JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

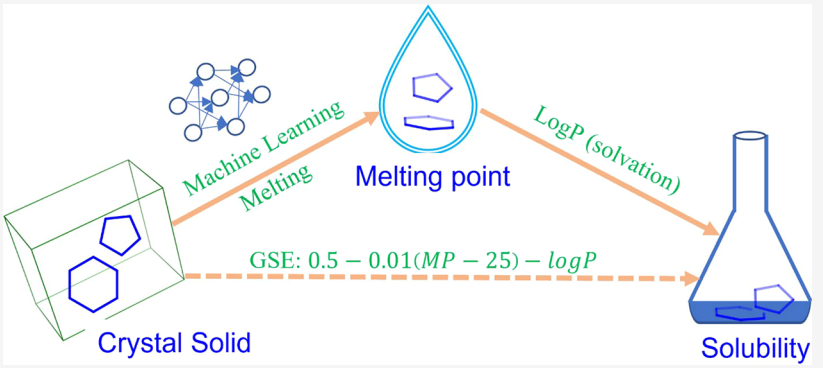
pubs.acs.org/jcim Article

Building Machine Learning Small Molecule Melting Points and Solubility Models Using CCDC Melting Points Dataset

Xiangwei Zhu, Valery R. Polyakov,* Krishna Bajjuri, Huiyong Hu, Andreas Maderna, Clare A. Tovee, and Suzanna C. Ward

Cite This: J. Chem. Inf. Model. 2023, 63, 2948−2959 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Predicting solubility of small molecules is a very difficult undertaking due to the lack of reliable and consistent experimental solubility data. It is well known that for a molecule in a crystal lattice to be dissolved, it must, first, dissociate from the lattice and then, second, be solvated. The melting point of a compound is proportional to the lattice energy, and the octanol−water partition coefficient (log *P*) is a measure of the compound's solvation efficiency. The CCDC's melting point dataset of almost one hundred thousand compounds was utilized to create widely applicable machine learning models of small molecule melting points. Using the general solubility equation, the aqueous thermodynamic solubilities of the same compounds can be predicted. The global model could be easily localized by adding additional melting point measurements for a chemical series of interest.

INTRODUCTION

Aqueous solubility is a key physicochemical attribute required for the characterization of an active pharmaceutical ingredient during drug discovery and development. A drug molecule needs to be sufficiently soluble to quantitate meaningful activities in in vitro assays as well as to achieve adequate absorption for safety and efficacy evaluation in preclinical models and clinical trials.[1] Hydrophilic payloads could facilitate their conjugation to the antibody in aqueous buffers and increase the solubility of antibody drug conjugates to improve their pharmacokinetics.[2] High throughput solubility measurements have been developed for quick determination of kinetic and thermodynamic (equilibrium) solubility.[3] However, using them to measure a new chemical's solubility with useful accuracy is still a practical problem.[4] The gold standard of saturation shake-flask measurement is low throughput, and it would take state-of-the-art techniques to achieve a low interlaboratory variability.[5] Factors like crystallinity, polymorphism, compound purity, particle size, and so forth all impact interlaboratory reproducibility. Quite a few small to medium size solubility datasets from different labs are available and much effort has been taken to curate solubility data measured with different methods from these sources.[6,7] The curated solubility datasets show large variance, for example, a set of 819 compounds compiled from multiple sources (≥3) in AqSolDB[6] have an average standard deviation (SD) of 0.36 (log base 10 transformed molar solubility) and 183 have an SD of at least 0.5. There is a significant variability of 9−30% in solubility values among some popular datasets.[7] Machine learning models have sometimes achieved high predictive performance on their own test sets,[8−13] but the intrinsic discrepancy of measurements, large interlaboratory variance, and limited domain applicability can make them untrustworthy to be implemented in practical drug discovery projects.

Dissolution of organic non-electrocytes in water involves breaking of crystal packing of the aforementioned molecules and solvation of them with water.[14] On the other hand, energy gained by solvation of the molecule is reciprocal to the octanol−

Received: February 27, 2023
Published: May 1, 2023

water partition coefficient (K_{ow}). The general solubility equation (GSE) describes the relationship between the melting point (MP) and solvation energy on the one side and solubility on the other side,[15] by assuming that the change of entropy on melting is constant for all organic non-electrolytes.

$$\log S = 0.5 - 0.01(\mathrm{MP} - 25) - \log P \quad (1)$$

where log S is the base 10 logarithm of molar intrinsic solubility, MP is the melting point in Celsius, and log P is the base 10 logarithm of K_{ow}. Note that this equation works for non-ionizable organic compounds, and it only predicts the intrinsic solubility instead of the apparent one. To compute the apparent solubility of an ionizable compound at a particular pH, say 7.4, log P in eq 1 needs to be replaced by log D at the same pH (refer to Intrinsic Solubility vs Apparent Solubility in the Supporting Information).

The term (MP-25) is set equal to zero for liquid solutes that melt below room temperature (25 °C).[16] This equation is quite robust and predictive for various datasets as shown in many studies.[15−17] log P can be accurately predicted using known literature methods.[18] However, careful measurement or prediction of MP is still required.

There were quite a lot of MP models built with relatively small datasets in the very early days. In one of the earliest studies, Bergström et al.[19] developed partial least squares models for a set of 277 molecules with both two-dimensional (2D) and/or three-dimensional (3D) descriptors, and their best model reached a root mean square error (RMSE) of 50. This study showed the physicochemical properties related to hydrophilicity/polarity, partial atom charge, and rigidity, which were found to increase the MP, whereas nonpolar descriptors and descriptors for molecular flexibility lowered the MP. Nigsch et al.[20] employed k-nearest neighbors to develop MP models for a set of 4119 diverse organic molecules with 2D and/or 3D descriptors and reached an RMSE of around 50 from their best model. A nice summary of MP models built previously can also be found in that paper.[20] Tetko et al.[21] collected the then largest dataset (more than 47 K compounds) from multiple sources, that is, OCHEM,[22] Enamine,[23] Bradley et al.,[24] and Bergström et al.,[19] and developed multiple models using 2D and/or 3D descriptors. The best of these models can achieve an RMSE of as low as 33 for molecules melting between [50, 250] °C. All those studies[19−21] showed that 2D representations of molecules were more successful in the prediction of MPs than descriptors generated from the computed 3D conformers, that is, the optimized geometry of molecules. Since the discrepancy between generated conformers and associated experimental crystal structures is proportional to the number of rotatable bonds,[25] it is reasonable to assume that 3D descriptors computed from inaccurate conformers were responsible for the underperformance of their machine learning models. Tetko et al.[21] suggested that the knowledge of the experimental crystal structure would likely be critical in predicting the MP. With those same large datasets, Sivaraman et al, in a follow up study,[26] reached a mean absolute error (MAE) of 28.9 with Gaussian process regression, a method including prediction uncertainties and graph convolutional neural networks. The authors indicated that prediction accuracy will likely be improved by considering the 3D molecular structure of the crystals. From those studies, it could be concluded that the key to develop predictive MP models is (1) to collect a large enough dataset that covers a vast chemical space and (2) to incorporate information about experimental crystal structures to MP models. With that in mind, the Cambridge Structural Database (CSD)[27] became the natural choice.

The Cambridge Crystallographic Data Centre (CCDC) is a non-profit organization specializing in structural chemistry data, software, and knowledge for materials and life sciences research and development. It was set up to collate and distribute the CSD.[27] Today, the CSD contains over 1 million organic and metal−organic experimental crystal structures. Most datasets are deposited by researchers worldwide pre-publication, and during deposition, scientists are encouraged to add additional metadata of relevance to the dataset including the MP. At the point of publication, each structure undergoes extensive curation at the CCDC using manual and automated methods to maintain the accuracy of the data. During curation, MPs are extracted from a variety of sources including the deposited files and associated scientific articles. The MPs in the CSD refer to the MP of the substance and the correct solid form. However, there can be inconsistencies in the MP data through incorrect associations, incorrect units, or inaccuracies in data reporting and the CCDC routinely undertake CSD improvement projects in this area to increase the accuracy of the data.

