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

Mario Lovrić [a, b, c, &, *], Kristina Pavlović [a, &], Petar Žuvela [d, &], Adrian Spataru [a], Bono Lučić [b], Roman Kern [a, c], Ming Wah Wong [d]

[a]Know-Center, Inffeldgasse 13/6, AT-8010 Graz
[b]NMR Centre, Ruđer Bošković Institute, Bijenička BB, HR-10000 Zagreb
[c]Institute of Interactive Systems and Data, TU Graz, Inffeldgasse 16c, AT-8010 Graz
[d]Department of Chemistry, National University of Singapore, 3 Science Drive 3, SG-117543
[&]these authors contributed equally

Keywords: intrinsic solubility, QSPR, LightGBM, random forests, PLS, LASSO, permutation importance

*Corresponding author: lovric@tugraz.at

Abstract

Here, we present a collection of publicly available intrinsic aqueous solubility data of 829 drug-like compounds. Four different machine learning algorithms (random forest, light GBM, partial least squares and LASSO) coupled with multi-stage permutation importance for feature selection and Bayesian hyper-parameter optimization were employed for the prediction of solubility based on chemical structural information. Our results have shown that LASSO yielded the best predictive ability on an external test set with and RMSE(test) of 0.70 *log* points and 105 features in the model. Taking into account the number of descriptors as well, an RF model achieved the best balance between complexity and predictive ability with an RMSE(test) of 0.72 with only 17 features. 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 involved and test set performance.

1. Introduction

Solubility is a critical topic in pharmaceutical development as it can be a limiting factor to drug absorption.[1] High attrition rate problem in drug development has been attributed to poor water solubility.[2] Predictive models - so-called 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 these challenging problematics. The first solubility challenge published by the same authors[4] demonstrated clear room for improvement in predicting solubility from (molecular) structural information. Palmer et al [5] concluded that there is still room for improvement with respect to predictive capabilities of QSPR rather quality experimental 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 afore-mentioned 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 (RMSE not published). Such a data size for solubility is rare amongst 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 authors also dealt with the solubility prediction problematics,[5, 7, 16–25, 8, 26–32, 9–15] predicting both logS_{w} and logS_0. A comparison with the previous work is difficult since authors often describe the results in different manners (train, test, cross-validation, out-of-fold) and a multitude of model metrics[33]. Specifically, for the intrinsic solubility, on external test sets, literature values of the predictive performance expressed through RMSE appear to vary between 0.7 and 1.05 log points[7, 11, 13, 23, 25, 27, 34] using a plethora of machine learning algorithms and datasets. The most recent study from Avdeef [7] with the largest curated database known (6355 logS_0 entries) employed the Random forest algorithm showing RMSE(test) in a range of 0.75-1.05 and with an R^2 value between 0.66-0.83 across several models. These results outperform studies with the aforementioned proprietary databases which signals the importance of careful data curation and chemical space consideration which Avdeef proposes. Within the aforementioned challenges, additional high-quality solubility data is published. Furthermore, powerful, and more efficient machine learning methods, as well as computing power, available in HPC environments are enabling more precise and faster learning.

Our goal in this work was to conduct a large-scale machine learning study to reveal 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 employed boosting and bagging ensemblers as well as PLS and LASSO. The last two being established machine learning modes which are often neglected over seemingly more powerful ensemble regressors.[35] Finally, we discussed the use of permutation importance for a multi-stage feature selection, the relevancy of commonly used feature preprocessing/preselection and data splitting paradigms.

2. Materials and methods

2.1. Data collection and processing

We have collected aqueous solubility data from the following literature sources.[4, 22, 40–49, 25, 50–54, 26, 27, 31, 36–39] The decision criteria on which literature to include for our study is initially based on the recommendations in the revisited solubility challenge.[3] Consecutively, we looked for additional literature sources where authors have included *pH*, temperature and inert gases in their measurements. Most of these sources refer to the intrinsic aqueous solubility (logS_0), while some of them refer to the aqueous solubility (logS_w). Most of the values were determined at 25 °C. 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 [55] and standardized to isomeric SMILES using the ChemAxon Standardizer (v18.28.0, ChemAxon, Budapest, Hungary) and the RDKit library[56]. We filtered compounds with the following properties: log$P$ 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] Some logS_{w} values in the extracted data were converted to logS_0 based on their formal charges as suggested in Ref. [52]. Since we had multiple values for intrinsic solubility per molecule, we removed values which were duplicated and averaged the rest. Our final dataset consists of intrinsic solubility values for 829 compounds available for download at http://doi.org/10.5281/zenodo.3968754 .The data preparation pipeline is depicted in **Figure 1**.

