Sudarshan Publications

Available online at https://s-epub.in/index.php/trp/index

Pharma Research

An International Peer Reviewed, Indexed Journal

10.62655/s-epub.2022.v14.i02.pp67-74

MACHINE LEARNING PREDICTION OF AQUEOUS SOLUBILITY FOR EARLY-STAGE DRUG CANDIDATES

FLORIAN HOFFMEISTER[1]*, VIKTOR NYBERG[2], LUCIE KUCERA[3]

Affiliation

[1]Department of Pharmaceutical Chemistry, Fichtelgebirge Centre for Drug Research, Hof, Germany
[2]Department of Pharmaceutical Chemistry, Norrland Institute of Clinical Pharmacology, Umea, Sweden
[3]Department of Medicinal Chemistry, Vysocina Institute of Pharmaceutical Chemistry, Jihlava, Czech Republic

*Corresponding author: Florian Hoffmeister, f.hoffmeister@fcdr-hof.de
ORCID iD: Florian Hoffmeister 0000-0001-1895-8629; Viktor Nyberg 0000-0001-4604-0786; Lucie Kucera 0000-0001-1495-759X

Article info

Received date:07-02-2022
Accepted date:28-05-2022
Published date:19-09-2022
ISSN;0975-8216

Keywords:

Aqueous solubility, Machine learning, QSPR, Drug discovery, Molecular descriptors

Abstract

Poor aqueous solubility is a frequent reason why drug candidates stall, and predicting it from structure before synthesis would save time and material. This narrative review summarises published work on computational solubility prediction, from the general solubility equation and linear descriptor models to random forests, graph neural networks and deep learning. It considers the quality of experimental data, benchmark exercises such as the Solubility Challenge, and practical issues of validation and applicability domain. Reported errors cluster around one log unit for new chemistry and are limited as much by noisy measurements as by algorithms. Model choice matters less than data curation and honest validation. The article is a review and presents no new data.

1. INTRODUCTION

A compound that will not dissolve is hard to develop. Poor solubility limits absorption after oral dosing, makes intravenous formulation difficult, and distorts biological assays, where precipitation can create false negatives or, less obviously, false positives. A large fraction of molecules emerging from modern discovery programmes are lipophilic and poorly soluble, partly because screening and optimisation tend to reward potency, which often travels with molecular size and lipophilicity [1, 2]. Chemists therefore look for ways to estimate solubility before a compound is made.

Computational prediction, usually framed as a quantitative structure-property relationship (QSPR) problem, has a long history. Early approaches used simple relationships between solubility, lipophilicity and melting point, and

later ones used regression on a handful of descriptors. In the last decade, machine learning methods applied to large descriptor sets and to molecular graphs have attracted most of the attention, helped by the growth of public datasets and by software that makes training models nearly effortless [3, 4, 5]. The question this article asks is how far that progress has really moved the field, and where limits remain.

A caution on expectations is appropriate at the outset. Machine learning methods are sometimes presented as if they remove the need for measurement, and that is not what the evidence supports. The more defensible claim is that models can narrow choices and flag likely problems. Because the practical uses of solubility predictions range from ranking virtual libraries of thousands of structures to checking a single analogue, the accuracy needed also varies, and a model that is adequate for one task may be unfit for another [2, 6, 7].

The text that follows reviews published work and reports no new data. It describes how solubility is defined and measured, outlines classical and machine learning prediction methods, discusses the data and benchmarks on which they are judged, and considers validation, applicability domain and use in early-stage discovery. Literature up to 2022 is considered. Attention is given to what a medicinal chemist needs to know when interpreting a predicted value for a novel structure.

2. SOLUBILITY AND ITS MEASUREMENT

Any prediction is only as meaningful as the quantity being predicted. Thermodynamic solubility is the concentration of dissolved compound in equilibrium with its most stable solid form in a given medium and temperature, usually measured by the saturation shake-flask method over a day or more [8, 9]. It depends on the crystal form, so polymorphs and solvates of one compound can give different values. Intrinsic solubility refers to the neutral species of an ionisable compound and is distinct from the apparent solubility at a particular pH, which varies by orders of magnitude for acids and bases [8].

