Pushing the limits of solubility prediction via quality-oriented data selection

Murat Sorkun
Dutch Institute for Fundamental Energy Research https://orcid.org/0000-0002-5531-0802
J. M. Koelman
DIFFER
Süleyman Er (s.er@differ.nl)
DIFFER https://orcid.org/0000-0002-5005-3894

Article

Keywords: Aqueous Solubility Predictions, Actual and Observed Performance, Statistical Validation, Consensus Machine Learning

Posted Date: October 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-84771/v1

Version of Record: A version of this preprint was published at iScience on January 1st, 2021. See the published version at https://doi.org/10.1016/j.isci.2020.101961.

Pushing the limits of solubility prediction via quality-oriented data selection

Murat Cihan Sorkun[1,2,3], J. M. Vianney A. Koelman[1,2,3], and Süleyman Er[1,2*]

Accurate prediction of the solubility of chemical substances in solvents remains a challenge. The sparsity of high-quality solubility data is recognized as the biggest hurdle in the development of robust data-driven methods for practical use. Nonetheless, the effects of the quality and quantity of data on aqueous solubility predictions have not yet been scrutinized. In this study, the roles of the size and the quality of datasets on the performances of the solubility prediction models are unraveled, and the concepts of actual and observed performances are introduced. In an effort to curtail the gap between actual and observed performances, a quality-oriented data selection method, which evaluates the quality of data and extracts the most accurate part of it through statistical validation, is designed. Applying this method on the largest publicly available solubility database and using a consensus machine learning approach, a top-performing solubility prediction model is achieved.

INTRODUCTION

The solubility of chemical compounds in water is of fundamental interest, besides being a key property in the design, synthesis, performance, and functioning of new chemical motifs for various applications, including but not limited to drugs, paints, coatings, and batteries. Due to time, cost, and feasibility constraints on experimental measurements [1], it is usually not straightforward to obtain the solubility data of compounds rapidly. Moreover, considering the vastness of chemical space, where the total number of small molecules (with up to 36 heavy atoms) is approximated to reach 10^{33} [2], it is necessary to find alternative routes for the accelerated screening of candidate molecules with intended solubility values. Data-driven modeling holds the promise of making solubility predictions in a tiny fraction of a second. A data-driven model development consists of three main steps:

- Collecting and choosing train and test data
- Extracting and choosing key molecular descriptors
- Training and testing the model

In recent years, there has been a burgeon of efforts that apply the above steps for the development of data-driven solubility prediction models. Although data-driven solubility prediction models cater for achieving results quickly, they haven't yet widely been adopted in the community due to accuracy issues [3]. The factors that affect the performances of prediction models can be basically grouped into four categories (Fig. 1a), where the first two pertain to the data and the latter two pertain to the model [4]:

- The size of data
- The quality of data
- The relevance of chemical descriptors
- The capability of algorithm

Depending on the physical domain of the problem, the above factors may vary in their significance. In the case of solubility, the paucity of measurement data, in addition to the internal errors that result from the uncertainties in experimental procedures, is well-known. Thus, the size and quality of data have priority interest when improving the performance of solubility prediction models [5, 6, 7, 8, 9, 10]. The latter is generally accepted as the accuracy threshold of a model. In this context, Jorgensen and Duffy stated that the accuracy of a model cannot exceed the accuracy of the experimental data [6]. Although this statement is correct, it can further be consolidated since machine learning (ML) algorithms are capable of dealing with errors in the training data [11]. To put it differently, the observed performance of a model cannot be better than the internal error of the test set. To improve the capability of solubility prediction algorithms, it is therefore important to distinguish the actual and the observed performances of a model and to comprehend the factors affecting them. Fig. 1b shows a decomposition of the factors that affect the actual and observed performances of a model. We define the actual performance as the accuracy of the model that would be observed on a test set with zero internal error. In contrast, the observed error is the accuracy of the model demonstrated on an available test set with internal error (Fig. 1b). Obviously, when testing a model one can obtain only the observed performance. For instance, testing a perfect model, which by definition should predict absolute true values, on a test set with internal error of ϵ, will result in observed error of ϵ, despite the true error being zero. Therefore, the test quality sets the theoretical limit for the observed performance of the model. In domains where high-quality data is accessible, the gap between the actual and observed performances is small enough to be ignored. However, for the case of solubility, this gap has decisive importance and should be carefully treated.

In the current work, to develop an accurate solubility prediction model, we focus on the effects of data size and data quality on the prediction performance of ML models. Starting with the design of a quality-oriented data selection method that extracts the most accurate part of the data, and applying it on AqSolDB [12] - the largest publicly available solubility dataset that has been curated by using multiple data sources - the Aqueous

[1]DIFFER - Dutch Institute for Fundamental Energy Research, De Zaale 20, 5612 AJ Eindhoven, The Netherlands.
[2]CCER - Center for Computational Energy Research, De Zaale 20, 5612 AJ Eindhoven, The Netherlands.
[3]Department of Applied Physics, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. *email: s.er@differ.nl

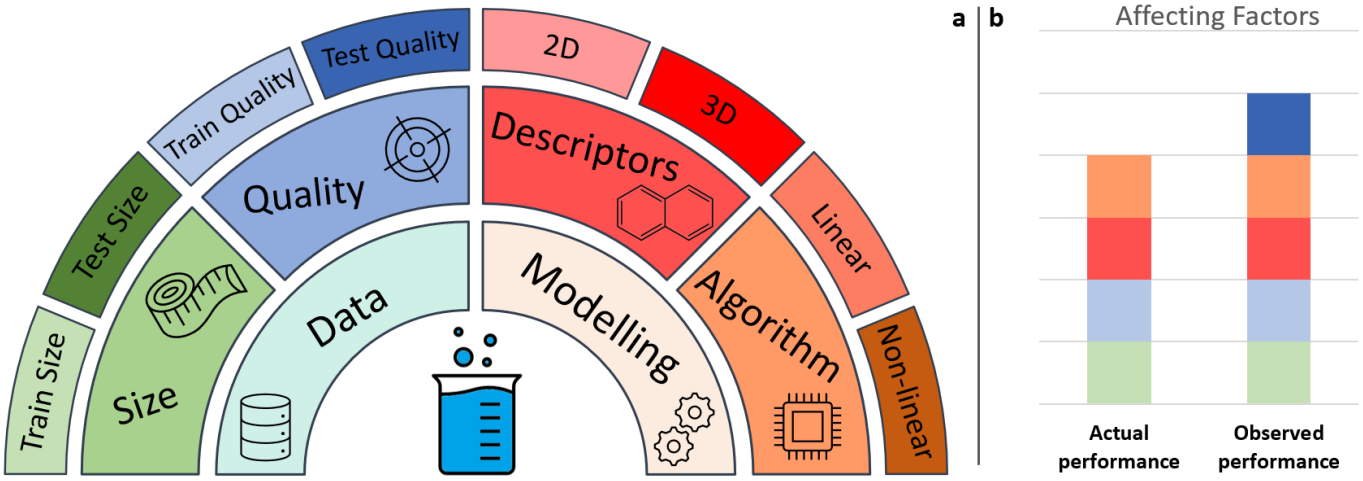

Fig.1 The categorization of the affecting factors for solubility predictions and their relationship with the actual and observed performances. a A three-layered structure showing the categorization of the affecting factors on the accuracy of solubility prediction ML models. **b** The representation of affecting factors shown by the colors in Fig. 1a on the actual and observed performances of solubility prediction models.