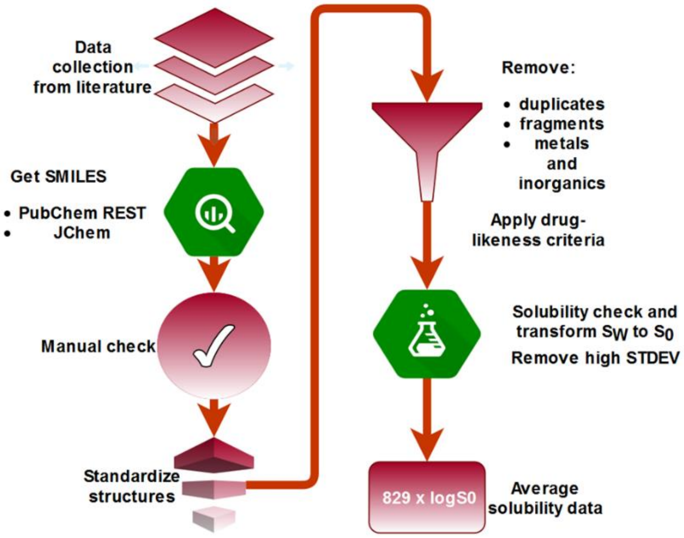

Figure 1. Data collection and preparation pipeline for the novel intrinsic solubility set.

We employed two types of predictive features: fingerprints (FPs)[57] and molecular descriptors (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 in Ref.[58]. From the available ~5000 DRAGON molecular descriptors, only a few groups of descriptors were selected based on chemical intuition. Specifically, constitutional, ring, topological descriptors, functional group counts and molecular properties. All descriptors with "NaN" values were removed. Such a pre-selection procedure yielded a total of 317 molecular descriptors. A combination of FPs and descriptors (FPDS) was also evaluated (5444 features in total).

2.2. Evaluated machine learning methods

In this work we employed four different regression algorithms, different in their paradigms: (i) Least absolute shrinkage and selection operator (LASSO), [59] (ii) Partial least squares (PLS), [60] (iii) Random forest, [61] and (iv) Light GBM [62]. All four are briefly summarized in the subsequent sub-sections.

2.2.1. LASSO

LASSO regression is a multivariate chemometric method, which employs the L1-penalty for regularization.[59] Given the multiple linear regression formulation with standardized predictors **X** and response values **y**, LASSO aims to solve the L1-penalized regression problem of finding β = {β_j}to minimize:

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

Because of the form of the L1-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 which 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) that explains the maximum multidimensional variance direction in the Y (target variable). Readers are referred to Ref. [60] for a more detailed overview.

2.2.3. Random forest

The Random forest (RF) algorithm, conceptualized by Breiman [63] creates a large collection of de-correlated decision trees by employing bootstrapping aggregation. The final prediction results are thereby averaged from a multitude of decision tree regressors, this reduces the bias in the models, while variance can be controlled by carefully optimizing weak learner hyperparameters, such as tree depth. Besides their good performance, Random forests 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. Being that trees in RF get trained in parallel, a significant advantage of RF is speed when comparing to 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 multi-thread 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 comparing to other boosting ensemblers.

2.3. Feature selection

In this work, we employed a multi-stage feature elimination. 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 to assess the relevance of the features with respect to the response vector (logS_0). The relative decrease in RMSE in a pre-trained model caused by a permuted feature is considered a "weight". The permutation procedure is repeated 10 times and averaged to a permutation importance vector. A cut-off value of 0.001 is chosen. The whole procedure is repeated in each stage during modeling. The features which were kept for the next stage with the same algorithm and dataset had either a permutation importance above the cut-off or the number of features used in the next stage were cut to one third of the number employed in the previous stage. The models from each stage are included in the performance evaluation.

