Journal of Solution Chemistry (2022) 51:1020–1055
https://doi.org/10.1007/s10953-022-01141-7

Predicting Solubility of Newly-Approved Drugs (2016–2020) with a Simple ABSOLV and GSE(*Flexible-Acceptor*) Consensus Model Outperforming Random Forest Regression

Alex Avdeef[1] · **Manfred Kansy**[2]

Received: 8 September 2021 / Accepted: 10 November 2021 / Published online: 7 February 2022

Abstract
This study applies the 'Flexible-Acceptor' variant of the General Solubility Equation, GSE(Φ,B), to the prediction of the aqueous intrinsic solubility, $\log_{10} S_0$, of FDA recently-approved (2016–2020) 'small-molecule' new molecular entities (NMEs). The novel equation had been shown to predict the solubility of drugs beyond Lipinski's 'Rule of 5' chemical space (bRo5) to a precision nearly matching that of the Random Forest Regression (RFR) machine learning method. Since then, it was found that the GSE(Φ,B) appears to work well not only for bRo5 NMEs, but also for Ro5 drugs. To put context to GSE(Φ,B), Yalkowsky's GSE(classic), Abraham's ABSOLV, and Breiman's RFR models were also applied to predict $\log_{10} S_0$ of 72 newly-approve NMEs, for which useable reported solubility values could be accessed (nearly 60% from FDA New Drug Application published reports). Except for GSE (classic), the prediction models were retrained with an enlarged version of the Wiki-pS_0 database (nearly 400 added $\log_{10} S_0$ entries since our recent previous study). Thus, these four models were further validated by the additional independent solubility measurements which the newly-approved drugs introduced. The prediction methods ranked RFR ~ GSE (Φ,B) > ABSOLV > GSE (classic) in performance. It was further demonstrated that the biases generated in the four separate models could be nearly eliminated in a consensus model based on the average of just two of the methods: GSE (Φ,B) and ABSOLV. The resulting consensus prediction equation is simple in form and can be easily incorporated into spreadsheet calculations. Even more significant, it slightly *outperformed* the RFR method.

Keywords Flexible-acceptor general solubility equation · Abraham solvation equation · Kier molecular flexibility index · Intrinsic solubility · Partial least squares · Random forest regression

✉ Alex Avdeef
alex@in-ADME.com

[1] in-ADME Research, 1732 First Avenue #102, New York, NY 10128, USA

[2] Freiburg im Breisgau, Germany

Abbreviations

S_0 Intrinsic aqueous solubility (i.e., the solubility of the *uncharged* form of the compound)

MPP Measure of prediction performance [128]. It refers to the percent of 'correct' predictions, as defined by the count of absolute residuals $|\log_{10} S^{obs}{}_0 - \log_{10} S^{calc}{}_0| \leq 0.5$ divided by n. MPP is represented as a pie chart in the correlation plots.

RMSE root-mean-square error, accounting for bias in the prediction of external test set solubility values: RMSE = $[\ (1/(n-1))\ \Sigma_i\ (y^{obs}{}_i - \text{bias} - y^{calc}{}_i)]^{1/2}$, where $y = \log_{10} S_0$, n = number of measurements of $\log_{10} S_0$.

r^2 coefficient of determination, accounting for bias in prediction of external test set solubility values [130]: $r^2 = 1 - \Sigma_i\ (y^{obs}{}_i - \text{bias} - y^{calc}{}_i)^2\ /\Sigma_i\ (y^{obs}{}_i - <y>)^2$, where $y = \log_{10} S_0$, and $<y>$ is the mean value of observed $\log_{10} S_0$.

bias intercept (a) in the regression fit: $y^{obs} = a + b\ y^{calc}$, where the slope factor (b) is fixed at unity.

SD standard deviation: SD = $[\ (1/n)\ \Sigma_i\ (y^{obs}{}_i - <y>)]^{1/2}$, where n = number of measurements, $<y>$ = mean value of $\log_{10} S_0$.

1 Introduction

In the 5-year period 2016–2020, 228 drugs were approved by the FDA, mostly for the treatment of cancer, infectious/viral diseases, and neurological disorders [1–7]. Of these drugs, 74% are 'small molecule' new molecular entities (NMEs). Many of the NMEs are larger, more lipophilic, and possess more H–bond acceptors, compared to older drugs in the Lipinski 'Rule of 5' (Ro5) chemical space [8, 9]. NMEs outside the Lipinski space are often dubbed 'beyond the Rule of 5' (bRo5) drugs [9–18]. Size inflation is not the only physicochemical characteristic of the NMEs. Some new drugs are relatively small.

Generally, large molecules may increase pharmacokinetic (PK) risks due to low solubility, possibly low cell permeability, increased efflux, and elevated metabolism. During drug discovery/early development, strategies to mitigate some of the risks have included: (i) selecting molecules which can dynamically form intramolecular H-bonds (IMHB) to shield polar groups, (ii) shielding polar groups by bulky side chains or by *N*-methylation, and (iii) selecting molecules with flexible rings structures [14–19]. Flexible molecules with the potential to form IMHBs have been of particular interest, since these may possess enhanced solubility in water, by adopting hydrophilic 'extended' conformations, as well as facilitated permeability across cell membranes, by adopting hydrophobic 'folded' conformations [17–19].

Solubility plays a central role in the fuller understanding of the PK risks. Reliable and actionable in silico models to predict solubility of NMEs and of promising molecules not yet synthesized, could be a valuable contribution to risk assessment [13]. We started to address this topic in a series of in silico studies [20–22]; the present contribution is a continuation of that effort.

In a recent study to predict the intrinsic solubility, $\log_{10} S_0$, of four standardized external test sets of mostly druglike molecules [20], three methods were critically examined: (i) Yalkowsky General Solubility Equation (GSE) [23], (ii) Abraham Solvation Equation (ABSOLV) [24], and (iii) Breiman Random Forest regression (RFR) machine learning method [25]. RFR was found to be most accurate: for a highly-curated external test set of 100 druglike molecules with consistently-determined solubility values

(average interlaboratory reproducibility, $SD_{avg} \sim 0.17$ $\log_{10}$ unit), the strength of the prediction was indicated by the coefficient of determination, $r^2 = 0.64$, and root-mean-square error, RMSE = 0.76 ($\log_{10}$ unit) [20]. However, the 'black-box' machine learning RFR method has some disadvantages: (i) it does not directly suggest how compounds could be altered to increase/decrease their solubility [26]; (ii) there is no obvious simple explicit equation to predict solubility which could be used in a spreadsheet calculation; (iii) the method 'learns' superlatively but 'teaches' tepidly. The linear ABSOLV model, based on Abraham's five solvation descriptors [24, 27], yielded poorer statistics: $r^2 = 0.26$ and RMSE = 1.10 for the same test set. The GSE, Eq. 1, was even slightly less successful compared to ABSOLV ($r^2 = 0.20$, RMSE = 1.13) [20]. Nevertheless, the simple classic GSE is particularly appealing since it requires no 'training.' Merely the melting point (*mp* in °C) and the calculated (or measured) octanol–water partition coefficient, log *P*, are required to predict solubility (in log molar units):

$$\log_{10} S_0^{\mathrm{GSE(classic)}} = 0.5 - 1.0 \log_{10} P - 0.01(mp - 25) \tag{1}$$

In a follow-up study [21], solubility prediction using the above three models was applied to large molecules (MW > 800 g·mol^{-1}). The novel aim was to explore to what extent Ro5 molecules could be used to predict the $\log_{10} S_0$ of molecules from the bRo5 space. For an external test set of 31 large molecules, RFR predicted solubility ($r^2 = 0.37$, RMSE = 1.07) better than the other two methods. The RFR results suggested that it was possible to develop a model trained on small Ro5 molecules to predict the solubility of large bRo5 molecules. Unfortunately, the 'how' was not explicitly obvious. Nevertheless, the RFR method could serve as a benchmark against which other more actionable models could be measured. Also, the study revealed that the traditional GSE systematically underpredicts solubility of poorly soluble ($S_0 < 50$ μmol·L^{-1}) large molecules and greatly overpredicts solubility of highly soluble large molecules. The regression analysis of the three coefficients in Eq. 1 (0.5, − 1.0, − 0.01), using data partitioned into small and large molecule sets, resulted in notable differences between the two sets of coefficients, particularly in the first two terms (solvation contributions): (i) the 0.5 intercept in Eq. 1 was found to be − 0.28 for small molecules and − 1.77 for large molecules, and (ii) the $\log_{10} P$ slope factor, − 1.0, changed from − 0.83 to − 0.40 in small to large molecules, respectively [21]. The ABSOLV equation (trained with small molecules) revealed a different pattern of large-molecule residuals from that of the GSE: the solvation equation underpredicted the solubility of *every* large molecule tested. This was especially evident for very flexible molecules (*e.g.,* gramicidin A, bryamycin, and vancomycin). The principal components analysis of the solubility database used to train the models revealed an asymmetric distribution in the data, resembling the shape of a 'comet', with small molecules symmetrically occupying the 'head' and large molecules (MW > 800 g·mol^{-1}) exclusively occupying the 'tail.' Two hallmarks of bRo5 chemical space reside in the tail [21]: large size and large number of H-bond acceptors (NHA).

The above study [21] and earlier investigations by Caron and coworkers [15–18, 28] suggested that the influence of flexibility of large molecules on their solubility and permeability characteristics could be substantial. The latter researchers recommended the use of the Kier Φ molecular flexibility index [29] in modeling the properties of bRo5 molecules.

In our most recent solubility prediction study of bRo5 drugs, we discovered a way to incorporate the Kier molecular flexibility index, Φ, plus the Abraham *B* descriptor (H-bond acceptor strength) into Yalkowsky's classic GSE to improve its performance substantially [22]. The three coefficients in Eq. 1 were empirically determined as smooth functions

of the sum descriptor, $\Phi + B$. The modified equation was named the 'Flexible-Acceptor' model, GSE(Φ,B). It was trained with small (Ro5) molecules to predict the solubility of large (bRo5) molecules (not used in the training). With just three coefficients in Eq. 1, each defined as a three-parameter exponential function of $\Phi + B$, the strength of prediction *nearly matched that of the RFR machine learning method.* The coefficient of $\log_{10} P$ (traditionally fixed at -1.0) changed smoothly from − 1.1 for rigid nonionizable molecules ($\Phi + B = 0$) to − 0.39 for typically flexible ($\Phi \sim 20$, $B \sim 6$) large molecules. The intercept (usually fixed at + 0.5) varied smoothly from + 1.9 for rigid small molecules to − 2.2 for flexible large molecules. The *mp* coefficient remained practically constant, slightly different from the traditional value (− 0.01) for most molecules. For a test set of 32 large molecules the GSE(Φ,B) predicted the intrinsic solubility with RMSE of 1.10 log unit, compared to 3.0 by GSE(classic), and 1.07 by RFR.

Since our last study, it was found that the GSE (Φ,B) appears to work well not only for large drugs, but across a wide range of sizes of molecules. This piqued our interest to direct the new solubility prediction equation to recently-approved drugs (2016–2020), which comprise both the bRo5 and Ro5 molecules. For comparison, the GSE(classic), ABSOLV, GSE(Φ,B), and RFR models were each applied to predict the intrinsic solubility of 72 new drugs, for which useable reported solubility values could be accessed, nearly 60% from FDA published New Drug Application (NDA) reports [1–7]. The method performances ranked: RFR ~ GSE (Φ,B) > ABSOLV > GSE(classic). The performance of the GSE (Φ,B) was almost as good as that of the RFR. However, when the GSE (Φ,B) and ABSOLV methods were averaged, the resulting consensus model slightly *outperformed* the RFR method.

2 Computational Methods and Data Sources

2.1 Thermodynamic Basis of the General Solubility Equation (GSE)

Yalkowsky and coworkers developed the General Solubility Equation (GSE), Eq. 1, to predict the solubility of liquid/solid nonelectrolytes (mostly industrial organic chemicals) in water [23, 30–35]. The thermodynamic basis of the equation posits that the dissolution of a crystalline substance in water comprises two main contributions: (a) crystal lattice effect (XTL), related to the energy needed to break down the lattice to form a hypothetical 'supercooled liquid' (SCL), and (b) solvation effect, related to the energy released as the SCL dissolves in water. The total solubility of the compound in water is the product of the above two contributions, which in logarithmic terms can be stated as the sum [33, 34]

$$\log_{10} S = \log_{10} S_W^{\mathrm{STL}} + \log_{10} S_W^{\mathrm{SCL}} \tag{2}$$

2.1.1 Crystal Lattice Effect

The lattice contribution, $\log_{10} S_{\mathrm{W}}^{\mathrm{XTL}} = -\Delta S_m \left(T_m - T\right) / 2.303RT$, arises from the application of the van't Hoff equation, where ΔS_m (kJ·mol^{-1}·K^{-1}) is the standard molar entropy of phase transformation and T_m is the melting point (K). For many small organic compounds, $\Delta S_m \approx 0.057$ kJ·mol^{-1}·K^{-1} [33, 34]. Since at 25 °C, 2.303 RT = 5.7 kJ·mol^{-1}·K^{-1}, Eq. 2 reduces to Eq. 3, where *mp* is the melting point in °C.

$$\log_{10} S = \log_{10} S_W^{\mathrm{SCL}} - 0.01(mp - 25) \tag{3}$$

2.1.2 Solvation Effect

Hansch and coworkers [36] demonstrated that $\log_{10} S$ of 156 simple liquid solutes correlated linearly with the octanol–water partition coefficients, $\log_{10} P \approx \log_{10}\left(S_{\mathrm{oct}}^{\mathrm{liq}} / S_{\mathrm{W}}^{\mathrm{liq}}\right)$. This led to the approximation:

$$\log_{10} S_W^{\mathrm{liq}} = \mathrm{a}_0 + \mathrm{a}_1 \log_{10} P \tag{4}$$

where $\mathrm{a}_0 = \log_{10} S_{\mathrm{oct}}^{\mathrm{liq}}$ (solubility of a liquid solute in octanol) and $\mathrm{a}_1 \approx -1$. For small alcohol, aromatic, and alkane solutes, the series-dependent a_0 intercepts were determined as: + 0.93, + 0.34, − 0.25, respectively. The a_1 slope factors varied less: − 1.1 (alcohols), − 1.0 (aromatics), and − 1.2 (alkanes).

Yalkowsky and coworkers surmised that $\mathrm{a}_0 = \log_{10} S_{\mathrm{oct}}^{\mathrm{SCL}} = 0.5$ [23]. The entropy of mixing favors complete miscibility of the two liquids (liquid solute and octanol); *i.e.*, the mole fraction = 0.5. Since the concentration of pure octanol is 6.32 mol·L^{-1}, then $\log_{10} S_{\mathrm{oct}}^{\mathrm{liq}} = \log_{10}(6.32 \times 0.5) = 0.5$. With this approximation (and with $\mathrm{a}_1 = -1$), Eq. 4 substituted into Eq. 3 reduces to Eq. 1.

