Advancing Aqueous Solubility Prediction: A Machine Learning Approach for Organic Compounds Using a Curated Dataset

Mushtaq Ali,[1] Sylvia Vanderheiden,[1] Christoph W. Grathwol,[1] Pascal Friederich,[2,3*] Nicole Jung,[1,4*] Stefan Bräse[1,5*]

[1]Institute of Biological and Chemical Systems (IBCS), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany. [2]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany; [3]Institute of Nanotechnology, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany; [4]Karlsruhe Nano Micro Facility (KNMFi), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany; [5]Institute of Organic Chemistry, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany.

[*]Corresponding author email: Nicole.jung@kit.edu, stefan.braese@kit.edu, pascal.friederich@kit.edu

Abstract

Aqueous solubility is one key property of a chemical compound that determines its possible use in different applications, from drug development to materials sciences. In this work, we present an aqueous solubility prediction study that leverages a curated dataset merged from four distinct sources. This unified dataset encompasses a diverse range of organic compounds, providing a robust foundation for our investigation of solubility prediction. Our approach involves employing

a variety of machine learning and deep learning models that combine an extensive array of chemical descriptors, fingerprints, and functional groups. This methodology is designed to address the complexities of solubility prediction, and it is tailored to achieve high accuracy and generalization. We tested the finalized model on a diverse dataset of 1282 unique organic compounds from the Husskonnen dataset. The results of our analysis demonstrate the success of our model, which, given an R2 value of 0.92 and an MAE value of 0.40, outperforms existing prediction methods for aqueous solubility on one of the most diverse datasets in the field.

Introduction

Aqueous solubility is a fundamental property essential for investigating and applying chemical compounds across various scientific disciplines, including chemistry and biology. Being able to dissolve a chemical substance plays a crucial role, e.g., in designing novel drugs, as the solubility has a high impact on the bioavailability of drugs and their distribution.[1] In materials sciences, the solubility of chemicals has an effect on how materials are built and makes the difference in how active functional components are applied. While the solubility of commercially available building blocks is usually known at least for a few solvents, the investigation of newly designed and synthesized compounds needs the determination of the compounds' solubility. The measurement of solubility can be done by different methods, e.g. using HPLC techniques[2], or gravimetrically[3]. The drawbacks of all techniques are that (1) the compound needs to be synthesized in a substantial amount, (2) different equipment needs to be available, and (3) the determination of solubility in different solvents needs a lot of expertise and time. The ability to predict solubility accurately could offer a lot of benefits to the current process of manually determining the solubility of new compounds. First of all, the laborious and time as well as resource-consuming measurements could

be replaced by suitable predictions (at least in parts), and secondly, predictions could help to design potentially interesting molecules - preventing the synthesis of compounds that do not obtain the expected properties.

A general understanding of solubility prediction was gained with the general solubility equation,[4] which includes solute-solvent interactions through parameters such as activity coefficients, and enables predictions under various conditions. Hildebrand and Hansen solubility parameters[5,6] describe the properties of solvents and solutes. Hansen developed this idea by introducing three solubility parameters that take into account hydrogen bonding, polarity, and dispersion forces, on which the different properties of the compound depend. In COSMOS-RS[75,68] quantum chemistry tools are used in the form of a conductor-like screening model for real solvents to forecast solubilities in complex systems. It takes into account the environment, electrostatic interactions, and molecular structure. Recently, significant progress has been made in the solubility prediction field, with a focus on the use of statistical methods supported by rigorous data analysis to uncover hidden patterns and correlations within solubility data. Statistical models leverage advanced algorithms to identify key factors influencing solubility, thereby improving predictive accuracy and machine learning methods[9]. These data-driven approaches harness the power of artificial intelligence to process vast datasets, learn complex relationships, and make accurate predictions. The majority of solubility predictions in the past have been made using overly simple models that only took into account a small number of molecular characteristics, such as molecular weight, log P (partition coefficient), atom counts, and ring counts[10]. These early models were insightful, but they frequently failed to account for the complex interactions that determine solubility.

In the case of aqueous solubility, the use of molecular descriptors such as diverse molecular descriptors characterizing the chemical structure or its properties has been shown to give the most

accurate prediction so far.[10,11,12,13,14,15] Sorkun et al.[9] achieved notable predictive accuracy by using a dataset with 4399 unique data points, a feature set with 123 descriptors, and a consensus model with Random Forest, XGBoost, and an artificial neural network.