2.4. Hyper-Parameter Optimization

Random and grid search commonly used for the hyper-parameter optimization suffer from a considerable computational cost even with parallel computing [66]. 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 employed Bayesian optimization (BO) [67] for hyper-parameter 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 improves, 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 ($\log S_0$ & the predictive sets) were split following two strategies: randomly and by means of diversity picking [68]. For both splits, the external test set was set to 20% of the whole data set, the remaining 80% are split further into a train (80%) and validation set (20%). We trained the models with three options for the predictive features, namely fingerprints (FP), descriptors (DS) and a joint data set of fingerprints and descriptors (FPDS); two splitting options; random or by diversity picking; four ML algorithms; with and without multi-stage feature selection. 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 is shown in (see Supplementary file 01). The models were trained on the train set and validated on the validation set during BO. Root mean square error (RMSE) computed out of the testing set was employed as a loss function for BO. The optimization experiment ran for ~2 days on a 24 x Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz with 30GB of RAM. We also followed per iteration results on the external test set, to 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, each time transferring the feature list to the next model sequentially. Such modelling pipeline is depicted in **Figure 2**. Finally, the best models were chosen based on a ranking schema, which we believe reflects an objective model evaluation. In equation (2), the weights were fixed 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.

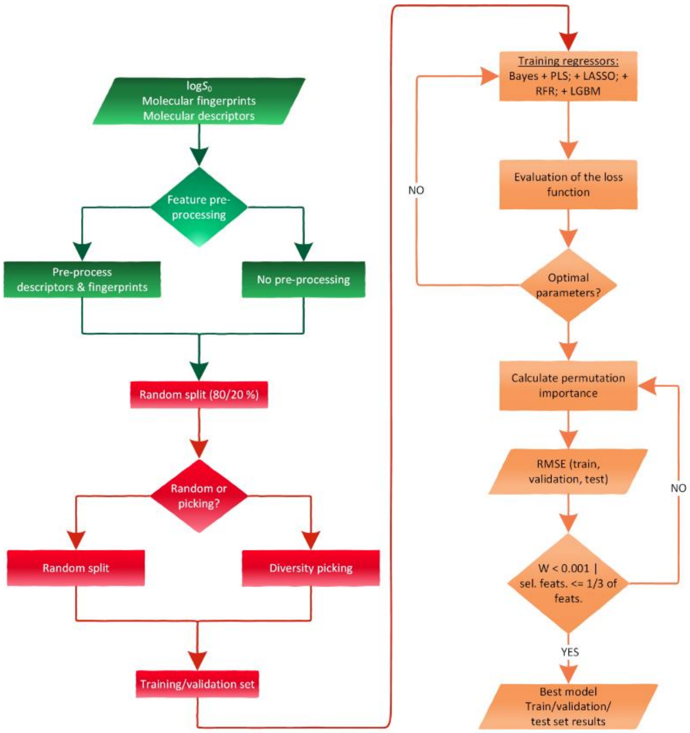
$\log S_0$
Molecular fingerprints
Molecular descriptors
Feature pre-processing
Pre-process descriptors & fingerprints
No pre-processing
Random split (80/20 %)
Random or picking?
Random split
Diversity picking
Training/validation set
Training regressors:
Bayes + PLS; + LASSO; + RFR; + LGBM
Evaluation of the loss function
NO
Optimal parameters?
Calculate permutation importance
RMSE (train, validation, test)
NO
W < 0.001 | sel. feats. <= 1/3 of feats.
YES
Best model Train/validation/ test set results

Figure 2. The model pipeline for the optimization experiment.

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

where, $R_{features}$ is the rank of the number of features employed in the model, $R_{RMSE(Test)}$ is the rank of RMSE of the respective test set whereas Δval and Δtrain are defined with equations (3) and (4), respectively. Both terms account for the generalizability of the models.

$$\Delta_{val} = \left| RMSE(test) - RMSE(val) \right| \qquad (3)$$

$$\Delta_{train} = \left| RMSE(train) - RMSE(val) \right| \qquad (4)$$

3. Results and discussion

3.1. Model optimization results

Models originating from all permutation importance stages are listed. Overall, just looking at the RMSE(test) is LASSO the best performing model (RMSE(test) = 0.69) with 105 features (fingerprints + descriptors, 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 RF models with some of them comprising as few as 16 features.