These fundamental considerations suggest that the traditional Eq. 1 could be adapted to compounds from the bRo5 chemical space, since Hansch's research hinted that the three coefficients in Eq. 1 could be optimized to different classes of compounds. If the 'supercooled liquid' form of a large polar solute is not fully miscible with octanol, then the $\log_{10} S_{\mathrm{oct}}^{\mathrm{SCL}}$ contribution could very well be a negative number. Hence, a large molecule with a decreased $S_{\mathrm{oct}}^{\mathrm{SCL}}$ (due to decreased miscibility) is expected to have an increased $S_{\mathrm{W}}^{\mathrm{SCL}}$. This, in effect, would lessen the contribution of lipophilicity to the predicted solubility.

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

In our earlier investigation [22] it was found that molecular flexibility (*Φ*) [29] could be incorporated into a *nonlinear* variant of the GSE to produce a promising trainable model with improved accuracy in predicting the solubility of large molecules (MW > 800 g·mol^{-1}). Further incremental improvements were achieved with an augmented second descriptor: Abraham's H-bond basicity (*B*), a measure of H-bond acceptor potential [24, 27]. The derived GSE(*Φ*,*B*) has the general form, with the three c-coefficients treated as three-parameter exponential functions of (*Φ* + *B*):

$$\log_{10} S_0^{\mathrm{GSE}(\Phi,B)} = c_0 + c_1 \cdot \log_{10} P + c_2 \cdot (mp - 25)/100 \tag{5}$$

$$\mathrm{c}_0 = \mathrm{b}_0 + \mathrm{b}_1 \exp(-\mathrm{b}_2 \cdot (\Phi + B)) \tag{6}$$

$$\mathrm{c}_1 = \mathrm{b}_3 + \mathrm{b}_4 \left[1 - \exp(-\mathrm{b}_\mathrm{s} \cdot (\Phi + B))\right] \tag{7}$$

$$\mathrm{c}_2 = \mathrm{b}_6 + \mathrm{b}_7 \left[1 - \exp(-\mathrm{b}_8 \cdot (\Phi + B))\right] \tag{8}$$

The c-coefficients at aggregated values 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 groupings of 209–1384 points. The details of the PLS procedure have been already described [22]. Since our database of solubility values has increased in size since our last study and since the focus now is on new drugs rather than specifically on big drugs, a new set of b-constants was determined in the current investigation, using druglike molecules as the training set, but excluding new drugs from the training.

Kier [29] constructed (considering structural attributes such as counts of chains, rings, branches, and heavy atoms) the molecular flexibility index, Φ, as the product of first and second order 'kappa' shape indices, $^1\mathbf{k}$ and $^2\mathbf{k}$, divided by the heavy atom count in the molecule. Here, 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 [37]. Table 1 lists these Φ values.

2.3 Abraham Descriptors and the ABSOLV Linear Model

To account for the thermodynamics of solute transfer from one phase to another, Abraham [24, 27] introduced five solvation descriptors: A, B, S_π, E, and V. Two of these constitute hydrogen bonding potential: A is the sum of H-bond acidity (donor strength) and B is the sum of H-bond basicity (acceptor strength) in the molecule. S_π is the dipolarity/polarizability (subscripted so as not to confuse it 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 $(cm^3{\cdot}mol^{-1})/100$. Since large molecules have greater number of H-bond acceptors, compared to small molecules [21], the Abraham B descriptor was selected to augment the Φ descriptor to further improve solubility prediction in the bRo5 chemical space [22]. Values of the Abraham descriptors were calculated from 2D structures using the ABSOLV algorithm [27] (*cf.*, www.acdlabs.com) and are listed in Table 1 for the new drugs.

Abraham and Le [24] amended the ABSOLV model to predict intrinsic solubility (log molar):

$$\log_{10} S_0^{\mathrm{ABSOLV}} = \mathrm{d}_0 + \mathrm{d}_1 A + \mathrm{d}_2 B + \mathrm{d}_3 S_\pi + \mathrm{d}_4 E + \mathrm{d}_5 V + \mathrm{d}_6 A \cdot B \tag{9}$$

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.

2.4 Statistical Machine Learning Random Forest Regression (RFR) Model

The implementation of the RFR open-source 'randomForest' library for the R statistical software has been described in our earlier solubility prediction studies [20–22]. The version used 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. The same procedure was applied in the current study. The method was re-trained with the enlarged database, excluding the newly-approved drugs.

Table 1 Physicochemical properties of newly approved drugs (2016–2020)

Approved drug	CAS	MW	mp (°C)	p$K^{a,b}$	S_0 (μg·mL^{-1})^c	t (°C)^d	clog$_{10}P^e$	NHA	NHD	Nrot	Φ^f	A	B	S_π	E	V	References
Abemaciclib	1,231,929–97-7	506.59	177	9.7	27	23	4.94	8	1	7	7.1	0.18	2.44	2.97	3.59	3.77	[38]
Acalabrutinib (Calquence)	1,420,477–60-6	465.51	215	**3.93, 5.60**	50	23	3.31	7	2	4	5.9	0.62	2.73	3.96	3.96	3.44	[39, 40]
Alpelisib (Piqray)	1,217,486–47-9	441.47	208	3.3, 9.4	19	37	3.84	5	2	4	5.8	0.80	1.80	2.86	2.27	2.99	[41]
Apalutamide (Erleada)	956,104–40-8	477.43	220	9.7	10	23	3.53	5	1	3	5.3	0.26	2.02	3.64	2.82	3.01	[42]
Amisulpride (Barhemsy)	71,675–85-9	369.48	182	9.37	723	25	1.29	6	2	7	6.3	0.50	2.18	3.16	1.70	2.81	[43]
Artesunate	88,495–63-0	384.43	143	**4.45**	199	25	2.60	7	1	4	4.7	0.57	1.73	1.82	1.24	2.73	[44]
Avapritinib (Ayvakit)	1,703,793–34-3	498.56	193	1.51, 3.50, 4.57, 6.97	0.63	23	2.61	10	1	5	5.6	0.21	2.74	2.96	4.06	3.62	[45]
Baloxavir Marboxil (Xofluza)	1,985,606–14-1	571.55	229	ni	20	37	3.39	10	0	4	6.8	0.00	3.28	4.03	3.47	3.70	[46, 47]
Benznidazole	22,994–85-0	260.25	189	ni	403	37	1.11	5	1	5	3.9	0.26	1.23	2.88	1.78	1.86	[48]
Brivaracetam (Briviact)	357,336–20-0	212.29	75	ni	800 mg·mL^{-1}	23	0.90	2	1	5	4.1	0.49	1.36	1.87	1.03	1.78	[49]
Capmatinib (Tabrecta)	1,029,712–80-8	412.42	197	*2.93*, 4.50	1.3	25	3.43	6	1	4	4.6	0.26	1.83	2.91	3.41	2.92	[50]
Cedazuridine (Inqovi)	1,141,397–80-9	268.21	164	*9.19*	47 mg·mL^{-1}	23	−1.57	5	4	2	3.4	1.06	1.96	1.98	1.34	1.64	[51]
Cenobamate (Xcopri)	913,088–80-9	267.67	98	ni	1.7 mg·mL^{-1}	25	1.16	6	1	4	3.7	0.44	1.22	2.31	1.87	1.78	[52]
Copanlisib (Aliqopa)	1,032,568–63-0	480.52	200	5.0, 8.5	2.5	25	0.66	11	2	7	6.7	0.61	3.21	3.55	3.34	3.45	[53]
Crisaborole (Eucrisa)	906,673–24-3	251.05	178	*8.95*	110	23	1.57	4	1	2	3.0	0.31	1.03	2.08	1.78	1.83	[54]
Dacomitinib (Vizimpro)	1,110,813–31-4	469.94	185	5.0, 8.5	0.022	23	5.16	6	2	7	7.4	0.68	1.99	3.10	3.17	3.38	[55]

Table 1 (continued)

Approved drug	CAS	MW	mp (°C)	p$K^{a,b}$	S_0 (μg·mL^{-1})^c	t (°C)^d	$c\log_{10}P^e$	NHA	NHD	Nrot	Φ^f	A	B	S_π	E	V	References
Darolutamide (Nubeqa)	1,297,538–32-9	398.85	180	**0.72**, 11.6	14	37	2.67	6	3	6	5.8	1.09	2.04	3.68	3.14	2.87	[56]
Decitabine (Inqovi)	2353–33-5	228.21	195	ni	10 mg·mL^{-1}	23	−2.14	8	3	2	2.9	0.71	2.32	1.83	1.90	1.52	[51]
Dolutegravir (GSK1349572A)	1,051,375–16-6	419.38	190	**7.26**	20	25	1.35	6	2	3	5.0	0.57	2.71	3.71	2.53	2.78	[57, 58]
Doravirine (Pifeltro)	1,338,225–97-0	425.75	191	*9.66*	120	23	2.65	7	1	4	5.2	0.16	1.93	3.14	2.17	2.60	[59]
Edaravone (Radicava)	89–25-8	174.20	130	ni	2.6 mg·mL^{-1}	25	1.80	2	0	1	1.8	0.00	0.97	1.21	1.36	1.34	[60–62]
Elbasvir (Zepatier)	1,370,468–36-2	882.02	242	**5.46, 8.77**	0.008	23	8.12	10	4	11	11.8	1.12	4.14	6.96	6.38	6.57	[63]
Enasidenib (Idhifa)	1,650,550–25-6	473.38	176	2.22, *11.37*	15	23	4.29	8	3	6	6.4	0.57	1.64	2.03	2.16	2.94	[64]
Entrectinib (Rozlytrek)	1,108,743–60-7	560.64	199	**1.72, 5.13**	12	37	5.03	6	3	7	7.8	1.01	2.29	3.49	3.85	4.10	[65]
Erdafitinib (Balversa)	1,346,242–81-6	446.54	141	1.9, 9.2	3	23	4.18	8	1	9	6.7	0.15	2.09	2.91	3.39	3.48	[66]
Fedratinib (Inrebic)	1,374,744–69-0	524.68	181	6.30, 9.50, *10.21*	0.013	37	4.82	8	3	10	8.3	0.62	2.33	3.27	3.21	4.03	[67]
Fosnetupitant (Akynzeo)	1,703,748–89-3	688.60	177	**4.83**, *6.14*	1.3	37	5.74	6	1	8	9.9	0.31	2.28	2.64	2.43	4.62	[68]
Gilteritinib (Xospata)	1,254,053–43-4	552.71	184	*3.96, 8.87*	0.37	20	2.70	10	3	9	9.2	0.69	3.09	3.35	3.45	4.33	[69]
Glasdegib (Daurismo)	1,095,173–27-5	374.44	214	1.7, 6.1	0.96	23	3.39	4	3	3	5.0	0.92	2.02	3.35	2.91	2.86	[70]
Grazoprevir (Zepatier)	1,350,514–68-9	766.90	174	*1.79*, 4.68	0.037	23	3.30	11	3	7	10.0	0.95	3.84	5.66	4.19	5.56	[63]
Istradefylline (Nourianz)	155,270–99-8	384.43	193	**0.35**	0.43	23	2.12	8	0	6	5.4	0.00	1.90	3.03	2.36	2.89	[71]
Lactitol	81,025–03-8	344.31	75	ni	1.57 mg·mL^{-1}	25	−5.76	11	9	8	8.2	2.33	3.37	2.80	2.23	2.34	[72]
Lifitegrast (Xiidra)	1,025,967–78-5	615.48	154	3.27	16	25	4.77	6	2	7	7.8	0.83	2.38	4.93	3.69	4.11	[73]

Table 1 (continued)

Approved drug	CAS	MW	mp (°C)	p$K^{a,b}$	S_0 (μg·mL^{-1})^c	t (°C)^d	$clog_{10}P^e$	NHA	NHD	Nrot	Φ^f	A	B	S_π	E	V	References
Lonafarnib (Zokinvy)	193,275–84-2	638.82	173	*3.28*	3	37	5.91	3	1	3	7.9	0.39	1.81	3.62	3.54	4.01	[74]
Macimorelin (Macrilen)	381,231–18-1	474.55	189	*8.34*	286	23	1.84	4	6	10	6.9	1.60	2.94	4.85	3.62	3.64	[75, 76]
Meropenem	96,036–03-2	383.46	196	*3.47, 9.39*	14 mg·mL^{-1}	25	−0.31	6	3	5	5.1	1.02	2.88	2.80	2.44	2.76	[77]
Moxidectin	113,507–06-5	639.82	150	ni	0.51	23	5.73	9	2	3	10.5	0.50	2.74	2.09	2.60	5.05	[78]
Naldemedine (Symproic)	916,072–89-4	570.64	195	*5.69, 9.07, 10.11*	2.6	23	3.48	9	4	6	5.3	1.07	2.83	3.78	4.32	4.07	[79]
Neratinib (Nerlynx)	698,387–09-6	557.04	186	4.66, 7.65	0.086	23	5.93	8	2	11	9.5	0.62	2.71	4.22	3.84	4.15	[80]
Nifurtimox (Lampit)	1,291,487–19-8	287.29	181	**0.79**	40	37	0.64	7	0	3	3.6	0.00	1.43	2.20	1.31	1.88	[81]
Niraparib (Zejula)	1,038,915–60-4	320.39	188	9.95	0.92	23	2.59	4	2	3	3.8	0.62	1.54	2.77	2.94	2.47	[82]
Obeticholic Acid (Ocaliva)	459,789–99-2	420.63	109	**4.76**	2.9	37	5.11	3	3	5	6.1	1.20	1.39	2.40	1.61	3.53	[83]
Ozanimod (Zeposia)	1,306,760–87-1	404.46	136	7.65	6.4	37	3.63	7	2	7	5.8	0.38	1.78	2.80	3.15	3.06	[84]
Pemigatinib (Pemazyre)	1,513,857–77-6	487.50	185	3.1, **5.15**	2.9	37	3.66	6	1	6	6.2	0.35	2.28	3.21	2.95	3.37	[85]
Pexidartinib (Turalio)	1,029,044–16-3	417.81	201	**1.50, 1.84, 5.75**	0.044	37	5.23	4	2	5	5.1	0.48	1.39	2.50	2.65	2.74	[86]
Pitolisant (Wakix)	362,665–56-3	295.85	192	9.35	120	23	4.17	2	0	8	7.4	0.00	0.83	1.18	1.13	2.44	[87]
Pralsetinib (Gavreto)	2,097,132–94-8	533.60	206	9.35	1.0	23	4.20	9	3	8	7.4	0.91	2.51	4.05	3.69	3.93	[88]
Pretomanid	187,235–37-6	359.26	150	ni	13	37	2.67	7	0	5	4.5	0.00	1.20	2.26	1.50	2.14	[89]
Relugolix (Orgovyx)	737,789–87-6	623.63	182	*7.96, 9.09*	26	25	3.75	11	2	9	8.8	0.51	3.01	4.19	3.97	4.24	[90]
Remdesivir (Veklury)	1,809,249–37-3	602.58	138	*3.21, 10.12*	18	37	2.31	13	4	13	9.8	0.89	3.57	4.09	2.94	4.32	[91]
Remimazolam (Byfavo)	308,242–62-8	439.31	170	*5.00, 6.44*	7.3	23	4.18	6	0	4	5.1	0.00	1.44	2.40	2.79	2.89	[92]

Table 1 (continued)

