Received: 23 December 2020 | Revised: 22 February 2021 | Accepted: 18 April 2021

DOI: 10.1002/cem.3349

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability?

Mario Lovrić[1,2] | **Kristina Pavlović**[1] | **Petar Žuvela**[3] | **Adrian Spataru**[1] | **Bono Lučić**[2] | **Roman Kern**[1,4] | **Ming Wah Wong**[3]

[1]Knowledge Discovery, Know-Center, Graz, Austria

[2]NMR Centre, Ruđer Bošković Institute, Zagreb, Croatia

[3]Department of Chemistry, National University of Singapore, Singapore

[4]Institute of Interactive Systems and Data, Graz University of Technology, Graz, Austria

Correspondence
Mario Lovrić, Knowledge Discovery, Know-Center, Inffeldgasse 13, 8010 Graz, Austria.
Email: mlovric@know-center.at

Funding information
European Union, Grant/Award Number: KK.01.1.1.01; European Union

Abstract

We present a collection of publicly available intrinsic aqueous solubility data of 829 drug-like compounds. Four different machine learning algorithms (random forests [RF], LightGBM, partial least squares, and least absolute shrinkage and selection operator [LASSO]) coupled with multistage permutation importance for feature selection and Bayesian hyperparameter optimization were used for the prediction of solubility based on chemical structural information. Our results show that LASSO yielded the best predictive ability on an external test set with a root mean square error (RMSE) (test) of 0.70 log points, an R^2(test) of 0.80, and 105 features. Taking into account the number of descriptors as well, an RF model achieves the best balance between complexity and predictive ability with an RMSE(test) of 0.72 log points, an R^2(test) of 0.78, and with only 17 features. On a more aggressive test set (principal component analysis [PCA]-based split), better generalization was observed for the RF model. We propose a ranking score for choosing the best model, as test set performance is only one of the factors in creating an applicable model. The ranking score is a weighted combination of generalization, number of features, and test performance. Out of the two best learners, a consensus model was built exhibiting the best predictive ability and generalization with RMSE(test) of 0.67 log points and a R^2(test) of 0.81.

KEYWORDS

consensus modeling, LASSO, LightGBM, PCA, permutation importance, QSAR, random forests

Mario Lovrić, Kristina Pavlović, and Petar Žuvela contributed equally.

1 | INTRODUCTION

Solubility is a critical topic in pharmaceutical development as it can be a limiting factor to drug absorption.[1] High attrition rate in drug development has been attributed to poor water solubility.[2] Predictive models such as quantitative structure–property relationships (QSPRs) can be useful tools to determine the solubility of a bioactive compound starting already in early development stages. Llinas and Avdeef[3] initiated the second solubility challenge in 2019 in order to engage the scientific community to address this challenging problem.

The first solubility challenge published by the same authors[4] demonstrated clear room for improvement in predicting solubility from (molecular) structural information. Palmer and Mitchell[5] concluded that there is still room for improvement with respect to predictive capabilities of QSPR rather than the lacking quality of data. Nevertheless, there is still a lack of public data available to develop quality models or at least cover a larger chemical space. In fact, it is the aforementioned solubility challenges that made quality data available. At the same time, pharmaceutical companies still own a large amount of unpublished data. Using such an unpublished dataset with experimental values of 38,841 compounds, Montanari et al.[6] tested multitask neural networks for solubility prediction. The authors built a model that yielded a cross-validated R^2 value of 0.59 (root mean square error [RMSE] not published). Such a data size for solubility is rare among publicly available datasets. Even though one cannot be sure about the quality of proprietary data, it might confirm Palmer's conclusion about limitations in modeling capabilities.

Many other research groups also dealt with the solubility prediction challenge,[5–30] attempting to predict both $\log S_w$ (aqueous solubility; measured at a certain pH) and $\log S_0$ (intrinsic solubility; solubility of a compound in its free acid or base form).[1] Key studies were summarized in Table S1. A comparison with previous studies is difficult because the authors often analyze the model quality in different manners (train, test, cross-validation, out-of-fold) and involved a multitude of model metrics.[31] Specifically, for the intrinsic solubility, literature values of the predictive performance of models on external test sets expressed by RMSE appear to vary between 0.7 and 1.05 log points[13,15,17,18,26,28,32] using a plethora of machine learning algorithms and datasets.

The most recent study from Avdeef[17] with the largest curated database known (6355 $\log S_0$ entries) applied the random forests (RF) algorithm yielded RMSE(test) in a range of 0.75–1.05 and with an R^2 values between 0.66 and 0.83 across several models. These results outperform studies with the aforementioned proprietary databases, which signals the importance of careful data curation and chemical space consideration that Avdeef advocated. Within the aforementioned challenges, additional high-quality solubility data were published. With the availability of efficient and reliable machine learning methods as well as the ever increasing in computing power in HPC environments, more precise and faster learning models are available nowadays. Our goal in this work was to conduct a large-scale machine learning study to investigate how one can achieve robust predictions while retaining minimum model complexity.

For this purpose, we curated a novel intrinsic solubility dataset from literature sources. For the machine learning tasks, we used boosting and bagging ensemblers as well as partial least squares (PLS) and least absolute shrinkage and selection operator [LASSO] methods. The last two being established machine learning modes that are often neglected over seemingly more powerful ensemble regressors.[33] Consensus modeling was employed to build a final QSPR model. Finally, we discussed the use of permutation importance for a multistage feature selection, the relationship of metrics within data splits, and the relevancy of commonly used feature preprocessing/preselection and data splitting paradigms. Furthermore, we present a more challenging test set to test the models' extrapolation capabilities.

2 | MATERIALS AND METHODS

2.1 | Data collection and processing

We have collected aqueous solubility data from the following literature sources.[4,12,15,16,18,22,34–52] The decision criteria on which literature to include for our study is initially based on the recommendations in the revisited solubility challenge.[3] Subsequently, we looked for additional literature sources where authors have included pH, which were measured between 22.5°C and 25°C temperature and used inert gases (argon, nitrogen) in their measurements. Most of the above-mentioned solubility data sources refer to the intrinsic aqueous solubility ($\log S_0$), while others refer to the aqueous solubility ($\log S_w$). For each compound, SMILES strings were retrieved from the name either through PubChem (https://pubchem.ncbi.nlm.nih.gov/), JChem (Marvin/JChem v20.9.0, ChemAxon, Budapest, Hungary), or via their CAS numbers (https://cactus.nci.nih.gov/translate/). SMILES strings were curated[53] and standardized to isomeric

SMILES using the ChemAxon Standardizer (v18.28.0, ChemAxon, Budapest, Hungary) and the RDKit library.[54] We filtered compounds with the following properties: log$P$[55] in [−3.6, 7.5], molecular weight larger than 88 g/mol, and structures with more than six heavy atoms. These ranges were determined according to the data published in the solubility challenges.[3] The obtained logS_w values in the extracted data were converted to logS_0 based on their formal charges as suggested by Abraham and Le[46] and Avdeef.[56] Because we had multiple values for intrinsic solubility per molecule, we removed the duplicated values and averaged the rest. In total, out of the 829 compounds in the final data set, 446 had originally logS_0 values, whereas for the other 383 compounds, we have calculated the values from logS_w.

The data preparation pipeline is depicted in Figure 1. We calculated and considered in modeling two types of predictive features: fingerprints (FPs)[57] and molecular descriptors (DPs) (calculated using DRAGON 6.0—Talete, Milano, IT). We chose FPs with a comparatively short radius of 3 bonds and large vector length of 5120 bits, to avoid bit collision as suggested by Landrum.[58] From the available ∼5000 DRAGON molecular DPs, only a few groups of DPs were selected based on chemical intuition, specifically, constitutional, ring, topological DPs, functional group counts, and molecular properties. All DPs with missing values were removed. Such a preselection procedure yielded a total of 317 molecular DPs. A combination of FPs and DPs (FPDS) was also evaluated (5444 features in total).