The first PLS model appeared on the 7th place comprising 33 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 fingerprint (FP)-based solubility models have generally underperformed when compared to the models built out of molecular descriptors or their combination. The models based on FP also exhibit a large spread in regard to RMSE(test). This outcome could have been expected being that none of the 4 algorithms (PLS, LASSO, LGBM, RF) creates metavariables out of the FPs like neural nets do in the hidden layers which contribute to their predictive ability.[69]

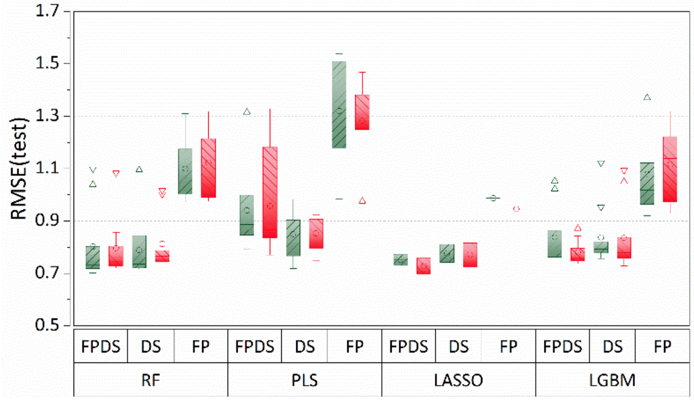

Figure 3. Distribution of testing set errors for the four evaluated machine learning algorithms. Differences between random train/test/validation split and diversity picking are depicted with green ascending, and red descending line patterns, respectively.

Also, with the addition of fingerprints to descriptors (FPDS), only marginal improvements can be observed. The PLS algorithm shows a small spread over all combinations (**Figure 3**), which can be explained by only one hyperparameter to be optimized (the number of latent features). LGBM shows a notably larger spread comparing to other algorithms (**Figure 4)**, which can be explained by evident overfitting on the train set, and lower predictive ability on the test set. Such performance decrease is not caused by the optimizer being stuck in local optima, which was evident from model results where optimal hyper-parameters 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 to the bagging ensemblers.[35, 70] Overall, the spreads per algorithm in **Figure 4** are larger for the FP and FPDS predictive sets. This might be explained by randomness which fingerprints can introduce by having a train or test bit with all zero values impeding convergence. Splitting the data either at random of via diversity picking did not exhibit notable differences in predictive ability on the test set. RMSE(val) values for models with datasets split via diversity picking, can be as low as 0.53. Nevertheless, the highest ratios of RMSE (Test / Validation) (above 1.2) are all originating from diversity-picked data splits.

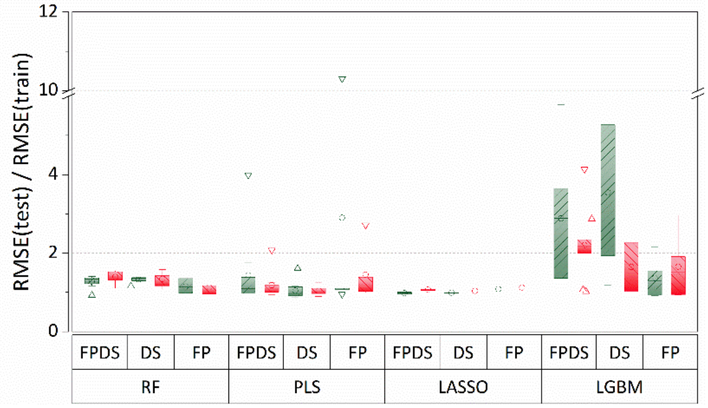

Figure 4. Generalization ability and robustness for all the models trained in this work. The color codes represent the three sets of predictive variables for the four employed methods. Differences between random train/test/validation split and diversity picking are depicted with green ascending, and red descending line patterns, respectively.