Here, we present the use of three machine learning methods and the largest ever MP dataset from the CSD to develop predictive models. The performance was systemically compared for Morgan fingerprints, 2D descriptors, and 3D descriptors generated from experimental crystal structures across different machine learning models. Combining the best machine learning MP model with the GSE, the solubility was predicted for compounds in six external datasets. The pros and cons of the models are discussed by comparing them systematically with other popular solubility prediction methods like ESOL[28] and solubility forecast index (SFI).[29] Many companies have internal measurements of MPs, and so the advantages of adding internal MP data to the CSD model, *a.k.a.*, localization, were also simulated by creating virtual series based on Bemis−Murcko clustering. Finally, the machine learning MP models were validated and tuned by using molecules with the experimental MP in practical drug discovery projects.

METHODS

Data Curation. The CSD[27] has over 170 K organic and metal−organic compounds with experimental MPs as of March 2022. Since the main purpose was to build models for virtual screening of organic small molecules in early drug discovery, only experimental MPs of organic compounds as well as their crystal structures were included. Metal−organic compounds were not included in this study. The workflow to curate the MP data includes (1) removing salts, mixtures, and hydrates from the dataset; (2) removing a few hundred records of qualitative descriptions, from which no specific MP can be extracted (some example descriptions are thermal decomposition, sublimes at 130 deg.C (dec.), above 240 dec., decomposes at 218 dec., less than 300 dec., and >684 K); (3) averaging measurements with the melting range such as 521−526 K, 521−522 K, 252−256 dec., and 252−257 dec.; (4) converting these MPs in Celsius to Kelvin; (5) deduplicating molecules with the same SMILES string and the same MP; (6) removing molecules with an MP lower than room temperature (298 K) or higher than 377 °C (650 K); and (7) removing molecules with MP ranges greater than 10 degrees. Because 3D descriptors generated from experimental crystal structures will be evaluated in this study, molecules with the same SMILES string but different MPs were kept, as some may have different 3D crystal structures. After

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

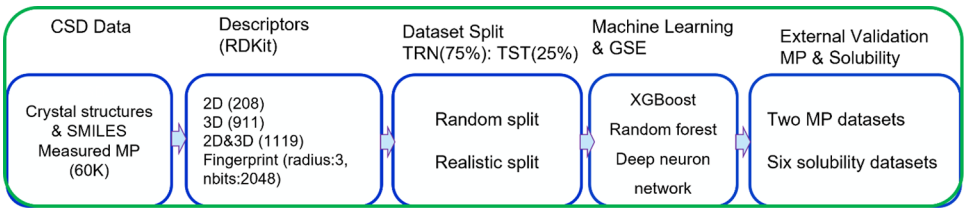


Figure 1. Workflow of the melting point and solubility modeling.

those steps, a set of 58,810 unique crystal structures with associated MPs were left in the dataset for model development. The average MP and molecular weight (MW) of this dataset are 150 ± 56.5 and 359.5 ± 139.7 °C, respectively. No drug-like filter was applied to the final dataset since the aim was to cover as much chemical space as possible. A remarkable 21,372 unique Bemis−Murcko scaffolds were identified within this set of 58,810 compounds. Despite their diversity, the majority of compounds exhibit physicochemical properties characteristic of drug-like molecules (Figure S1). The detailed summary of some physicochemical properties of this CSD dataset is shown in Table S1. Figure 1 shows the workflow of modeling, which will be explained in detail.

Molecular Descriptors. The melting of a crystalline solid involves the breaking of intermolecular forces, which are highly related to the size, constitution, configuration, and conformation of compounds. Thus, it is a natural way to evaluate the performance of 2D and/or 3D descriptors, especially when the experimental crystal structures are all available. A set of 2D descriptors (208) were computed based on SMILES strings using Descriptors in RDKit[30] for each molecule. Another set of 3D descriptors (911) was computed based on the crystal structures of molecules using Descriptors3D in RDKit. Structures (both 2D SMILES and 3D crystal structures) were used as they were retrieved from the CSD database while computing those descriptors. As a benchmark, RDKit Morgan radius 3 substructure fingerprints of 2048 bits were also computed. Models were built using RDKit 2D, 3D, and Morgan fingerprint descriptors (radius 3, 2048 bits) individually. Then, RDKit 2D and 3D descriptors were combined (1119 of them) to build MP models with the purpose of improving model performance. A simpler Morgan fingerprint (radius 2, 1024 bits) was used to compute the similarity among molecules.

Random vs Realistic Split. Two ways of selecting a held-out test set were used to evaluate the quality of the models. The first one was achieved by randomly selecting 25% of compounds as the test set which would be structurally similar to the training set. The second one was a so-called realistic training/test set split generated via the Butina algorithm,[31] as stated in previous studies.[32,33] The realistic split got its name, for it could mirror the novelty of new compounds synthesized in real drug discovery projects. After clustering the whole set of compounds via the Butina algorithm in terms of their structural similarities calculated based on Morgan radius 2 substructure fingerprints of 1024 bits, the training set is collected starting with the largest cluster and proceeding to successively smaller clusters until 75% of the compounds have been gathered. The remaining singletons and small clusters make up the "realistic" test set. Figure 2 shows that the test set compounds in the realistic split are structurally less like corresponding training set compared to those in the random split. Even though there are huge structural differences between the two ways of splitting, the distribution of MP shows no obvious difference between them (Figure S1).

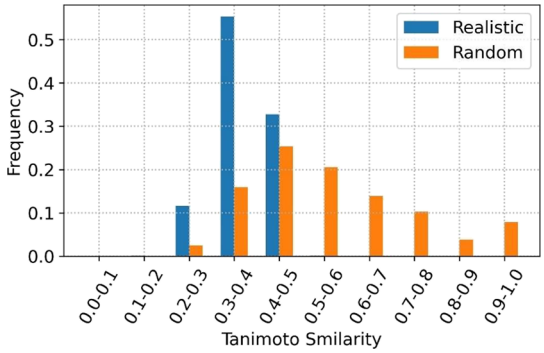


Figure 2. Realistic split test set is structurally less similar to its associated training set compared to the random split. Test set compounds from the realistic split are quite dissimilar to associated training set ones with an average Tanimoto similarity of 0.39. The test set compounds from the random split share higher Tanimoto similarity (an average value of 0.58) with the associated training set.

Modeling Methods. Three widely used machine learning methods for regression were evaluated in this study. A tree-based ensemble learning, that is, Random Forest (RF)[34] was first considered for its robustness to outliers and non-linear features. RF modeling was performed using the scikit-learn RandomForestRegressor (v1.0.2), and the following hyperparameters were tuned to achieve the best performance: the number of trees (n_estimators), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples required at a leaf node (min_samples_leaf), and the number of features to consider when looking for the best split (max_features).

The eXtreme Gradient Boosting (XGBoost),[35] an implementation of gradient-boosted decision trees, was also considered for its demonstrated speed and performance in many machine learning studies. XGBoost models were developed using the XGBoost package (v1.6.1) in python. The following four hyperparameters were tuned: the number of trees (n_estimators), the maximum depth of each tree (max_depth), the learning rate, and the subsample ratio of columns when constructing each tree (colsample_bytree).