2.2 | Evaluated machine learning methods

For development of intrinsic solubility models of chemical based on their structure, four regression algorithms different in their paradigms were applied: (i) LASSO,[59] (ii) PLS,[60] (iii) RF,[61] and (iv) LightGBM.[62] All four are briefly summarized in the subsequent subsections.

2.2.1 | Least absolute shrinkage and selection operator

LASSO regression is a multivariate chemometric method, which involves the L_1-penalty for regularization.[59] Given the multiple linear regression formulation with standardized features/predictors X (N, p) and response variable (N, 1) y, LASSO aims to solve the L_1-penalized regression problem of finding a set of p model coefficients $\beta = \{\beta_j\}$ to minimize:

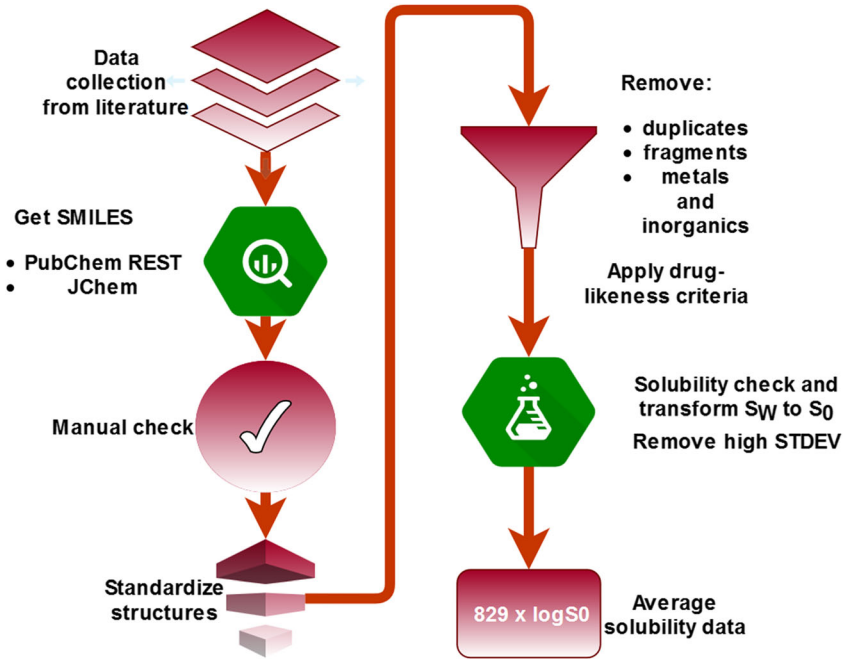

FIGURE 1 Data collection and preparation pipeline for the novel intrinsic solubility set

$$\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|, \tag{1}$$

where N is the total number of cases (compounds) in the training set and λ is the penalty term. In a linear regression model having constant term, the number of predictors (features, DPs) involved is equal to $p-1$. Because of the form of the L_1-penalty, LASSO inherently performs feature selection and shrinkage at the same time returning an extremely sparse coefficient matrix.

2.2.2 | Partial least squares

PLS regression is a chemometric method that aims to reduce the dimension of both the predictors (X-space) and the dependent variables (Y-space) by compressing them into latent variables (LVs). LVs are constructed in the direction of maximum correlation between X- and Y-spaces, where one wants to find the multidimensional direction in the X-space (predictive variables [N, p]) that explains the maximum multidimensional variance direction in the y (target variable [N, 1]). Readers are referred to Bro[60] for a more detailed overview.

2.2.3 | Random forests

The RF algorithm, conceptualized by Breiman,[63] creates a large collection of decorrelated decision trees by using bootstrapping aggregation. The final prediction results are thereby averaged from a multitude of decision tree regressors; this reduces the bias in the models, whereas variance can be controlled by carefully optimizing weak learner hyperparameters, such as tree depth. Besides their good performance, RF and other decision tree-based learners accept many feature representations and are associated with reduced preprocessing efforts, making them convenient for use in many applications, including manufacturing. Because trees in RF get trained in parallel, a significant advantage of RF is the speed when compared with boosting ensemblers.

2.2.4 | LightGBM

Light Gradient Boosting Machine (LGBM)[62] is a framework using the decision tree as a base algorithm. LGBM uses the first-order derivative information when optimizing the loss function. The leaf growth strategy with depth limitation and multithread optimization in LGBM contributes to solve the excessive memory consumption with respect to other boosting-ensemble machine learning methods. LGBM was selected to reduce the computational cost of calculations compared with other boosting ensemblers.

2.3 | Feature selection

In this work, we applied a multistage post hoc feature selection. The strategy is based on permutation importance[64] for eliminating features.[65] Using each of the trained models, the method permutes the values of individual features (one-by-one) to assess the relevance of the features with respect to the response vector ($\log S_0$). The relative decrease in RMSE in a pretrained model caused by a permuted feature is considered a "weight." The permutation procedure was repeated 10 times for the feature matrix and averaged to a permutation importance vector. A cut-off value of 0.001 for the average weight was chosen. The feature elimination procedure was conducted in multiple stages. Models were trained, and then a set of features was eliminated either by having an average weight above the cut-off or the number of features used in the next stage were reduced to one third of the number of features, whichever was smaller. The models from each stage were included in the performance evaluation.

2.4 | Hyperparameter optimization

For hyperparameter optimization in machine learning, random and grid searches over hyperparameter spaces are used very often.[66] Because hyperparameter space can be large either by means of number of parameters or grid-points included, the procedure can suffer from large computational cost even with parallel computing.[67] Local optima in the parameter space are difficult to avoid if the grid is not dense enough with properly set parameter ranges. In this work, we applied Bayesian optimization (BO)[68] for hyperparameter optimization with RMSE (Validation) as a loss function. BO aims to construct a posterior distribution of functions (Gaussian process) that best describes the loss function. As the number of observations grows, the posterior distribution becomes narrower, and the algorithm becomes more certain of which regions in the parameter space are worth exploring and which are not. In the process of parameter optimization, the model is continuously trained, and the regression results obtained by each parameter combination are evaluated. Finally, the optimal parameter combination is obtained when a stopping criterion is reached (predefined number of iterations).

2.5 | Model training

To train the models, the datasets ($\log S_0$ and the predictive sets) were split following two strategies: randomly and by means of diversity picking (a method of picking diverse molecules into subsets by means of their FP similarity).[69] For both splits, the external test set was set to 20% of the whole data set a priori (Table S2; previously published at Lovrić et al.[70]), and the remaining 80% were further split by one of the two strategies into training (80%) and validation (20%) sets. We trained the models with (i) three options for the predictive features, namely, FP, DS, and a joint data set of FPDS; (ii) two splitting options: random or by diversity picking; (iii) four ML algorithms; (iv) with and without multistage feature selection; and (v) with and without feature preprocessing. The code for the preprocessing method (available at https://github.com/mariolovric/solubility) comprises the following sequential steps: removing features with any missing values, removal of correlated features (Pearson correlation > 0.85), separation of categorical features (from binary and continuous) and their conversion to binary features (based on binning to four "dummy" bins), and removal of low variance binary features (lower than 1% variance). The parameters of the ML models were tuned using BO for each of the named combinations. The available parameter space (upper and lower bounds) per algorithm can be found in the code repository. The models were trained on the training set and validated on the validation set during BO. RMSE computed out of the validation set was used as a loss function for BO. The optimization experiment ran for ~48 h on a virtual machine with 24× Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz with 30 GB of RAM. We also followed per iteration results on the external test set, to later on report the estimated generalization performance. Apart from LASSO, which has an internal regularization of the feature space, the models were trained iteratively with the permutation importance feature selection strategy multiple times, with each time transferring the feature list to the next model sequentially. Such modeling pipeline is depicted in Figure 2.