Approved drug	CAS	MW	mp (°C)	p$K^{a,b}$	S_0 (μg·mL^{-1})^c	t (°C)^d	$clog_{10}P^e$	NHA	NHD	Nrot	Φ^f	A	B	S_π	E	V	References
Ribociclib (Kisqali)	1,211,441–98-3	434.54	198	5.44, 8.27	21	37	2.80	8	2	5	5.7	0.29	2.45	3.47	3.39	3.32	[93, 94]
Ripretinib (Qinlock)	1,442,472–39-0	510.36	185	4.35	6.0	37	5.67	5	3	5	6.4	0.78	1.90	3.63	3.36	3.35	[95]
Risdiplam (Evrysdi)	1,825,352–65-5	401.46	209	*5.63, 8.83*	0.026	20	1.96	8	1	2	3.6	0.16	2.59	2.50	3.21	2.93	[96]
Safinamide (Xadago)	133,865–89-1	302.34	142	7.17	71	37	2.37	3	2	7	5.5	0.62	1.44	2.25	1.71	2.32	[97]
Secnidazole (Solosec)	3366–95-8	185.18	76	**0.78**	35 mg·mL^{-1}	27	0.48	5	1	3	2.6	0.31	0.89	1.74	1.13	1.33	[98, 99]
Selinexor (Xpovio)	1,393,477–72-9	443.31	177	**1.96**	3.0	37	3.39	7	2	5	5.9	0.39	1.63	2.83	2.26	2.61	[100]
Selpercatinib	2,152,628–33-4	525.60	199	*0.95, 4.58,* **5.75**	15	25	3.28	10	1	8	6.3	0.23	3.02	3.60	3.79	3.90	[101]
Selumetinib (Koselugo)	606,143–52-6	457.68	219	*5.31, 10.27*	3.4	23	3.53	6	3	6	6.0	0.62	1.86	3.22	3.15	2.72	[102, 103]
Sofosbuvir (Epclusa)	1,190,307–88-0	529.45	99	*9.19, 10.07*	2.0 mg·mL^{-1}	37	1.66	10	3	10	8.5	0.68	3.10	4.03	1.99	3.63	[104]
Stiripentol (Diacomit)	49,763–96-4	234.29	75	ni	5.0 mg·mL^{-1}	23	2.84	3	1	2	3.1	0.31	0.93	1.23	1.24	1.87	[105]
Talazoparib (Talzenna)	1,207,456–01-6	380.35	137	**0.80, 11.34**	16	37	2.63	6	2	2	3.5	0.45	1.62	2.88	2.94	2.51	[106]
Tazemetostat (Tazverik)	1,403,254–99-8	572.74	177	5.15, 6.73	12	37	4.73	6	2	9	9.8	0.51	2.77	3.78	3.43	4.56	[107]
Tecovirimat (TPOXX)	869,572–92-9	376.33	195	ni	26	23	2.40	3	1	2	3.2	0.26	1.65	2.40	2.15	2.37	[108]
Tenapanor (Ibsrela)	1,234,423–95-0	1145.05	150	*6.84, 9.86*	0.032	23	6.16	12	6	29	23.8	1.52	5.07	6.93	5.52	8.10	[109]
Tezacaftor (Trikafta)	1,152,311–62-0	520.50	81	ni	82	23	3.40	7	4	8	5.9	1.26	2.18	3.05	2.91	3.49	[110]
Tucatinib (Irinitinib, Tukysa)	937,263–43-9	480.52	230	3.07, 4.19, 6.15	6.8	23	5.09	10	2	5	5.1	0.26	2.24	3.25	4.20	3.48	[111]
Upadacitinib (Rinvoq)	1,310,726–60-3	380.37	186	**4.70**	185	37	2.91	4	2	3	4.2	0.47	1.52	2.23	2.03	2.52	[112]
Velpatasvir (Epclusa)	1,377,049–84-7	883.00	167	3.2, 4.6, *11.44*	1.3	25	7.73	10	4	11	11.9	1.12	4.18	6.91	6.37	6.53	[101, 104]

Table 1 (continued)

Approved drug	CAS	MW	mp (°C)	p$K^{a,b}$	S_0 (μg·mL^{-1})^c	t (°C)^d	$c\log_{10}P^e$	NHA	NHD	Nrot	Φ^f	A	B	S_π	E	V	References
Venetoclax (Venclexta)	1,257,044–40-8	868.44	139	*3.51*, *4.29*, 7.95	0.0038	27	8.66	11	3	13	12.4	0.91	3.39	5.82	5.59	6.29	[113]
Vibegron (Gemtesa, MK-4618)	1,190,389–15-1	444.53	198	1.6, 8.7	135	23	2.77	6	3	6	6.1	0.79	2.60	3.25	3.19	3.37	[114]
Zanubrutinib (Brukinsa)	1,691,249–45-2	471.55	189	**2.64**	5.7	23	4.22	6	2	6	6.3	0.57	2.27	3.90	3.75	3.57	[115]

Drug names in parentheses are brand names, and/or alternative names

A, *B*, S_π, *E*, *V* Abraham solvation descriptors; NHA, NHD, Nrot number of H-bond acceptors, donors, and rotatable groups, resp Underlined brand names are dual API compounds: each API is treated separately

[b]Underlined values refer to acid groups (otherwise the pK_a is of a basic group); bold values determined in this study from published solubility data; values in *italics* calculated by ChemAxon MarvinSketch v5.3.7 program (ChemAxon Ltd., https://www.chemaxon.com)

[c]Intrinsic solubiltiy in μg·mL^{-1} units, unless otherwise indicated

[d]Room temperature or unreported temperature assumed to be 23 °C

[e]Calculated octanol–water partition coefficient (RDKit)

[f]Kier molecular flexibility index, Φ

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

The annual mini-reviews of FDA drug approvals by Mullard [1–5] were convenient starting points to identify the new drugs and to begin the search for their solubility values. The data for the newly-approved drugs were wearisome to locate. Since the drugs are relatively new, there are not many journal publications reporting properties of the compounds. Most of the data were found in FDA 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 properties of compounds under consideration (review documents: Product Quality, Quality Assessment, Multi-Discipline, Clinical Pharmacology and Biopharmaceutics, and Other). Unfortunately, sometimes the information about solubility is redacted in these reports. Other useful sources include Product Monographs, Highlights of Prescription Information, and Safety Data Sheets. Some solubility data were found in patents. The European Medicines Agency (EMA) publishes Assessment Reports. The Australian regulatory agency publishes Australian Public Assessment Reports (AUSPAR), as well as Australian Product Information documents. These potential sources of measured solubility data were searched with the 'solub' key.

Generally, there was virtually no experimental detail about the measurements in the published regulatory reports. Most of the reported solubility values are of drugs in water (S_W), without mention of the saturation pH. The temperature was assumed to be 23 °C when not stated or when reported as 'room temperature' (Table 1). In the dearth of experimental detail, it is a challenge to assess the quality of the reported measurements in most of the FDA/EMA/AUSPAR reports. Still, 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 169 small-molecule NMEs approved in the 5-year period, 98 *quantitative* solubility measurements were found for only 72 NMEs [38–115]. The reported values were transformed into the intrinsic solubility scale, S_0, using known (or predicted) pK_a values, and adjusted to 25 °C [116] using the program *p*DISOL-X (*in-ADME* Research) [117–122]. Table 1 lists the normalized solubility data, along with the pK_a values used in the data analysis.

The *Wiki-p*S_0 (*in-ADME* Research) intrinsic aqueous solubility database of mostly drug-like molecules (currently with 7190 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 [20–22, 116–126], 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 72 new drugs considered here (along with the year of approval) are shown in the Appendix (Fig. 9). In dual-API drug products, each API was treated as a separate 'drug' in the data analysis.

2.6 Sources of Octanol–Water Partition Coefficients ($\log_{10} P$) and Melting Points (*mp*)

Originally in Eq. 1, *mp* and $\log_{10} P$, were taken to be experimental values. However, it has become a common practice to use calculated values, *clogP*, in place of measured $\log_{10}$

P. In this study, *clogP* values were in all cases calculated by the Wildman–Crippen sum of atomic contributions method in the open-source RDKit chemoinformatics library [37]. Experimental *mp* values were employed where available and were calculated otherwise [127]. Values of *mp* are difficult to predict accurately. Prediction studies suggest root-mean-square error of about 35 °C. From this, the *mp* contribution to calculated log *S* could be uncertain by ~ 0.4 $\log_{10}$ unit. Some uncertainty lingers even with tabulated experimental values, as it is sometimes unclear whether a particular *mp* value refers to a salt form or a free-acid/base form of the compound.

3 Results and Discussion

3.1 Data Reduction

For half of the new drugs, $\log_{10} S_0$ values in Table 1 were determined from reported S_W values, using *p*DISOL-X. The program also calculated the pH of the saturated solution, as though the Henderson–Hasselbalch (HH) equation were valid. When aggregates/complexes form or when supersaturation persists in the suspension, the HH equation does not accurately predict the shape of the $\log_{10} S$–pH curve [118–122]. There is no simple way to recognize such anomalies just from a single S_W measurement.

The remaining $\log_{10} S$ data were sourced at two or more values of pH, which generally allowed for more confident determinations of $\log_{10} S_0$. These 'raw' $\log_{10} S$–pH measurements required further data reduction and normalization. For ionizable molecules, the pK_a values are required for such analysis. 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 $\log_{10} S$–pH profiles (bold values in Table 1).

Examples of quality experimental $\log_{10} S$–pH profiles reported for some of the new drugs are shown in Fig. 1. Frames a–c are of bases (acalabrutinib, pexidartinib, upadacitinib); frame d is that of an acid (dolutegravir); frame e is that of an ampholyte (talazoparib). The data from these five drugs appeared to follow shapes predicted by the Henderson–Hasselbalch equation: it was possible not only to determine the best-fit $\log_{10} S_0$, but also the values of pK_a (in five cases) and the pK_{sp} (in two cases). When profiles deviate from expected 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 [118–122]. Figure 1f (safinamide) shows such an example of anomaly, where at pH 4.5, the solubility is higher than that expected for a solution saturated in the free base. Since the solubility values at pH 1.2 and 4.5 are nearly the same, the suspension at pH 4.5 may have been supersaturated with respect to the charged form of the base during the measurement. Had only a single measurement been reported at pH 4.5, the intrinsic solubility might have been determined at an order of magnitude too high.

3.2 Comparing Properties of the Newly-Approved Drugs to Those in the Database Training Set

Figure 2 shows the distribution of intrinsic solubility values for the database training set and the NMEs test set. The new drugs on the average are nearly an order of magnitude less

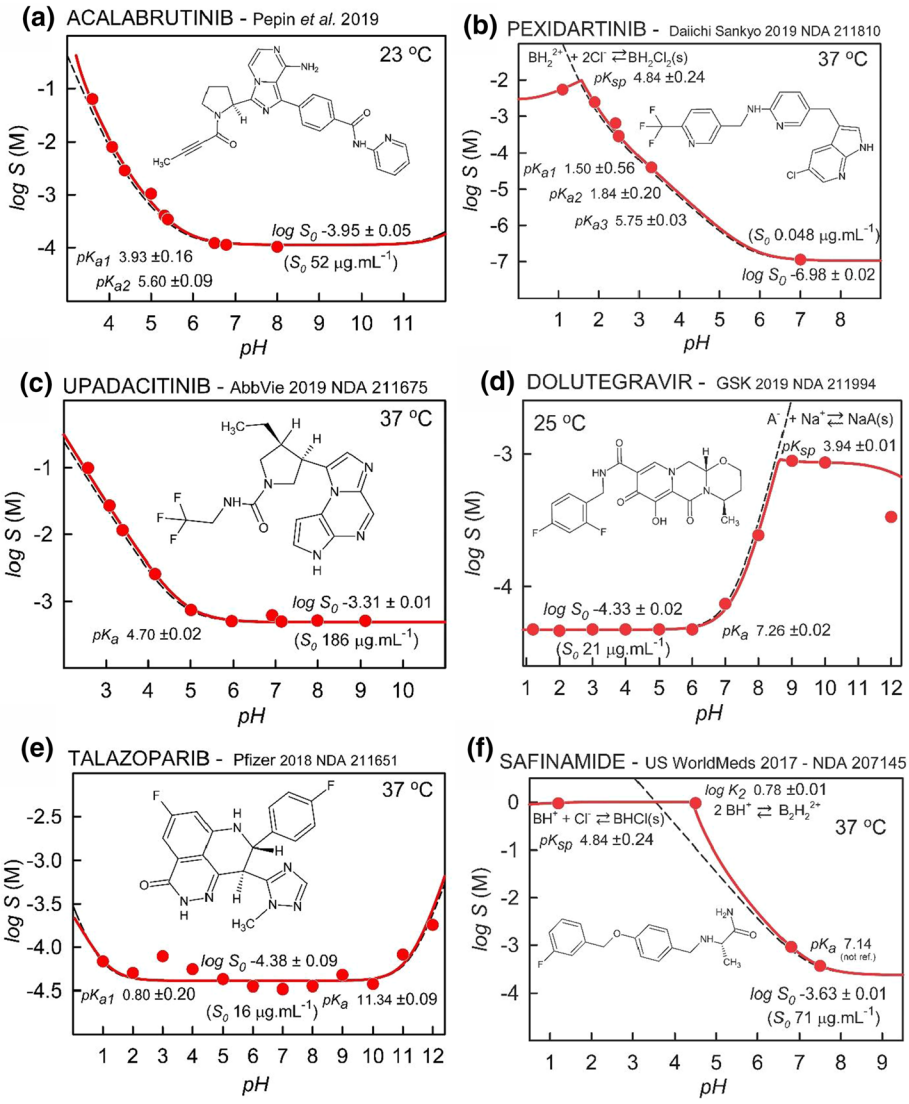

Fig. 1 New-drug examples of $\log_{10} S$–pH profiles of good disposition. The solid red curves are the best fit to the measured data (circles), using the regression analysis program pDISOL-X. It was also possible to determine the pK_a values in cases (**a–e**). The dashed curves were calculated using the Henderson–Hasselbalch equation, incorporating the pK_a used and the refined $\log_{10} S_0$. In cases (**b**), (**d**), and (**f**), it was possible to determine the salt solubility products (Color figure online)

soluble. Figure 3 compares the properties used to evaluate whether a compound falls into Lipinski's 'Rule of 5' chemical space. The lipophilicities (as indicated by *clogP*) of the new drugs on the average are nearly an order of magnitude higher than those of the older drugs (Fig. 3a). The mean molecular weight of the older drugs is just under 300 g·mol^{-1}; it is 450 g·mol^{-1} for the new drugs (Fig. 3b). Whereas the distribution of H-bond donors is nearly the same in the two sets (Fig. 3c), the distribution of H-bond acceptors is quite

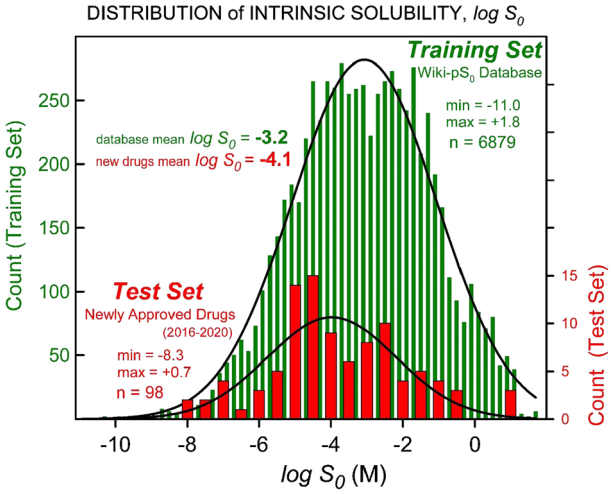

Fig. 2 Distribution of intrinsic solubility values, $\log_{10} S_0$, in the training database set (green upper trace, scaled to the left vertical axis) and the test set (red lower trace, scaled to the right vertical axis) of newly approved drugs. The bell-shaped curves illustrate that the newly-approved drugs are about an order of magnitude less soluble (mean $S_0 = 79$ μmol·L^{-1}) than the training-set molecules (mean $S_0 = 631$ μmol·L^{-1}) (Color figure online)