Solubility Prediction Model (AqSolPred) is developed. AqSolPred shows superior test performance when compared to available models on a conventionally used benchmark dataset [13]. In addition to quality-oriented data selection, AqSolPred comprises a consensus of three different ML algorithms, namely Artificial Neural Network (ANN), Random Forest (RF), and Extreme Gradient Boosting (XGB). Below we provide a detailed description of the development process, alongside the links to open-source codes and the data.

In the following paragraphs, we briefly review the principal factors that affect the accuracy of solubility predictions.

The size of data

It is a well-known fact that increasing the number of data instances in the training set has a positive effect on the accuracy of data-driven models. For instance, Lusci *et al.* trained four different UG-RNN models by using datasets with 1144, 1026, 74, and 125 instances, and obtained the respective Root Mean Squared Errors (RMSE) of 0.58, 0.60, 0.96, and 1.14 [14]. It should be noted that the size of the train and test sets yield different impacts. While the size of the training set affects the accuracy of the model, the size of the test set affects the reliable evaluation of the model's accuracy. A proper test set should be both large and diverse enough to cover the chemical space of the training set and to be minimally affected by outliers. Moreover, the solubility values of the test set should have a distribution similar to that of the training set. For example, one of the test sets [15] commonly used in the literature [5, 16, 17] consists of only 21 instances, which is not large enough for reliable testing. Since there had been very few solubility data publicly available, studies on solubility prediction have been limited with a few thousands of compounds for training and a few hundreds of compounds for testing [8, 17]. With an increase in public data resources, such as AqSolDB [12] consisting of a diverse set of $\sim 10^4$ compounds, it is becoming more feasible to conduct reliable testing studies to improve the accuracies of the data-driven models.

The quality of data

Performing high-quality solubility measurements is a difficult task due to uncertainties in experimental procedures, as explained in detail in Ref. [18]. Additionally, unintentional misprints, such as the erroneous conversions of values or units while carrying them from one source to another, cause deterioration in the quality of data. Unfortunately, not all solubility data sources provide uncertainty information on individual compounds or on the complete dataset. The generally accepted SD of public datasets is between 0.5 and 0.6 LogS [6, 8]. Recently, Avdeef has determined the average SD of 870 molecules from the Wiki-pS_0 database as 0.17 LogS [19], which is quite distant from the conceded values in literature. Therefore, we should keep in mind that the SD values are specific to data and they may differ significantly depending on the uncertainty of the measurement methods and the types of chemical compounds they contain. For example, lowly soluble compounds are extremely difficult to measure [9], thus the experimental errors in their measurements can be high. Accordingly, one expects that the datasets that contain many lowly soluble compounds to have high SDs. Therefore, it is essential to determine the quality of the datasets prior to the development of supervised ML models.

Similar to data size, the quality of the train and the test sets have distinct effects on the performance, and therefore on the assessment of the model. Test set quality regulates the theoretical limit of observed performance (Fig. 1). Therefore, to correctly evaluate the performance of a model, it is vital to use a high-quality test set. For instance, in a recent solubility prediction challenge [20], two test sets of different qualities: high quality (SD:0.17 LogS) and low quality (SD:0.62 LogS), have been shared and the participants were invited to predict the solubility of compounds by using their own training datasets and methods. From a total of 37 different methods, the average RMSE for the low- and the high-quality datasets were 1.62 and 1.14 LogS, respectively. All the prediction models performed worse on the low-quality data and better on the high-quality data. This result shows the importance of test set quality on the observed performance of the models. While the test set quality affects only the

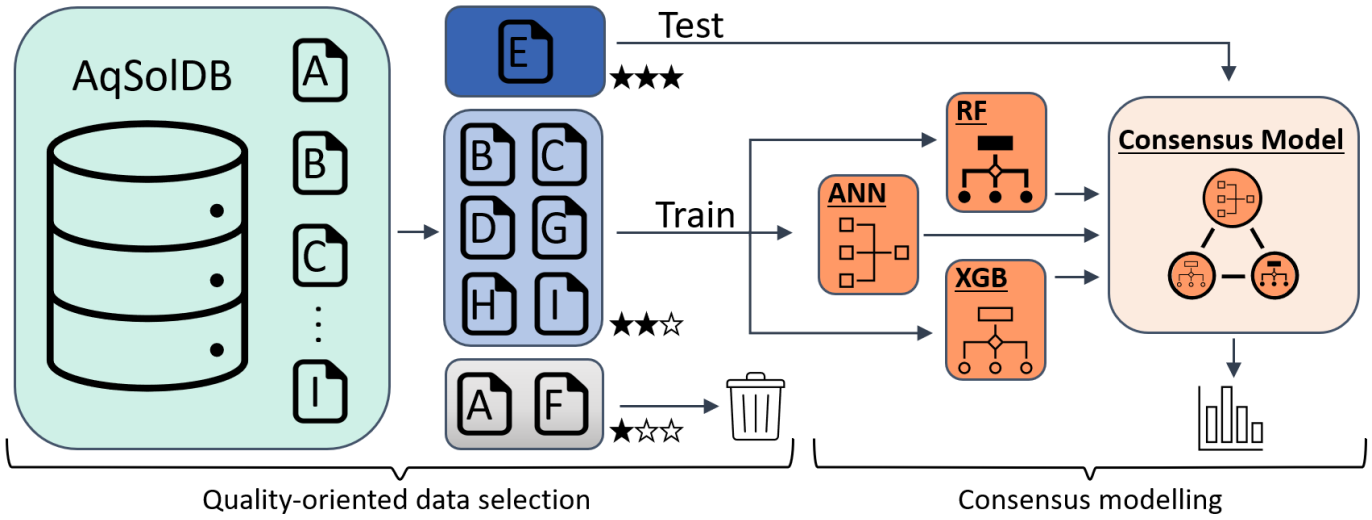

Fig.2 The development phases of AqSolPred. The quality-oriented data selection method that selects the test and training data based on their quality levels as indicated by stars (*left*). The development of the consensus model based on ANN, RF, and XGB, and its processes of training and testing (*right*).

observed performances of the models, the training set quality affects both the actual and observed performances. However, the internal errors of the training sets are partly compensated by capable ML algorithms depending on the size and the diversity of data. Thus, the effects of the internal errors of the training sets on the models' performances are usually smaller than the internal errors themselves.