Finally, the best models were chosen based on a ranking schema, which we believe it reflects an objective model evaluation. In Equation 2, the weights were chosen in such a manner that performance on the test is given the largest importance, followed by complexity expressed through the number of features and two terms representing generalization all combined in the average rank Rk_M. All ranks are sorted ascending.

$$Rk_M = 0.5R_{\mathrm{RMSE(test)}} + 0.3R_{\mathrm{features}} + 0.1R_{\Delta\mathrm{val}} + 0.1R_{\Delta\mathrm{train}}, \tag{2}$$

where R_{features} is the rank based on the total number of features involved in the model and $R_{\mathrm{RMSE(test)}}$ is the rank of RMSE of the respective test set, whereas Δ_{val} and Δ_{train} are defined with Equations 3 and 4, respectively. Both terms account for the generalizability of the models.

$$\Delta_{\mathrm{val}} = |\mathrm{RMSE(test)} - \mathrm{RMSE(val)}|, \tag{3}$$

$$\Delta_{\mathrm{train}} = |\mathrm{RMSE(train)} - \mathrm{RMSE(val)}|. \tag{4}$$

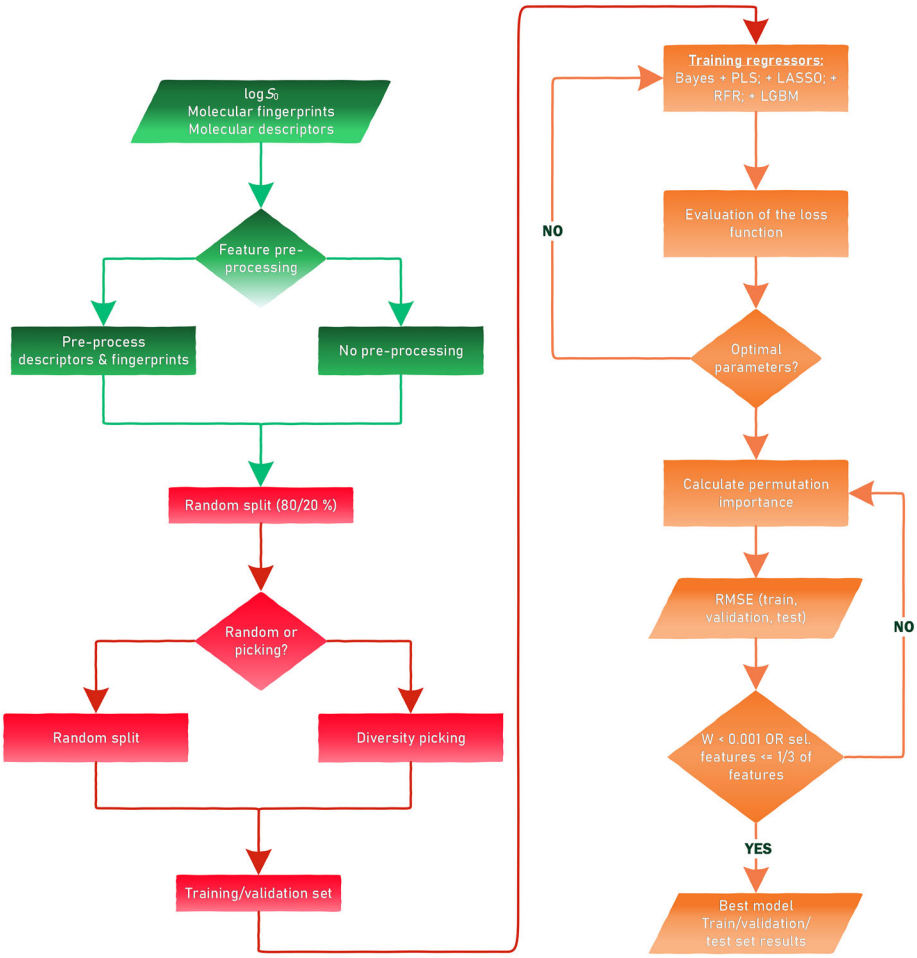

FIGURE 2 The model pipeline for the optimization experiment

3 | RESULTS AND DISCUSSION

In this work, we have compared four machine learning methods for prediction of aqueous solubility. Namely, PLS, LASSO, RFs, and LGBMs. PLS is an LV method in which predictors X are correlated to the dependent variable y by compressing both into LVs. The LVs are extracted by maximizing the variance in both X and y, as well as correlation between them. PLS is suitable for very intercorrelated data such as spectral information, has generally good generalization ability, and is able to deal with datasets with larger number of features than observations. LASSO is a method in which an L_1-penalty is introduced for regularization with inherent feature selection, which makes it robust and also able to handle high-dimensional data. However, LASSO and PLS can be quite sensitive to outliers. RFs belong to a family of ensemble (nonlinear) methods where a series of weak learners are trained and aggregated with the aim of building strongly predictive models. The fourth algorithm is LGBM, a gradient boosting algorithm in which first-derivative information is used while computing the loss function for generation of the ensemble model. It possesses similar regression features as RFs. Finally, consensus models can be built out of the best regressors to further improve predictive ability, generalization, and robustness.

3.1 | Model optimization results

Detailed results of all trained models are summarized in Table S3. Based on the RMSE(test) values, LASSO is the best performing model (RMSE(test) = 0.69) with 105 features (FPDS) involved. RMSE(train) and RMSE(val) for LASSO were 0.66 and 0.96, respectively.

This model, ranked by RMSE(test), was followed by five (second to sixth position in Table S3) RF models with some of them comprising as few as 16 features. The first PLS model appeared on the seventh place comprising 33 original features (10 latent features). The best LGBM model by means of RMSE(test) was ranked 15th comprising 47 features. Figure 3 depicts the contributions of the choice of predictors, algorithm, and the splitting method.

It can be observed that the FP-based solubility models have generally underperformed when compared with the models built out of molecular DPs or their combination. The models based on FP also exhibit a large spread in regard to RMSE(test). This outcome could have been expected because none of the four algorithms (PLS, LASSO, LGBM, and RF) creates metavariables (hidden layer abstract molecular representations) out of the FPs like deep neural networks do in the hidden layers that contribute to their predictive ability.[71] Furthermore, with the addition of FPDS, only marginal improvements can be observed. LGBM shows a notably larger spread compared with other algorithms (Figure 4), which can be explained by evident overfitting on the train set and lower predictive ability on the test set.

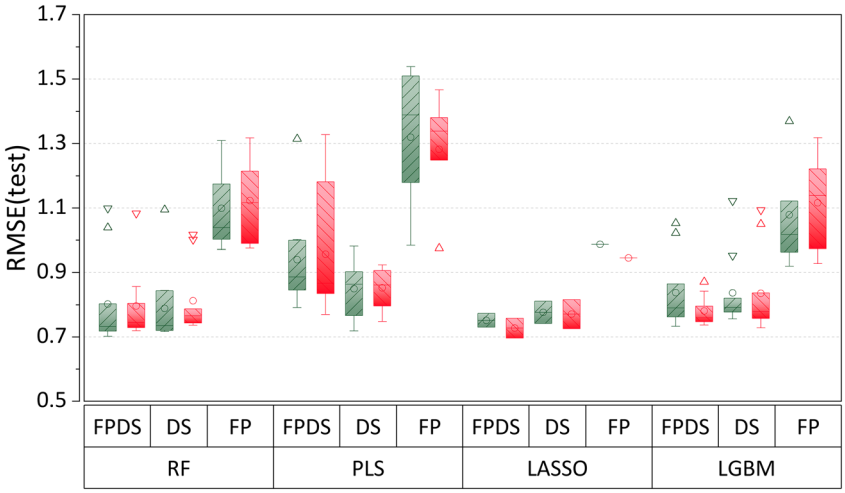

FIGURE 3 Distribution of testing set errors for the four evaluated machine learning algorithms in cases when two algorithms are used for training/test set partition. Differences between random train/test/validation split and diversity picking are depicted with green ascending and red descending line patterns, respectively. Mean values of the testing errors are depicted with green and red circles, whereas the outliers are depicted with green and red upwards-facing triangles, for random train/test/validation split and diversity picking, respectively

