 413 log S_0 measurements. For each bin (represented by a point in the plots), the three constants in Eq. 1 (*cf.*, Eq. 5) were determined by PLS regression to best fit the intra-bin solubility data. Quaternary ammonium compounds and newly approved drugs were excluded from the analysis.

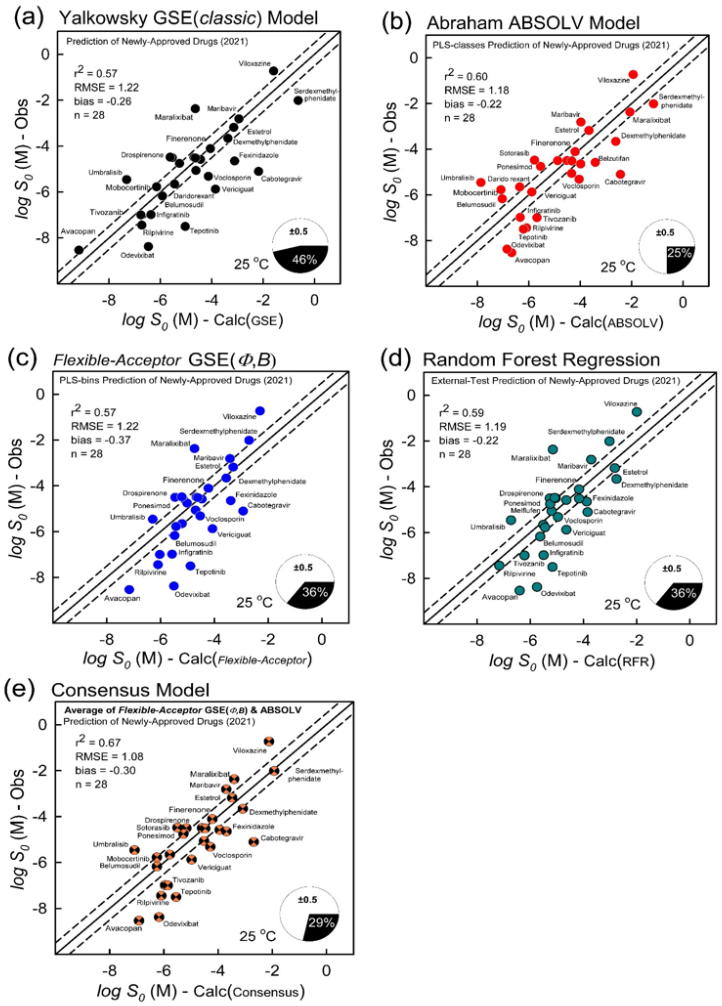

Figure 5

Test set predictions of the four models considered: measured log S_0 of newly-approved drugs vs. calculated log S_0. **(a)** GSE(classic) model, according to Eq. 1 (untrained). **(b)** ABSOLV model, Eq. 6, with coefficients determined by PLS regression (see text). **(c)** Flexible-Acceptor GSE(Φ,B) model, according to Eqs. 5a-c, with (Φ+B)-dependent c-coefficient functions determine by PLS regression (see text), and **(d)** Random Forest Regression (RFR) external test set of newly approved drugs. **(e)** Consensus model: average of GSE(Φ,B) and ABSOLV model predictions applied to newly-approved drugs.