Deep neuron network (DNN)[36] models were built using the PyTorch package (v1.11.0). Three layers of network were used and the number of neurons in each layer, learning rate, and batch size were tuned. Adam optimizer was used for gradient descent and early stopping based on empirical criteria to avoid overfitting. The RDKit 2D/3D descriptors show huge variance in their magnitude (Figure S2). Unlike tree-based methods (e.g., RF and XGBoost), DNN is sensitive to the range as well as the collinearity of input data. The workflow for data preprocessing is as follows: (1) removing descriptors with SD lower than 0.1 or larger than 200; (2) keeping only one descriptor if two or more of them showed squared Pearson's r larger than 0.95; and (3) scaling the descriptors in an appropriate way. After the first two

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

steps, several conventional data scaling techniques were evaluated such as normalization and standardization, as implemented in sklearn. Then, a customized scaling method was proposed, as implemented in the open-source code published with the paper, which is specific to datasets with a huge difference in SD like the RDKit 2D/3D descriptors.

Models were trained using the training set, and their performance was evaluated by the test set. Coarse tuning followed by a fine optimization were performed for all hyperparameters in those three methods. After training the models on 75% of the training data and fine-tuning the parameters, we opted to construct models using the entire dataset in the production environment. This approach aims to maximize the benefits derived from the available MP data.

After predicting the MP of new molecules, their solubility was computed with the GSE (eq 1), in which the octanol−water partition coefficient (log P) was calculated using Crippen's approach implemented in the RDKit.

External Solubility Datasets. We compiled six external datasets with experimental solubility (the log base 10 transformation, log S) as well as some experimental MPs from the literature to validate the MP models as well as the methodology of solubility prediction. Dataset 1 contains 72 small molecule drugs approved during 2016 and 2020.[37] With an average MW of 464.6 ± 170, 49 of these drugs are insoluble in aqueous water (log S < −4) and the averaged solubility of the whole dataset is −4.50 ± 1.86. Dataset 2 has 132 compounds with an average solubility of −4.32 ± 1.62 compiled from a recent solubility challenge proposed by Llinas et al.[38] It consists of 100 regular compounds with an average interlaboratory uncertainty of 0.17 log unit and 32 drug-like challenging ones with even larger averaged uncertainty of 0.62 log unit. In addition to the solubility, it also has the experimental MP for all the compounds. Dataset 3 contains both experimental solubility and MP for a set of 148 compounds (Ran & Yalkowsky, 2001), most of which are soluble (log S > −4), with an average solubility of −2.35 ± 2.05. Dataset 4 has the experimental solubility of 900 compounds with an average solubility of −3.01 ± 2.44.[8] Dataset 5 contains a set of 8613 compounds curated from AqSolDB,[6] by removing salts and mixtures, which was compiled by consolidating a total of nine different aqueous solubility datasets. Compounds may have either intrinsic solubility or apparent solubility measured at an unknown pH using either kinetic or thermodynamic methods. To provide more choices for users, we treated those measurements with an unknown type of solubility as apparent solubility. We then converted these values to intrinsic solubility using the equation provided in the section "Intrinsic Solubility vs. Apparent Solubility" in the Supporting Information. Further details about this conversion process can be found in the section "Solubility Conversion" in the Supporting Information. Readers need to be cautious to interpret the prediction results since the GSE only predicts the intrinsic solubility. Dataset 6 contains experimental solubility for a set of 21 proteolysis targeting chimeric (PROTAC) compounds.[39] These are apparent solubility at a pH of 7.0 measured in a thermodynamic way. The dataset also includes intrinsic solubility converted from apparent solubility based on the calculated log D at pH of 7.0 in ACD Labs. Details can be found in the section "Intrinsic Solubility vs Apparent Solubility" in the Supporting Information. On average, those PROTAC compounds show extremely large MW (958.7 ± 111.9) and substantially low solubility (−5.16 ± 1.10). Detailed physicochemical properties of those six datasets can be found in Table S2.

Simulation of Model Localization. To simulate the advantages of a localized model, a.k.a. localization, compared to the global model, Bemis−Murcko clustering was used to select clusters of the same scaffolds from the CSD dataset to mimic compound series in drug discovery projects. For a particular cluster of N molecules, a global model was built based on a reduced CSD dataset with N molecules in that cluster and their corresponding nearest neighbors with Tanimoto similarity larger than 0.25 were removed. A localized model was simulated by adding $N - 1$ molecules in the aforementioned cluster to the global model. The one compound kept as a leave-one-out validation set was then predicted by both the global and localized models. This process was repeated N times until all those N molecules in that cluster serve as a validation set once. The prediction accuracy in terms of MAE for the N molecules in the cluster was evaluated between the global and localized models.

RESULTS

A series of models were developed based on the combination of three machine learning methods (DNN, XGBoost, and RF) and four sets of descriptors (2D, 3D, 2D&3D, and Fingerprint) for both random and realistic split. Table 1 shows the performance

Table 1. Mean Absolute Error of Machine Learning Models on the Random and Realistic Test Sets

split	method	2D	3D	2D&3D	fingerprint
random	DNN	29.2	33.3	29.1	33.9
	XGBoost	28.1	32.3	28.1	33.8
	Random Forest	30.1	35.3	30.9	35.4
realistic	DNN	30.3	34.3	30.2	38.0
	XGBoost	30.3	34.2	30.0	37.6
	Random Forest	32.5	37.1	33.0	39.0

of these models in terms of MAE on the random and realistic test sets. Contrary to our expectations, the RDKit 2D outperformed the 3D descriptors in all scenarios. Furthermore, the addition of conformational information did not improve the models, as the combination of 2D and 3D showed performance similar to that of 2D alone. It is no surprise that Fingerprint was not comparable to 2D or 3D descriptors in all models.

Three machine learning methods showed similar performance for the same descriptors. XGBoost might be the best, even though it did not show a substantial advantage over the DNN and RF. Generally, the impact of the machine learning method on the model performance was not as much as that of descriptors.

Compared to the random split, the realistic split showed slightly worse prediction, around 1−4 °C higher in terms of MAE across all the models, which was expected with the lack of similarities between the training and test sets (Figure 2). However, the performance on the realistic set might be a real reflection of the predictivity of our models on novel molecules in drug discovery projects.

The prediction error of the three machine learning models built with RDKit 2D descriptors provides more information about model performance and bias. As shown in Figure 3A,B, all three models have MAE lower than 25 °C, the error limit of experimental measurements; for compounds with the measured MP between 100 and 200 °C, a range covers up to 62% of compounds, for both random and clustering test sets. For the drug-like region, that is, molecules with the measured MP