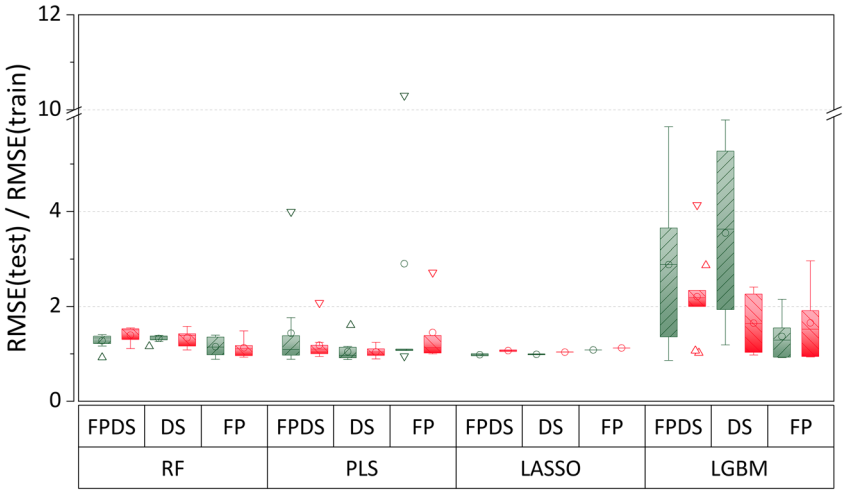

FIGURE 4 Generalization ability and robustness for all the models trained in this work. The RMSE(test) /RMSE(train) ratio depicted in this figure was grouped based on the method used (RF, PLS, LASSO, LGBM) for model development and three sets of predictive variables. Differences between random train/test/validation split and diversity picking are depicted with green ascending, and red descending line patterns, respectively. Mean values of the testing errors are depicted with green and red circles, whereas the outliers are depicted with green and red upwards-facing triangles, for random train/test/validation split and diversity picking, respectively

Such performance decrease is not caused by the optimizer being stuck in local optima, as evident from Table S4 where optimal hyperparameters of LGBM vary considerably in each run.

Even though the LGBM is a powerful algorithm, it has a large variety of hyperparameters, and finding the right set of those can appear troublesome. The spread of RF tends to be smaller than LGBM, which can be explained by the bagging + decorrelation paradigms, which can help in avoiding any local optima during BO. In our previous work, we observed boosting ensemble methods also underperforming when compared with the bagging ensemblers.[33,72] Overall, the spreads per algorithm in Figure 4 are larger for the FP and FPDS predictive sets. This might be explained by randomness that FPs can introduce by having a train or test bit with all zero values impeding convergence.

Herein, we also evaluate the contribution of the data-splitting strategies. RMSE(val) values for models with datasets split via diversity picking can be as low as 0.53 (Table S3). Nevertheless, the highest ratios of RMSE(test/val) (above 1.2) are all originating from diversity-picked data splits. Diversity-picking leads to similar train and validation set that points to an overestimation of the model quality on any external test set. Therefore, the validation or other cross-validation metrics for models with diversity-picking-based splitting can point to lower generalization/robustness. Based on Δ_{train} (Equation 4), LASSO is overall the best performer. PLS performs well in terms of both generalization metrics. RF models exhibited overfit but in a lesser extent than LGBM. Table 1 summarizes the 10 best models according to the Rk_M metric only for random splits, because we have shown that diversity-picking can deviate the impression in generalization. Even though the LASSO model has the best score by RMSE(test), it has a high number of features, which is deteriorating its Rk_M score. Because the LASSO algorithm is penalizing the coefficients, it can perform well with a high number of features if it sets the coefficients close to zero, which was the case with this model. The coefficients are in a range −0.38 to 0.29 with ∼42% coefficients being in the range from −0.01 to 0.01. A coefficient plot is given in Figure S1.

The Rk_M metric was chosen in such a manner as to create a simple model by means of the number of features and a good result on the (external) test set but still taking into account generalization/robustness (see Equation 1).

By means of Rk_M, a RF model using 17 features was ranked as best. The predictive ability of the two best models based on RMSE(test) and Rk_M is depicted in Figure 5A,B, respectively. Out of the 10 best models by Rk_M, four are RF and four are LGBM, the rest being LASSO. Interestingly, there are two LGBM models using two and three features for training. Even though not ranked as the best, they exhibited reasonable generalization. Eight out of these 10 models are not using preprocessing (part of the grid search), which shows that ensemble methods work well with the original data as preprocessing can remove valuable information. None of the best models was based on FPs. The models in Table 1 were based either on DPs or the combined with FPs. The R^2 values for the two best models are 0.80 (LASSO) and 0.78 (RF).

It is worth pointing that out of the two best models (LASSO and RFs), a consensus model was built outperforming all the evaluated models with RMSE(test) of 0.67 log points (R^2 of 0.81).

TABLE 1 Results across models sorted by the scoring method Rk_M

Algorithm	Data set	Preprocessed	RMSE(test)	# features	RMSE(train)	RMSE(val)	Rk_M score
*RF	FPDS	FALSE	0.72	17	0.47	0.94	21.6
LGBM	FPDS	FALSE	0.84	2	0.82	1.01	25.1
RF	FPDS	TRUE	0.74	8	0.57	0.98	25.2
LGBM	DS	FALSE	0.74	15	0.46	0.96	25.8
LASSO	DS	FALSE	0.73	92	0.70	0.97	26.2
RF	FPDS	FALSE	0.72	51	0.47	0.94	26.5
*LASSO	FPDS	FALSE	0.69	105	0.66	0.96	26.6
RF	DS	FALSE	0.74	19	0.53	0.96	26.7
LGBM	DS	FALSE	0.73	47	0.30	0.92	27.1
LGBM	DS	TRUE	0.84	3	0.82	1.04	28.0

Note: A lower Rk_M means better performance. An asterisk assigns the two chosen winners, one based on scoring the other on RMSE(test).
Abbreviations: DS, descriptors; FPDS, fingerprints and descriptors; LASSO, least absolute shrinkage and selection operator; LGBM, Light Gradient Boosting Machine; RF, random forests; RMSE, root mean square error.

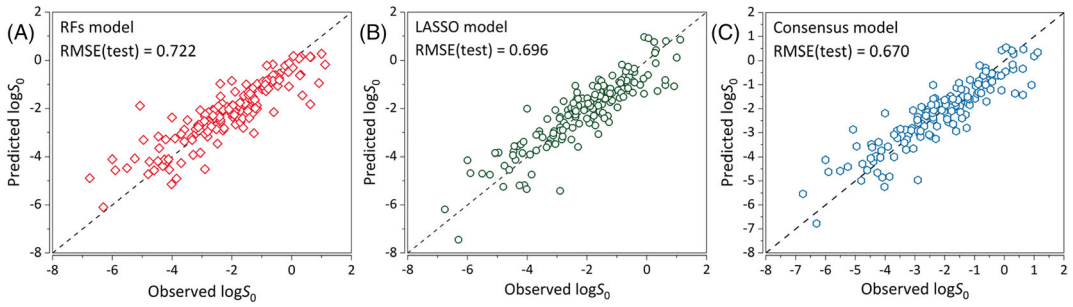

FIGURE 5 Predictive ability of the two best intrinsic solubility QSPR models from Table 1. (A) LASSO model, (B) RF model, and (C) the consensus model

TABLE 2 Pearson correlation coefficients of RMSE scores across 158 trained models (78 randomly and 80 diversity picked)

Data splitting		RMSE(test)	RMSE(train)	RMSE(val)	RMSE(train, val)
Random	RMSE(test)	1	0.85	0.87	**0.92**
	RMSE(train)	0.85	1	0.71	0.96
	RMSE(val)	0.87	0.71	1	0.88
	RMSE(train, val)	0.92	0.96	0.88	1
Diversity picking	RMSE(test)	1	0.77	0.87	0.82
	RMSE(train)	0.77	1	0.87	0.99
	RMSE(val)	0.87	0.87	1	0.94
	RMSE(train, val)	0.82	0.99	0.94	1

Note: The bolded number indicates the best performance.