Some major challenges consist of the existence of measurement noise and data quality issues, the diversity and scale of the molecular space, and the broad range of solubility values.

Methods

Data Curation and Preprocessing

In order to maximize the training dataset, we collected data from four different sources, which are further referenced as datasets A (Boobier *et al.*[17]), B (Panapitiya[18]), C (Cui, Q. *et* [20]), and D (Sorkun[9]) which have already been used in the literature. An additional dataset E (Huuskonen[19]) was reserved as a dataset for testing the model. Merged the four datasets, resulting in a total of 28,859 samples for training and 1,291 samples for testing. Table 1 and Figure 1 give additional information about the process of data collection and data curation for the data that was used as a training dataset to train the model and the test dataset on which the model has been evaluated. To ensure the consistency of the combined dataset A-D, in particular to prevent having included duplicates, the representation of chemical structures is unified by converting the molecular representations into the canonical SMILES format. This standardization step eliminates inconsistencies caused by disparate chemical notations or representations. In parallel, the information on the solubility was harmonized by converting the solubility values of all data, e.g. from g/L into logS. The obtained datasets with harmonized logS properties were searched for equal SMILES codes, indicating duplicates in the dataset. Molecules with two or more records were sorted into two groups: molecules appearing multiple times having the same solubility values, and

those where the solubility values differ. In the first case, we kept only one of the equal examples. In the second case, an average of the solubility values was calculated for all molecules that have a difference lower than 0.5 in logS. 1383 examples with more than a 0.5 difference for the given logS in comparison to the closest other logS value were removed from the dataset (see SI). After the removal and/or harmonization of the duplicates from the training and test datasets, the test dataset was compared to the training dataset. The molecules with matching SMILES in the test and training datasets were removed from the training dataset to ensure that the training and test datasets have no overlapping data and to keep the test dataset complete for validation purposes in direct comparison to other results in the literature. With this process, we finally got 17937 training samples and 1282 test samples (Figure 1).

Table 1: Details of the 5 different datasets used to generate a unique dataset to train and test the model

Dataset	Name	Dataset Size	Use	Duplicate	Unique	Reference
A	BNN Lab	900	Training	4	898	[17]
B	Gihan	11862	Training	261	11724	[18]
C	Xian Zeng	9942	Training	347	9750	[20]
D	Sorkun	6154	Training	471	5907	[9]
E	Huuskonen	1291	Testing	18	1282	[19]