The relevance of chemical descriptors

Descriptors provide a mathematical representation of the chemical information contained in a compound. They are valuable inputs for data-driven models aimed at the prediction of chemical properties. Descriptors can be classified into two groups; 2D and 3D. Basically, all the descriptors that require 3D optimization of the structure are considered as 3D descriptors while the remaining are considered as 2D descriptors. There are several publicly available resources to calculate molecular descriptors [21, 22, 23]. Most 2D descriptors are calculated with absolute accuracy while the 3D descriptors carry the errors of the methodological approximations they have been calculated with [24]. Admitting that the 3D descriptors provide more detailed information, such as atomic distances and energy data of the compounds, there is yet no clear evidence about their impacts on the solubility predictions [8, 25, 26, 27]. Although a large number of chemical descriptors are available, it's usually preferred to use a modest number of relevant descriptors to avoid redundancy and overfitting issues during the training of ML models [10].

The capability of the algorithms

The earlier methods for solubility prediction were based on simple linear regression (LR) methods [16, 28, 29, 30] and used only a few descriptors, such as lipophilicity (LogP), melting point, and molecular weight. While these methods are easy to apply and interpret, their predictive power is rather limited since the LR works only for linear dependencies. In the last years, ML algorithms, such as the variations of ANNs and tree-based ensembles, proved their ability on solving complex problems in various research fields, also including the solubility predictions [5, 13, 14, 31, 32, 33]. Due to their black-box nature, these algorithms are hard to interpret by humans. Moreover, they require large datasets and expert domain knowledge to circumvent the overfitting issues. As ML algorithms are properly configured and fed with sufficient amount of data, they become more competent in solubility predictions. Compared to the individual models, consensus modelling that combines the predictions of different models [34] with an aim to compensate the weaknesses of each model, show improved performances [7, 35, 36, 37]. Additionally, the variances in the predictions of the constituting models provide valuable information about the prediction uncertainties.

RESULTS

Quality assessment of the solubility datasets

The data selection and model development phases of the AqSolPred are shown in Fig. 2. For train and test purposes, AqSolDB that merges nine different sub-datasets, named from *A* to *I*, is used (Table 1). A more detailed information about the sub-datasets is provided in Ref. [12], alongside the publicly accessible database [38] and the source code including the steps for data curation [39].

As explained above, the train and test data affect the actual and observed performance of the models differently. Therefore, instead of using all available data directly, we applied a quality-oriented selection procedure for the training and test data. We determined the quality of each sub-dataset in terms of the SD of multi-lab measurements as described in the Methods. The total number of multi-lab measurements (N(SD)) and the calculated SDs are shown in Table 1. The SDs of the nine sub-datasets vary significantly, with numerical values between 0.274 and 0.717 LogS. The dataset E has the lowest SD and therefore is considered to contain the highest quality data. Adversely, the datasets *A* and *F* have the largest SDs. The SDs of the remaining datasets are close to each other and all are < 0.4 LogS.

Fig.3 Visualization of the chemical space covered by the training and test data. The chemical space is visualized by the t-SNE dimensionality reduction technique. Blue hexagons show the chemical space that is covered by the training data, whereas the red dots show the test instances in the chemical space. The color scale on the right shows the density of molecules found in the hexagons.

Table 1 | The SD of AqSolDB and its sub-datasets

Dataset	Size	Filtered Size	*N*(SD)	SD
A	6110	3266	3093	0.717
B	4651	3185	1215	0.372
C	2603	1798	668	0.380
D	2115	1054	179	0.361
E	1291	1290	337	0.274
F	1210	1011	202	0.582
G	1144	363	170	0.392
H	578	148	100	0.383
I	94	62	46	0.338
All	9982	6937	-	0.495
non-AF	6154	4399	-	0.356

Size: number of instances before pre-processing, **Filtered size:** number of instances after pre-processing, ***N*(SD):** total number of multiple values used to calculate SD, **SD:** standard deviation

Selection of the test and the training datasets

For a proper evaluation of the model, the observed performance of the model should approach the actual performance as explained above. Therefore, the test data should be of the highest possible quality. Additionally, it should be large enough to cover the chemical space of the training set. We selected dataset *E* as the test set, since it has the highest quality among the sub-datasets. It is important to note that, dataset *E* is also known as the Huuskonen dataset, which is commonly used in literature as a benchmark dataset. Using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique [40], we validated that dataset *E* largely covers a reduced chemical space of the training data (Fig. 3). We also validated that the distribution of the solubility values of dataset *E* is compatible with the training set (Supplementary 1 Fig.S1). After reserving dataset *E* as the test set, we also removed the two sub-datasets, *A* and *F* that have large SDs, and used the remaining datasets to generate a high-quality training set, namely the *non-AF*. The SD of the *non-AF* dataset has been calculated by incorporating the SDs of the constituent sub-datasets into Eq. 2. For comparison, we also calculated the SD of the entire AqSolDB, namely the *All*, using the same procedure (Table 1).

Effect of quality and size of the training set

As discussed above, both the size and the quality of training set are positively correlated with a model's accuracy. However, quality-oriented data selection decreases the size of the data while increasing the quality. To analyze the trade-off between size and quality, we developed separate models for each solubility sub-dataset. For a fair comparison, we trained sub-datasets with the same combinations of feature selection methods and ML algorithms explained in the Methods. We selected the best configurations based on 10-fold cross-validation performances

Table 2 | Comparison of AqSolPred to literature results

Year	Model	Method	Total size	Test size/method	MAE	RMSE	R2	Reference
2000	Huuskonen	ANN	1294	413	-	0.600	0.92	[13]
2000	Huuskonen	MLR	1294	413	-	0.710	0.88	[13]
2001	Tetko	ANN	1291	412	-	0.620	0.91	[5]
2003	Yan	MLR	1294	496	0.680	0.790	0.82	[31]
2003	Yan	ANN	1294	496	0.490	0.590	0.92	[31]
2004	Delaney*	MLR	1290	1290	0.685	0.876	0.71	[16]
2004	Hou	MLR	1294	412	0.520	0.630	0.90	[41]
2007	Schroeter	GP	1290	3 fold CV	0.412	0.579	-	[32]
2007	Schroeter	RR	1290	3 fold CV	0.586	0.996	-	[32]
2007	Schroeter	SVM	1290	3 fold CV	0.431	0.600	-	[32]
2007	Schroeter	RF	1290	3 fold CV	0.485	0.660	-	[32]
2012	Ali*	MLR	1290	1290	0.728	0.940	0.73	[42]
2013	Lusci	UG-RNN	1026	10 fold CV	0.460	0.600	0.91	[14]
2016	Filter-it*	MLR	1290	1290	0.893	1.154	0.68	[43]
2018	Bjerrum	ANN	1297	10 fold CV	-	0.650	0.90	[44]
2020	Tang	MPN	1310	10 fold CV	-	0.661	-	[33]
2020	AqSolPred	Consensus	1290	1290	**0.397**	**0.539**	**0.93**	-
2020	AqSolPred	Consensus	1290	LOO	**0.348**	**0.483**	**0.94**	-