3.2 | Comparison of model scores

The aim of modeling is to develop a model by which we will be able to predict a modeled activity of an external (unseen) set of molecules. For this reason, it is of utmost importance to estimate generalization based on known performance of model obtained in training and validation procedures. We have therefore compared here the RMSE values for 158 models (separately for splitting methods), that is, the scores on train, validation, and test set. Additionally, we have calculated the average of RMSE(train) and RMSE(val) as RMSE(train, val). The results are shown in Table 2.

The comparison of results for randomly split data shows a correlation of 0.85 for train test and 0.87 for val test. The same comparison of models where train and val were diversity picked shows a somehow lower correlation of 0.77 for RMSE(train) – RMSE(val) with RMSE(val) – RMSE(test) being the same at 0.87. The reader is reminded here that the test set is a true external set that was split a priori. Only train – val splits were tested by the splitting strategies. Generally, a better prediction of intrinsic solubility for external set of compounds can be achieved if the model is validated (during model optimization and development) by means of cross-validation or a validation set in which the training set was split randomly into validation subsets. We propose that the diversity picking, another splitting algorithm applied in this study, can lead to overly optimistic results. A correlation of 0.87 in RMSE(train) – RMSE(val) within the diversity picking split supports this further, compared with the correlation of RMSE(train) – RMSE(val) in random split, which is at 0.71 (a lower correlation) meaning the distribution of the train and validation sets differs slightly. It is shown here that a drift in the distributions of train and validation sets can lead to a better generalization (on the true test set). Even more interesting is that in random splitting, the average of the train and validation by means of RMSE(train, val) delivers a good overview of the generalization of the model because they show a correlation as high as 0.92 to RMSE(test). Therefore, we suggest the use of RMSE(train, val) after they were randomly split for choosing models with good generalization on external unseen data.

3.3 | Feature importance

Careful analysis of the involved features for all the models in this study showed some interesting patterns (Table S3). First, the PLS models in general did not reduce to as few features during the feature selection as RF or LGBM. Second, LASSO mostly converged to subsets of 50–100 features. The multistage feature selection was not used in the case of LASSO as feature selection is inherent to this technique. Third, RF models have overall exhibited a reasonable model quality with a smaller number of features. This points to a fact that RF seems more efficient in removing features due to its bagging and decorrelation paradigms. The best model by means of Rk_M was refitted with the resulting features and the resulting parameters. The retrained model was subjected to permutation importance, the results of which are depicted in Figure 6.

Table 3 summarizes the descriptions of DPs involved in the best final RF model (Table 1), selected using the permutation importance strategy.

Detailed descriptions of all the DPs can be found in Todeschini and Consonni.[73] The analysis of the permutation importance of the DPs in Figure 6 shows that the best RF model is most sensitive to the order of values of the SCBO

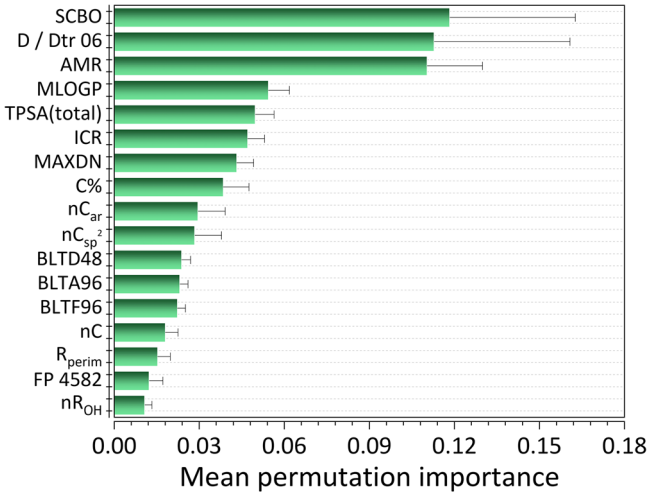

FIGURE 6 Mean permutation importance for 1000 random resampling runs of the best model with 17 features (RF model from Table 1)

TABLE 3 Full names of descriptors selected into the final/best RF model from Table 1

Descriptor	Description
SCBO	Sum of conventional bond orders (H-depleted)
D/Dtr 06	Distance/detour ring index of order 6
AMR	Ghose–Crippen molar refractivity
MLOGP	Moriguchi octanol-water partition coefficient
TPSA (tot)	Topological polar surface area using N, O, S, P polar contributions
ICR	Radial centric information index
MAXDN	Maximal electrotopological negative variation
C%	Percentage of C atoms
nC_{ar}	Number of aromatic carbons
nC_{sp2}	Number of sp^2 hybridized Carbon atoms
BLTD48	Verhaar Daphnia base-line toxicity from MLOGP (mmol/L)
BLTA96	Verhaar algae base-line toxicity from MLOGP (mmol/L)
BLTF96	Verhaar fish base-line toxicity from MLOGP (mmol/L)
nC	Number of carbon atoms
R_{perim}	Ring perimeter
FP 4582	Fingerprint 4582
nR_{OH}	Number of hydroxyl groups

descriptor. The largest decrease of RMSE of the best RF model from Table 1 is caused by the permutation of values of the SCBO descriptor. This result could be an evidence that SCBO descriptor interacts the most with other 16 DPs involved in the RF model. Thus, by permutation of values of the SCBO descriptor within the RF model, a large number of model coefficients become suboptimal, and the quality of the model decreases the most. Additionally, results of permutation importance analysis surely depend on the distribution of values of DPs. If the descriptor has more diverse values, then the permutation can result in significantly different sequences of its values, causing the largest drop in model quality measured by RMSE. On the other hand, if the descriptor has monotonic values, the total number of different possible sequences of descriptor values will be smaller, as well as the change of RMSE of the actual model compared with RMSE of the model containing permuted values of one descriptor.

3.4 | Physical interpretation of the relation between molecular DPs and aqueous solubility

The above analysis clearly established certain quantitative structure–aqueous solubility relationships for drug compounds. However, the question is: What is the physical interpretation of the correlating between the molecular (structural) parameters and the aqueous solubility? Here, we attempt to provide the physical interpretation of the top five important molecular DPs (Figure 6).

1. SCBO: The sum of conventional bond orders in a molecule is related to the size (or molecular weight) of a compound, as well as to the total number of hydrogens in it. In general, larger (organic) molecules are less soluble in aqueous medium because it is more difficult for water solvent molecules to surround the larger molecules. Therefore, this strong correlation is expected.
2. D/Dtr 06: This descriptor describes the cyclic character of the evaluated molecules in terms of topological patterns that allow one to compare the cyclic complexity of structures, namely, the number of molecule cycles and the manner in which the cycles are connected. As the cyclic character also relates to the size of a solute, negative correlation with aqueous solubility is also anticipated.
3. AMR: Molecular refraction, a measure of the total polarizability, is often used as a solubility parameter, for example, *Abraham* solvation parameter model. Good correlations between solubility parameters and refractive indices have been reported. Hence, AMR is believed to be a good molecular descriptor of aqueous solubility.
4. MLOGP: The octanol-water partition coefficient (log*P*) is a ratio of the solubilities of a solute in a two-phase octanol/water system, which is an important index in measuring solubility. This is an obvious parameter correlates well with aqueous solubility of drug molecule.
5. TPSA (total): The polar surface area (surface sum over all polar atoms) represents potential area of a molecule that interacts with water molecule as a solvent. A large total polar surface area of a solute indicates stronger solvation in an aqueous medium. Thus, it is an important molecular descriptor to quantify the solute–solvent interaction of a drug molecule in aqueous environment.