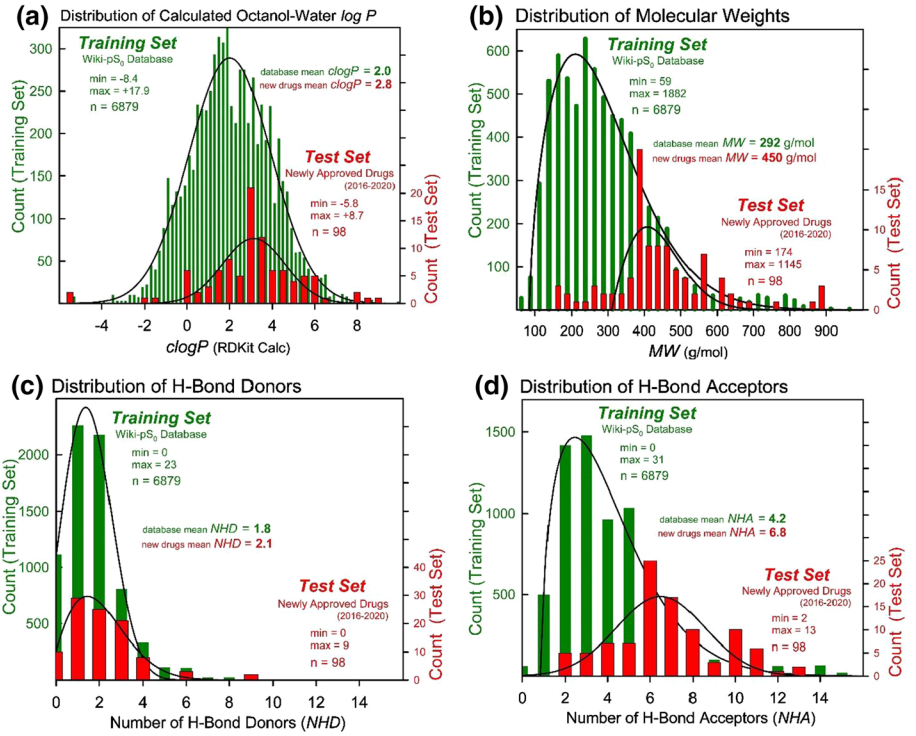

Fig. 3 Lipinski's Ro5 property distributions: **a** *clogP* (RDKit-calculated $\log_{10} P$, Wildman-Crippen type) **b** molecular weight (*MW*), **c** number of H-bond donors (*NHD*), and **d** number of H-bond acceptors (*NHA*). The training set databases distributions are in the upper traces (with counts scaled to right vertical axis) and the test set new drug distributions are in the lower traces (with counts scaled to right vertical axis). On the average, compared to the training-set molecules, the newly-approved drugs are about 1.4 times more lipophilic, have greater molecular weights by about 1.5 times, have similar distributions of H-bond donors (2/molecule), but have higher numbers of H-bond acceptors (7/molecule, compared to 4/molecule in the training set) (Color figure online)

different (Fig. 3d). On the average, the number of H-bond acceptors, NHA, is about 4 per molecule for older drugs and nearly 7 per molecule in the newly-approved drugs.

The new drugs violate the boundary conditions of Lipinski's Ro5 more often than those in the training set. There are relatively more molecules with $clogP > 5$ in the new drugs set (15% of NMEs), compared to that of the training set (6%). For the new drugs, 28% of the substances have MW > 500 g·mol^{-1}, compared to 7% in the training set. The relative number of NHD > 5 in the new drugs set (5%) is higher than in the training set molecules (2%). The relative number of NHA > 10 for the new drugs (9%) is greater than in the training set (3%).

The distributions of Φ values for the training and test sets are shown in Fig. 4. On the average, the new drugs are more flexible (mean $\Phi = 6.2$) than the molecules in the training set (mean $\Phi = 4.3$). The training set spans a wide range of Φ values, from 0.4 to 43. The newly-approved drugs subtend that space, with Φ values ranging from 1.9 to 24.

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

The training set solubility data were sorted by $\Phi + B$ into ten bins of increasing values. For a narrow range of $\Phi + B$ values in each bin, the three GSE coefficients in Eq. 5 were determined by linear PLS regression, in the way that Hansch et al. [36] had trained the GSE for different chemical classes of compounds. Table 2 lists the set of determined c-constants for each of the bins. The resultant c-constants are depicted by the points on the three curves in Fig. 5, displayed as a function of the average values of $\Phi + B$ from each bin. It is possible to recognize trends for the substantially decreasing c_0, the steadily increasing c_1, and the very slightly increasing c_2 coefficients with increasing values of $\Phi + B$. Apparently, *crystal lattice* contributions are not appreciably affected by molecular flexibility and H-bond acceptor character, and trend near the traditional value (−0.01) in Eq. 1. Evidently, solubility dependence on flexibility and H-bond acceptor strength are mediated by solution-phase interactions [26]. The bin analysis results are summarized in Table 2.

From the thermodynamics considerations, the c_0 coefficient may be viewed as a measure of the solubility of the 'supercooled' liquid solute in octanol ($c_0 \approx \log_{10} S_{oct}^{sliq}$). Increasingly flexible molecules with strong H-bond acceptor character appear to be less

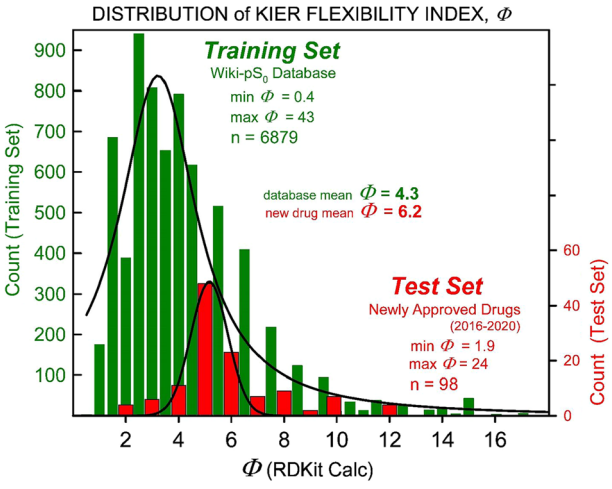

Fig. 4 Distribution of the Kier flexibility index, Φ, in the training database set (upper trace, left axis scale) and the test set (lower trace, right axis scale) of newly approved drugs. The index was calculated using RDKit 'kappa' descriptors (see text). The newly-approved drugs are more flexible than those in the training set (6.2 vs. 4.3) and span a narrower range of values

Table 2 PLS analysis in bins ordered by $\Phi+B$: $\log_{10} S_0^{GSE} = c_0 + c_1\, clogP + c_2\, (mp-25)/100$

Bin	$\Phi+B$[a]	Range[b]	c_0	c_1	c_2	r^2	RMSE	n
1	1.7	0.8–1.9	1.14	−1.24	−0.92	0.70	0.94	209
2	2.2	1.9–2.4	0.51	−1.04	−0.82	0.72	0.91	426
3	3.0	2.4–3.6	0.36	−0.95	−0.88	0.62	1.08	1196
4	4.0	3.6–4.4	−0.11	−0.90	−0.77	0.70	0.96	1187
5	5.0	4.4–5.7	−0.76	−0.77	−0.70	0.64	1.04	1384
6	6.0	5.7–6.4	−0.88	−0.78	−0.61	0.55	1.01	703
7	7.0	6.4–7.8	−1.05	−0.77	−0.59	0.66	1.14	760
8	8.5	7.8–9.5	−1.67	−0.62	−0.63	0.63	1.13	556
9	10.9	9.5–13.6	−1.81	−0.54	−0.94	0.50	1.45	400
10	19.9	13.6–55.9	−3.06	−0.33	−0.14	0.39	1.21	244

[a]Average $\Phi+B$ in bin (Φ Kier molecular flexibility index, B Abraham H-bond acceptor)

[b]Range of $\Phi+B$ values in bin

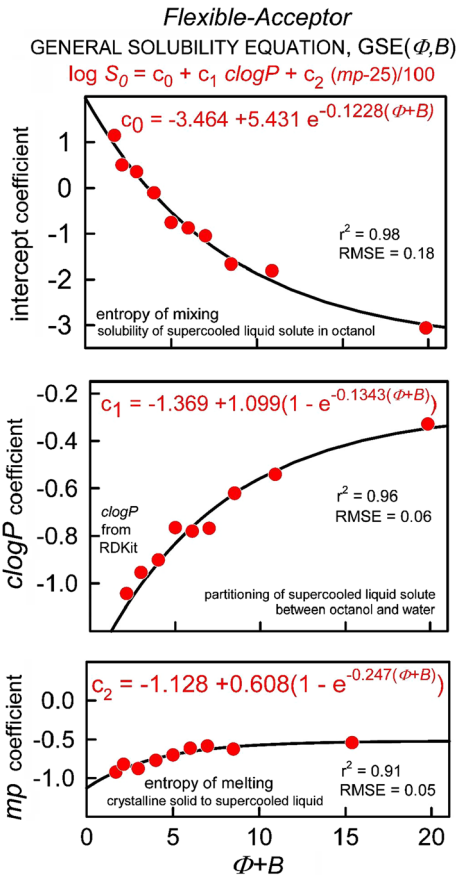

Fig. 5 Training the Flexible-Acceptor GSE (Φ,B) model. The solubility data in the training set were sorted on $\Phi+B$ and then divided into ten practically constant ($\Phi+B$) bins (Table 2). On the average, each bin contained about 700 $\log_{10} S_0$ measurements. For each bin (represented by a point in the plots), the three constants in Eq. 1 were determined by PLS regression to best fit the intra-bin solubility data. The aggregate intercept coefficient, $c_0(\Phi,B)$ in Eq. 6, is described by an exponential decay function spanning from + 1.1 (rigid, octanol-miscible molecules) to − 3.4 (flexible, octanol-immiscible molecules). In the next two frames, the coefficient functions, $c_1(\Phi,B)$ in Eq. 7 and $c_2(\Phi,B)$ in Eq. 8, are depicted by exponential rises to the limiting values − 0.27 (= − 1.369 + 1.099)) and − 0.52 (= − 1.128 + 0.608), respectively

miscible with octanol, as suggested by the decreasing c_0 coefficients with increasing $\Phi + B$ (*cf*., Table 2 and Fig. 5). Between bins 1 and 10, S_{oct}^{sliq} decreases by four orders of magnitude. 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 [23, 33, 34] where c_1 is constant.

The points in Fig. 5 were fitted to exponential forms as functions of $\Phi + B$ to determine the b-parameters (Eqs. 6–8), using standard nonlinear least-squares methods. The resultant best-fit curves in Fig. 5 define the aggregated form of the GSE(Φ,B), in the final form with the nine b-parameters determined, as shown below.

$$c_0 = 3.464 + 5.431 \exp(-0.1228 \cdot (\Phi + B) \tag{10}$$

$$c_1 = 1.369 + 1.099\left[1 - \exp(-0.1343 \cdot (\Phi + B))\right] \tag{11}$$

$$c_2 = 1.128 + 0.608\left[1 - \exp(-0.247 \cdot (\Phi + B))\right] \tag{12}$$

3.4 ABSOLV Training

The d-coefficients in Eq. 9 were determined by PLS regression using the $\log_{10} S_0$ values from the Wiki-pS_0 database, excluding those of the NMEs: $r^2 = 0.65$, RMSE = 1.16, n = 7092.

$$\log_{10} S_0^{\mathrm{ABSOLV}} = -0.640 + 0.128A + 1.751B + 0.083S_\pi - 1.526E - 1.223V + 0.065A \cdot B \tag{13}$$

3.5 Solubility Prediction Results for the Newly-Approved Drugs

3.5.1 Model Training

Figure 6 shows the results of the training of the four models, as measured $\log_{10} S_0$ vs. calculated $\log_{10} S_0$ correlation plots. The solid diagonals are identity lines. The dashed diagonals are ± 0.5 $\log_{10}$ 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 $\log_{10}$ unit of the observed values [128]. In the first three frames, the symbols represent the predominant charge states of molecules at pH 7.4: black diamonds represent uncharged molecules, blue squares represent bases (positive charged), red circles represent acids (negative charged), and yellow diamonds represent zwitterions. The zwitterions are less well predicted in the GSE model, compared to the ABSOLV model [20]. 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 database, it is expected that their $\log_{10} S_0$ could be predicted with $r^2 = 0.90$, RMSE = 0.62, with 76% of the molecules 'correctly' predicted (Fig. 6d). Generally, the other three methods are less precise, with MPP values ranging around half of the RFR value.

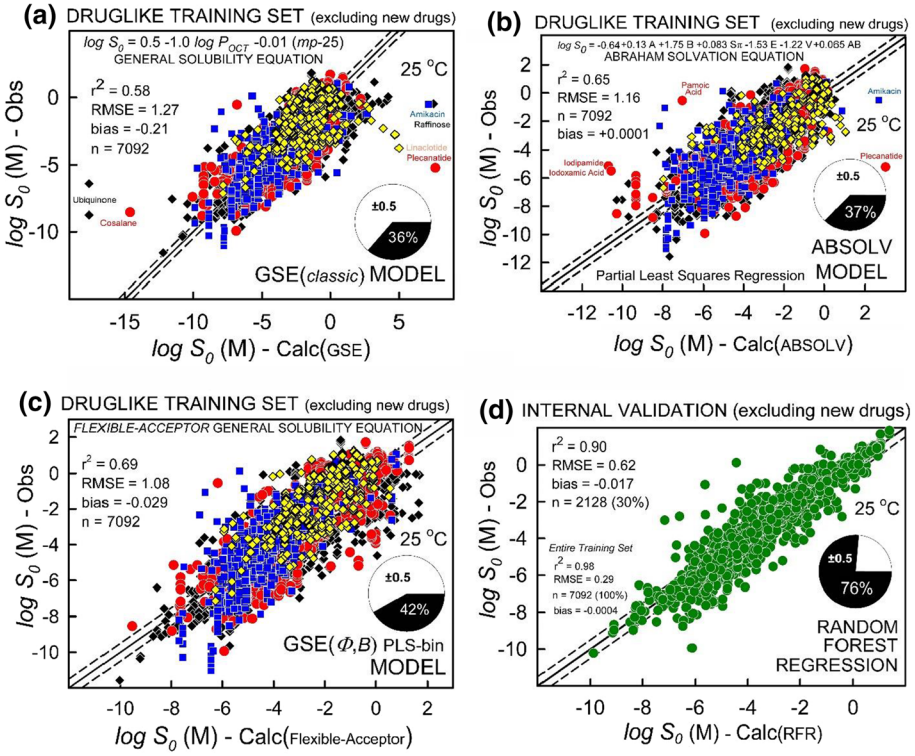

Fig. 6 Training set predictions of the four models considered: measured $\log_{10} S_0$ vs. calculated $\log_{10} S_0$. The solid diagonals are identity lines. The dashed diagonals are ± 0.5 $\log_{10}$ unit displaced from the identity lines. The pie-charts indicate the percentage of 'correctly' predicted values (see text). **a** GSE(classic) model, according to Eq. 1 (untrained). **b** ABSOLV model, Eq. 13, with coefficients determined by PLS regression. **c** Flexible-Acceptor GSE(Φ,B) model, according to Eqs. 10–12 (see text), and **d** Random Forest Regression (RFR) internal validation applied to randomly-selected 30% of the database, trained using the other 70% of the database

3.5.2 Model Testing

Figure 7 shows the results of the predictions of the solubility of the newly-approved drugs (external test sets) by the four models. Table 3 summarizes the results. Briefly, the four results look similar, as MPP values range from 28 to 39%. Note that the horizontal scale in Fig. 7a is quite different from those of the other frames. The GSE(classic) underperformed compared to the other methods. The Flexible-Acceptor model produced prediction metrics nearly equal to those of RFR. None of the methods produced RMSE < 1, which may be indicative of the uncertain quality of half of the new drugs solubility data reported as single-point values in water. On the other hand, RFR uncharacteristically overpredicted the solubility of drugs with $\log_{10} S_0 < 7$, which may hint that those molecules possessed structural features not found in the Wiki-pS_0 training database. The GSE(Φ,B) also shows similar overpredictions.