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

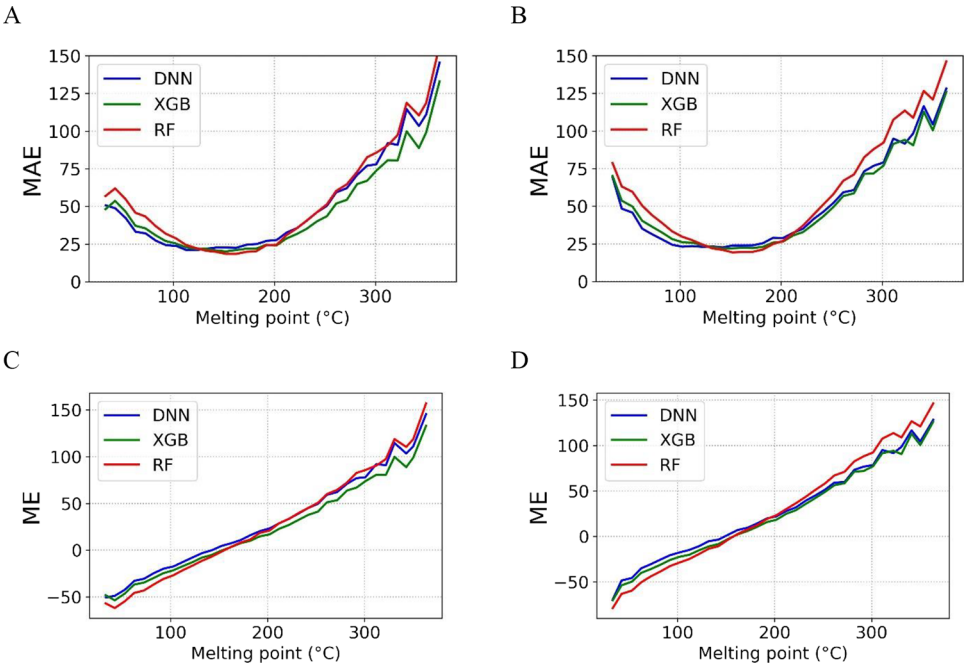


Figure 3. Mean absolute error (MAE) and mean error (ME) of machine learning models with RDKit 2D descriptors over the experimental melting point binned by 10 degrees. All models show high prediction accuracy for compounds with the melting point in the drug-like region (50−250 °C) for both random and realistic split. Models on average overpredicted compounds with the melting point below 150 °C and underpredicted those with the melting point higher than that value. (A) MAE of the random test; (B) MAE of the realistic test; (C) ME of the random test; and (D) ME of the realistic test.

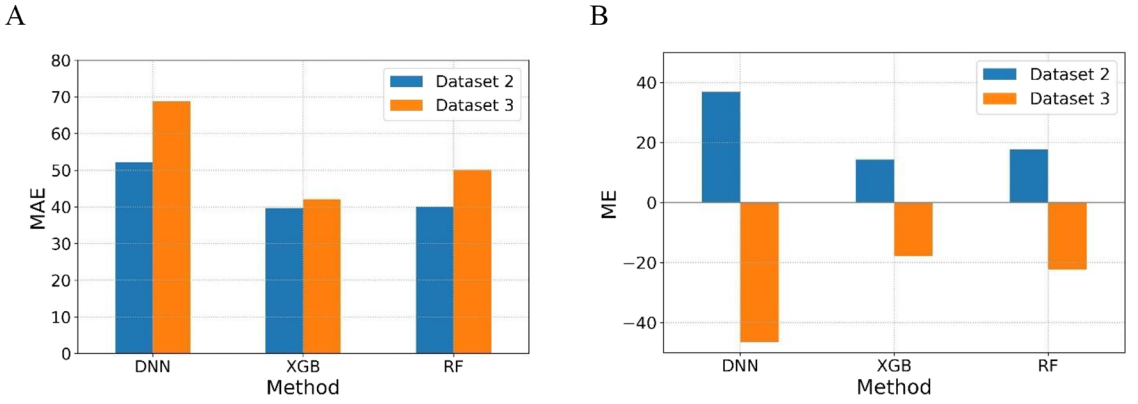


Figure 4. External validation of machine learning MP models built with the RDKit 2D descriptors based on the whole CSD data. XGB and RF show similar performance, and DNN is worse than them, for there are some outliers in both datasets. (A) Mean absolute error (MAE) of DNN, XGBoost, and RF for Dataset 2 were 52.2, 39.6, and 41.0, respectively. The corresponding values for Dataset 3 were 68.7, 42.1, and 50.7, respectively. (B) Mean error (ME) of DNN, XGBoost, and RF for Dataset 2 were 36.8, 14.3, and 17.7, respectively. The corresponding values for Dataset 3 were −46.5, −17.8, and −22.4.

between 50 and 250 °C, a range covers up to 93% of compounds and has prediction errors (MAE) between 25 and 27 °C for DNN, XGB, and RF. Compounds with high experimental MPs (>300 °C) tend to have high prediction errors, for example, MAE larger than 75 °C. It is no surprise as the experimental error at that range is usually a lot higher. Previous studies[25,26] also observed a much higher prediction accuracy on the drug-like region than the extremely low and high MP regions with their own models. The model bias, which has not been revealed in any previous studies, indicates that all models tend to overpredict at the low MP range and underpredict at the high MP range. As indicated by mean error (ME) in Figure 3C,D, an MP of 150 °C is the turning point, which is equals to the median MP of the CSD dataset (Table S1). In a sense, the model bias has nothing to do with the descriptors and machine learning methods but a reflection of regression to the average MP of a dataset.

Evidently, RDKit 2D descriptors were the best choice to further develop machine learning models using 100% of the CSD MP data. Thus, three models (DNN, XGBoost, and RF) were developed with the whole dataset by reusing associated tuned hyperparameters based on the realistic training set (Tables S3−S5). Their performance will be further evaluated by six external datasets.

Validation of MP Models. Datasets 2 and 3 have measured MPs for 132 and 148 compounds, respectively. Excluding those that overlap with the CSD dataset, a set of 93 compounds in Dataset 2 and 122 in Dataset 3 are unique. Their MPs were

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

predicted by using DNN, XGBoost, and RF models based on the RDKit 2D descriptors.

XGBoost showed the lowest MAE on both datasets, while DNN showed the highest (Figure 4A). The average Tanimoto similarity of Datasets 2 and 3 to the CSD dataset are 0.53 and 0.51, respectively, in between the realistic split (0.39) and random one (0.58). However, the distribution of measured MPs for Datasets 2 and 3 were 190 ± 66 and 102 ± 92 °C, respectively, showing much discrepancy to the CSD data of 155 ± 57 °C (Table S1). It is within expectations that all models underpredicted Dataset 2 and overpredicted Dataset 3 (Figure 4B) in general by looking up the ME based on where their averaged MPs located in the *x*-axis of Figure 3C or D.

Validation of GSE Models with External Solubility Datasets. After predicting the MP of all six external datasets by those three machine learning models, the GSE was then used to calculate the solubility with the predicted MP and Crippen log *P* in RDKit. Table 2 shows the detailed performance of DNN, XGBoost, and RF models in combination with GSE on those six external datasets. Here, DNN-GSE, XGB-DSE, and RF-GSE were used to represent solubility models based on the MP predicted by DNN, XGB, and RF, respectively, and ML-GSE to represent machine learning-based GSE models, which include all three methods. The square of Pearson correlation coefficient (r^2) generally indicates how strong the correlation is between the predictions and the measured solubility. Solubility models with high r^2 are not really indicative of the prediction accuracy, but they are very useful in virtual screening to prioritize a top list of molecules or can be used as one of the functions in multiobjective optimization of drug discovery projects based on their relativity. All three models have similar r^2, but XGB-GSE is slightly higher than the other two for Datasets 2, 3, 4, and 5. The coefficient of determination (R^2) tells how well the machine learning and GSE models fit the measured solubility. Solubility models with high R^2 indicate their prediction accuracy and are of key importance to inform drug designers of the possible solubility of their new designs. XGB-GSE again shows slightly higher R^2 for four out of six datasets (Datasets 2, 3, 4, and 5). As an example, Figure 5 shows the experimental solubility versus the XGB-GSE prediction.

XGB-GSE shows slightly higher prediction accuracy (i.e., lower MAE) for Datasets 1, 2, 3, and 4 than the other two methods. XGB-GSE and RF-GSE are consistent with each other on overpredicting Dataset 3 and underpredicting the others in terms of ME. DNN-GSE shows a very different trend on it. Here, tree-based algorithms like XGB-GSE and RF-GSE are more reliable and robust than DNN-GSE in these models. Dataset 3 has an average experimental solubility of −2.35 ± 2.05, which is the highest among the six (Table S2). One trend is that models perform well on datasets with more soluble molecules by comparing statistics (e.g., r^2, R^2, or MAE) of tree-based models (Table 2) with the average experimental solubility of six datasets (Table S2). For example, the Spearman correlation coefficient is 0.89, 0.94, and 0.71 for r^2, R^2, and MAE of the XGBoost model, respectively, versus the averaged experimental solubility for six datasets. Those six validation datasets with predicted MPs and solubility were attached as Supplemental Materials.