Discovery laboratories rarely have time for the full procedure. Kinetic solubility assays dilute a dimethyl sulfoxide (DMSO) stock into aqueous buffer and detect precipitation by turbidity or by analysis after filtration. They are fast and need little material, but the result reflects the amorphous or metastable precipitate and generally exceeds the thermodynamic value, sometimes by a large margin [10, 11]. Because the two measures are not interchangeable, datasets that mix them introduce hidden noise, and a model trained on one may not transfer to the other.

Experimental conditions matter in ways that are easy to overlook. Temperature, buffer composition and ionic strength, equilibration time, the method of separating solid from solution, and the analytical readout can all alter the number. Even for a single compound under nominally identical conditions, different laboratories have reported results that disagree by half a log unit or more [6, 9]. The unit commonly used in modelling is the base-ten logarithm of molar solubility, logS, and an error of one unit corresponds to a ten-fold difference in concentration, which is large for formulation purposes.

For oral drugs, the link between solubility and absorption is made explicit in the Biopharmaceutics Classification System (BCS), which groups drugs by solubility and permeability. A drug is considered highly soluble when its highest dose dissolves in a small volume of aqueous medium across the physiological pH range [12]. This classification has regulatory consequences, including waivers of in vivo bioequivalence studies, and it explains why early estimates of solubility are used alongside predictions of permeability and dose when ranking candidates [1, 2, 12].

The consequences for development are concrete. A poorly soluble drug may need a salt, a co-crystal, particle size reduction, a lipid-based system or an amorphous solid dispersion to reach adequate exposure, and each route adds cost and time. In biological assays, compounds that precipitate at the test concentration give flat or erratic dose-response curves and can be wrongly discarded or, worse, wrongly pursued [10, 11]. Early information about solubility lets

teams deprioritise a series, or at least plan the formulation work, before significant resources are committed [2, 12].

3. FROM EMPIRICAL EQUATIONS TO MACHINE LEARNING

The general solubility equation (GSE) of Yalkowsky and colleagues relates logS to the octanol-water partition coefficient and the melting point, reflecting the energy needed to disrupt the crystal and to create a cavity in water [13]. It is physically reasonable and needs only two inputs, but the melting point is unavailable for an unmade compound, and the equation performs poorly for highly polar or ionised molecules. Variants have replaced experimental inputs with predicted ones, and one study added topological polar surface area to improve the fit [14, 15].

Descriptor-based thinking owes much to the rule of five. Lipinski observed that poor absorption or permeation is more likely when a compound has more than five hydrogen-bond donors, more than ten acceptors, a molecular weight above 500 or a calculated logP above five [2]. The rules were not derived as a solubility model, but they established the idea that a few computed properties carry information about developability. Later solubility models developed this idea quantitatively, and logP in particular is the descriptor with the strongest and most consistent negative relationship to logS [1, 3].

Regression models built from calculated descriptors followed. The ESOL model estimated solubility from four easily computed quantities, namely calculated logP, molecular weight, number of rotatable bonds and the proportion of aromatic atoms, and its simplicity and openly available dataset made it a popular reference point [3]. Fragment and atom-contribution methods assigned a solubility contribution to each atom type and summed them [16, 17], and later work combined atom counts with solvent-accessible surface areas [18]. Such models are transparent and quick, with typical errors around one log unit on test sets of similar chemistry.

Non-linear machine learning arrived with random forests and support vector machines applied to descriptor sets of hundreds of variables. Random forests are an ensemble of decision trees trained on resampled data, are robust to irrelevant descriptors and rarely require fine tuning, which explains their continued popularity [19, 20]. Molecular fingerprints such as extended-connectivity fingerprints encode substructures as bit vectors and are the usual input when no physical model is intended [21]. Comparisons in the literature generally find that non-linear methods give modest improvements over linear ones in-domain, but not a step change [7, 22].