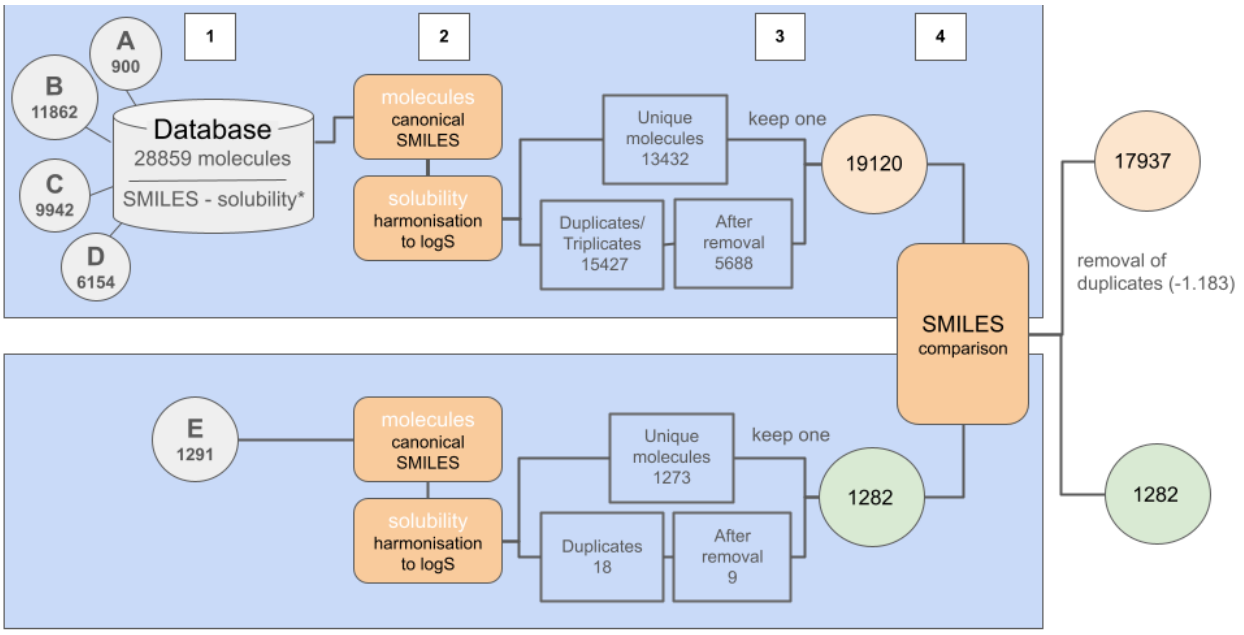


Figure 1. Schematic presentation of the preprocessing pipeline for training and test data consisting of (1) data collection, (2) generation of canonical SMILES and solubility harmonization, (3) removal of duplicates and data cleaning in the training and test data sets, (4) comparison of the test to the training dataset and removal of duplicates from the latter one.

Generation of Descriptors

Following the idea of using molecules and their descriptors for the training of neural networks, we created an initial set of four fundamental descriptors representing essential molecular characteristics by RDkit (exact molecular weight, water-octanol partition coefficient logP, aromatic proportion, and rotatable bonds), forming the foundation of our feature representation. To systematically investigate the impact of including additional descriptors on the model's performance, we incrementally expanded the descriptor set. This stepwise augmentation allowed us to assess how the inclusion of a higher number of descriptors influences the accuracy and predictive power of our model. As we progressed, we continually introduced new descriptors into our feature space, each chosen to capture specific chemical attributes and properties. The step-by-step expansion of our descriptor set provided insights into the optimal feature space for our

predictive modeling task, allowing us to refine and enhance our solubility predictions. Within the extended descriptor set, four different descriptor types used at different stages of our study can be described: (1) 125 classical descriptors, including topological, physicochemical, and electronic properties. These descriptors offer a comprehensive characterization of molecular structures and properties (SI, chapter 2.1). (2) Different molecular fingerprints with varying bit lengths, ranging from 128 to 1024 bits (radius of 2) were considered in our model. Molecular fingerprints enable the capture of fine-grained structural information at various levels of granularity. (3) We included binary representations of the presence or absence of specific functional groups as descriptors, as we expected them to be a potentially critical property of the molecular structures referring to their solubility. An overview of the prevalence and diversity of functional groups present in the training dataset and a summary of the influence of functional groups on the solubility of molecules is included in the Supplemental Information (SI, chapter 2.2). (4) Along with these descriptors, we added 38 molecular descriptors, including charge, double bonds, valence electrons, hybridization types, bond types, and chirality features. These descriptors aim to capture detailed structural and electronic characteristics of the molecules for improved predictive accuracy (SI, chapter 2.1).