The bias in the predicted results is near zero (−0.06) in the RFR method. Both GSE methods show negative bias (−0.33 and −0.25), whereas the ABSOLV method produces a positive bias (+0.29). A consensus model was suggested by averaging the ABSOLV

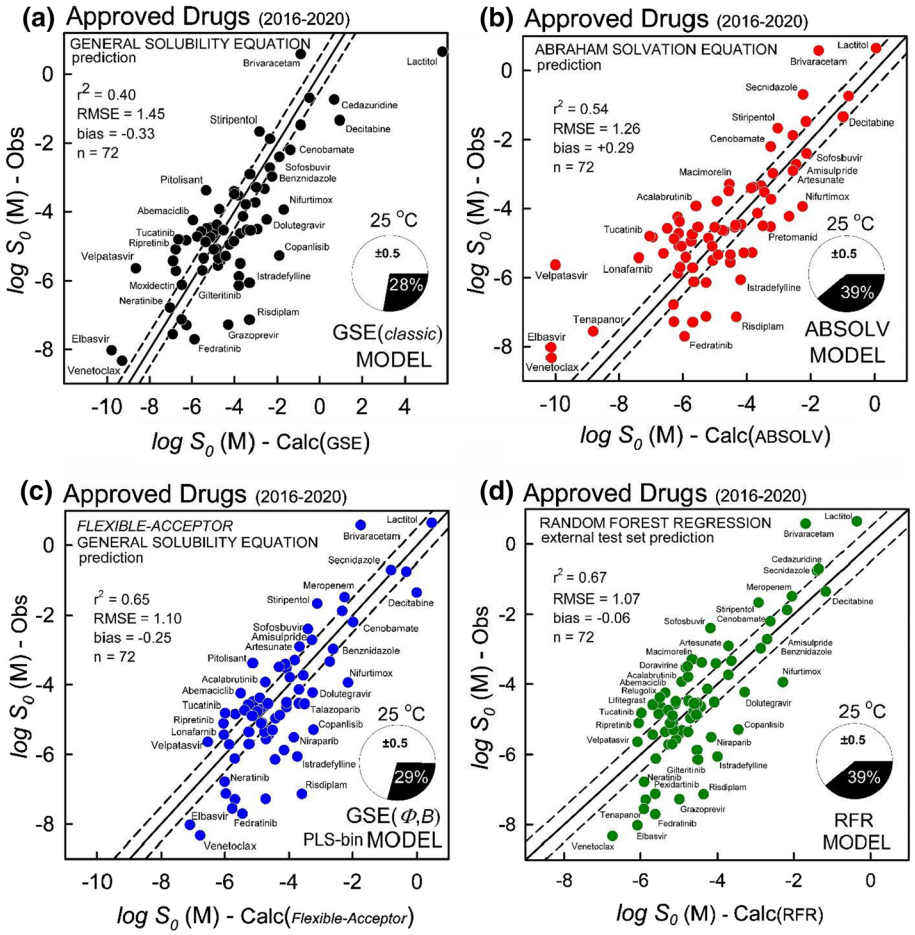

Fig. 7 Test set predictions of the four models considered: measured log S_0 of newly-approved drugs vs. calculated log S_0. See Fig. 6 caption for definitions of common features. **a** GSE(classic) model, according to Eq. 1 (untrained). **b** ABSOLV model, Eq. 13, with coefficients determined by PLS regression. **c** Flexible-Acceptor GSE(Φ,B) model, according to Eqs. 10–12, with ($\Phi+B$)-dependent c-coefficient functions determine by PLS regression (see text), and **d** Random Forest Regression (RFR) external test set of newly approved drugs

and GSE(Φ,B), to minimize the method bias. Figure 8 shows the results of the consensus model. Although the r^2 (0.67) and RMSE (1.07) values in the consensus method match those of the RFR method, the MPP value (40%) and the bias (+0.02) in the consensus model are slight improvements.

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

The *Wiki-p*S_0 database of druglike molecules has steadily grown over the last 10 years. Lately, it has been our observation that this alone has not proportionately improved its ability to predict the solubility of drugs. Metrics such as those in Fig. 6 have remained largely

Table 3 Predicted intrinsic solubility ($\log_{10} S_0$) of newly approved drugs (2016–2020)

Approved Drug	$\log_{10} S_0$[a] (mol·L^{-1})	SD[b]	n[c]	GSE (classic)[d]	ABSOLV [e]	GSE (Φ,B)[f]	RFR[g]	Consensus[h]	Residual[i]
Abemaciclib	−4.25		1	−5.96	−6.16	−5.50	−5.35	−5.83	1.6
Acalabrutinib	−3.93	0.01	2	−4.71	−5.59	−4.74	−4.92	−5.17	1.2
Alpelisib	−4.47	0.14	1	−5.17	−4.17	−5.01	−4.74	−4.59	**0.1**
Apalutamide	−4.65		1	−4.98	−4.72	−4.87	−5.46	−4.79	**0.1**
Amisulpride	−2.71	0.08	2	−2.36	−2.46	−3.28	−2.69	−2.87	**0.2**
Artesunate	−2.90	0.17	9	−3.28	−2.55	−3.67	−3.71	−3.11	**0.2**
Avapritinib	−5.87	0.48	1	−3.79	−6.15	−4.15	−4.53	−5.15	−0.7
Baloxavir Marboxil	−4.55	0.02	2	−4.93	−4.38	−4.93	−5.08	−4.66	**0.1**
Benznidazole	−2.97		1	−2.25	−3.18	−2.61	−2.86	−2.90	**−0.1**
Brivaracetam	0.58		1	−0.90	−1.75	−1.74	−1.69	−1.74	2.3
Capmatinib	−5.42	0.12	2	−4.65	−5.90	−4.62	−4.99	−5.26	**−0.2**
Cedazuridine	−0.75		1	0.69	−0.82	−0.33	−1.39	−0.58	**−0.2**
Cenobamate	−2.20		1	−1.39	−3.25	−1.98	−2.61	−2.62	**0.4**
Copanlisib	−5.29	0.10	1	−1.91	−3.83	−3.23	−3.45	−3.53	−1.8
Crisaborole	−3.33		1	−2.60	−3.56	−2.72	−3.63	−3.14	**−0.2**
Dacomitinib	−7.29	0.27	1	−6.26	−5.69	−5.68	−5.86	−5.69	−1.6
Darolutamide	−4.64		1	−3.72	−4.78	−4.08	−4.44	−4.43	**−0.2**
Decitabine	−1.34		1	0.95	−0.99	0.00	−1.16	−0.49	−0.8
Dolutegravir	−4.23	0.14	2	−2.50	−2.67	−3.26	−3.28	−2.97	−1.3
Doravirine	−3.53		1	−3.81	−3.45	−4.08	−4.82	−3.76	**0.2**
Edaravone	−1.87	0.25	3	−2.35	−2.56	−2.33	−2.17	−2.44	0.6
Elbasvir	−8.03		1	−9.79	−10.14	−7.10	−6.08	−8.62	0.6
Enasidenib	−4.49		1	−5.30	−4.36	−5.11	−5.08	−4.74	**0.2**
Entrectinib	−4.84	0.04	1	−6.27	−6.95	−5.67	−5.53	−6.31	1.5
Erdafitinib	−5.09	0.23	1	−4.84	−6.13	−4.84	−5.15	−5.49	**0.4**
Fedratinib	−7.71	0.27	1	−5.88	−5.94	−5.45	−5.62	−5.69	−2.0
Fosnetupitant	−5.71		1	−6.76	−5.70	−5.86	−5.26	−5.78	**0.1**
Gilteritinib	−6.14	0.42	1	−3.79	−5.28	−4.44	−4.51	−4.86	−1.3
Glasdegib	−5.56		1	−4.78	−4.52	−4.72	−5.07	−4.62	−0.9

Table 3 (continued)

Approved Drug	$\log_{10} S_0$[a] (mol·L^{-1})	SD[b]	n[c]	GSE (classic)[d]	ABSOLV [e]	GSE (Φ,B)[f]	RFR[g]	Consensus[h]	Residual[i]
Grazoprevir	−7.28	0.64	1	−4.29	−6.28	−4.73	−4.98	−5.51	−1.8
Istradefylline	−6.06	0.04	1	−3.30	−4.20	−3.74	−4.00	−3.97	−2.1
Lactitol	0.65	0.12	2	5.76	0.04	0.47	−0.36	0.25	**0.4**
Lifitegrast	−4.58		1	−5.56	−6.48	−5.27	−5.68	−5.88	1.3
Lonafarnib	−5.43	0.41	1	−6.89	−7.38	−6.03	−5.68	−6.70	1.3
Macimorelin	−3.29	0.15	2	−2.98	−4.55	−3.82	−4.65	−4.19	0.9
Meropenem	−1.48	0.06	6	−0.90	−2.14	−2.25	−2.04	−2.20	0.7
Moxidectin	−6.12		1	−6.48	−5.66	−5.69	−5.60	−5.67	**−0.4**
Naldemedine	−5.31		1	−4.68	−6.61	−4.71	−4.91	−5.66	**0.3**
Neratinib	−6.78	0.08	1	−7.04	−6.29	−6.00	−5.91	−6.15	−0.6
Nifurtimox	−3.94	0.33	1	−1.70	−2.25	−2.14	−2.28	−2.20	−1.7
Niraparib	−5.50		1	−3.72	−5.08	−3.85	−4.16	−4.46	−1.0
Obeticholic Acid	−5.35	0.36	1	−5.45	−4.52	−5.25	−5.35	−4.89	**−0.5**
Ozanimod	−4.96	0.40	1	−4.24	−5.75	−4.42	−4.69	−5.09	**0.1**
Pemigatinib	−5.35		1	−4.76	−4.91	−4.78	−4.75	−4.84	**−0.5**
Pexidartinib	−7.13	0.08	1	−6.49	−5.29	−5.97	−5.61	−5.63	−1.5
Pitolisant	−3.38	0.13	1	−5.34	−3.80	−5.13	−4.41	−4.46	1.1
Pralsetinib	−5.70	0.07	1	−5.51	−6.08	−5.25	−5.17	−5.66	**0.0**
Pretomanid	−4.53	0.30	1	−3.42	−3.26	−3.69	−4.54	−3.47	−1.1
Relugolix	−4.38		1	−4.82	−6.10	−4.92	−5.49	−5.51	1.1
Remdesivir	−4.51	0.09	2	−2.94	−3.50	−4.08	−4.09	−3.79	−0.7
Remimazolam	−4.76		1	−5.13	−5.71	−5.00	−4.61	−5.36	0.6
Ribociclib	−4.86	0.53	2	−4.03	−5.21	−4.29	−4.54	−4.75	**−0.1**
Ripretinib	−5.10		1	−6.77	−6.04	−6.04	−6.05	−6.04	0.9
Risdiplam	−7.14		1	−3.30	−4.33	−3.59	−4.36	−3.96	−3.2
Safinamide	−3.74	0.04	1	−3.04	−3.24	−3.55	−3.71	−3.39	**−0.3**
Secnidazole	−0.70	0.07	2	−0.49	−2.23	−0.79	−1.34	−1.51	0.8

Table 3 (continued)

Approved Drug	$\log_{10} S_0$[a] (mol·L^{-1})	SD[b]	n[c]	GSE (classic)[d]	ABSOLV[e]	GSE (Φ,B)[f]	RFR[g]	Consensus[h]	Residual[i]
Selinexor	−5.29		1	−4.41	−4.10	−4.52	−5.14	−4.31	−1.0
Selpercatinib	−4.54		1	−4.52	−5.53	−4.66	−4.60	−5.10	0.6
Selumetinib	−5.10		1	−4.97	−5.09	−4.88	−5.23	−4.99	**−0.1**
Sofosbuvir	−2.40	0.04	1	−1.90	−2.13	−3.40	−4.18	−2.77	**0.4**
Stiripentol	−1.66		1	−2.84	−3.03	−3.11	−2.92	−3.07	1.4
Talazoparib	−4.55	0.35	1	−3.25	−5.02	−3.51	−4.60	−4.26	**−0.3**
Tazemetostat	−4.73	0.32	1	−5.75	−6.13	−5.38	−5.28	−5.75	1.0
Tecovirimat	−4.14		1	−3.60	−3.67	−3.69	−4.26	−3.68	**−0.5**
Tenapanor	−7.56	0.11	1	−6.91	−8.82	−5.76	−5.91	−7.29	**−0.3**
Tezacaftor	−3.79	0.14	1	−3.46	−4.94	−3.97	−4.75	−4.45	0.7
Tucatinib	−4.81		1	−6.64	−7.04	−5.99	−5.98	−6.52	1.7
Upadacitinib	−3.41	0.03	1	−4.02	−3.87	−4.12	−4.04	−3.99	0.6
Velpatasvir	−5.64	0.28	2	−8.65	−10.00	−6.54	−6.08	−8.27	2.6
Venetoclax	−8.33		1	−9.30	−10.13	−6.77	−6.74	−8.45	**0.1**
Vibegron	−3.49		1	−4.00	−4.57	−4.32	−4.77	−4.44	0.9
Zanubrutinib	−4.89		1	−5.36	−6.27	−5.15	−5.16	−5.71	0.8

[a]Negative logarithm of the intrinsic solubility, based on analysis of reported literature solubility data (the 'observed' value)

[b]Estimated standard deviation in the determined $\log_{10} S_0$ value, with the average value of 0.20 $\log_{10}$

[c]Number of independently reported solubility data sources for the determinations of $\log_{10} S_0$

[d]Yalkowsky's General Solubility Equation: $\log_{10} S = 0.5 - \log P_{OCT} - 0.01\,(mp - 25)$; $\log_{10} P_{OCT}$ calculated in RDKit (Wildman − Crippen type)