Table 1. Results of model optimization 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	0.70	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

Diversity-picking leads to similar train and validation set which points to an overestimation of the model quality on any external 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} (Eq. (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. Overall LASSO and PLS appear to be algorithms with much better generalization capabilities for QSPR modelling of intrinsic solubility. **Table 1** summarizes the ten best models according to the Rk_M metric only for random splits, since we have shown that diversity-picking can deviate the impression in generalization. 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 Random forest 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 Figures **5A**, and **5B**, respectively.

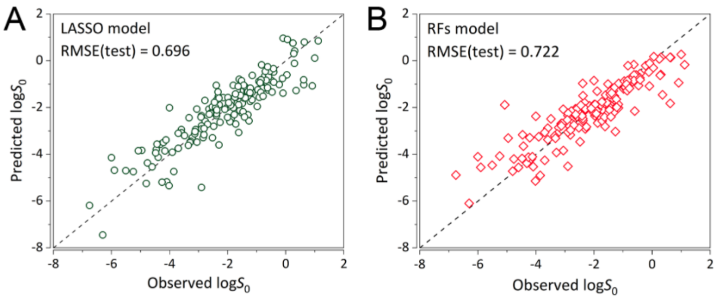

Figure 5. Predictive ability of the two best intrinsic solubility QSPR models. **A)** LASSO model, **B)** RF model.

Out of the ten 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 ten models are not using preprocessing which shows that ensemble methods work well with the original data as preprocessing can remove valuable information. None of the best models were based on FPs. The models in **Table 1** were either based on descriptors or the combined with FPs.

3.2. Feature importance

The analysis of the employed features for all the models in this study showed some interesting patterns. The PLS models in general did not reduce to as few features during the feature selection as RF or LGBM. LASSO mostly converged with subsets of 50-100 features. The multi-stage feature selection was not used in the case of LASSO as feature selection is inherent to this technique. 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 re-fitted with the resulting features and the resulting parameters. The re-trained model was subjected to permutation importance, the results of which are depicted in **Figure 6**. Detailed descriptions of all the descriptors found in Todeschini and Consonni.[71]

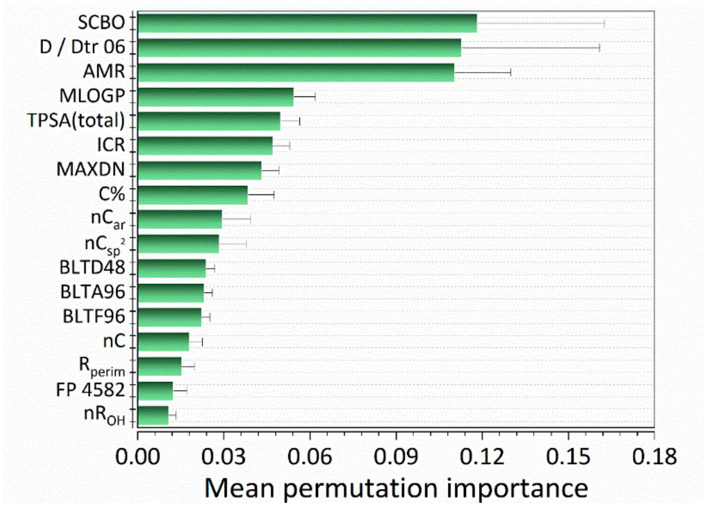

Figure 6. Mean permutation importance for 1000 random resampling runs of the best model with 17 features.

4. Conclusions

In this work, we tested multiple factors affecting machine learning outcomes in order to return the best prediction for intrinsic solubility. Besides the four regressors; LASSO, Random forest, LightGBM and Partial least squares, we tested the effect of feature selection by means of permutation importance, the data set (fingerprint and molecular descriptors), Bayesian optimization and data splitting options. The intrinsic solubility data employed here is a novel collection of curated values and structures obtained from literature with 829 “drug-like” compounds. The best result by means of performance was a LASSO regressor.

Nevertheless, we propose a ranking schema for choosing the best models by not solely the measure's performance on a fixed test set, but also by taking the number of features and the estimated generalization performance into account. The rankings reveal a clear dominance of the Random forest algorithm since it can predict well with less features involved. Even though LightGBM is a powerful algorithm, it has complex hyperparameter space hard to optimize and overfits most of the time. We show that there is no single criterion, data set nor algorithm which can cover it all, but rather a multiverse of possibilities and decision to be embraced for building robust models with strong generalizability.