Model Building

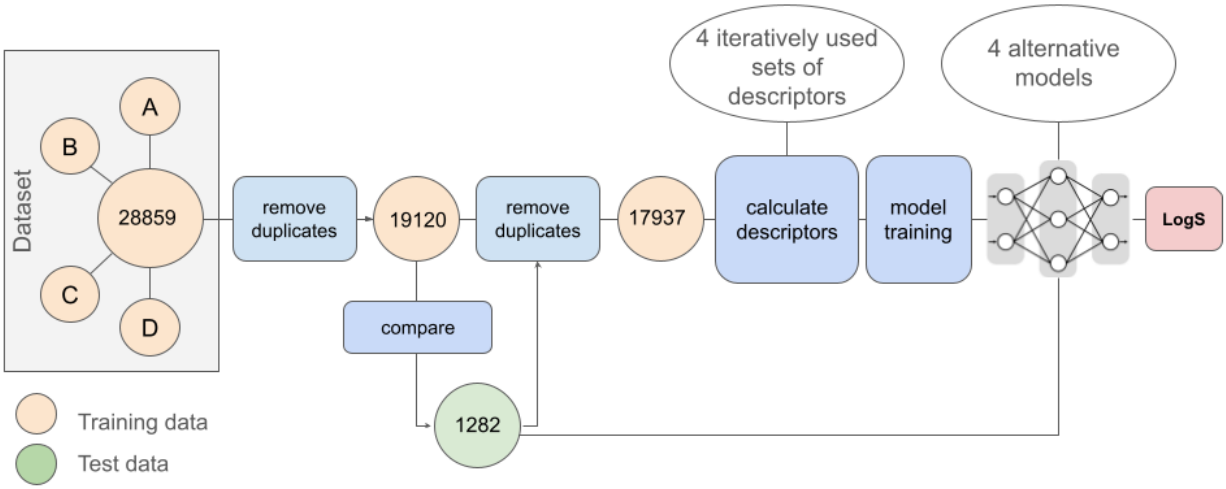


Figure 3. Schematic description of the workflow followed in this work consisting of dataset preparation for training and testing, calculation of descriptors, training of the model (according to four different basic models), and the prediction of the solubility in LogS.

We used four different machine learning models in this study, each designed to capture particular nuances in our solubility prediction task (see SI). (1) The ensemble learning technique **Random Forest** was chosen because of its ability to handle a variety of data types and identify non-linear relationships. Hyperparameters like tree depth and forest size were given special consideration when building our random forest model. (2) **XGBoost** was used because of its excellent prediction abilities. It excels at modeling intricate data relationships. Our method for optimizing crucial parameters, such as the learning rate, tree depth, and number of estimators (trees in the ensemble), included hyperparameter tuning. (3) **Artificial Neural Networks** (ANNs) are renowned for their capacity to recognize intricate data relationships and patterns. Multiple hidden layers with movable sizes and activation mechanisms make up our ANN architecture [2]. To ensure the network's ideal learning and generalization, hyperparameter tuning included learning rates, weight regularization, and dropout. (4) **Message-Passing Neural Networks** employed both a standard Message-Passing Neural Network (MPNN) and a hybrid MPNN architecture for solubility

prediction. The hybrid MPNN comprises seven layers, including a combination of message-passing layers, fully connected layers, and integration mechanisms for additional physical property features. This architecture was designed to enhance the predictive accuracy by capturing both molecular graph-level information and physical property data. Comparative evaluations were conducted to assess the performance improvements achieved with the hybrid MPNN over the standard MPNN approach. The selection of message-passing functions, layer configurations, and other crucial parameters were addressed during the hyperparameter tuning phase. The optimal hyperparameters of all models can be found in the SI.

To improve the predictive performance and interpretability of our machine learning model, we employed the Least Absolute Shrinkage and Selection Operator (LASSO) regularization technique, which applies L1 regularization to penalize and shrink less important feature coefficients towards zero. In order to evaluate the model, we perform five-fold cross-validation, which checks that our models generalize well and do not overfit the training data. One-fold was reserved for the validation set for each iteration, and the remaining folds served as the training set. For each fold, the model's performance was assessed using a specific metric: mean absolute error (MAE), root mean squared error (RMSE), and R2 coefficient of determination. The model's performance was then evaluated across the entire dataset by aggregating the performance metrics across all five cross-validation splits. Cross-validation was used for hyperparameter tuning and performance evaluation. In order to find the configuration that maximizes model performance, various sets of hyperparameters were tested throughout the hyperparameter tuning process.

Results

In our work, we adopt the concept of using molecular descriptors as described previously by others. To improve the currently available models with respect to the accuracy of the model and the suitability for a wide range of chemical substance classes, we collected additional key aspects that we expected to improve the solubility prediction. The main changes included: (1) the application of traditional machine learning algorithms like XGBoost and Random Forest, along with neural networks and (hybrid) message-passing neural networks from graph neural networks. (2) We further extended the molecular descriptors used in a stepwise approach, (3) We included altogether four datasets from different established sources as a training set and compared it to the previously used test dataset published by Husskonen[19].