Model Localization. 10 clusters each with a unique scaffold were selected from the CSD (Table S6), and 10 corresponding global models were developed. These 10 clusters show various intracluster pairwise Tanimoto similarity (Figure S3A and Table S6), and their median MPs (Figure S3B and Table S6) were around or above 150 °C. Each global model was built by using

Table 2. Mean Absolute Error (MAE) of Solubility Models on External Validation Sets

		r^{2a}				R^{2b}				MAE[c]				ME[d]				% unit[e]			
dataset	#Cpds	ESOL	RF	XGB	DNN	ESOL	RF	XGB	DNN	ESOL	RF	XGB	DNN	ESOL	RF	XGB	DNN	ESOL	RF	XGB	DNN
1	72	0.63	0.6	0.6	0.61	0.06	0.36	0.28	0.27	1.39	1.14	1.13	1.18	0.86	−0.24	−0.19	−0.54	0.43	0.6	0.57	0.46
2	132	0.58	0.58	0.59	0.58	0.41	0.45	0.46	0.36	0.92	0.89	0.88	0.99	0.13	−0.39	−0.37	−0.6	0.65	0.63	0.67	0.64
3	148	0.8	0.81	0.82	0.78	0.78	0.80	0.82	0.73	0.8	0.72	0.69	0.85	0.3	0.2	0.17	0.44	0.7	0.74	0.72	0.64
4	900	0.77	0.77	0.78	0.76	0.76	0.77	0.77	0.75	0.93	0.93	0.93	0.95	0.08	−0.13	−0.15	0.08	0.62	0.61	0.61	0.60
5	8613	0.64	0.64	0.64	0.63	0.55	0.58	0.58	0.55	1.01	1.0	1.01	1.05	0.31	−0.02	−0.06	0.22	0.64	0.63	0.62	0.60
6	21	0.22	0.55	0.54	0.50	−19.91	−0.83	−0.63	−0.18	4.75	1.19	1.1	0.98	4.75	1.03	0.92	0.67	0	0.43	0.52	0.48

[a]The square of the Pearson correlation coefficient (higher number is better). [b]Coefficient of determination (higher number is better). [c]Mean absolute error (lower number is better). [d]Mean error (lower absolute value is better). [e]Percentage of prediction errors in 1 log unit (higher number is better).

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

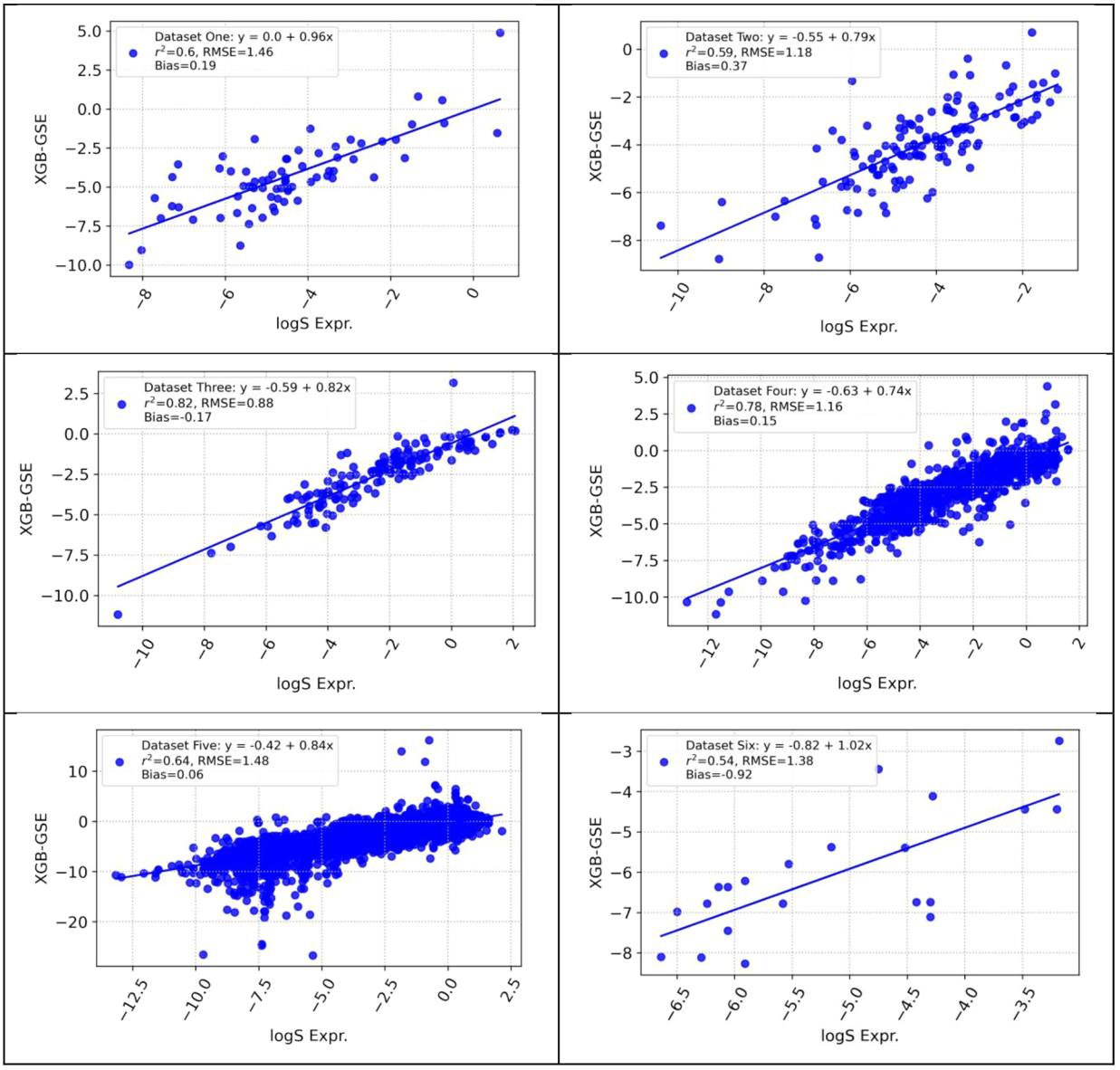


Figure 5. Experimental aqueous solubility (log *S* Expr.) plotted against the solubility calculated using the general solubility equation, which is based on the melting point predicted by the XGBoost model (XGB-GSE). Each graph displays the fitting equation for experimental versus predicted values, along with the associated squared Pearson's correlation coefficient (r^2), root mean squared error (RMSE), and bias value.

XGBoost based on a curated CSD dataset with an individual cluster and corresponding nearest neighbors removed. Except for cluster 6, 9 out of 10 clusters show reduced MAE to different extents, that is, from 2.1 to 10.1 °C, for the localized models compared to corresponding global models (Figure 6).

DISCUSSION

MP Modeling. With the largest ever CSD dataset, all three machine learning MP models built with RDKit 2D descriptors achieved quite a good performance, for example, MAE between 25 and 27 °C for the realistic test set molecules in the drug-like region, that is, 50−250 °C. To the best of our knowledge, these might be the most accurate models with the largest coverage in chemical space. Two previous massive modeling studies,[21,26] developed using different machine learning methods on public datasets, observed similar prediction errors for their random test set molecules in the drug-like region only for models built with high-quality datasets. It also emphasizes the importance of data

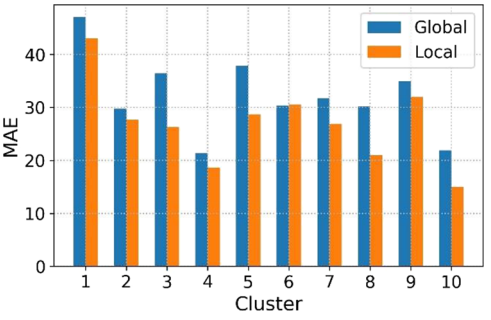


Figure 6. Comparison of mean absolute error (MAE) of global and localized models across 10 clusters. Comparing to global models, except cluster 6, localized models show the largest improvement (ΔMAE of 10.1 °C) for cluster 3 and the slightest improvement (ΔMAE of 2.1 °C) for cluster 2.