ANN: Artificial neural networks, **MLR:** Multiple linear regression, **GP:** Gaussian processes, **RR:** Ridge regression, **SVM:** Support vector machine, **RF:** Random forest, **UG-RNN:** Undirected graph-recursive neural networks, and **MPN:** Message parsing neural network, and **Consensus:** an ensemble of ANN, RF, and XGB. ***:** Results collected from SwissADME web tool [43]

of each of the sub-datasets. We trained the final models using their best configurations and the entire training data. After ensuring that no test compounds were used in the training process (see Methods), we tested the performances of the final models against the test dataset *E* (Fig. 4). To understand the effect of data quality in predictions, we compared the datasets of similar size, *A*-*B* and *D*-*F*, and found that those having higher quality perform significantly better than those having lower quality. To understand the effect of size, we compared the datasets of similar quality. First, we compared datasets *B*, *C*, and *D*, with 3185, 1798, and 1054 instances, respectively. The test performances of these three datasets are very close, within ~ 0.1 LogS (Fig. 4). Secondly, we compared datasets *G*, *H*, and *I*, whose qualities are similar but the sizes are 363, 148, and 62, respectively. This time the size effect is more obvious, as the accuracy decreases when the size of the data becomes smaller (Fig. 4). Despite having the lowest SD within the group of training sub-datasets, *I* shows the lowest accuracy due to its small size. According to these results, we conclude that the data size is more influential on small-sized datasets with a few hundred or fewer instances, while the data quality is more effective on large-sized datasets with thousands of instances.

The quality-oriented data selection dataset, *non-AF*, shows superior performance among all datasets by virtue of its quality, despite the fact that this dataset has 2617 fewer instances than the largest dataset *All*. So far all the models have been developed without using any compounds from dataset *E*. To quantify the impact of including this high-quality data, in a new experiment we included dataset *E* into the training process. We applied the leave-one-out (LOO) cross-validation method and left out a single compound at a time from dataset *E* for validation and included the remaining compounds in the training data. This process was repeated for each molecule in dataset *E*. As expected, the inclusion of dataset *E* improved the performance as shown by the bottom two rows in Table 2. Furthermore, we conducted experiments by oversampling the highest quality data, but since this did not result in noteworthy improvements we have not included them here.

These results show that both the quality and the size have major impacts on the solubility prediction performances of the ML models. Moreover, instead of a direct use of all the available data for training, a quality-oriented data selection method empowers the model.

Effect of descriptors and algorithms

We used a total of 123 2D descriptors for which the groupings, sizes, and use cases from literature are shown in Table 3.

To pick out a minimum number of relevant descriptors, we independently applied the LASSO and PCC feature selection methods as described in the Methods. The definitions and the correlation matrix of these descriptors are shown in Supplementary 1 Table S1 and Fig S1, respectively. The cross-validation results of the various configurations show that the LASSO performs slightly better than the PCC. Using the former method, a total of 58 descriptors have been selected.

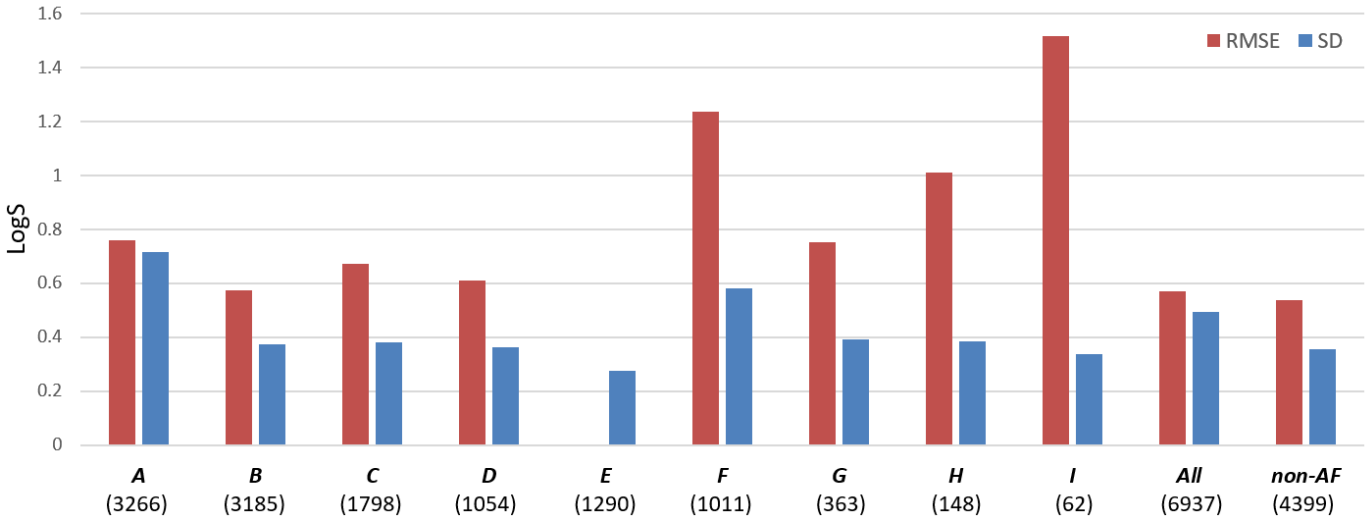

Fig.4 The quality and accuracy comparison of the sub-datasets. Blue bars show the SD of sub-datasets, whereas the red bars show the test performances (RMSE) on dataset *E* of the models that have been trained by that sub-dataset. Both the SD and RMSE are given units of LogS. The total number of data instances that have been used to train the models are shown for each sub-dataset.

Table 3 | The sets of chemical descriptors

Descriptor sets	Size	References
Atom-based	19	[14, 18, 26, 32, 33, 41]
Ring-based	6	[18, 26, 33]
Bond-based	9	[6, 14, 16, 18, 26, 33, 37]
LogP	1	[14, 16, 18, 26, 32, 37, 42]
Topological	18	[6, 13, 18, 26, 32, 37, 42]
E-state indices	70	[5, 13, 18]

Trained on each of the datasets, a consensus model that combines three different ML algorithms (ANN, RF, and XGB) as described in Methods, exceeds the performance of any of the singular models that have been trained by a single algorithm. Also importantly, using a consensus model it is possible to collect additional uncertainty information, whereas using the individual algorithms independently does not provide this information. This is because the SDs from different model predictions are good indicators for the uncertainties observed in the final predictions. The configurations of the different ML models and their results are shown in Supplementary 2.