Comparative analysis with different sets of features

Using XGBoost, Random Forest regression, ANNs, and MPNNs, four models using a particular set of descriptors were trained (Figure 3, Table 2). Our analysis focused on MAE and RMSE as important metrics to compare the performance of the models and descriptors (see SI). We evaluated the models' cross-validation results to determine how well they generalize to unseen data. We further investigated the subtleties of each model's configuration, taking into account how the quantity and variety of descriptors affected the complexity of the model. We started our investigation with XGBoost, due to its ability to capture complex, non-linear relationships and mitigate overfitting through built-in regularization. Its robustness, efficiency, and adaptability, combined with extensive hyperparameter tuning, make it an ideal choice for our dataset's characteristics. Using XGBoost with 4 descriptors (molecular weight, logP, aromatic proportion, and rotatable bonds) gave an R2 score of 0.87, in combination with an RMSE of 0.72 and MAE

of 0.56. The extension of the descriptor set to 17 descriptors (XGB-17) increased the performance of the model to a level that is comparable to the best models currently available referring to the same test data from Huuskonen (see Table 2 and Table 3). Further increase in the number of descriptors yielded a systematic improvement of the performance of the XGBoost model, however, the improvement from 17 to 125 descriptors and the following ones were lower than the improvement made in the first step (increase from 4 to 17). The last improvements given by the introduction of the 128-bit fingerprint descriptor and the addition of 7 functional groups are very small (not represented in the number of digits). The increase in fingerprint size to 512 and 1024 bits (Table 2 entries 7,8) led to a decrease in the performance on the test dataset, likely due to the introduction of redundant or irrelevant features. This causes overfitting. The best results were finally achieved with XGB in combination with 298 descriptors (Table 2). This model (further named XGB-298), yielding an MAE value of 0.40, an RMSE value of 0.55, and an R2 score of 0.92, was used for further investigation and comparison with three other known models in the literature. Neither Random Forest (MAE = 0.49, RMSE = 0.64, R2 = 0.90) nor ANN (MAE = 0.52, RMSE = 0.70, R2 = 0.88) could compete with the model XGBoost-298 even when the same set of descriptors was used. Attempts using MPNN with 6 layers included were even worse, giving an R2 of 0.82, MAE of 0.68, and RMSE of 0.76, a hybrid MPNN model.

Table 2: Comparison of test set performance on different models and combinations of descriptors given by the metrics MAE, RMSE, and R2.

Entry	Model Name	MAE	RMSE	R2	No. of descriptors/fingerprint version				
					No descr[a]	FP[b]	FGs[c]	feat[d]	layers
1	XGB-4	0.56	0.72	0.87	4	-	-	-	-
2	XGB-17	0.44	0.58	0.91	17	-	-	-	-
3	XGB-125	**0.40**	**0.55**	**0.92**	125	-	-	-	-
4	XGB-253	**0.40**	**0.55**	**0.92**	125	128	-	-	-
5	XGB-260	**0.40**	**0.55**	**0.92**	125	128	7	-	-
6	XGB-298	**0.40**	**0.55**	**0.92**	125	128	7	38	-
7	XGB-682	0.41	**0.55**	**0.92**	125	512	7	38	-
8	XGB-1194	0.41	**0.55**	**0.92**	125	1024	7	38	-
9	RANDOM FOREST	0.49	0.64	0.90	125	128	7	38	-
10	MPNN	0.68	0.76	0.82	-	-	-	-	6
11	Hybrid MPNN	0.46	0.63	0.90	125	-	7	38	7
12	ANN	0.52	0.70	0.88	125	128	7	38	6

[a]Model includes the given number of descriptors; [b]Type of Fingerprint: [c]FGs = number of functional groups included; [d]Additional selected descriptors included. Details of the type functional groups as well as the uncertainties for MAE are described in the SI.

To the best of our knowledge, there are currently 11 studies dealing with the prediction of aqueous solubility and using the Huuskonen dataset as a reference test dataset (Table 3). We compared the results of our model with results from the literature using the same test dataset.[19] In our first

comparison, we found our model to be comparatively powerful as the best model from the literature, which was published in 2020 by Sorkun *et al.* (Table 3). We then compared the models and workflows in more detail and found one main difference in the preparation of the training and test dataset. While we harmonized the data sets used for training and testing by the transformation of all molecules into canonical smiles without stereochemical information, Sorkun *et al.* used the InChIKey from stereochemical SMILES in the training set and SMILES without stereochemical information in the test set. This difference in the preparation of the data has an important effect on identifying possible overlaps in training and test data. In the canonical SMILES approach without stereochemical information, molecules with defined stereochemistry in either of the training datasets yield the same canonical SMILES code as the same molecule in the test set without specific stereochemical annotation - and duplicates are consequently removed. In this way, e.g. a double bond which is given in either Z or E annotation (or a mixture of both) in one dataset is to be considered as the same molecule in another dataset if the double bond is not annotated specifically and just given as any double bond. In contrast to this, the InChIKey approach gives different InChIKey codes for molecules with and without assigned isomer details. Consequently, the approach of Sorkun *et al.* includes molecules that might be the same but are described with fewer isomer details in the test data than in the training data (and vice versa). Our review of the dataset used by Sorkun *et al.* gave 133 duplicate samples where such a correlation might be an issue, therefore, we reproduced the model once with and without these 133 data points in the training set. With the data included, we could reproduce the published results and gain exactly the values given in Table 3. When removing the 133 data from the training dataset, we could not reproduce the former results and came to an R2 value of 0.87 in combination with an MAE value of 0.54 and an RMSE value of 0.73. We can think of two possible explanations for the results of