[e]Abraham Solvation Equation, Eq. 13

[f]'Flexible − Acceptor' enhanced GSE calculation, with the three traditional constant coefficients expanded as functions of Kier flexibility index, Φ, and Abraham's H-bond acceptor parameter, B. Equations 10–12

[g]Random Forest Regression model, trained on druglike database of $\log_{10} S_0$ values, not including those of recently approved drugs

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

[i]Residual = observed $\log_{10} S_0$ minus consensus value. The drug names are in bold when the residual values are between −0.5 and +0.5

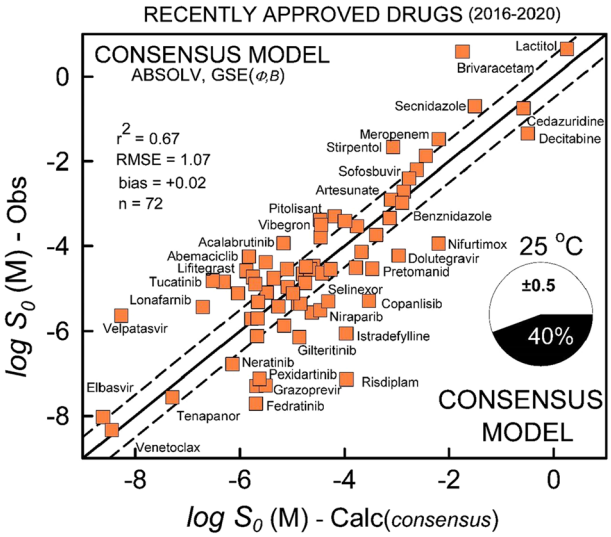

Fig. 8 Consensus model: average of GSE(*Φ,B*) and ABSOLV model predictions applied to newly-approved drugs

unchanged [20–22]. Solubility prediction depends on multi-dimensional factors (quality of measurements of both 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, etc.), with some factors yet to be recognized. Simply increasing the size of the solubility training set may not lead to improved predictions. Lipinski has suggested that compiling a large physicochemical property database aimed at maximizing chemical diversity may be an inefficient strategy for predicting the properties of novel molecules, given the enormous size of the chemical space, and since drugs appear to exist there as small tight clusters [129]. However, small improvement in solubility prediction can be expected as the training set acquires additional measurements of regulatory newly-approved molecules on a regular basis—*i.e.*, drawing from the "tight cluster" space. It would be helpful if the quality of such measurements were to improve with time. 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

Many of the new drugs are large and fall outside of the Lipinski Ro5 chemical space, as depicted in Fig. 3. It would have been helpful to have access to more quantitative solubility measurements of the newly-approved drugs than provided in the regulatory agency reports. The experimental uncertainty of nearly half of the new measurements could not be directly verified. If better practices in solubility measurement were adhered to, as detailed in the recent data-quality 'white paper' by experts from six countries [121], and the experimental details were more openly shared, newly-reported measurements could achieve results with interlaboratory SD < 0.2 $\log_{10}$ unit. But apparently this is work still in progress. The data quality in the curated database (SD < 0.2 log unit) used here as the training set is not the limiting factor in prediction, given that the best root-mean-square error achieved in this study was 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 (2016–2020). A similar near-match had been previously reported by us in the prediction of the solubility of large (bRo5) drugs, supporting the general applicability of the Flexible-Acceptor model. A 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. The relatively-simple consensus model can be readily incorporated into spreadsheet calculations.

Appendix

The structures of the 72 newly-approved drugs, along with the year of approval, are shown in Fig. 9

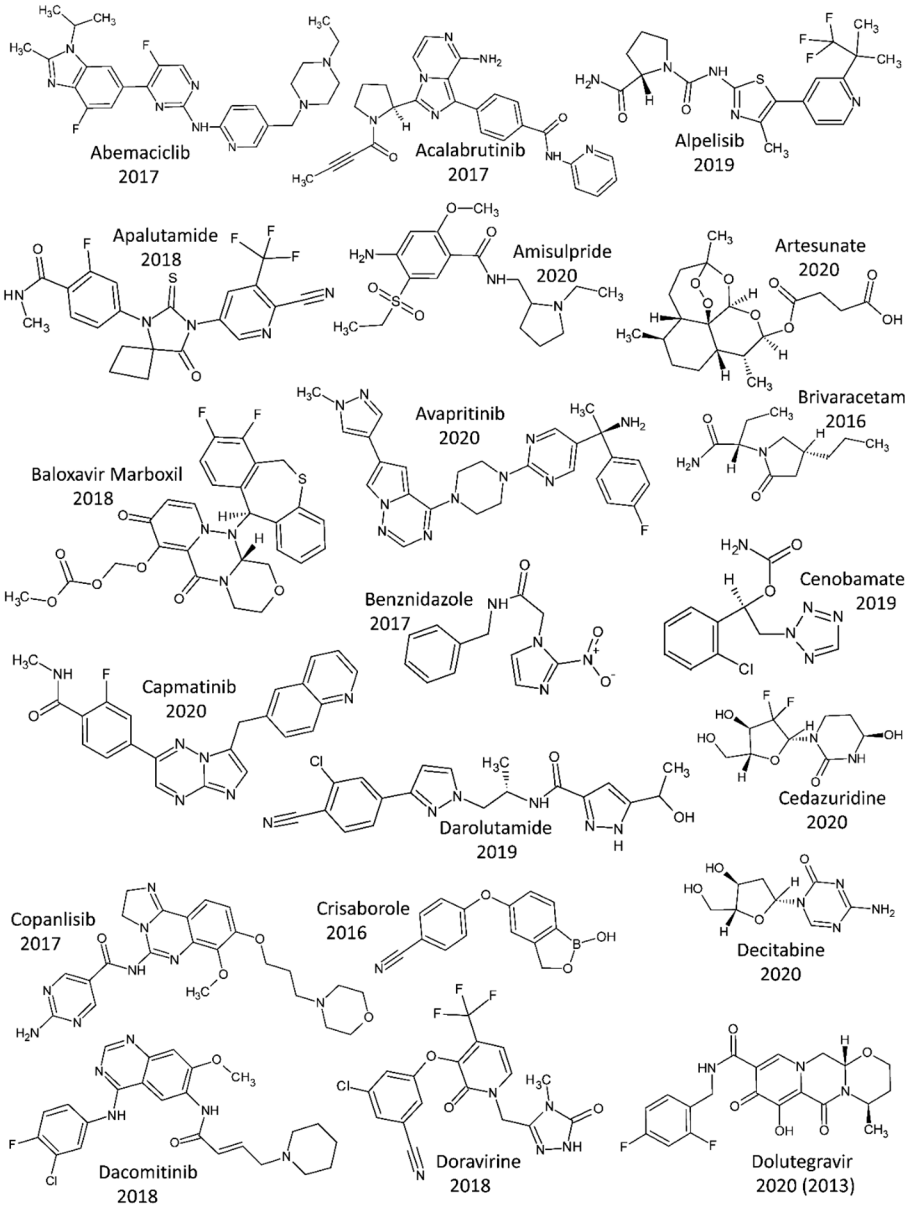
Abemaciclib
2017
Acalabrutinib
2017
Alpelisib
2019
Apalutamide
2018
Amisulpride
2020
Artesunate
2020
Avapritinib
2020
Brivaracetam
2016
Baloxavir Marboxil
2018
Benznidazole
2017
Cenobamate
2019
Capmatinib
2020
Darolutamide
2019
Cedazuridine
2020
Copanlisib
2017
Crisaborole
2016
Decitabine
2020
Dacomitinib
2018
Doravirine
2018
Dolutegravir
2020 (2013)

Fig. 9 .

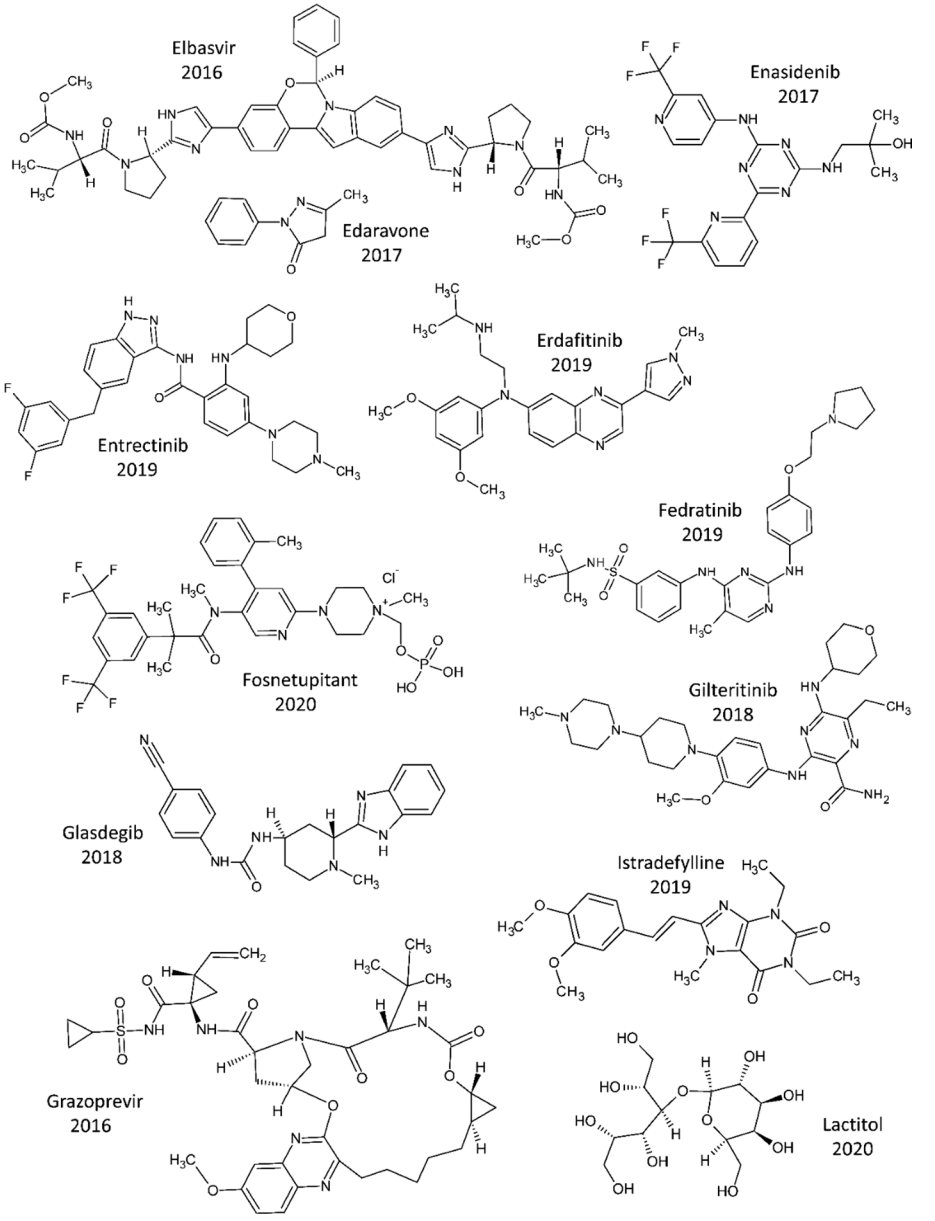

Fig. 9 (continued)

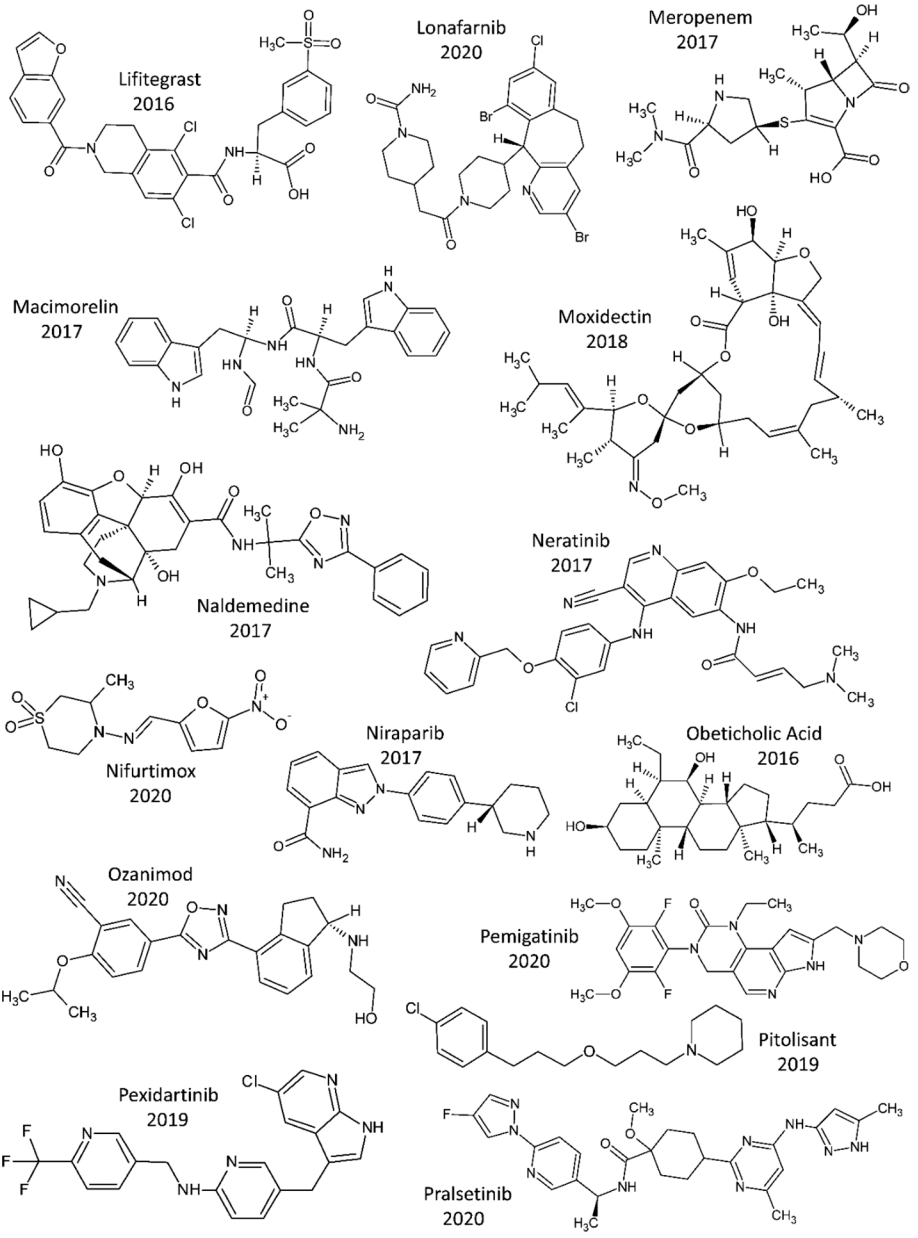

Fig. 9 (continued)