Our result here confirms the valuable roles of constitutional, topological, geometrical, and electronic DPs to predict the aqueous solubility. Some of the selected DPs were utilized in many solubility prediction studies, for example, MLOGP (log*P*) as the most frequent by appearance in the literature,[16,17,74–77] as well as other top DPs from Table 3 like the total number of carbon atoms (*n*C),[78] TPSA,[32,76,77] SCBO,[75] AMR (MR),[17,32] and the number of aromatic atoms (here nC_{ar}).[32]

3.5 | Evaluation of the models' extrapolation capabilities on a more challenging test set

In order to test the extrapolation capabilities of the models, we have introduced a more challenging test set, that is, an extreme-case scenario. For this purpose, we have principal component analysis (PCA)-transformed FP data to LVs (principal components). The three components explain only ∼28.35% of total variance. This however does not affect the research issue at hand, which is the creation of a different train–test split which should reveal extrapolation capabilities of the winning models. Prior to PCA, low-variance FPs were removed (below 0.05). The centroid of the PCA space (three dimensions) was calculated as well as the Euclidean distances of all compounds to the centroid. The Euclidean

distances to the centroid were sort and split at 80 percentiles of distance. All compounds below 80 percentiles were set as a new train set (PCA-train) and above as a new test set (PCA-test). The validation set is subtracted, so the number of compounds corresponds to the other splits (529 in PCA-train, 167 in PCA-test). The PCA-split space is depicted in Figure 7. The splits were subjected to the two winning models presented in Figure 5 (LASSO and RF); that is, the same hyperparameters and features were utilized, but the model was retrained on the PCA-train set and evaluated on the PCA-test set.

The LASSO model (Figure 5A) had here an RMSE of 1.31 log-points on PCA-test, which is a large increase of error compared with the random split that results in a RMSE of 0.69. The RF model scores an RMSE of 0.89 on PCA-test compared with 0.72 obtained on the randomly split test set (Figure 5B). It is interesting that the RF model performs better in such an extreme-case scenario. The LASSO model was chosen in the first place based on its test set performance, while the RF model was chosen based on Rk_M, which is including also performance on train and validation sets and therefore present a better tool for estimating generalization. This confirms the appropriateness of using quality estimation by means of Rk_M but also the importance of challenging the models with extreme-case scenarios such as this. It could be expected that similar descriptors utilized in models (which are presented in Figure 6) and a worse RMSE(test) of RF comparing to LASSO (see Table 1) would deteriorate the extrapolation capability, which was not the case since RF performed better in this more challenging task. This supports our discussion that the ensemblers can stabilize the models in case of descriptor redundancy.

3.6 | Limitations of the machine learning approaches for prediction of solubility

This study was designed as a multifactor evaluation for training machine learning models for the prediction of solubility. Some conventions like removal of collinear features were varied as a segment of a grid search to evaluate whether that might have an influence on model performance. The top models summarized in Table 1 have shown that eight of 10 models were those run without extensive preprocessing. Interestingly, the models with redundant features included fared better and had better predictive performance than the models that involved extensive preprocessing. This points to the fact that some machine learning models do profit from redundancy, at least those with intrinsic feature prioritization such as ensemble learners.

The best RF model in our evaluations has shown a slight bias on the test set. One potential cause of such bias can be attributed to the chosen metric for evaluation (RMSE—on the testing set in our case) because the research community still did not fully agree on the model quality metrics to be used.[31,79] This also makes comparison of research works and models published in literature challenging. Some biases can be avoided by using other or

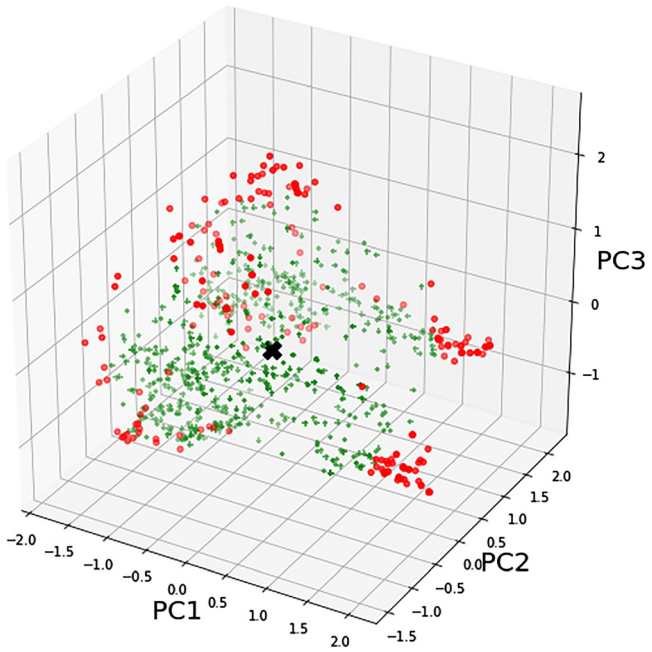

FIGURE 7 PCA scores plot for the intrinsic solubility data. Molecular fingerprints are transformed onto three axes (PC1, PC2, PC3). Black X marker is the centroid of the space. Data points are colored by means of Euclidean distance from the centroid. Molecules that are in the 80 percentiles closest to the centroid are colored in green, whereas those further apart are colored in red

weighted metrics. In this work, we limited ourselves to the RMSE, to avoid ambiguity in the decision-making process. Furthermore, bias can also be introduced by the experimental data itself. Literature suggests that the standard deviation in solubility laboratory measurements increases with decreasing intrinsic solubility.[17] Even though we limited ourselves to data where measurements are well described, there is a lack of coherence within data sources, which is described previously in literature.[17,56] Nevertheless, the RF model showed better extrapolation in the extreme-case scenario on the PCA-test set, which supports the use of proposed ranking methods and the stability of model regardless of the redundant features.

Even though we suggest two winners, the LASSO model by RMSE(test) and the RF model by Rk_M, both are arbitrarily chosen criteria because (a) our ranking approach is a heuristic reasoned by weighting and (b) RMSE is chosen due to the its higher robustness comparing to the correlation coefficient,[17] but there are also other model quality metrics that can be used. It is important to note that the winning models have marginal improvements over follow-up models in the ranks. Presented results appear to converge to the values of RMSE(test) ~0.7, which may suggest low structure-based information content involved in the calculated molecular features involved or certain limitations of quantitative structure–activity relationship (QSAR) predictive approaches that were used in this study. Even though the two predictive challenges[4,80] were 10 years apart and the second had an improved data quality, but very poor (or no) improvement has been achieved by means of the use of advanced machine learning models. Our approach with the ensemble models led us to be among the top performers (MLKC team) in the 2019 solubility challenge.[80] However, we acknowledge that limitations were reached for predictive capabilities QSAR models and the most popular chemical representations such as molecular DPs and FPs, which are also utilized in this work. Furthermore, we have curated our own data set to increase the size of the data, which can lead to error propagation because not all data sources have the same reliability.

4 | CONCLUSIONS

In this work, we tested the effects of multiple factors affecting machine learning outcomes in order to obtain the best prediction for intrinsic aqueous solubility. Besides the four regressors, namely, LASSO, RF, LightGBM, and PLS, we tested the effects of feature selection by means of permutation importance, the type and size of chemical representation (FP and molecular DPs), Bayesian optimization, and two data splitting options. The intrinsic solubility data used here is a novel collection of curated values and structures obtained from literature with 829 drug-like compounds. The best model by means of predictive performance on external test set is a LASSO regressor based on 105 features giving a RMSE of 0.7 (log units) in prediction on an external test set of organic compounds. Nevertheless, we proposed a ranking schema for choosing the best models based not solely on the measure's performance on a fixed test set but also by taking into account the number of features and the estimated generalization performance estimated on the training and validation sets. The rankings reveal a clear dominance of the RF algorithm because it can predict well with less features involved but has also a better performance on the more challenging PCA-split test set. Even though LightGBM is a powerful algorithm, it has a complex hyperparameter space, which is hard to optimize and was working in the overfitting regime in most cases. We show that there is no single criterion, data set, nor algorithm that can cover it all but rather a multiverse of possibilities and decisions to be embraced for building robust models with strong generalizability.