Choice of features is as consequential as choice of algorithm. Descriptor pools easily reach thousands of variables, many of them correlated, and training on all of them invites over-fitting when the number of compounds is a few thousand. Feature selection, regularised regression and the inherent selection in tree ensembles help, as does simply using a smaller set of chemically interpretable descriptors [19, 21, 22]. Fingerprints avoid the selection problem but obscure the physical meaning, so it is difficult to say why a particular structure is predicted as insoluble.

Deep learning offers another route. Recursive and convolutional architectures operating directly on molecular graphs learn their own features instead of relying on predefined descriptors, and early applications to aqueous solubility reported lower errors than fingerprint-based models on large collections [23, 24]. Message-passing neural networks formalised the idea of aggregating information over atoms and bonds [25], and a directed message-passing implementation became a widely used baseline in property prediction [26]. Large networks trained on bigger datasets have since claimed further reductions in error [4, 27].

Physics-based routes exist but have not displaced statistical ones. Computer simulation of the free energy of solvation, combined with estimates of crystal lattice energy, can in principle predict solubility from first principles. In practice, accurate crystal structure prediction for an unmade compound is a hard problem, and errors in the lattice term dominate. Monte Carlo

simulation approaches have shown reasonable correlation for small sets of compounds, but the cost and expertise they demand limit routine use [14, 15, 28]. Their value lies more in understanding trends than in screening large libraries.

Hybrid approaches combine physics with learning. One study used machine learning to predict solubility in organic solvents and water from physicochemical relationships such as the thermodynamic cycle linking solubility to melting properties and solvation, with descriptors informed by those relationships [28]. The motivation is to reduce the amount of data needed and improve extrapolation, because a model embedded in physical reasoning should degrade more gracefully on unfamiliar structures than a purely statistical one. These hybrids are still few, and independent validation on prospective compounds is limited.

4. DATA QUALITY AND BENCHMARKS

Data are the limiting factor in many analyses. Public collections assembled from the literature merge measurements from different laboratories, methods and temperatures, with unit errors and duplicates inevitably creeping in. A curated reference set named AqSolDB combined several sources and removed duplicates, giving nearly ten thousand compounds with 2D descriptors [5]. Curated sets of this kind raise the baseline for reproducibility, and the Wiki-pS0 database, with careful attention to measurement conditions and intrinsic values, has been used to train models with more attention to data provenance [29].

The Solubility Challenge offered a direct test. Participants received a training set of 100 compounds with carefully measured intrinsic solubilities and were asked to predict values for 32 further molecules [30]. When the results were compared, the organisers noted that no method stood out, that predictions by some groups were poor for compounds that most models handled easily, and that success was not clearly related to model complexity [31]. A second challenge using a different data source repeated the exercise and again showed that predictive accuracy was limited by the variability of the data.

A key follow-up asked whether experimental error sets a ceiling on achievable accuracy. Analysis of inter-laboratory differences for the same compounds suggested a standard deviation of roughly half a log unit, so that models achieving errors of 0.6 to 0.7 log units are approaching the noise in the data [6]. Another study asked human experts to predict solubilities and found that their predictions were no better than those from computational models on the same compounds, which again indicated that the limit lay in the problem, not the algorithm [32].

Benchmark collections have done much for reproducibility. MoleculeNet gathered standard datasets, including a solubility set derived from the ESOL data, with recommended splits and metrics, so that new models can be compared on a like-for-like basis [4]. The downside is that popular benchmarks can be over-fitted by the community, since they are small, chemically narrow and often random-split. A method that wins on a benchmark of about a thousand compounds may fail to generalise to proprietary collections, whose chemistry is both different and evolving.

The composition of public data deserves scrutiny. The ESOL set contains many pesticides and industrial chemicals alongside drugs, while AqSolDB gathers compounds from several sources that were assembled for different purposes [3, 5]. Neither is a faithful sample of the lead-like and drug-like molecules that discovery chemists make today. Models trained on such sets may perform well on the test portion yet poorly on a new series with larger molecular weight, more stereocentres and different heteroatom patterns, a mismatch described as distribution shift and recognised as a major threat to deployment of machine learning in chemistry [4, 5].