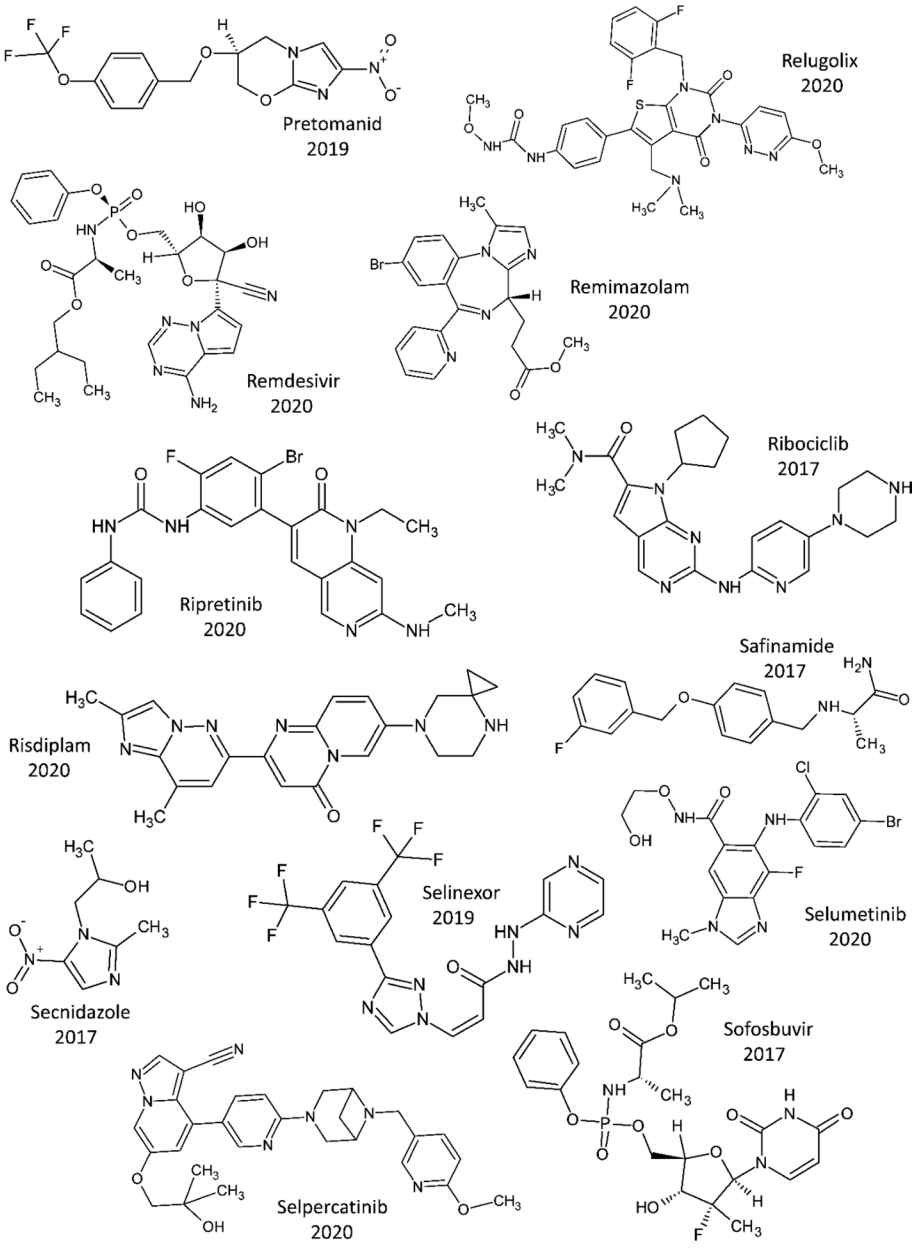

Fig. 9 (continued)

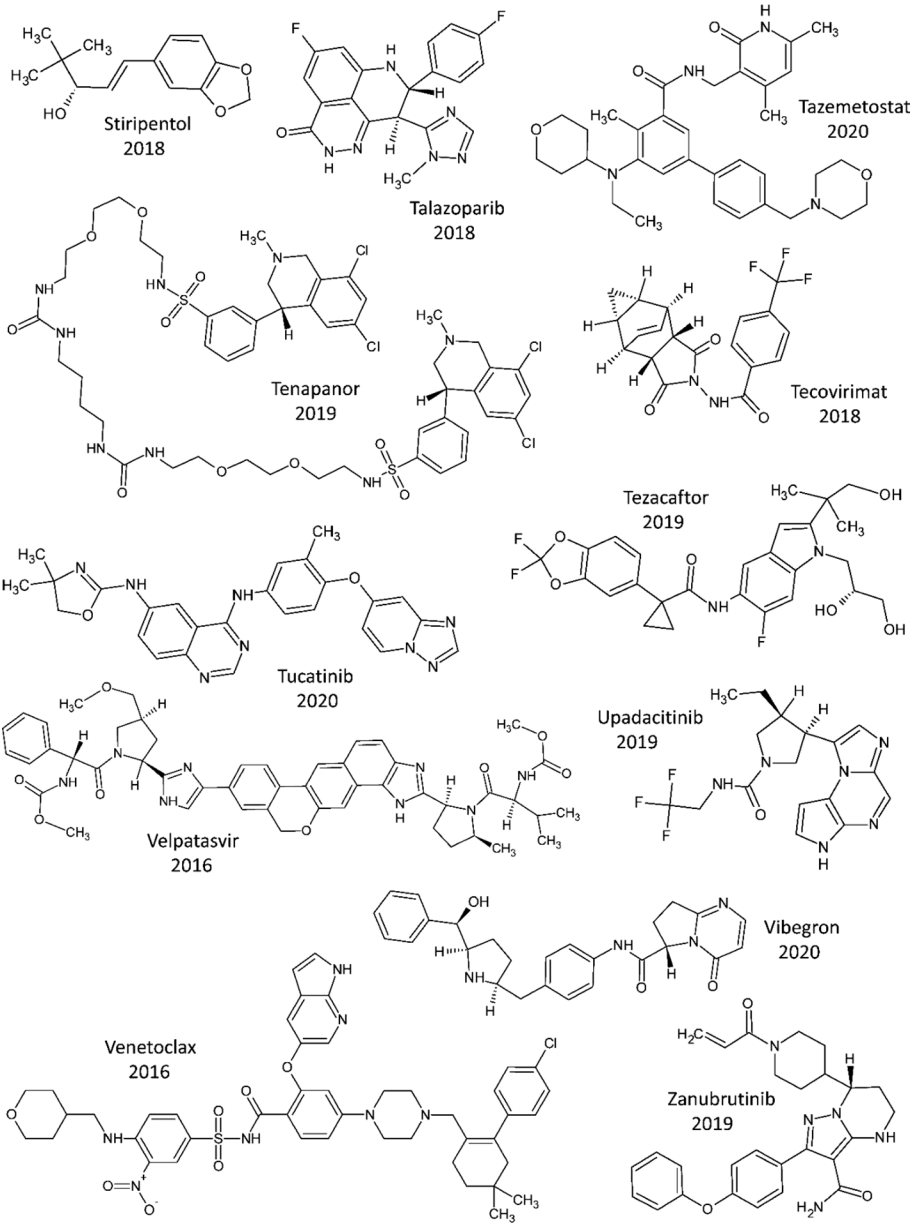

Fig. 9 (continued)

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. He is remembered as a teacher and a friend. 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).

Funding This study was self-funded. The authors declare that they have no known competing financial interests that could have appeared to influence the work reported in this paper.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. Kinch, M.S., Griesenauer, R.H.: 2017 in review: FDA approvals of new molecular entities. Drug Discov. Today. **23**, 1469–1473 (2018)
7. Roskoski, R., Jr.: Properties of FDA-approved small molecule protein kinase inhibitors. Pharmacol. Res. **144**, 19–50 (2019)
8. 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 Deliv. Rev. **23**, 3–25 (1997)
9. Leeson, P.D.: Molecular inflation, attrition & the rule of five. Adv. Drug Deliv. Rev. **101**, 22–33 (2016)
10. Doak, B.C., Over, B., Giordanetto, F., Kihlberg, J.: Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chem. Biol. **21**, 1115–1142 (2014)
11. Matsson, P., Doak, B.C., Over, B., Kihlberg, J.: Cell permeability beyond the rule of 5. Adv. Drug Deliv. Rev. **101**, 42–61 (2016)
12. DeGoey, D.A., Chen, H.-J., Cox, P.B., Wendt, M.D.: Beyond the rule of 5: lessons learned from AbbVie's drugs and compound collection. J. Med. Chem. **61**, 2636–2651 (2018)
13. 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)
14. 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)
15. Ermondi, G., Vallaro, M., Goetz, G., Shalaeva, M., Caron, G.: Experimental lipophilicity for beyond Rule of 5 compounds. Future Drug. Discov. (2019). https://doi.org/10.4155/fdd-2019-0002
16. 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
17. 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.org/10.1002/med.21562
18. 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
19. 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)
20. 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). https://doi.org/10.5599/admet.766
21. Avdeef, A., Kansy, M.: Can small drugs predict the intrinsic aqueous solubility of 'beyond rule of 5' big drugs? ADMET & DMPK (2020). https://doi.org/10.5599/admet.794
22. 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
23. Yalkowsky, S.H., Valvani, S.C.: Solubility and partitioning I: Solubility of nonelectrolytes in water. J. Pharm. Sci. **69**, 912–922 (1980)

24. 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)
25. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
26. 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)
27. 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)
28. Ermondi, G., Poongavanam, V., Vallaro, M., Kihlberg, J., Caron, G.: Solubility prediction in the bRo5 chemical space: where are we right now? ADMET & DMPK (2020). https://doi.org/10.5599/admet.834
29. Kier, L.B.: An index of molecular flexibility from kappa shape attributes. Quant. Struct.-Act. Relat. **8**, 221–224 (1989)
30. Yalkowsky, S.H., Banerjee, S.: Aqueous Solubility: Methods of Estimation for Organic Compounds, p. 142. Marcel Dekker Inc, New York (1992)
31. Alantari, D., Yalkowsky, S.: 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**, 739–740 (2018)
32. Ran, Y., Yalkowsky, S.H.: Prediction of drug solubility by the general solubility equation. J. Chem. Inf. Comput. Sci. **41**, 354–357 (2001)
33. Jain, N., Yalkowsky, S.H.: Estimation of the aqueous solubility I: application to organic nonelectrolytes. J. Pharm. Sci. **90**, 234–252 (2001)
34. 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)
35. Jain, N., Yang, G., Machatha, S.G., Yalkowsky, S.H.: Estimation of the aqueous solubility of weak electrolytes. Int. J. Pharm. **319**, 169–171 (2006)
36. 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)
37. 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/. Accessed 18 Jan 2022
38. Eli Lilly and Company. Prod. Monogr. Incl. Patient Med. Info. VERZENIO® (Abemaciclib mesylate). http://pi.lilly.com/ca/verzenio-ca-pm.pdf. Accessed 31 Jul 2020
39. Blatter, F.; Ingallinera, T.; Barf, T.; Aret, E.; Krejsa, C.; Evarts, J.: Crystal forms of (S)-4-(8-Amino-3-(1-(but-2-ynoyl)pyrrolidin-2-yl)imidazo[1,5-A]pyrazin-1-yl)-*N*-(pyridin-2-yl)benzamide. US 9,796,721 B2.
40. Pepin, X.J.H., Sanderson, N.J., Blanazs, A., Grover, S., Ingallinera, T.G., Mann, J.C.: Bridging in vitro dissolution and in vivo exposure for acalabrutinib. Part I. Mechanistic modelling of drug product dissolution to derive a P-PSD for PBPK model input. Eur. J. Pharm. Biopharm. **142**, 421–434 (2019)
41. Food and Drug Administration (USA): Alpelisib (Piqray), Novartis pharmaceuticals NDA 212526Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212526Orig1s000MultidisciplineR.pdf. Accessed 18 Jan 2022
42. European Medicines Agency: Apalutamide (Erleada®), CHMP assessment report, Procedure No. EMEA/H/C/004452/0000. https://www.ema.europa.eu/en/documents/assessment-report/erleada-epar-public-assessment-report_en.pdf. Accessed 15 Nov 2018
43. Zhang, W.-P., Chen, D.-Y.: Crystal structures and physicochemical properties of amisulpride polymorphs. J. Pharm. Biomed. Anal. **140**, 252–257 (2017). https://doi.org/10.1016/j.jpba.2017.03.030
44. Xu, R., Han, T., Shen, L., Zhao, J., Lu, X.: Solubility determination and modeling for artesunate in binary solvent mixtures of methanol, ethanol, isopropanol, and propylene glycol + water. J. Chem. Eng. Data **64**, 755–762 (2019)
45. Food and Drug Administration (USA): Avapritinib (Ayvakit), Blueprint meds. Corp. 2019 NDA 212608Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/212608Orig1s000ChemR.pdf. Accessed 18 Jan 2022
46. Genentech, Inc. Safety Data Sheet: Baloxavir marboxil (Xofluza). https://www.gene.com/download/pdf/XOFLUZATablets40mgSAPSDS2.pdf. Accessed 16 Oct 2018
47. Food and Drug Administration (USA): Baloxavir marboxil (Xofluza). NDA 210854Orig1s000, CDER Quality Assessment Review. Applicant: Shionogi. Accessed 18 Sep 2018.