the reproduction of the literature-known work: Either the overlap in test and training sets caused an improvement of the performance of the model and their removal gives a more reliable and unbiased estimation of the generalization performance of the model, or the decrease in training set size due to the removal of the 133 samples caused the model's performance drop. The first conclusion is more likely, taking into account that the 133 data points are only a small portion of the overall training dataset. Consequently, a comparison of the models according to Table 3 should be done based on the data that can be obtained with the reproduced results of the model by Sorkun *et al.* without the 133 data points. Taking this into account, we were able to improve the currently best results gained in previous work (by Yan and Gasteiger, as well as Lusci *et al.*) with respect to the gained MAE and RMSE values and can compete with the currently highest R2 value of 0.92 gained.

Table 3: Comparison of our model XGB-298 with previous work in the literature

Entry	Method	test data Size	MAE	RMSE	R2	Year	Reference
1	ANN	1294	-	0.71	0.88	2000	Huuskonen[21]
2	ANN	1291	-	0.62	0.91	2001	Tetko et al[22]
3	ANN	1294	0.68	0.59	0.92	2003	Yan and Gasteiger[23]
4	MLR	1290	0.68	0.87	0.71	2004	Delaney[13]
5	MLR	1294	0.52	0.63	0.90	2004	Hou et al.[24]
6	SVM	1290	0.43	0.60	-	2007	Schroeter et al. [25]
7	MLR	1290	0.72	0.94	0.73	2012	Ali et al.[26]
8	UG-RNN	1026	0.46	0.60	0.91	2013	Lusci et al.[10]
9	MLR	1290	0.93	1.15	0.68	2017	Daina et al. [27]
10	ANN	1297	-	0.65	0.90	2018	Bjerrum and Sattarov[27,28]
11a	Consensus	1290	(0.39)	(0.53)	(0.93)	2020	Sorkun et al.[9]

11b*			0.54*	0.73*	0.87*	2024*	rework from Sorkun et al*
13	**XGB-298**	**1282**	**0.40**	**0.55**	**0.92**	**2024**	**this work**

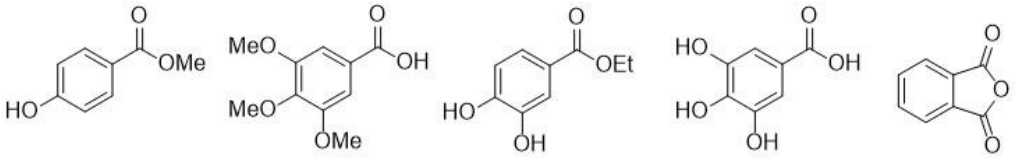
*Results obtained with the model and test data of Sorkun *et al.* after the exclusion of data referring to the same SMILES code in training and test datasets (see SI, chapter 4). **ANN**, artificial neural network, **MLR,** multiple linear regression, SVM **support** vector machine **UG RNN,** Undirected graph recursive neural networks, **RF,** random forest; **XGB** Extreme Gradient Boosting; **Consensus,** an ensemble of ANN, RF, and XGB. Best results are given in bold; best results in past developments are underlined.