5. References

1. Hörter D, Dressman JB (2001) Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. Adv Drug Deliv Rev 46:75–87. https://doi.org/10.1016/S0169-409X(00)00130-7
2. Kalepu S, Nekkanti V (2015) Insoluble drug delivery strategies: Review of recent advances and business prospects. Acta Pharm. Sin. B 5:442–453
3. Llinas A, Avdeef A (2019) 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 59:3036–3040. https://doi.org/10.1021/acs.jcim.9b00345
4. Hopfinger AJ, Esposito EX, Llinàs A, et al (2009) Findings of the challenge to predict aqueous solubility. J Chem Inf Model 49:1–5. https://doi.org/10.1021/ci800436c
5. Palmer DS, Mitchell JBO (2014) Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? Mol Pharm 11:2962–2972. https://doi.org/10.1021/mp500103r
6. Montanari F, Kuhnke L, Ter Laak A, Clevert D-A (2019) Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. Molecules 25:44. https://doi.org/10.3390/molecules25010044
7. Avdeef A (2020) Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. ADMET DMPK 8:29. https://doi.org/10.5599/admet.766
8. Montanari F, Kuhnke L, Ter Laak A, et al (2020) Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks. Molecules 25:44. https://doi.org/10.3390/molecules25010044
9. Tang B, Kramer ST, Fang M, et al (2020) A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. J Cheminform 12:1–9. https://doi.org/10.1186/s13321-020-0414-z
10. Deng T, Jia G zhu (2020) Prediction of aqueous solubility of compounds based on neural network. Mol Phys 118:1–8. https://doi.org/10.1080/00268976.2019.1600754
11. Boobier S, Osbourn A, Mitchell JBO (2017) Can human experts predict solubility better than computers? J Cheminform 9:1–14. https://doi.org/10.1186/s13321-017-0250-y
12. Zang Q, Mansouri K, Williams AJ, et al (2017) In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. J Chem Inf Model 57:36–49. https://doi.org/10.1021/acs.jcim.6b00625
13. McDonagh JL, Nath N, De Ferrari L, et al (2014) Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. J Chem Inf Model 54:844–856. https://doi.org/10.1021/ci4005805
14. Lusci A, Pollastri G, Baldi P (2013) Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. J Chem Inf Model 53:1563–1575. https://doi.org/10.1021/ci400187y
15. Salahinejad M, Le TC, Winkler DA (2013) Aqueous solubility prediction: Do crystal lattice interactions help? Mol Pharm 10:2757–2766. https://doi.org/10.1021/mp4001958
16. Cao DS, Xu QS, Liang YZ, et al (2010) Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. J Chemom 24:584–595. https://doi.org/10.1002/cem.1321
17. Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA (2008) New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorganic Med Chem 16:7944–7955. https://doi.org/10.1016/j.bmc.2008.07.067
18. Louis B, Agrawal VK, Khadikar P V. (2010) Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. Eur J Med Chem 45:4018–4025. https://doi.org/10.1016/j.ejmech.2010.05.059
19. Schäfer RB, Pettigrove V, Rose G, et al (2011) Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. Environ Sci Technol 45:1665–1672. https://doi.org/10.1021/es103227q
20. Wichard Ö, Kühne R PREDICTING AQUEOUS SOLUBILITY FROM STRUCTURE. 1–3
21. Wang J, Hou T, Xu X (2009) Aqueous solubility prediction based on weighted atom type counts and solvent accessible surface areas. J Chem Inf Model 49:571–581. https://doi.org/10.1021/ci800406y
22. Palmer DS, Llinàs A, Morao I, et al (2007) Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. Mol Pharm xxx:545–556. https://doi.org/10.1021/mp7000878
23. Chen XQ, Cho SJ, Li Y, Venkatesh S (2002) Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. J Pharm Sci 91:1838–1852. https://doi.org/10.1002/jps.10178
24. Bergström CAS, Luthman K, Artursson P (2004) Accuracy of calculated pH-dependent aqueous drug solubility. Eur J Pharm Sci 22:387–398. https://doi.org/10.1016/j.ejps.2004.04.006
25. Bergström CAS, Wassvik CM, Norinder U, et al (2004) Global and local computational models for aqueous solubility prediction of drug-like molecules. J Chem Inf Comput Sci 44:1477–1488. https://doi.org/10.1021/ci049909h
26. Delaney JS (2004) ESOL: Estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 44:1000–1005. https://doi.org/10.1021/ci034243x
27. Bergström CAS, Norinder U, Luthman K, Artursson P (2002) Experimental and computational screening models for prediction of aqueous drug solubility. Pharm Res 19:182–188. https://doi.org/10.1023/A:1014224900524
28. Engkvist O, Wrede P (2002) High-throughput, in silico prediction of aqueous solubility based on one- and two-dimensional descriptors. J Chem Inf Comput Sci 42:1247–1249. https://doi.org/10.1021/ci0202685
29. McElroy NR, Jurs PC (2001) Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. J Chem Inf Comput Sci 41:1237–1247. https://doi.org/10.1021/ci010035y
30. Huuskonen J, Salo M, Taskinen J (1996) Neural Network Modeling for Estimation of the Aqueous Solubility of Structurally Related Drugs
31. Mitchell BE, Jurs PC (1998) Prediction of aqueous solubility of organic compounds from molecular structure. J Chem Inf Comput Sci 38:489–496. https://doi.org/10.1021/ci970117f
32. Sutter JM, Jurs PC (1996) Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure-property relationship. J Chem Inf Comput Sci 36:100–107. https://doi.org/10.1021/ci9501507
33. Lučić B, Batista J, Bojović V, et al (2019) Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges. Croat Chem Acta 92:. https://doi.org/10.5562/cca3551
34. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO (2007) Random forest models to predict aqueous solubility. J Chem Inf Model 47:150–158. https://doi.org/10.1021/ci060164k
35. Šimić I, Lovrić M, Godec R, et al (2020) Applying machine learning methods to better understand, model and estimate mass concentrations of