ACKNOWLEDGMENT

Bono Lučić is partially supported by the Croatian Government and the European Union through the Programme KK.01.1.1.01—The Scientific Centre of Excellence for Marine Bioprospecting—BioProCro.

ORCID

Mario Lovrić https://orcid.org/0000-0002-3541-9624
Kristina Pavlović https://orcid.org/0000-0001-6069-7020
Petar Žuvela https://orcid.org/0000-0001-6481-2241
Adrian Spataru https://orcid.org/0000-0002-6195-2385
Roman Kern https://orcid.org/0000-0003-0202-6100
Ming Wah Wong https://orcid.org/0000-0003-2162-1220

REFERENCES

1. Hörter D, Dressman JB. Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. *Adv Drug Deliv Rev*. 2001;46(1-3):75-87. https://doi.org/10.1016/S0169-409X(00)00130-7
2. Kalepu S, Nekkanti V. Insoluble drug delivery strategies: review of recent advances and business prospects. *Acta Pharm Sin B*. 2015;5(5):442-453. https://doi.org/10.1016/j.apsb.2015.07.003
3. Llinas A, Avdeef A. Solubility challenge revisited after ten years, with multilab shake-flask data, using tight (SD ∼0.17 log) and loose (SD ∼0.62 log) test sets. *J Chem Inf Model*. 2019;59(6):3036-3040. https://doi.org/10.1021/acs.jcim.9b00345
4. 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):1-5. https://doi.org/10.1021/ci800436c
5. Palmer DS, Mitchell JBO. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? *Mol Pharm*. 2014;11(8):2962-2972. https://doi.org/10.1021/mp500103r
6. Montanari F, Kuhnke L, Ter Laak A, Clevert DA. Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. *Molecules*. 2020;25(1):1-13. https://doi.org/10.3390/molecules25010044
7. Cao DS, Xu QS, Liang YZ, Chen X, Li HD. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J Chemometr*. 2010;24(9):584-595. https://doi.org/10.1002/cem.1321
8. Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA. New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorganic Med Chem*. 2008;16(17):7944-7955. https://doi.org/10.1016/j.bmc.2008.07.067
9. Louis B, Agrawal VK, Khadikar PV. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. *Eur J Med Chem*. 2010;45(9):4018-4025. https://doi.org/10.1016/j.ejmech.2010.05.059
10. Schäfer RB, Pettigrove V, Rose G, et al. Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. *Environ Sci Technol*. 2011;45(4):1665-1672. https://doi.org/10.1021/es103227q
11. 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(3):571-581. https://doi.org/10.1021/ci800406y
12. Palmer DS, Llinàs A, Morao I, et al. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol Pharm*. 2008;5(266-279):545-556. https://doi.org/10.1021/mp7000878
13. Chen XQ, Cho SJ, Li Y, Venkatesh S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J Pharm Sci*. 2002;91(8):1838-1852. https://doi.org/10.1002/jps.10178
14. Bergström CAS, Luthman K, Artursson P. Accuracy of calculated pH-dependent aqueous drug solubility. *Eur J Pharm Sci*. 2004;22(5):387-398. https://doi.org/10.1016/j.ejps.2004.04.006
15. 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(4):1477-1488. https://doi.org/10.1021/ci049909h
16. Delaney JS. ESOL: Estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci*. 2004;44(3):1000-1005. https://doi.org/10.1021/ci034243x
17. Avdeef A. Prediction of aqueous intrinsic solubility of druglike molecules using random forest regression trained with Wiki-pS0 database. *Admet Dmpk*. 2020;8(1):29-77. https://doi.org/10.5599/admet.766
18. Bergström CAS, Norinder U, Luthman K, Artursson P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm Res*. 2002;19(2):182-188. https://doi.org/10.1023/A:1014224900524
19. Engkvist O, Wrede P. High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. *J Chem Inf Comput Sci*. 2002;42(5):1247-1249. https://doi.org/10.1021/ci0202685
20. McElroy NR, Jurs PC. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J Chem Inf Comput Sci*. 2001;41(3-6):1237-1247. https://doi.org/10.1021/ci010035y
21. Huuskonen J, Salo M, Taskinen J. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J Pharm Sci*. 86(4):450-454. https://doi.org/10.1021/js960358m
22. Mitchell BE, Jurs PC. Prediction of aqueous solubility of organic compounds from molecular structure. *J Chem Inf Comput Sci*. 1998;38(3):489-496. https://doi.org/10.1021/ci970117f
23. Sutter JM, Jurs PC. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure-property relationship. *J Chem Inf Comput Sci*. 1996;36(1):100-107. https://doi.org/10.1021/ci9501507
24. Tang B, Kramer ST, Fang M, Qiu Y, Wu Z, Xu D. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J Chem*. 2020;12(1):1-9. https://doi.org/10.1186/s13321-020-0414-z
25. Deng T, Jia GZ. Prediction of aqueous solubility of compounds based on neural network. *Mol Phys*. 2020;118(2):1-8. https://doi.org/10.1080/00268976.2019.1600754
26. Boobier S, Osbourn A, Mitchell JBO. Can human experts predict solubility better than computers? *J Chem*. 2017;9(1):1-14. https://doi.org/10.1186/s13321-017-0250-y
27. Zang Q, Mansouri K, Williams AJ, et al. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J Chem Inf Model*. 2017;57(1):36-49. https://doi.org/10.1021/acs.jcim.6b00625
28. McDonagh JL, Nath N, De Ferrari L, Van Mourik T, Mitchell JBO. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *J Chem Inf Model*. 2014;54(3):844-856. https://doi.org/10.1021/ci4005805
29. 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(7):1563-1575. https://doi.org/10.1021/ci400187y