Performance comparisons of the model with the literature

The AqSolPred shows the highest accuracy on the Huuskonen dataset (i.e. dataset *E*), when compared to the available results from the literature on solubility predictions (Table 2). Due to the differences in pre-processing steps, the total number of data instances that have been used by each method differs slightly as shown in Table 2. Furthermore, some studies have used cross-validation techniques while the others have divided data into train and test sets (Table 2).

DISCUSSION

A cardinal result of the current study is the differentiation of actual and observed performances of the solubility models. Because the observed performance is highly sensitive to the quality of the test set, when the test data contains high uncertainty, the difference between actual and observed performances becomes more pronounced. Therefore, it is imperative to use high-quality data in testing to obtain an observed performance close to the actual performance of a model. For this reason, the quality assessment prior to training and testing experiments constitutes a vital step. The generally employed assumptions on the SDs of experimental datasets (e.g. such as up to 0.6 LogS error) are fuzzy and they do not necessarily reflect the true quality of datasets (see Table 1). Instead, comparing multi-lab measurement data of compounds provides a way to estimate the solubility data quality. For instance, in the current study, we collected a total of 6010 multi-lab measurements on 2236 unique compounds from nine different sources. We identified duplicates based on their InChIKeys, a safe way to match the exact compounds. Considering that the different datasets may contain compounds from the same source, as an early procedure, the duplicates should be identified to ensure the usage of the same information only once. An added value of comparisons between multi-lab values, next to that of determining the quality of the datasets, is the detection of outliers in data, such as the ones caused by misprints.

A second conclusion is the impact of training size on the accuracy of data-driven models. We found that, regardless of their quality, the small-sized datasets do not include the generic information to address the solubility problem and they do not adequately cover the chemical space of the test data. Therefore, we recommend that extra care should be taken when reaching conclusions based on models that have been trained with small-sized datasets.

Data diversity is another important concept that designates the applicability domain of ML models. In addition to being sufficiently large as explained above, a good training set should also have a high ratio of the data size over the chemical diversity of compounds. In the case of the test data, it should cover the chemical domain defined by the training set. Visualizing the data in two-dimensions allows for inspecting to what extend the test set covers the chemical compound space of the training set. Dimensionality reduction methods (e.g. t-SNE [40] and UMAP [45]) provide interpretable 2D graphs by clustering the

chemical compounds based on their local similarities. Defining the chemical space based on tailored similarities and using only the relevant descriptors of target properties, provides a better representation than using arbitrary similarities such as the predefined fingerprints [46].

During the prediction of aqueous solubility data of compounds here, the observed superior performance of a consensus model over the singular models promises that there is still room for algorithmic improvements to further improve the accuracies in solubility predictions of the compounds. When building a consensus model, increasing the number of constituent algorithms would generate more accurate predictions by facilitating the elimination of the outliers before merging the prediction results. Moreover, the uncertainty information obtained from multiple predictions would be more reliable. Lastly, since they are modeling the problem from different aspects, bringing fundamentally diverse algorithms into play would provide better results compared to using the same stochastic algorithm multiple times with different initializations.

In summary, applying a quality-oriented data selection method, employing 58 LASSO-selected 2D descriptors and an ensemble of advanced ML algorithms, we developed the AqSolPred, a high-calibre solubility prediction model.

METHODS

Quality-oriented data selection

Quality-oriented data selection identifies the quality of datasets by calculating the deviations in the multi-lab experimental measurements of the compounds. Using the quality information, the highest quality dataset is reserved as the test set and the poor quality datasets are removed from the training set. To assess the quality of each dataset, the following steps have been applied:

- Compounds that have multi-lab measurement data have been identified.
- The average of the measured solubility values of compounds have been calculated.
- The deviations of measurement data from the average values have been calculated.
- The SDs of the constituting datasets have been calculated.

The SDs for each dataset (from *A* to *I*) have been calculated using Eq. 1:

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}} \tag{1}$$

where n is the total number of compounds that have multi-lab measurement data, x_i is the experimentally measured solubility value of compound i, and $\bar{x}$ is the average of multi-lab solubility values of the compound.

The SDs of the combinatorial datasets (i.e. "*non-AF*" and "*All*") have been calculated using Eq. 2:

$$SD = \frac{1}{N}\sum_{j=1}^{Z} SD_j T_j \tag{2}$$

where N is the total number of compounds in the dataset, Z is the total number of constituent datasets, SD_j is the SD of dataset j, and T_j is the total number of compounds that have been included from dataset j.

Data pre-processing

To prepare the datasets for training, we removed the compounds from datasets when they met any of the following criteria:

- The compound exists in the test set (dataset *E*).
- The compound does not contain carbon atom.
- The compound contains adjoined mixtures.
- The compound contains charged atoms.

The remaining numbers of compounds found in each training sub-dataset, obtained after the completion of data preprocessing, have been shown in Table 1 (Filtered Size).

Descriptor selection

To generate the molecular descriptors, we used the Mordred Python package [23]. Currently, there are more than 1800 2D and 3D descriptors in the Mordred catalog. To determine the most relevant descriptors, we applied the following feature selection methods:

- **Least absolute shrinkage and selection operator (LASSO):** A regression analysis method that enhances the prediction accuracy and interpretability of the statistical model. To learn the best descriptors (i.e. variables) the LASSO regularization eliminates the irrelevant descriptors by forcing their coefficients to zero.
- **Pearson correlation coefficient (PCC):** Selects the descriptors that have PCC with LogS higher than a defined threshold parameter.

For both methods, we tested different parameter sets that change the strictness of selections. The results of these different configurations are provided in Supplementary 2.

Machine learning algorithms

We employed the following ML algorithms in combination with the scikit-learn and xgboost Python packages.

- Artificial neural network (ANN)
- Random forest (RF)
- Extreme gradient boosting (XGB)

ANN is a network consisting of several layers that are connected to each other through the neurons it contains. ANN learns non-linear functions by modifying the coefficients between neurons via a back-propagation algorithm. In the current work, the ANN configuration employs single hidden layer with 500 neurons and a *tanh* activation function. RF is an ensemble of decision trees that use bootstrap aggregating of the instances and a random sampling of the features. Our RF configuration consists of 1000 trees with the maximum depth. XGB is a regularized gradient boosting algorithm that creates a strong learner from an ensemble of many weak trees that are trained sequentially. Our XGB configuration consists of 1000 trees with

a maximum depth of six. Other parameters of the models are used with their default values. Lastly, our consensus model is based on a combination of the above three ML models and an arithmetic averaging of the predictions by these models.

Configuration of the AqSolPred
The best performing AqSolPred model has been achieved by using the following configuration:

- **Training set:** *non-AF* (4399 data instances)
- **Features:** 58 2D descriptors as selected by LASSO with $\alpha = 0.01$
- **ML Algorithm:** A consensus of ANN, RF, and XGB models

Chemical space visualization
We used tailored similarity for the visualization of the chemical space based on 58 LASSO-selected descriptors. We applied t-SNE from scikit-learn Python package to reduce the data into two-dimensions with the following two parameters, while the remaining parameters are used with their default values:

- **Perplexity:** 50
- **Random state:** 1

Code Availability
The reproducibility of the AqSolPred can be verified by executing the provided scripts on Code Ocean [47].