traffic-related pollutants at a typical street canyon. Environ Pollut 263:114587. https://doi.org/10.1016/j.envpol.2020.114587

36. Avdeef A, Berger CM (2001) pH-metric solubility.: 3. Dissolution titration template method for solubility determination. Eur J Pharm Sci 14:281–291. https://doi.org/10.1016/S0928-0987(01)00190-7

37. Rytting E, Lentz KA, Chen XQ, et al (2005) Aqueous and cosolvent solubility data for drug-like organic compounds. AAPS J 7:. https://doi.org/10.1208/aapsj070110

38. Sköld C, Winiwarter S, Wernevik J, et al (2006) Presentation of a structurally diverse and commercially available drug data set for correlation and benchmarking studies. J Med Chem 49:6660–6671. https://doi.org/10.1021/jm0506219

39. Wassvik CM, Holmén AG, Bergström CAS, et al (2006) Contribution of solid-state properties to the aqueous solubility of drugs. Eur J Pharm Sci 29:294–305. https://doi.org/10.1016/j.ejps.2006.05.013

40. Llinàs A, Glen RC, Goodman JM (2008) Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? J Chem Inf Model 48:1289–1303. https://doi.org/10.1021/ci800058v

41. Baek K, Jeon S Bin, Kim BK, Kang NS (2018) Method Validation for Equilibrium Solubility and Determination of Temperature Effect on the Ionization Constant and Intrinsic Solubility of Drugs. J Pharm Sci Emerg Drugs 06:1–6. https://doi.org/10.4172/2380-9477.1000125

42. Stuart M, Box KJ (2005) Chasing equilibrium: Measuring the intrinsic solubility of weak acids and bases. Anal Chem 77:983–990. https://doi.org/10.1021/ac048767n

43. Bergström CAS, Strafford M, Lazorova L, et al (2003) Absorption classification of oral drugs based on molecular surface properties. J Med Chem 46:558–570. https://doi.org/10.1021/jm020986i

44. McFarland JW, Avdeef A, Berger CM, Raevsky OA (2001) Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. J Chem Inf Comput Sci 41:1355–1359. https://doi.org/10.1021/ci0102822