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

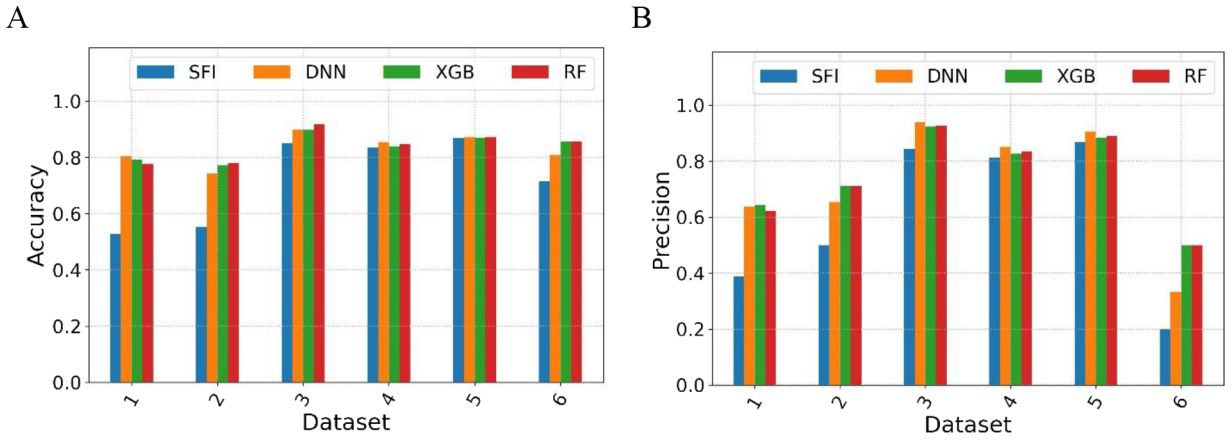


Figure 7. Classification performance of the solubility forecast index (SFI) and GSE models. (A) Accuracy is the ratio of correctly predicted soluble and insoluble compounds; (B) precision is the number of truly soluble compounds divided by the number of predicted soluble ones. The higher the better for both metrics.

over algorithms in developing machine learning models for tabular datasets.

Quite contrary to the assumption of many researchers that 3D descriptors based on experimental crystal structures would outperform 2D descriptors, our modeling showed higher accuracies with RDKit 2D descriptors than 3D ones across the combination of the three machine learning models and two training/test set split strategies. After checking the correlation between 2D descriptors and MPs, 46 out of 209 showed absolute Pearson's coefficient higher than 0.20 with the MP, most of which were related to MW, the number of rings (both aromatic rings and heterocycles), the number of hydrogen bond donors, and molecular surface area (LabuteASA and TPSA). There are also quite a few 3D descriptors showing high correlation with the MP; however, most of them also show collinearity with each other and have no specific physiochemical explanations. Since the prediction of 3D conformation with the lowest energy is yet another long-standing challenge, building MP models using 2D descriptors could save researchers' time and reduce uncertainties in applying those models to new molecules.

Compared with tree-based algorithms like RF and XGBoost, the DNN shows compromised performance in predicting the MP of external compounds (Figure 4) and hence solubility in terms of accuracy and correlation (Table 2). The fully connected deep networks are known to be sensitive to outliers, multicollinearity, and non-linear features. While tree-based algorithms are relatively insensitive to those factors. During minmax standardization, it was observed that the features of many compounds in the external datasets were out of the scope of those in the CSD dataset. These potential outliers could be handled well by tree-based algorithms but may affect the performance of DNN greatly. Many studies using other representations of a molecule, such as weave module,[40] SMILES string,[41] and molecular graph[41,42] achieved generally a performance similar to that of Morgan fingerprint or 2D/3D descriptors on the prediction of solubility or other physicochemical properties. These new representations of molecules were not applied in this study. However, since the MP data, external test datasets are available, and the source code for the MP and solubility modeling (OpenSOL) is open source on GitHub, we welcome readers to contribute their code to GitHub here.

GSE Predictability. Besides the MP, log *P* is another important term in the GSE that influences solubility. The Crippen algorithm,[43] an atom-based contribution method, as implemented in RDKit, was used to compute the Clog *P* of molecules. GSE models generally underpredict the solubility of PROTAC compounds in Dataset 6 (an average log *S* of −5.16), for example, ME of 0.92 for XGB-GSE. Since these machine learning models usually underpredict the MP of large molecules, the ME could be even larger if the MP was predicted correctly. Two reasons are suggested that might explain the failure of the ML-GSE modeling in predicting Dataset 6. First, Clog *P* shows much larger prediction errors for compounds with high experimental log *P* (>5.0) than those with low experimental log *P* (<4.0).[18] Most compounds in Dataset 6 are extremely insoluble, so they could have quite high log *P*. Second, GSE assumes that the entropy of melting is constant (ΔSm = 56.5 J/degree mole) for all the compounds.[16] This assumption may hold true for small rigid molecules, but less so for larger more flexible molecules like PROTACs, which on average have substantially larger MW and way more rotatable bonds (e.g., the average number of rotatable bonds is 19.6 for PROTACs) than compounds in other datasets (Table S2). Users must be careful when interpreting prediction results while using GSE for large molecules.

Comparison with ESOL. ESOL[28] is one simple linear model proposed by Delaney that is widely used to predict solubility and is also widely used for its simplicity. As shown in eq 2, it consists of four parameters: Clog *P*, MW, the number of rotatable bond (RB), and aromatic proportion (AP).

$$\log S = 00.16 - 0.63\ \mathrm{Clog}\ P - 0.0062\ \mathrm{MW} + 0.066\ \mathrm{RB} - 0.74\ \mathrm{AP} \quad (2)$$

These physicochemical parameters needed for ESOL all can be easily computed, and no measurement is required. It was applied to those six external datasets just for comparison. ESOL shows a comparable prediction accuracy in terms of all statistics for Datasets 2, 3, and 4 (Table 2). However, it tremendously underestimated the solubility for Datasets 1 and 6 with ME of 0.86 and 4.75, respectively, which is much larger than ML-GSE models. Datasets 1 and 6 on average show much larger molecule weight, lower solubility, and more rotatable bonds than the other four (Table S2). Overall, ESOL showed disadvantage in predicting insoluble compounds.