Finally, negative and borderline results are underreported. Papers and databases favour compounds that dissolved well enough to be measured, and very insoluble compounds are often recorded only as less than some limit, then excluded from training. The resulting datasets under-represent the tail of poorly soluble structures that are of greatest concern,

and regression models trained on them tend to compress predictions towards the middle of the range. Methods for handling censored data are available in statistics, but are rarely applied to solubility [5, 6].

5. VALIDATION AND APPLICABILITY

How a model is validated determines how much to trust it. Random splitting of a dataset places structurally similar compounds in both training and test sets, so test error is optimistic for genuinely novel chemistry. Scaffold-based splits, cluster-based splits and time-split validation, in which models are trained on older compounds and tested on later ones, give more realistic estimates for prospective use [33]. Time-split studies in industrial datasets typically find larger errors than random cross-validation, sometimes by a substantial margin [33].

Reporting also affects interpretation. Root mean square error and mean absolute error in log units are more informative than the coefficient of determination, which depends strongly on the spread of the test data and can look high even when the absolute error is too large to be useful. Reporting confidence intervals, repeating splits with several random seeds, and comparing against a trivial baseline such as the training mean or a simple linear model are low-cost steps that many papers omit. Without them, claims of marginal improvement between methods cannot be assessed [33, 34, 35].

The Organisation for Economic Co-operation and Development (OECD) has set out principles for validation of quantitative structure-activity relationship models: a defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and a mechanistic interpretation where possible [35]. Methodological work in the academic literature broadly agrees, with emphasis on external validation rather than reliance on internal cross-validation statistics, and on the avoidance of selecting models solely on test-set performance [34]. Many published solubility models meet some of these criteria but not all.

The applicability domain is the region of chemical space in which a model's predictions are reliable. Distance to training compounds, fingerprint similarity, descriptor range and ensemble variance have all been used as indicators. For early-stage candidates, which by definition may differ from known compounds, the domain issue is central, since a prediction with a narrow domain is unhelpful precisely when it is most wanted. Models that supply an uncertainty estimate with the prediction are more useful than those that give a single number, and ensemble approaches provide this at modest cost [20, 22, 35].

Adoption depends on trust, and trust depends on explanation. Medicinal chemists are more likely to act on a prediction if they can see which features drove it, whether a lipophilic fragment or an aromatic ring count, and to compare the case with similar measured compounds. Feature importance for tree ensembles, substructure attribution methods for graph networks, and simple nearest-neighbour displays all serve this purpose [7, 20, 26]. A model whose explanation contradicts chemical sense should be treated sceptically, even if its overall statistics look strong.

Ionisation and solid-state form create specific blind spots. Most models predict intrinsic or unspecified solubility and do not distinguish ionisation states, so a prediction for a basic amine may be far from the value at gastric or intestinal pH unless it is corrected using the pKa [8, 12]. Crystal packing effects are not captured by 2D descriptors at all, which is why compounds with high melting points are often under-predicted in solubility and why seemingly minor structural changes can give large jumps in observed solubility. Including predicted melting point or lattice energy has been explored with limited success [14, 15].

Solubility in biorelevant media is a distinct question from solubility in water. Bile salts and phospholipids in fasted and fed intestinal fluids can raise the apparent solubility of lipophilic drugs substantially, and for such compounds the aqueous value may seriously understate what is available for absorption. Predictive models for these media are few and trained on far smaller datasets, so most discovery teams still use aqueous predictions only as a first screen and measure biorelevant

solubility experimentally for shortlisted compounds [11, 12]. Expect this to change as data accumulate.

6. LIMITATIONS AND FUTURE DIRECTIONS

The main findings from the literature are presented in table 1, and the distribution of entries by theme is shown in fig. 1. The table illustrates a pattern that runs through the whole field: the methods are many, but the evidence about data limits and validation is more decisive than any comparison of algorithms. Most studies also report retrospective performance on random or benchmark splits, so statements about expected accuracy on new candidates should be read with caution.