30. Salahinejad M, Le TC, Winkler DA. Aqueous solubility prediction: do crystal lattice interactions help? *Mol Pharm*. 2013;10(7):2757-2766. https://doi.org/10.1021/mp4001958
31. Lučić B, Batista J, Bojović V, et al. Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges. *Croat Chem Acta*. 2019;92(3):379-391. https://doi.org/10.5562/cca3551
32. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO. Random forest models to predict aqueous solubility. *J Chem Inf Model*. 2007;47(1):150-158. https://doi.org/10.1021/ci060164k
33. Šimić I, Lovrić M, Godec R, Kröll M, Bešlić I. Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon. *Environ Pollut*. 2020;263:1-9. https://doi.org/10.1016/j.envpol.2020.114587
34. 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(7):1289-1303. https://doi.org/10.1021/ci800058v
35. Baek K, Jeon SB, Kim BK, Kang NS. Method validation for equilibrium solubility and determination of temperature effect on the ionization constant and intrinsic solubility of drugs. *J Pharm Sci Emerg Drugs*. 2018;06(01):1-6. https://doi.org/10.4172/2380-9477.1000125
36. Stuart M, Box KJ. Chasing equilibrium: measuring the intrinsic solubility of weak acids and bases. *Anal Chem*. 2005;77(4):983-990. https://doi.org/10.1021/ac048767n
37. Bergström CAS, Strafford M, Lazorova L, Avdeef A, Luthman K, Artursson P. Absorption classification of oral drugs based on molecular surface properties. *J Med Chem*. 2003;46(4):558-570. https://doi.org/10.1021/jm020986i
38. McFarland JW, Avdeef A, Berger CM, Raevsky OA. Estimating the water solubilities of crystalline compounds from their chemical structures alone. *J Chem Inf Comput Sci*. 2001;41(3-6):1355-1359. https://doi.org/10.1021/ci0102822
39. Avdeef A. Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. *Admet Dmpk*. 2019;7(3):210-219. https://doi.org/10.5599/admet.698
40. Box KJ, Comer J. Using measured pKa, LogP and solubility to investigate supersaturation and predict BCS class. *Curr Drug Metab*. 2008;9(9):869-878. https://doi.org/10.2174/138920008786485155
41. Etherson K, Halbert G, Elliott M. Determination of excipient based solubility increases using the CheqSol method. *Int J Pharm*. 2014;465(1-2):202-209. https://doi.org/10.1016/j.ijpharm.2014.02.007
42. Fornells E, Fuguet E, Mañé M, et al. Effect of vinylpyrrolidone polymers on the solubility and supersaturation of drugs; a study using the Cheqsol method. *Eur J Pharm Sci*. 2018;117(February):227-235. https://doi.org/10.1016/j.ejps.2018.02.025
43. Llinàs A, Burley JC, Box KJ, Glen RC, Goodman JM. Diclofenac solubility: independent determination of the intrinsic solubility of three crystal forms. *J Med Chem*. 2007;50(5):979-983. https://doi.org/10.1021/jm0612970
44. Schönherr D, Wollatz U, Haznar-Garbacz D, et al. Characterisation of selected active agents regarding pKa values, solubility concentrations and pH profiles by SiriusT3. *Eur J Pharm Biopharm*. 2015;92:155-170. https://doi.org/10.1016/j.ejpb.2015.02.028
45. Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci*. 2000;40(3):773-777. https://doi.org/10.1021/ci9901338
46. Abraham MH, Le J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J Pharm Sci*. 1999;88(9):868-880. https://doi.org/10.1021/js9901007
47. Shareef A, Angove MJ, Wells JD, Johnson BB. Aqueous solubilities of estrone, 17β-estradiol, 17α-ethynylestradiol, and bisphenol A. *J Chem Eng Data*. 2006;51(3):879-881. https://doi.org/10.1021/je050318c
48. Narasimham L, Barhate VD. Kinetic and intrinsic solubility determination of some b-blockers and antidiabetics by potentiometry. *J Pharm Res*. 2011;4(2):532-536.
49. Avdeef A, Berger CM. pH-metric solubility: 3. Dissolution titration template method for solubility determination. *Eur J Pharm Sci*. 2001;14(4):281-291. https://doi.org/10.1016/S0928-0987(01)00190-7
50. Rytting E, Lentz KA, Chen XQ, Qian F, Venkatesh S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS j*. 2005;7(1):E78-E105. https://doi.org/10.1208/aapsj070110
51. Sköld C, Winiwarter S, Wernevik J, et al. Presentation of a structurally diverse and commercially available drug data set for correlation and benchmarking studies. *J Med Chem*. 2006;49(23):6660-6671. https://doi.org/10.1021/jm0506219
52. Wassvik CM, Holmén AG, Bergström CAS, Zamora I, Artursson P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur J Pharm Sci*. 2006;29(3-4):294-305. https://doi.org/10.1016/j.ejps.2006.05.013
53. Mansouri K, Kleinstreuer N, Abdelaziz AM, et al. CoMPARA: collaborative modeling project for androgen receptor activity. *Environ Health Perspect*. 2020;128(2):1-17. https://doi.org/10.1289/EHP5580
54. Landrum G. RDKit: open-source cheminformatics. 2006.
55. Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci*. 1999;39(5):868-873. https://doi.org/10.1021/ci990307l
56. Avdeef A. *Absorption and Drug Development*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2012. https://doi.org/10.1002/9781118286067
57. Lovrić M, Molero JM, Kern R. PySpark and RDKit: moving towards big data in cheminformatics. *Mol Inform*. 2019;38(6):1-5. https://doi.org/10.1002/minf.201800082
58. Landrum G. RDKit: Colliding Bits III. http://rdkit.blogspot.com/2016/02/colliding-bits-iii.html. Accessed December 23, 2019.
59. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B*. 1996;58(1):267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

60. Bro R. Multiway calibration. Multilinear PLS. *J Chemometr*. 1996;10(1):47-61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1%3C47::AID-CEM400%3E3.0.CO;2-C
61. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intel Lab Syst*. 2001;58:109-130. https://doi.org/10.1016/S0169-7439(01)00155-1
62. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3147-3155. https://doi.org/10.5555/3294996.3295074
63. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324
64. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340-1347. https://doi.org/10.1093/bioinformatics/btq134
65. Han H, Guo X, Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*. Vol 0. IEEE Computer Society; 2016:219-224. https://doi.org/10.1109/ICSESS.2016.7883053
66. Lerman PM. Fitting segmented regression models by grid search. *Appl Stat*. 1980;29(1):77-84. https://doi.org/10.2307/2346413
67. Pontes FJ, Amorim GF, Balestrassi PP, Paiva AP, Ferreira JR. Design of experiments and focused grid search for neural network parameter optimization. *Neurocomputing*. 2016;186:22-34. https://doi.org/10.1016/j.neucom.2015.12.061
68. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *Nips 2012*. 2012;4:2951-2959. https://doi.org/10.5555/2999325.2999464
69. Dudgeon T. RDKit: revisting the MaxMinPicker. http://rdkit.blogspot.com/2017/11/revisting-maxminpicker.html. Published 2017. Accessed December 23, 2019.
70. Lovrić M, Pavlović K, Kern R, et al. 829 drug-like molecules intrinsic solubility dataset. July 2020. https://doi.org/10.5281/ZENODO.3968754
71. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci*. 2016;3(FEB):1-15. https://doi.org/10.3389/fenvs.2015.00080
72. Žuvela P, Lovric M, Yousefian-Jazi A, Liu JJ. Ensemble learning approaches to data imbalance and competing objectives in design of an industrial machine vision system. *Ind Eng Chem Res*. 2020;59(10):4636-4645. https://doi.org/10.1021/acs.iecr.9b05766
73. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. WileyVCH, Weinheim. Vol 11. Mannhold R, Kubinyi H, Timmerman H (Eds.). New York: Wiley; 2000. https://doi.org/10.1002/9783527613106
74. Gozalbes R, Doucet JP, Derouin F. Application of topological descriptions in QSAR and drug design: history and new trends. *Curr Drug Targets - Infect Disord*. 2002;2(1):93-102. https://doi.org/10.2174/1568005024605909
75. Hemmateenejad B, Baumann K. Screening for linearly and nonlinearly related variables in predictive cheminformatic models. *J Chemometr*. 2018;32(4):1-14. https://doi.org/10.1002/cem.3009
76. Raevsky OA, Polianczyk DE, Grigorev VY, Raevskaja OE, Dearden JC. In silico prediction of aqueous solubility: a comparative study of local and global predictive models. *Mol Inform*. 2015;34(6-7):417-430. https://doi.org/10.1002/minf.201400144
77. Lovrić M. Molekulsko modeliranje odnosa strukturnih svojstava i aktivnosti molekula s pomoću programskog jezika Python (prvi dio). *Kem U Ind*. 2018;67(9-10):409-419. https://doi.org/10.15255/KUI.2017.052
78. Gozalbes R, Pineda-Lucena A. QSAR-based solubility model for drug-like compounds. *Bioorganic Med Chem*. 2010;18(19):7078-7084. https://doi.org/10.1016/j.bmc.2010.08.003
79. Avdeef A. Do you know your r2? *Admet Dmpk*. 2020;1:69-74. https://doi.org/10.5599/admet.888
80. Llinas A, Oprisiu I, Avdeef A. Findings of the second challenge to predict aqueous solubility. *J Chem Inf Model*. 2020;60(10):4791-4803. https://doi.org/10.1021/acs.jcim.0c00701

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lovrić M, Pavlović K, Žuvela P, et al. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *Journal of Chemometrics*. 2021;e3349. https://doi.org/10.1002/cem.3349