Comparative analysis with experimental values and prediction

Besides the standard analysis of predictive models and the comparison with models that refer to the same test dataset of Husskonen as described above, we performed a comparative evaluation of our model against well-established solubility prediction tools. To this aim, we selected five standard compounds that are easily available, for which literature solubility values are available, and for which solubility data can be determined in our labs. The combination of solubility values from the literature and our labs (experimental part given in SI, Section 5) as reference values allowed us to gain additional confidence in the correctness of the literature values and allowed us to identify potentially problematic compounds, for that the measurement details such as the pH might be crucial for the results but perhaps not included to the literature data. We selected the models of VCC labs,[29,30] Sorkun,[29,30] and Chembcpp[31] for the comparative analysis as they are well-known (e.g. the VCC labs model is included in DrugBank) and available in the form of an online service. The comparison was achieved by statistical analysis and by determining those models (Table 4, green) that are closest to either the literature value or the experimentally found values. Across the five compounds, our model achieved an average mean absolute error (MAE) of 0.88, compared to 2.04 for VCCLAB, 1.56 for Sorkun, and 1.30 for Chembcpp. Although this

statistical analysis only considers a small number of compounds, the values confirm the superior performance of XGBoost-298 in comparison to the other models at least with respect to the given compound scope. In addition, in the identification of the models that fit best to literature or experimental results, XGBoost-298 gave better results (4 values fit best) than Chembcpp (2 values fit best), VCC (1 value fits best), and the Sorkun model (0 values fit best).

Table 4: Comparison of solubility values gained from previous models, the literature, and experiments



1 2 3 4 5

Structure	VCC[32] [g/L][a]	Sorkun[9] [g/L][a]	Chembcpp [g/L][a]	XGB-298 [g/L][b]	Exp. Lit [g/L][c]	Exp. lab [g/L][d]
1	3.64	5.39	**3.18**	**2.26**	2.99[33]	2.42
2	2.07	1.68	2.12	**2.97**	2.58[34]	3.10
3	3.55	2.88	1.63	**2.32**	2.50[35]	2.12
4	4.90	12.61	**7.21**	**11.85**	11.90[36]	7.20
5	**5.01**	2.57	2.41	2.64	6.00[37]	12.0

[a] predicted solubility in g/L from a reference model (value calculated from the given information in logS); [b]results gained from our model; [c]experimental values extracted from different literature sources; [d]experimental values determined in our labs (see supplemental information for details on the used method). Highlighted in bold: Values for that the given model performs best in comparison to the other predictive models either referring to the experimental data from the literature or the experimental data determined in our labs.

Conclusion

With XGBoost-298, we introduce a model for the prediction of solubility values for chemical compounds in water. The model was built using different sets of descriptors allowing us to determine the most efficient combination of available descriptors in a step-by-step approach and yielded 298 descriptors as the most successful approach. The model based on XGBoost was shown to be superior to other models such as Random Forest, ANN, and MPNN even if they were used with the same descriptors. The results were obtained with a curated dataset collected from four different sources for training purposes; an additional dataset was reserved as a test dataset for comparison with literature-known models. A further comparison with these other literature models showed that XGBoost-298 can compete with the most successful results in the literature and gives even better results with respect to the MAE and MRSE. We further compared our results to data that can be retrieved from online resources of other projects dealing with aqueous solubility. We could see that for four of the selected five examples, our model gave the closest prediction referring to either the literature data or data that was experimentally determined in our labs. The model is available as an online resource and can be used as a service.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The source code and data used for model development are available for free and can be obtained on Github (https://github.com/ComPlat/water-solubility-prediction).

Competing interests

There is no competing interest.

Funding

The project was funded by the Helmholtz research field Information at the Karlsruhe Institute of Technology (KIT) and the assigned project VirtMat (Virtual Materials Design).

Authors' contributions

AM designed and developed the herein-described model. NJ, SB, and PF contributed to the conceptual work of this project. CG contributed to the design of the solubility measurements and SV supported the measurements to obtain the solubility via HPLC. All authors contributed to writing the manuscript.

Acknowledgements

We are thankful for the support of the Ministry of Science, Research, and the Arts of Baden-Württemberg (MWK Baden-Württemberg, project MoMaF). We further acknowledge the support

of the Helmholtz research field information and the Karlsruhe Nano Micro Facility, which support the maintenance of the software Chemotion ELN. We acknowledge support by the Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of the Karlsruhe Institute of Technology. We thank Pierre Tremouilhac for supporting the project IT-wise.

Authors' information

Supporting information:

The supplementary information includes (1) details describing the generation of the training data set referring to the removal of duplicates, (2) lists of and information about the used descriptors, e.g. an overview of functional groups selected as descriptors in the dataset and their influence on the solubility of molecules, (3) model parameters of XGB-298, (4) a comparative analysis of the model XGB-298 with training and test data, (5) details on the duplicates that were removed from the training data of previous work to gain the results given in Table 3, entry 11b, (6) the experimental description of the solubility experiments performed in our labs, and (7) a short summary of the functionalities of the web service, that's provided based on the therein described model. Further figures, tables, and datasets are provided in the GitHub repository (https://github.com/ComPlat/water-solubility-prediction).