References

[1] Murdande, S. B., Pikal, M. J., Ravi, M. & Bogner, R. H. Aqueous solubility of crystalline and amorphous drugs: challenges in measurement. *Pharmaceutical development and technology* **16**, 187–200 (2011).

[2] Polishchuk, P. G., Madzhidov, T. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* **27**, 675–679 (2013).

[3] Jouyban, A. *Handbook of solubility data for pharmaceuticals* (CRC Press, Boca Raton, 2010).

[4] Haghighatlari, M. *et al.* Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **6**, 1527-1542 (2020).

[5] Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *Journal of chemical information and computer sciences* **41**, 1488–1493 (2001).

[6] Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Advanced drug delivery reviews* **54**, 355–366 (2002).

[7] Bergström, C. A. S., Carola, M., Norinder, U., Luthman, K. & Artursson, P. Global and local computational models for aqueous solubility prediction of drug-like molecules. *Journal of chemical information and computer sciences* **44**, 1477–1488 (2004).

[8] 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. *Current medicinal chemistry* **13**, 223–241 (2006).

[9] Hewitt, M. *et al.* In silico prediction of aqueous solubility: the solubility challenge. *Journal of chemical information and modeling* **49**, 2572–2587 (2009).

[10] Wang, J. & Hou, T. Recent advances on aqueous solubility prediction. *Combinatorial chemistry & high throughput screening* **14**, 328–338 (2011).

[11] Kordos, M. & Rusiecki, A. Reducing noise impact on MLP training. *Soft computing* **20**, 49–65 (2016).

[12] Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific data* **6**, 1–8 (2019)

[13] Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of chemical information and computer sciences* **40**, 773–777 (2000).

[14] Lusci, A., Pollastri, G. & Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling* **53**, 1563–1575 (2013).

[15] Yalkowsky, S. H. & Banerjee, S. *Aqueous solubility: Methods of estimation for organic compounds*. (Marcel Dekker, New York, 1992).

[16] Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences* **44**, 1000–1005 (2004).

[17] Dearden, J. C. In silico prediction of aqueous solubility. *Expert opinion on drug discovery* **1**, 31–52 (2006).

[18] Avdeef, A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. *ADMET and DMPK* **8**, 29–77 (2020).

[19] Avdeef, A. Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. *ADMET and DMPK* **7**, 210–219 (2019).

[20] Llinas, A., Oprisiu, I. & Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility Dedicated to the memory of Anton J. Hopfinger and Oleg A. Raevsky. *Journal of chemical information and modeling* **–**, – (2020).

[21] Landrum, G. RDKit: open-source cheminformatics. (2006).

[22] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* **32**, 1466–1474 (2011).

[23] Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *Journal of cheminformatics* **10**, 4 (2018).

[24] Raevsky, O. A., Grigorev, V. Y., Polianczyk, D. E., Raevskaja, O. E. & Dearden, J. C. Aqueous Drug Solubility: What Do We Measure, Calculate and QSPR Predict?. *Mini reviews in medicinal chemistry* **19**, 362–372 (2019).

[25] Gao, K. Are 2D fingerprints still valuable for drug discovery?. *Physical chemistry chemical physics* **22**, 8373–8390 (2020).

[26] Yan, A., Gasteiger, J., Krug, M. & Anzali, S. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *Journal of computer-aided molecular design* **18**, 75–87 (2004).

[27] Salahinejad, M., Le, T. C. & Winkler, D. A. Aqueous solubility prediction: do crystal lattice interactions help?. *Molecular pharmaceutics* **10**, 2757–2766 (2013).

[28] Hansch, C., Quinlan, J. E. & Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *The journal of organic chemistry* **33**, 347–350 (1968).

[29] Yalkowsky, S. H. & Valvani, S. C. Solubility and partitioning I: solubility of nonelectrolytes in water. Journal of pharmaceutical sciences. *Journal of pharmaceutical sciences* **69**, 912-922 (1980).

[30] Meylan, W. M., Howard, P. H. & Boethling, R. S. Improved method for estimating water solubility from octanol/water partition coefficient. *Environmental toxicology and chemistry: An International Journal* **15**, 100–106 (1996).

[31] Yan, A. & Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *Journal of chemical information and computer sciences* **43**, 429–434 (2003).

[32] Schroeter, T. S. *et al.* Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *Journal of Computer-aided molecular design* **21**, 485–498 (2007).

[33] Tang, B. *et al.* A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics* **12**, 1–9 (2020).

[34] Todeschini, R, Consonni, V., Ballabio, D. & Grisoni, F. Chemometrics for QSAR Modeling. *In: Comprehensive Chemometrics (Second Edition)* 599 - 634 (Elsevier, Oxford, 2020).

[35] Abshear, T., Banik, G. M., D'Souza, M. L., Nedwed, K. & Peng, C. A model validation and consensus building environment. *SAR and QSAR in environmental research* **17**, 311–321 (2006).

[36] Chevillard, F. *et al.* In silico prediction of aqueous solubility: a multimodel protocol based on chemical similarity. *Molecular pharmaceutics* **9**, 3127–3135 (2012).

[37] Raevsky, O. A., Polianczyk, D. E., Grigorev, V. Y., Raevskaja, O. E. & Dearden, J. C. In silico prediction of aqueous solubility: A comparative study of local and global predictive models. *Molecular informatics* **6-7**, 417–430 (2015).

[38] Sorkun, M. C., Khetan, A. & Er, S. AqSolDB: A curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Harvard dataverse* https://doi.org/10.7910/DVN/OVHAW8 (2019).

[39] Sorkun, M. C., Khetan, A. & Er, S. AqSolDB (Aqueous Solubility Data Curation). *Code ocean* https://doi.org/10.24433/CO.1992938.v1 (2019).

[40] Maaten, L. V. D. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579–2605 (2008).

[41] 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. *Journal of chemical information and computer sciences* **44**, 266–275 (2004).

[42] 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. *Journal of chemical information and modeling* **52**, 2950–2957 (2012).

[43] Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports* **7**, 42717 (2017).

[44] Bjerrum, E. J. & Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **8**, 131 (2018).

[45] McInnes, L., Healy, J., & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2020).

[46] Gute, B. D., Basak, S. C., Mills, D. & Hawkins, D. M. Tailored similarity spaces for the prediction of physicochemical properties. *Internet electronic journal of molecular design* **1**, 374–387 (2002).

[47] Sorkun, M. C. AqSolPred (Aqueous Solubility Prediction model based on quality-oriented data selection method). *Code ocean* https://codeocean.com/capsule/1699487/tree/v1 (2020).