48. Sobrinho, J.M.S., Soares, M.F.R., Labandeira, J.J.T., Alves, L.D.S., Neto, P.J.R.: Improving the solubility of the antichagasic drug benznidazole through formation of inclusion complexes with cyclodextrins. Quim. Nova **34**, 1534–1538 (2011)
49. Australian Public Assessment Report for brivaracetam. Proprietary Product Name: Briviact. Sponsor: UCB Australia Pty Ltd. 2017. https://www.tga.gov.au/sites/default/files/auspar-brivaracetam-170307.pdf. Accessed 18 Jan 2022
50. Food and Drug Administration (USA): Capmatinib (Tabrecta), Novartis Ringaskiddy Pharma Ltd. 2017 NDA 213591Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213591Orig1s000ChemR.pdf. Accessed 18 Jan 2022
51. Otsuka Pharmaceutical Co., Ltd. Decitabine + Cedazuridine (Inqovi). Product monograph. 3 Jul 2020; https://www.taihopharma.ca/documents/31/INQOVI_Product_Monograph.pdf. Accessed 18 Jan 2022
52. Food and Drug Administration (USA): Cenobamate (Xcopri), SK Life Science, Inc. NDA 212839Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212839Orig1s000OtherR.pdf. Accessed 18 Jan 2022
53. Freundlieb, J.; Jacobs, T.: Formulations of copanlisib. Bayer Pharma AG Patent Application Publication: US 2020/0281932 A1. 10 Sep 2020. https://uspto.report/patent/app/20200281932. Accessed 18 Jan 2022
54. Fantini, A., Demurtas, A., Nicoli, S., Padula, C., Pescina, S., Santi, P.: In vitro skin retention of crisaborole after topical application. Pharmaceutics **12**, 491 (2020). https://doi.org/10.3390/pharmaceutics12060491
55. Pfizer Canada ULC. Product Monograph: Dacomitinib (Vizimpro). Accessed 22 Feb 2019.
56. Food and Drug Administration (USA): Darolutamide (Nubeqa), Bayer HealthCare Pharmaceuticals Inc. NDA 212099Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212099Orig1s000ChemR.pdf. Accessed 18 Jan 2022
57. Food and Drug Administration (USA): Dolutegravir <GSK1349572A>, GSK. 211994Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211994Orig1s000OEList.pdf. Accessed 18 Jan 2022
58. Gigante, V., Pauletti, G.M., Kopp, S., Xu, M., Gonzalez-Alvarez, I., Merino, V., McIntosh, M.P., Wessels, A., Lee, B.-J., Rezende, K.R., Scriba, G.K.E., Jadaun, G.P.S., Bermejo, M., M.: Global testing of a consensus solubility assessment to enhance robustness of the WHO biopharmaceutical classification system. ADMET and DMPK (2020). https://doi.org/10.5599/admet.850
59. Bleasby, K., Fillgrove, K.L., Houle, R., Lu, B., Palamanda, J., Newton, D.J., Lin, M., Chan, G.H., Sanchez, R.I.: In vitro evaluation of the drug interaction potential of doravirine. Antimic. Agents Chemother. **63**, 1–12 (2019)
60. Rong, W.-T., Lu, Y.-P., Tao, Q., Guo, M., Lu, Y., Ren, Y., Yu, S.-Q.: Hydroxypropyl-sulfobutyl-β-cyclodextrin improves the oral bioavailability of edaravone by modulating drug efflux pump of enterocytes. J. Pharm. Sci. **103**, 730–742 (2014)
61. Parikh, A., Kathawala, K., Tan, C.C., Garg, S., Zhou, X.-F.: Development of a novel oral delivery system of edaravone for enhancing bioavailability. Int. J. Pharm. **515**, 490–500 (2016)
62. Zeng, J., Ren, Y., Zhou, C., Yu, S., Chen, W.-H.: Preparation and physicochemical characteristics of the complex of edaravone with hydroxypropyl-?-cyclodextrin. Carbohydr. Polym. **83**, 1101–1105 (2011)
63. European Medicines Agency: Zepatier® (elbasvir / grazoprevir) CHMP assessment report. Procedure No. EMEA/H/C/004126/0000. https://www.ema.europa.eu/en/documents/assessment-report/zepatier-epar-public-assessment-report_en.pdf. Accessed 26 May 2016
64. Celgene Inc. (Canada). Product Monograph: Enasidenib mesylate (Idhifa). Accessed 5 Feb 2019.
65. Food and Drug Administration (USA): Entrectinib(Rozlytrek), Genentech NDA 212725Orig1s000 & 212726Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212725Orig1s000,%20212726Orig1s000MultidisciplineR.pdf
66. Food and Drug Administration (USA): Erdafitinib(Balversa), Jansseb Biotech Inc. NDA 212018Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212018Orig1s000ChemR.pdf. Accessed 18 Jan 2022
67. Food and Drug Administration (USA): Fedratinib (Inrebic), Impact Biomedicines, Inc. NDA 212327Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212327Orig1s000ChemR.pdf. Accessed 18 Jan 2022
68. Helsinn Healthcare SA. Highlights of Prescribing Information: Fosnetupitant Dihydrochloride (in AKYNZEO). NDA 210–493 (2018). https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/210493s000lbl.pdf. Accessed 18 Jan 2022

69. Food and Drug Administration (USA): Gilteritinib (Xospata). NDA 211349Orig1s000, CDER Quality Assessment Review. Applicant: Astellas Pharma US, Inc. 26 Nov 2018.
70. Pfizer, Inc. Highlights of prescribing information: Glasdegib (Daurismo). Nov 2018.
71. Food and Drug Administration (USA): Istradefylline (Nourianz), Kyowa Kirin, Inc. NDA 022075Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/022075Orig1s000ChemR.pdf. Accessed 18 Jan 2022
72. O'Neil, M.J., Heckelman, P.E., Dobbelaar, P.H., Roman, K.J. (eds.): The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals, 15th edn. The Royal Society of Chemistry (2013)
73. Crasto, A.M.: Drug Approvals International. http://drugapprovalsint.com/lifitegrast/. Accessed 18 Jan 2022
74. Food and Drug Administration (USA): Lonafarnib (Zokinvy), Eiger Biopharmaceuticals Inc. NDA 213969Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213969Orig1s000ChemR.pdf. Accessed 18 Jan 2022
75. Food and Drug Administration (USA): Macimorelin acetate (Macrilen). NDA 205598Orig1s000, CDER Quality Assessment Review. Applicant: Aeterua Zentaris. 19 Oct 2017.
76. Garcia, J.M., Swerdloff, R., Wang, C., Kyle, M., Kipnes, M., Biller, B.M.K., Cook, D., Yuen, K.C.J., Bonert, V., Dobs, A., Molitch, M.E., Merriam, G.T.: Macimorelin (AEZS-130)-stimulated growth hormone (GH) test: validation of a novel oral stimulation test for the diagnosis of adult GH deficiency. J. Clin. Endocrinol. Metab. **98**, 2422–2429 (2013)
77. Zhou, Z., Du, S., Wang, T., Wu, S., Guo, Z., Wang, Z., Zhou, L.: Measurement and correlation of solubility of meropenem trihydrate in binary (water + acetone/tetrahydrofuran) solvent mixtures. Chin. J. Chem. Eng. **25**, 1461–1466 (2017)
78. Medicines Development for Global Health (Australia) (2018). Highlights of prescribing information: Moxidectin (Daurismo). https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/210867lbl.pdf. Accessed 18 Jan 2022
79. Food and Drug Administration (USA): Naldemedine (Symproic). NDA 208854Orig1s000, CDER Quality Assessment Review. Applicant: Shionogi Inc. 13 Jan 2017.
80. Puma Biotechnology, Inc. Highlights of prescribing information: Neratinib (Nerlynx). Jul 2017.
81. Food and Drug Administration (USA): Nifurtimox (Lampit), Bayer Healthcare Pharmaceuticals, Inc. NDA 213464Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213464Orig1s000ChemR.pdf. Accessed 18 Jan 2022
82. Tesaro, Inc. Highlights of prescribing information: Niraparib (Zejula). Mar 2017.
83. Lancaster, R.G.; Olmstead, K.K.; Kagihiro, M.; Matono, M.; Taoka, I.; Pruzanski, M.; Shapiro, D.; Hooshmand-Rad, R.; Pencek, R.; Sciacca, C.; Eliot, L.; Edwards, J.; MacConell, L.A.; Marmon, T.K.: Compositions of obeticholic acid and methods of use. PTO US 2020/0054650 A1. Feb. 20, 2020.
84. Food and Drug Administration (USA): Ozanimod(Zeposia), Celgene Corp. 2019 NDA 209899Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/209899Orig1s000ChemR.pdf. Accessed 18 Jan 2022
85. Food and Drug Administration (USA): Pemigatinib(Pemazyre), Incyte Corp. 2019 NDA 213736Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213736Orig1s000ChemR.pdf. Accessed 18 Jan 2022
86. Food and Drug Administration (USA): Pexidartinib (Turalio), Daiichi Sankyo Inc. NDA 211810Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211810Orig1s000MultidisciplineR.pdf. Accessed 18 Jan 2022
87. Food and Drug Administration (USA): Pitolisant (Wakix), Bioprojet Pharma NDA 211150Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211150Orig1s000ChemR.pdf. Accessed 18 Jan 2022
88. Food and Drug Administration (USA): Pralsetinib (Gavreto). Blueprint Medicines Corp. NDA 213721. Product Quality Review(s). 5 Aug 2020. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213721Orig1s000ChemR.pdf. Accessed 18 Jan 2022
89. Food and Drug Administration (USA): Pretomanid, The Global Alliance for TB Drug Development. NDA 212862Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212862Orig1s000ChemR.pdf. Accessed 18 Jan 2022
90. Food and Drug Administration (FDA). Highlights of prescribing information. Relugolix (Orgovyx®) Myovant Sciences, Inc. label (Dec 2020). https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/214621s000lbl.pdf. Accessed 18 Jan 2022
91. Yu, K.; Chen, S.; Amgoth, C.; Tang, G.; Bai, H.; Hu, X.: Two polymorphs of remdesivir: crystal structure, solubility, and pharmacokinetic study. Cryst. Eng. Comm. (2021). https://www.rsc.org/suppdata/d1/ce/d1ce00175b/d1ce00175b1.pdf. Accessed 18 Jan 2022

92. Huang, H.; Zhou, G.; Shang, G.; et al. Hydrobromate of benzodiazepine derivative, preparation method and use thereof. Chengdu Brilliant Pharmaceutical Co., Ltd. Eur. Patent Applic. EP 3 553 059 A1 (2019).
93. Sirvent, J.A., Lücking, U.: Novel pieces for the emerging picture of sulfoximines in drug discovery: synthesis and evaluation of sulfoximine analogues of marketed drugs and advanced clinical candidates. Chem. Med. Chem. **12**, 487–501 (2017). https://doi.org/10.1002/cmdc.201700044
94. Samant, T.S., Dhuria, S., Lu, Y., Laisney, M., Yang, S., Grandeury, A., Mueller-Zsigmondy, M., Umehara, K., Huth, F., Miller, M., Germa, C., Elmeliegy, M.: Ribociclib bioavailability is not affected by gastric *p*H changes or food intake: in silico and clinical evaluations. Clin. Pharmacol. Ther. **104**, 374–383 (2018)
95. Food and Drug Administration (USA): Ripretinib (Qinlock), Deciphera Pharmaceuticals, LLC. 2020 NDA 213973Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213973Orig1s000RiskR.pdf
96. Genentech, Inc. Safety Data Sheet: Risdiplam (Evrysdi). https://www.gene.com/download/pdf/EVRYSDIRisdiplam0.75mgpermlSDS.pdf. Accessed 16 June 2020
97. Food and Drug Administration (USA): Safinamide (Xadago). NDA 207145Orig1s000, CDER Quality Assessment Review. Applicant: Newron Pharmaceuticals. 29 Dec 2014.
98. Food and Drug Administration (USA): Secnidazole (Solosec). NDA 209363Orig1s000, CDER Quality Assessment Review. Applicant: Symbiomix Therapeutics, LLC. 27 Jul 2017.
99. Rivera, A.B., Hernández, R.G., de Armas, H.N., Elizástegi, D.M.C., Losada, M.V.: Physico-chemical and solid-state characterization of secnidazole. Il Farmaco **55**, 700–707 (2000)
100. Food and Drug Administration (USA): Selinexor (Xpovio), Karyopharm Therapeutics Inc. NDA 212306Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212020Orig1s000ChemR.pdf. Accessed 18 Jan 2022
101. Mogalian, E., German, P., Kearney, B.P., Yang, C.Y., Brainard, D., Link, J., McNally, J., Hane, L., Ling, J., Mathias, A.: Preclinical pharmacokinetics and first-in-human pharmacokinetics, safety, and tolerability of velpatasvir, a pangenotypic HCV NS5A inhibitor, in healthy subjects. Antimicrob. Agents Chemother. **61**, e02084-e2116 (2017). https://doi.org/10.1128/AAC.02084-16
102. Food and Drug Administration (USA): Selumetinib <Koselugo>, AstraZeneca NDA213756Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213756Orig1s000ChemR.pdf. Accessed 18 Jan 2022
103. Leijen, S., Soetekouw, P.M.M.B., Evans, T.R.J., Nicolson, M., Schellens, J.H.M., Learoyd, M., Grinsted, L., Zazulina, V., Pwint, T., Middleton, M.: A phase I, open-label, randomized crossover study to assess the effect of dosing of the MEK 1/2 inhibitor Selumetinib (AZD6244; ARRY-142866) in the presence and absence of food in patients with advanced solid tumors. Cancer Chemother. Pharmacol. **68**, 1619–1628 (2011). https://doi.org/10.1007/s00280-011-1732-7
104. European Medicines Agency: Epclusa? (sofosbuvir/velpatasvir) CHMP Assessment Report. EMA/399285/2016. Procedure No. EMEA/H/C/004210/0000 https://www.ema.europa.eu/en/documents/assessment-report/epclusa-epar-public-assessment-report_en.pdf. Accessed 18 Jan 2022
105. Target Molecule Corp.: Stiripentol (Diacomit). https://www.targetmol.com/compound/Stiripentol. Accessed 18 Jan 2022
106. Food and Drug Administration (USA): Talazoparib (Talzenna). NDA 211651Orig1s000, CDER Quality Assessment Review. Applicant: Pfizer. 3 Oct 2018.
107. Food and Drug Administration (USA): Tazemetostat (Tazverik), Epizyme 2019 NDA 211723Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/211723Orig1s000ChemR.pdf. Accessed 18 Jan 2022
108. Drug Approvals International: Tecovirimat (TPOXX). http://drugapprovalsint.com/tecovirimat/. Accessed 14 Jul 2018
109. Food and Drug Administration (USA): Tenapanor(Ibsrela), Ardelyx, Inc. NDA 211801Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211801Orig1s000ChemR.pdf. Accessed 18 Jan 2022
110. Food and Drug Administration (USA): Tezacaftor (Trikafta), Vertex Pharmaceuticals, Inc. NDA 210491Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2018/210491Orig1s000ChemR.pdf. Accessed 18 Jan 2022
111. Food and Drug Administration (USA): Tucatinib <Tukysa>, Seattle Genetics. NDA213411Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213411Orig1s000MultidisciplineR.pdf. Accessed 18 Jan 2022
112. Food and Drug Administration (USA): Upadacitinib (Rinvoq), AbbVie Inc. NDA 211675Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211675Orig1s000ChemR.pdf. Accessed 18 Jan 2022

113. Food and Drug Administration (USA): Venetoclax <Venclexta>, AbbVie Inc. NDA 208573Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2016/208573Orig1s000ClinPharmR.pdf. Accessed 18 Jan 2022
114. Food and Drug Administration (USA): Vibegron (Gemtesa). Urovant Sciences, Inc. NDA 213006. Product Quality Review(s). https://www.accessdata.fda.gov/drugsatfda_docs/nda/2020/213006Orig1s000ChemR.pdf. Accessed 24 Nov 2020
115. Food and Drug Administration (USA): Zanubrutinib (Brukinsa), BioGene USA. NDA 213217Orig1s000. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/213217Orig1s000MultidisciplineR.pdf
116. Avdeef, A.: Solubility temperature dependence predicted from 2D structure. ADMET & DMPK **3**, 298–344 (2015)
117. 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)
118. Avdeef, A.: Anomalous solubility behavior of several acidic drugs. ADMET & DMPK **2**, 33–42 (2014)
119. Avdeef, A.: Phosphate precipitates and water-soluble aggregates in re-examined solubility–pH data of twenty-five basic drugs. ADMET & DMPK **2**, 43–55 (2014)
120. 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)
121. 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)
122. Bergström, C.A.S., Avdeef, A.: Perspectives in solubility measurement and interpretation. ADMET & DMPK **7**, 88–105 (2019)
123. Avdeef, A.: Absorption and Drug Development, 2nd edn. Wiley-Interscience, Hoboken NJ (2012)
124. Avdeef, A.: Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. ADMET & DMPK **7**, 210–219 (2019). https://doi.org/10.5599/admet.698
125. 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.org/10.1021/acs.jcim.9b00345
126. 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.org/10.1021/acs.jcim.0c00701
127. 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. Accessed 18 Jan 2022
128. 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)
129. Lipinski, C.A.: Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Tox. Meth. **44**, 235–249 (2000)
130. Avdeef, A.: Do you know your r^2? ADMET& DMPK (2021). https://doi.org/10.5599/admet.888