Comparison with SFI. SFI, expressed as Clog $D_{\mathrm{pH7.4}}$ + [Ar] (the number of aromatic ring), serves as an important metric for

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

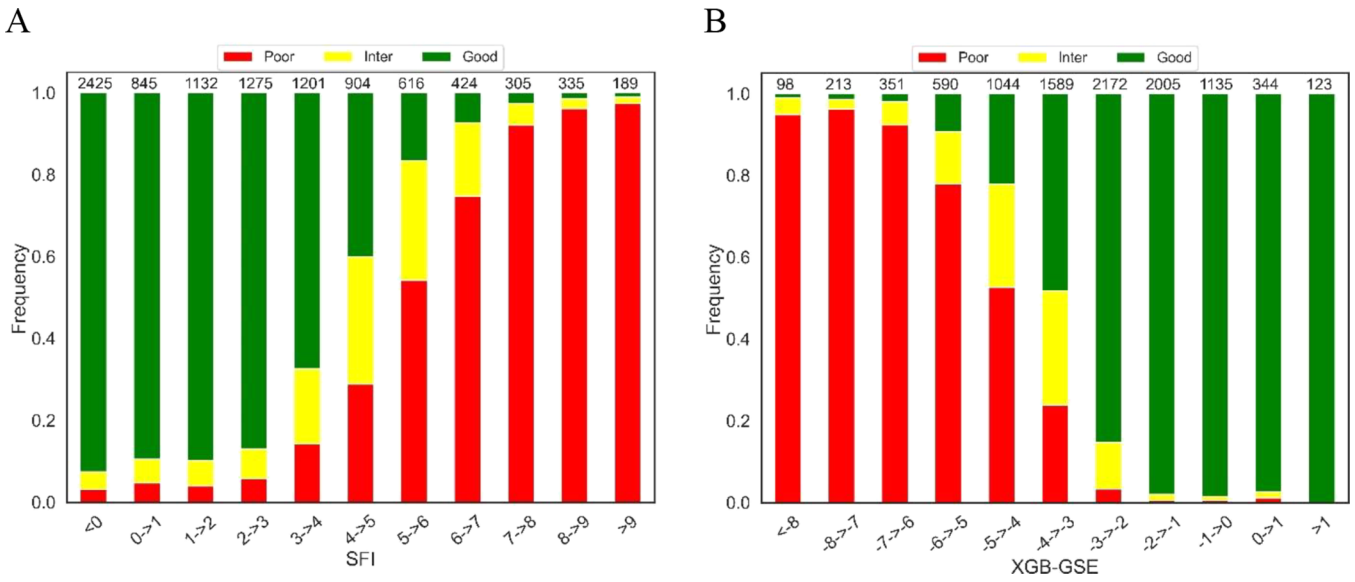


Figure 8. Distribution of solubility category as a function of SFI (A) and XGB-GSE (B). Poor: poorly soluble (<30 μM, or log S < −4.5); Inter: intermediate solubility (30−200 μM, or − 4.5 < log S < −3.7); Good: good solubility (>200 μM, or log S > −3.7).

Table 3. Internal Validation of Three Machine Learning Melting Point Models

project	Cpds	MW	Clog P	expr. (°C)	XGBoost (°C)	RF (°C)	DNN (°C)
TLR4	1	523.7	4.64	256.8 ± 0.6	204.7	193.5	160.9
	2	541.7	4.77	232.5	203.1	195.0	180.1
	3	541.7	4.81	252.6 ± 0.2	195.3	194.3	172.6
	4	541.7	4.98	328.8	197.1	193.7	189.3
P2	5	692.8	5.93	191.2 ± 1.4	195.7	178.9	191.7
	6	686.6	2.91	181.7 ± 4.0	172.2	172.4	150.5
	7	672.6	2.44	183.8 ± 3.4	163.1	173.2	207.6

solubility.[29] It penalizes a molecule's solubility by subtracting an extra log unit of hydrophobicity for each aromatic ring in addition to its intrinsic hydrophobicity value. Even though it is continuous, users often categorize it based on a threshold of 5.0, less than which a compound is more likely to have good solubility (>100 uM, or log S > −4.0). A solubility of at least 100 uM is a typical desired value for drug molecules.[44] More often, the SFI is adjusted by subtracting 3 if any ionizable organic functional groups exist at the physiological pH as they make the molecule more polar and more soluble in water (eq 3).

$$\mathrm{SFI} = \begin{cases} \mathrm{Clog}\, D_{\mathrm{pH7.4}} + [\mathrm{Ar}], & \text{if no ionizable} \\ \mathrm{Clog}\, D_{\mathrm{pH7.4}} + [\mathrm{Ar}] - 3, & \text{if ionizable} \end{cases} \quad (3)$$

The SFIs for all six external datasets were computed based on eq 3 with the Clog $D_{\mathrm{pH7.4}}$ calculated with ACD Labs (*v2021.2*).

A compound was defined as soluble or insoluble if its log S was larger or smaller than −4.0. Based on that, the classification performance was compared between the SFI and the ML-GSE models. Compared to the ML-GSE models, SFIs show similar accuracy for Datasets 3, 4, 5, and 6, but compromised accuracy for Datasets 1 and 2 (Figure 7A). Looking at the precision, the ratio of true soluble compounds out of all those predicted as soluble is important in virtual screening of early drug discovery.[45] The SFI shows comparable performance to ML-GSE models for Datasets 3, 4, and 5. However, it gives less precision compared to the ML-GSE models for Datasets 1 and 2 (Figure 7B), showing on average much lower solubility than the other datasets except 6 (Table S2), which was excluded from this comparison for the lack of statistical significance with a few soluble compounds.

Hill and Young[29] demonstrated that the probability of having poor (<30 μM), intermediate (30−200 μM), or good solubility (>200 μM) by a binned absolute SFI could provide even higher resolution to understand the predictability of the SFI. The compounds from the six datasets were combined together to show the distribution of solubility category as a function of SFI, as shown in Figure 8A. Figure 8B shows a similar graph to show the distribution of solubility category as a function of XGB-GSE, that is, our best solubility model. Comparing the grade bars with SFI <5 (Figure 8A) with those with log S larger than −4 (Figure 8B), it can be seen that the latter has fewer compounds with poor or intermediate solubility being predicted as soluble.

Model Localization. Structures in practical drug discovery projects may differ tremendously from those in the CSD dataset. Adding internal MP data to the training set could improve the prediction accuracy for novel compounds within a project or compound series. Molecules in three drug discovery projects at Sutro showed Tanimoto similarity between 0.20 and 0.40 to their nearest neighbors from the CSD dataset (Figure S4). In the simulation, a Tanimoto similarity of 0.25 was used as a cutoff to remove nearest neighbors of individual clusters from the CSD to create datasets to build global models. Generally, the localized models showed higher prediction accuracy compared to the corresponding global models due to a much higher similarity between the predicted molecules and their nearest neighbors in

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959

the training sets (Table S6). However, it did not hold true for cluster 6. Also a few others showed only slight improvement (Figure 6). The reasons could vary from the presence of polymorphism to the quality of the measurement. We are confident that adding more high-quality internal data could improve the MP prediction of novel molecules and their solubility.

Internal Validation. Seven compounds from two projects in Sutro were measured (refer to Melting Point Measurement in the Supporting Information for detailed methods) to validate these three machine learning MP models (Table 3). XGBoost, RF, and DNN presented MAEs of 46.3, 53.0, and 76.1 °C, respectively, for the four Toll-like receptor 4 (TLR4) agonist compounds. For another project, not ready to disclose (P2), those three machine learning methods showed similar MAEs (i.e., 18.4 °C for XGBoost, 17.6 °C for RF, and 19.9 °C for DNN) for three tested compounds. Considering all evaluations, XGBoost was selected to build our production model for implementation.

Implementation. Without the need for calculating 3D descriptors, which is quite sensitive to mixtures, hydrates, and salts, over 100 K CSD molecules were used, instead of the current dataset of nearly 60 K, as well as dozens of molecules from several of Sutro's own drug discovery projects, to build machine learning MP models using XGBoost with RDKit 2D descriptors. It was incorporated into our internal molecule management platform to calculate solubility instantly upon compound's registration. It can also be used for virtual screening and prioritizing a top list of molecules based on their predicted solubility.