Acknowledgements
The authors acknowledge funding from the initiative "Computational Sciences for Energy Research" of Shell and the Netherlands Organisation for Scientific Research (NWO) grant no 15CSTT05. SE acknowledges funding from NWO, through the COLORFLOW project partnership of DIFFER and Green Energy Storage, in the framework of the Materials for Sustainability programme and from the Ministry of Economic Affairs in the framework of the "PPS-Toeslagregeling" grant no 739.017.013. This work was sponsored by NWO Exact and Natural Sciences for the use of supercomputer facilities.

Author contributions
M.C.S. developed all the codes of AqSolPred and performed the experiments. S.E. supervised the project. All authors contributed to the analysis of results and the writing of the manuscript.

Competing interests
The authors declare no competing interests.

Additional Information
Two separate supplementary files are available for this paper.

Figures

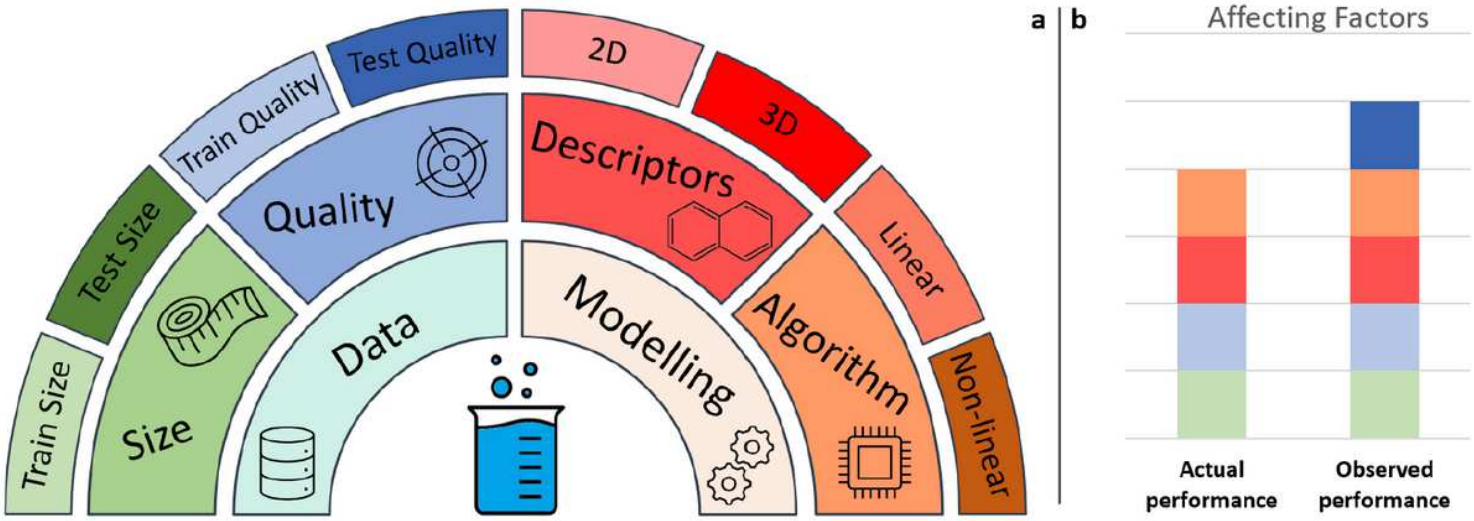

Figure 1

The categorization of the affecting factors for solubility predictions and their relationship with the actual and observed performances. a A three-layered structure showing the categorization of the affecting factors on the accuracy of solubility prediction ML models. b The representation of affecting factors shown by the colors in Fig. 1a on the actual and observed performances of solubility prediction models.

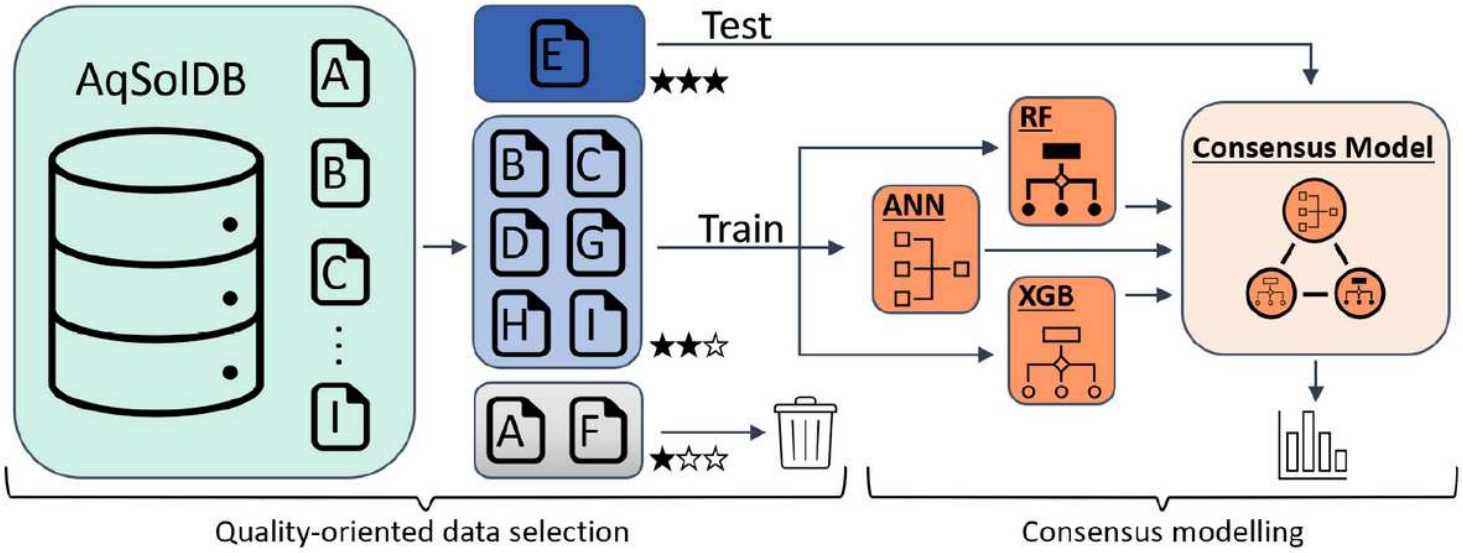

Figure 2

The development phases of AqSolPred. The quality-oriented data selection method that selects the test and training data based on their quality levels as indicated by stars (left). The development of the consensus model based on ANN, RF, and XGB, and its processes of training and testing (right).