45. Avdeef A (2019) Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. ADMET DMPK 7:210–219. https://doi.org/10.5599/admet.698

46. Box KJ, Comer J (2008) Using Measured pKa, LogP and Solubility to Investigate Supersaturation and Predict BCS Class. Curr Drug Metab 9:869–878. https://doi.org/10.2174/138920008786485155

47. Etherson K, Halbert G, Elliott M (2014) Determination of excipient based solubility increases using the CheqSol method. Int J Pharm 465:202–209. https://doi.org/10.1016/j.ijpharm.2014.02.007

48. Fornells E, Fuguet E, Mañé M, et al (2018) Effect of vinylpyrrolidone polymers on the solubility and supersaturation of drugs; a study using the Cheqsol method. Eur J Pharm Sci 117:227–235. https://doi.org/10.1016/j.ejps.2018.02.025

49. Llinàs A, Burley JC, Box KJ, et al (2007) Diclofenac solubility: Independent determination of the intrinsic solubility of three crystal forms. J Med Chem 50:979–983. https://doi.org/10.1021/jm0612970

50. Schönherr D, Wollatz U, Haznar-Garbacz D, et al (2015) Characterisation of selected active agents regarding pKa values, solubility concentrations and pH profiles by SiriusT3. Eur J Pharm Biopharm 92:155–170. https://doi.org/10.1016/j.ejpb.2015.02.028

51. Huuskonen J (2000) Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. J Chem Inf Comput Sci 40:773–777. https://doi.org/10.1021/ci9901338

52. Abraham MH, Le J (1999) The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. J Pharm Sci 88:868–880. https://doi.org/10.1021/js9901007

53. Shareef A, Angove MJ, Wells JD, Johnson BB (2006) Aqueous solubilities of estrone, 17β-estradiol, 17α- ethynylestradiol, and bisphenol A. J Chem Eng Data 51:879–881. https://doi.org/10.1021/je050318c

54. Lakshmi Narasimham Y, Barhate VD (2011) Kinetic and intrinsic solubility determination of some b-blockers and antidiabetics by potentiometry. J Pharm Res 4:532–536

55. Mansouri K, Kleinstreuer N, Abdelaziz AM, et al (2020) CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. Environ Health Perspect 128:027002. https://doi.org/10.1289/EHP5580

56. Landrum G (2006) RDKit: Open-source cheminformatics

57. Lovrić M, Molero JM, Kern R (2019) PySpark and RDKit: Moving towards Big Data in Cheminformatics. Mol Inform 38:. 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 23 Dec 2019

59. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B 58:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

60. Bro R (1996) Multiway calibration. Multilinear PLS. J Chemom 10:47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C

61. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: A basic tool of chemometrics. In: Chemometrics and Intelligent Laboratory Systems. Elsevier, pp 109–130

62. Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017-Decem:3147–3155

63. Breiman L (2001) Random Forests

64. Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: A corrected feature importance measure. Bioinformatics 26:1340–1347. https://doi.org/10.1093/bioinformatics/btq134

65. Han H, Guo X, Yu H (2016) 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. IEEE Computer Society, pp 219–224

66. Pontes FJ, Amorim GF, Balestrassi PP, et al (2016) Design of experiments and focused grid search for neural network parameter optimization. Neurocomputing 186:22–34. https://doi.org/10.1016/j.neucom.2015.12.061

67. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian Optimization of Machine Learning Algorithms. In: NIPS 2012

68. Dudgeon T (2017) RDKit: Revisting the MaxMinPicker. http://rdkit.blogspot.com/2017/11/revisting-maxminpicker.html. Accessed 23 Dec 2019

69. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) DeepTox: Toxicity prediction using deep learning. Front Environ Sci 3:. https://doi.org/10.3389/fenvs.2015.00080

70. Žuvela P, Lovric M, Yousefian-Jazi A, Liu JJ (2020) Ensemble Learning Approaches to Data Imbalance and Competing Objectives in Design of an Industrial Machine Vision System. Ind Eng Chem Res 59:4636–4645. https://doi.org/10.1021/acs.iecr.9b05766

71. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. WileyVCH, Weinheim. Wiley, New York