Table 1: Solubility prediction approaches and evidence on data and validation

Approach or issue	Reported finding	Ref.
General solubility equation logP and melting point	Physically grounded but needs melting point; weak for ionised or very polar compounds	[13, 14, 15]
ESOL four descriptors	Simple linear model; widely used reference dataset and baseline	[3]
Atom-contribution models fragment methods	Transparent summation of atom type terms; errors near one log unit	[16, 17, 18]
Random forest descriptors and fingerprints	Robust, little tuning; modest gains over linear models in domain	[19, 20, 21]
Graph neural networks learned features	Lower errors reported on large sets; gains depend on data size	[23, 24, 25, 26]
Deep learning large datasets	Claimed improved prediction of novel compounds with deeper architectures	[27]
Solubility Challenge blind test	No method stood out; accuracy limited by data variability	[30, 31]
Inter-laboratory error data quality	About half a log unit noise suggests ceiling near 0.6 to 0.7 log	[6, 32]
Curated databases AqSolDB and Wiki-pS0	Duplicate removal and provenance improve reproducibility	[5, 29]
Time-split validation prospective use	Gives larger and more realistic errors than random cross-validation	[33]
OECD principles applicability domain	Defined endpoint, algorithm, domain and validation measures	[35]

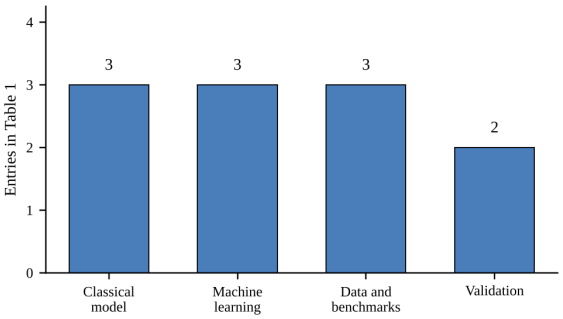


Fig. 1: Number of entries in table 1 by theme

Regulatory and organisational factors also bear on how these tools are used. In regulatory toxicology, QSAR predictions are accepted under defined conditions, with documentation of the domain and the validation, and similar reporting discipline would benefit discovery models [35]. Inside companies, models are often retrained on proprietary measurements, and local models built from a single series can beat global models for that series, a finding that goes back to early comparisons of global and local approaches [22]. Teams therefore benefit from combining a broad public model with a locally fitted correction.

Several directions appear promising. Better and larger collections of consistent thermodynamic measurements, including temperature, pH and solid form, would address the problem at the root. Models that combine physical descriptors with learned representations, and that report calibrated uncertainty, are likely to be more reliable than either approach alone. Transfer learning from large unlabelled molecular sets and multi-task learning across related properties have been proposed but need systematic testing for solubility. Open sharing of high-quality industrial data, perhaps through federated approaches, could change what is achievable [4, 5, 6, 29].

7. CONCLUSION

Solubility prediction has progressed from two-parameter equations to graph neural networks, but the improvement in error has been much smaller than the growth in model complexity. For compounds resembling the training data, errors of about one log unit or a little less are achievable with several methods. For new chemotypes, the error is larger and often unknown, because

applicability domain and prospective validation are seldom reported [3, 13, 23, 27, 33].

The most effective investments are therefore likely to lie outside the algorithm: consistent, well-annotated experimental data, validation schemes that mimic prospective use, and uncertainty estimates that tell the chemist when to ignore the number. Models are best used for triage and for ranking ideas before synthesis, and measured solubility should remain the arbiter once material exists [1, 6, 12, 30, 35].

8. ABBREVIATIONS

GSE-General solubility equation, QSPR-Quantitative structure-property relationship, QSAR-Quantitative structure-activity relationship, logS-Base-ten logarithm of molar solubility, logP-Logarithm of the octanol-water partition coefficient, ESOL-Estimated solubility, OECD-Organisation for Economic Co-operation and Development, BCS-Biopharmaceutics Classification System, DMSO-Dimethyl sulfoxide

9. FUNDING

Nil

10. AUTHORS CONTRIBUTIONS

All the authors have contributed equally.