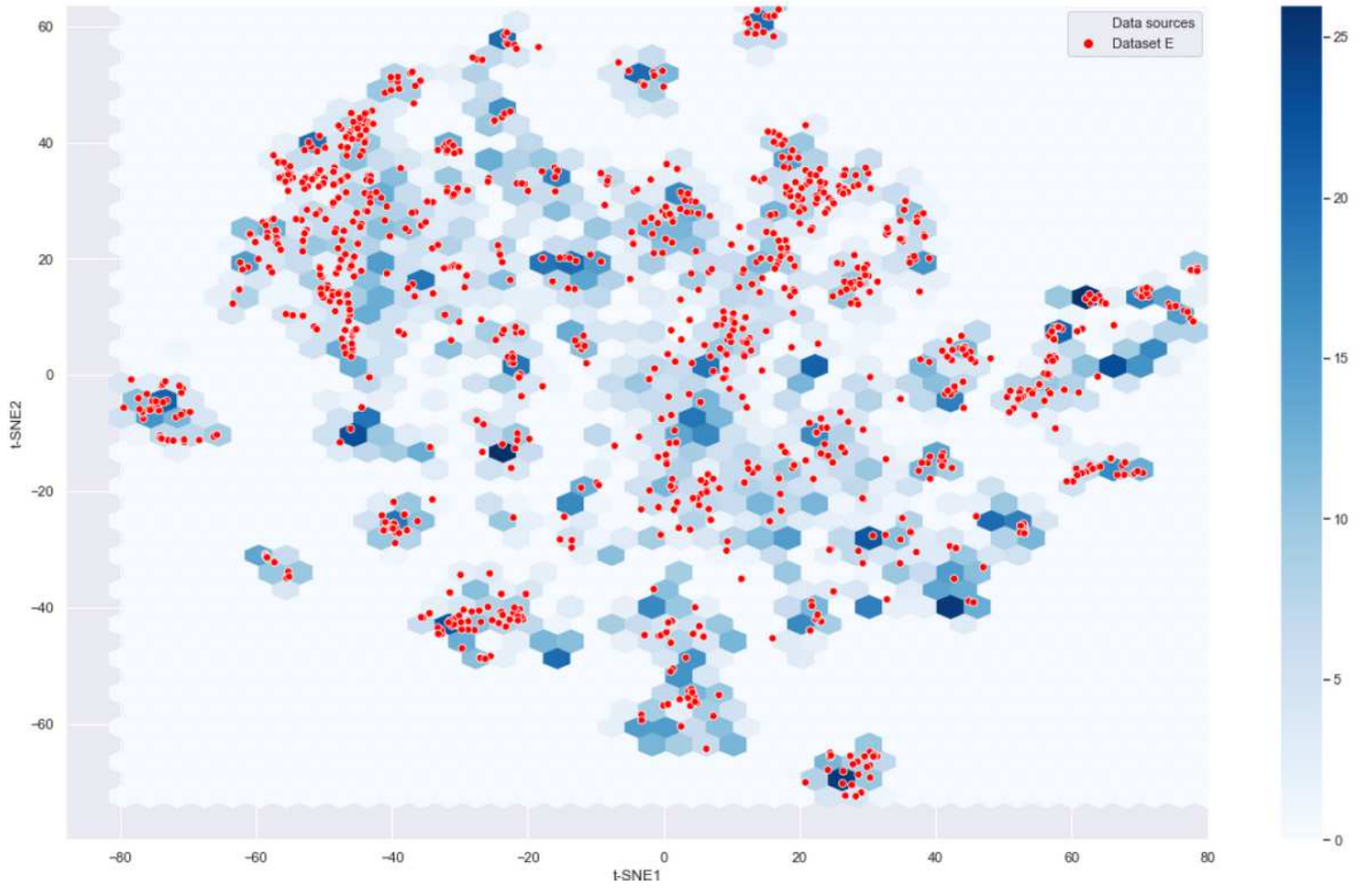

Figure 3

Visualization of the chemical space covered by the training and test data. The chemical space is visualized by the t-SNE dimensionality reduction technique. Blue hexagons show the chemical space that is covered by the training data, whereas the red dots show the test instances in the chemical space. The color scale on the right shows the density of molecules found in the hexagons.

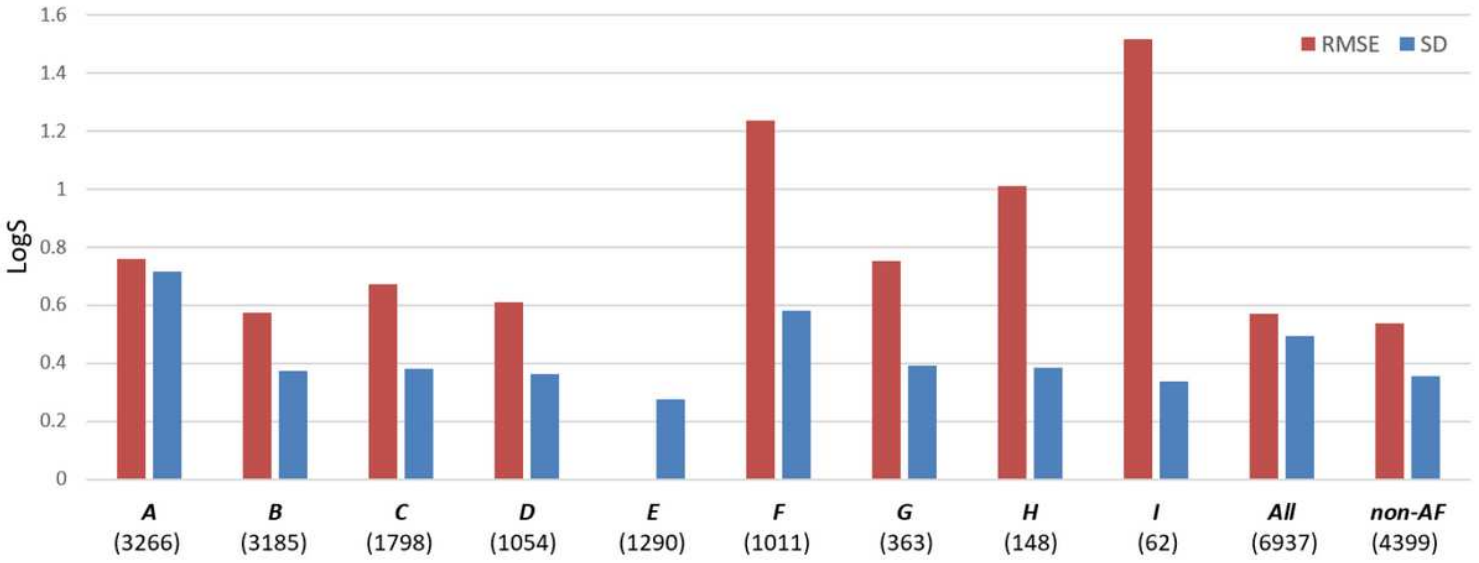

Figure 4

The quality and accuracy comparison of the sub-datasets. Blue bars show the SD of sub-datasets, whereas the red bars show the test performances (RMSE) on dataset E of the models that have been trained by that sub-dataset. Both the SD and RMSE are given units of LogS. The total number of data instances that have been used to train the models are shown for each sub-dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20200928SI1.pdf
- 20200928SI2.pdf