List of the solubility references from the literature

1. Llompart, P. *et al.* Will we ever be able to accurately predict solubility? *Sci Data* **11**, 303 (2024).
2. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv.*

Drug Deliv. Rev. **46**, 3–26 (2001).

3. An improved gravimetric method with anti-solvent addition to measure the solubility of d-allulose in water. *J. Food Eng.* **355**, 111582 (2023).
4. Jain, N. & Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **90**, 234–252 (2001).
5. Hückel, W. Solubility of non-electrolytes. Von Prof. Joel H. Hildebrand. 203 Seiten. Reinhold Publishing Corporation, New York 1936. Preis geb. $4,50. *Angew. Chem. Weinheim Bergstr. Ger.* **49**, 703–704 (1936).
6. Hansen, C. M. *Hansen Solubility Parameters: A User's Handbook, Second Edition*. (Taylor & Francis, 2007).
7. Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).
8. Klamt, A., Eckert, F., Hornig, M., Beck, M. E. & Bürger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **23**, 275–281 (2002).
9. Sorkun, M. C., Koelman, J. M. V. A. & Er, S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **24**, 101961 (2021).
10. 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.* **53**, 1563–1575 (2013).
11. Avdeef, A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. *ADMET DMPK* **8**, 29–77 (2020).
12. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **54**, 355–366 (2002).
13. Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).
14. Llinas, A., Oprisiu, I. & Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **60**, 4791–4803 (2020).

15. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).
16. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* **6**, 143 (2019).
17. Boobier, S., Hose, D. R. J., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **11**, 5753 (2020).
18. Panapitiya, G. *et al.* Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **7**, 15695–15710 (2022).
19. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777 (2000).
20. Cui, Q. *et al.* Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **10**, 121 (2020).
21. Huuskonen, J., Rantanen, J. & Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **35**, 1081–1088 (2000).
22. Balakin, K. V., Savchuk, N. P. & Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr. Med. Chem.* **13**, 223–241 (2006).
23. Yan, A. & Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **43**, 429–434 (2003).
24. Hou, T. J., Xia, K., Zhang, W. & Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **44**, 266–275 (2004).
25. Schroeter, T. S. *et al.* Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.* **21**, 485–498 (2007).

26. Ali, J., Camilleri, P., Brown, M. B., Hutt, A. J. & Kirton, S. B. In silico prediction of aqueous solubility using simple QSPR models: the importance of phenol and phenol-like moieties. *J. Chem. Inf. Model.* **52**, 2950–2957 (2012).
27. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717 (2017).
28. Bjerrum, E. J. & Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **8**, 131 (2018).
29. Website. https://aqsolpred.streamlit.app/.
30. Interactive ALOGPS Calculations at VCCLAB site. https://vcclab.org/web/alogps/.
31. Index-Home-ChemBCPP. http://chembcpp.scbdd.com/home/index/.
32. Tetko, I. V. *et al.* Virtual computational chemistry laboratory--design and description. *J. Comput. Aided Mol. Des.* **19**, 453–463 (2005).
33. Human Metabolome Database: Showing metabocard for Methylparaben (HMDB0032572). https://hmdb.ca/metabolites/HMDB0032572.
34. Human Metabolome Database: Showing metabocard for Eudesmic acid (HMDB0033839). https://hmdb.ca/metabolites/HMDB0033839.
35. Ethyl 3,4-dihydroxybenzoate (Ethyl protocatechuate). *MedchemExpress.com* https://www.medchemexpress.com/ethyl-3-4-dihydroxybenzoate.html.
36. Human Metabolome Database: Showing metabocard for Gallic acid (HMDB0005807). https://hmdb.ca/metabolites/HMDB0005807.
37. PubChem. Hazardous Substances Data Bank (HSDB) : 4012. https://pubchem.ncbi.nlm.nih.gov/source/hsdb/4012.
38. ChemBCPP: A freely available web server for calculating commonly used physicochemical properties. *Chemometrics Intellig. Lab. Syst.* **171**, 65–73 (2017).