11. CONFLICT OF INTERESTS

Declared none

12. REFERENCES

1. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 2007;6:881–90.
2. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 2000;44:235–49.
3. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 2004;44:1000–5.
4. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, *et al.* MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9:513–30.
5. Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Sci Data 2019;6:143.
6. Palmer DS, Mitchell JBO. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? Mol Pharm 2014;11:2962–72.
7. Bergström CAS, Larsson P. Computational prediction of drug solubility in water-based systems: qualitative and quantitative approaches used in the current drug discovery and development setting. Int J Pharm 2018;540:185–93.
8. Avdeef A. Solubility of sparingly-soluble ionizable drugs. Adv Drug Deliv Rev 2007;59:568–90.
9. Baka E, Comer JEA, Takács-Novák K. Study of equilibrium solubility measurement by saturation shake-flask method using hydrochlorothiazide as model compound. J Pharm Biomed Anal 2008;46:335–41.
10. Di L, Kerns EH. Biological assay challenges from compound solubility: strategies for bioassay optimization. Drug Discov Today 2006;11:446–51.
11. Alsenz J, Kansy M. High throughput solubility measurement in drug discovery and development. Adv Drug Deliv Rev 2007;59:546–67.
12. Yu LX, Amidon GL, Polli JE, Zhao H, Mehta MU, Conner DP, *et al.* Biopharmaceutics classification system: the scientific basis for biowaiver extensions. Pharm Res 2002;19:921–5.
13. Ran Y, Yalkowsky SH. Prediction of drug solubility by the general solubility equation (GSE). J Chem Inf Comput Sci 2001;41:354–7.
14. Jorgensen WL, Duffy EM. Prediction of drug solubility from structure. Adv Drug Deliv Rev 2002;54:355–66.
15. Ali J, Camilleri P, Brown MB, Hutt AJ, Kirton SB. Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. J Chem Inf Model 2012;52:420–8.
16. Hou TJ, Xia K, Zhang W, Xu XJ. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. J Chem Inf Comput Sci 2004;44:266–75.
17. Tetko IV, Tanchuk VY, Kasheva TN, Villa AE. Estimation of aqueous solubility of chemical compounds using E-state indices. J Chem Inf Comput Sci 2001;41:1488–93.

18. Wang J, Hou T, Xu X. Aqueous solubility prediction based on weighted atom type counts and solvent accessible surface areas. J Chem Inf Model 2009;49:571–81.
19. Breiman L. Random forests. Mach Learn 2001;45:5–32.
20. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43:1947–58.
21. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.
22. Bergström CAS, Wassvik CM, Norinder U, Luthman K, Artursson P. Global and local computational models for aqueous solubility prediction of drug-like molecules. J Chem Inf Comput Sci 2004;44:1477–88.
23. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model 2013;53:1563–75.
24. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, *et al.* Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst 2015;28:2224–32.
25. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. Proc Int Conf Mach Learn 2017;70:1263–72.
26. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, *et al.* Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59:3370–88.
27. Cui Q, Lu S, Ni B, Zeng X, Tan Y, Chen YD, *et al.* Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. Front Oncol 2020;10:121.
28. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun 2020;11:5753.
29. Avdeef A. Prediction of aqueous intrinsic solubility of druglike molecules using random forest regression trained with Wiki-pS0 database. ADMET DMPK 2020;8:29–77.
30. Llinàs A, Glen RC, Goodman JM. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? J Chem Inf Model 2008;48:1289–303.
31. Hopfinger AJ, Esposito EX, Llinàs A, Glen RC, Goodman JM. Findings of the challenge to predict aqueous solubility. J Chem Inf Model 2009;49:1–5.
32. Boobier S, Osbourn A, Mitchell JBO. Can human experts predict solubility better than computers? J Cheminform 2017;9:63.
33. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 2013;53:783–90.
34. Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model 2015;55:1316–22.
35. Organisation for Economic Co-operation and Development. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD Series on Testing and Assessment No. 69. Paris: OECD; 2007.