Data and Software Availability. The OpenSOL program for downloading CSD MPs, data curation, descriptor generation, model development, solubility prediction, and one XGBoost MP model based on RDKit 2D descriptors are open source on GitHub: https://github.com/sutropub/OpenSOL

Sutro MP training data include two binary files of RDKit 2D descriptors and Morgan fingerprint with radius 2 and 1024 bits, as well as those curated MPs, and realistic and random train/test splits for 58,110 compounds are available under the Validation Test Sets section of the CCDC website https://www.ccdc.cam.ac.uk/support-and-resources/downloads/

CONCLUSIONS

In this study, three machine learning models were successfully developed (DNN, RF, and XGBoost) for MP prediction using the largest ever CSD dataset. RDKit 2D descriptors outperformed 3D and Morgan Fingerprint across all three machine learning methods. The models reached an MAE of 25−27 °C for molecules with the measured MP between 50 and 250 °C, a range which contains up to 93% of all compounds. Five out of six external datasets validated the effectiveness of solubility prediction by using GSE with MP predicted by our machine learning models. One limitation of GSE, also common to all other models, was its failure in predicting molecules with extremely large molecule weight, high number of rotatable bonds, and extremely low solubility, as exemplified by those PROTAC compounds. Theoretically, the ML-GSE model could be more accurate if the MP models were fine-tuned by adding more compounds in specific drug discovery projects.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00308.

Six validation datasets (SI_dataset_one.csv, SI_dataset_two.csv, SI_dataset_three.csv, SI_dataset_four.csv, SI_dataset_five.csv, SI_dataset_six.csv) with predicted melting points and solubility by three machine learning models based on RDKit 2D descriptors; SI_dataset_two.csv and SI_dataset_three.csv; melting points predicted by three machine learning models (random_forest_MP, xgboost_MP, and dnn_MP); solubility predicted by ML-GSE (random_forest_logS, xgboost_logS, and dnn_logS); solubility (esol_logS) predicted by ESOL (eq 2); SFI predicted by eq 3; maximum Tanimoto similarity (knn_max) of each compound to its nearest neighbor in the CSD melting point dataset; average Tanimoto similarity (knn_mean) of each compound to its five nearest neighbors in the CSD melting point dataset; and Clog *P* (rdkit_ClogP) computed in RDKit (ZIP)

Summary of the physicochemical properties of the melting point dataset for modeling; physicochemical properties of six validation datasets; fine-tuned hyperparameters of random forest, XGBoost, and DNN for realistic split based on RDKit 2D descriptors; molecular and modeling information of the 10 Clusters; distribution of melting points in the training and test sets; distribution of the standard deviation of RDKit 2D and 3D descriptors; intracluster Tanimoto similarity and melting point across 10 clusters each with unique scaffold; Tanimoto similarity between small molecules in three of Sutro's drug discovery projects and their five nearest neighbors from the CSD melting point dataset (PDF)

AUTHOR INFORMATION

Corresponding Author

Valery R. Polyakov − *Sutro Biopharma, South San Francisco, California 94080, United States;* orcid.org/0000-0001-5135-4376; Email: valery.polyakov@gmail.com

Authors

Xiangwei Zhu − *Sutro Biopharma, South San Francisco, California 94080, United States;* orcid.org/0000-0002-1894-7679

Krishna Bajjuri − *Sutro Biopharma, South San Francisco, California 94080, United States*

Huiyong Hu − *Sutro Biopharma, South San Francisco, California 94080, United States*

Andreas Maderna − *Sutro Biopharma, South San Francisco, California 94080, United States*

Clare A. Tovee − *Cambridge Crystallographic Data Centre, Cambridge CB2 1EZ, U.K.*

Suzanna C. Ward − *Cambridge Crystallographic Data Centre, Cambridge CB2 1EZ, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00308

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors express their gratitude to the thousands of researchers who deposited their melting point data to the CSD database over many decades that represent the real effort behind the success of the melting point and solubility models.

■ ABBREVIATIONS

ADC, Antibody drug conjugate; CCDC, the Cambridge Crystallographic Data Centre; CSD, the Cambridge Structural Database; DNN, deep neuron network; GSE, general solubility equation; log *S*, the log base 10 transformation of solubility; log *D*, the log base 10 transformation of distribution constant; log *P*, the octanol−water partition coefficient; MAE, mean absolute error; ME, mean error; ML, machine learning; MP, melting point; PROTAC, proteolysis targeting chimeric; RF, random forest; RMSE, root-mean-square error; SFI, solubility forecast index; XGBoost, extreme gradient boosting

■ REFERENCES

(1) Barrett, J. A.; Yang, W.; Skolnik, S. M.; Belliveau, L. M.; Patros, K. M. Discovery Solubility Measurement and Assessment of Small Molecules with Drug Development in Mind. *Drug Discovery Today* **2022**, *27*, 1315−1325.

(2) McCombs, J. R.; Owen, S. C. Antibody Drug Conjugates: Design and Selection of Linker, Payload and Conjugation Chemistry. *AAPS J.* **2015**, *17*, 339−351.

(3) Alsenz, J.; Kansy, M. High Throughput Solubility Measurement in Drug Discovery and Development. *Adv. Drug Delivery Rev.* **2007**, *59*, 546−567.

(4) Murdande, S. B.; Pikal, M. J.; Shanker, R. M.; Bogner, R. H. Aqueous Solubility of Crystalline and Amorphous Drugs: Challenges in Measurement. *Pharm. Dev. Technol.* **2011**, *16*, 187−200.

(5) Gigante, V.; Pauletti, G. M.; Kopp, S.; Xu, M.; Gonzalez-Alvarez, I.; Merino, V.; McIntosh, M. P.; Wessels, A.; Lee, B.-J.; Rezende, K. R.; Scriba, G. K. E.; Jadaun, G. P. S.; Bermejo, M. Global Testing of a Consensus Solubility Assessment to Enhance Robustness of the WHO Biopharmaceutical Classification System. *ADMET DMPK* **2021**, *9*, 23−39.

(6) 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* **2019**, *6*, 143.

(7) Meng, J.; Chen, P.; Wahib, M.; Yang, M.; Zheng, L.; Wei, Y.; Feng, S.; Liu, W. Boosting the Predictive Performance with Aqueous Solubility Dataset Curation. *Sci. Data* **2022**, *9*, 71.

(8) 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.* **2020**, *11*, 5753.

(9) Deng, C.; Liang, L.; Xing, G.; Hua, Y.; Lu, T.; Zhang, Y.; Chen, Y.; Liu, H. Multi-Channel GCN Ensembled Machine Learning Model for Molecular Aqueous Solubility Prediction on a Clean Dataset. *Mol. Divers.* **2022**, DOI: 10.1007/s11030-022-10465-x.

(10) Francoeur, P. G.; Koes, D. R. SolTranNet−A Machine Learning Tool for Fast Aqueous Solubility Prediction. *J. Chem. Inf. Model.* **2021**, *61*, 2530−2536.

(11) Falcón-Cano, G.; Cabrera-Pérez, M. Á.; Molina, C. ADME Prediction with KNIME: In Silico Aqueous Solubility Consensus Model Based on Supervised Recursive Random Forest Approaches. *ADMET DMPK* **2020**, *8*, 251−273.

(12) Cui, Q.; Lu, S.; Ni, B.; Zeng, X.; Tan, Y.; Chen, Y. D.; Zhao, H. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **2020**, *10*, 121.

(13) Wu, K.; Zhao, Z.; Wang, R.; Wei, G.-W. TopP−S: Persistent Homology-Based Multi-Task Deep Neural Networks for Simultaneous Predictions of Partition Coefficient and Aqueous Solubility. *J. Comput. Chem.* **2018**, *39*, 1444−1454.

(14) Bergström, C. A. S.; Larsson, P. Computational Prediction of Drug Solubility in Water-Based Systems: Qualitative and Quantitative Approaches Used in the Current Drug Discovery and Development Setting. *Int. J. Pharm.* **2018**, *540*, 185−193.

(15) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.

(16) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(17) Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. *J. Chem. Inf. Model.* **2012**, *52*, 420−428.

(18) Plante, J.; Werner, S. JPlogP: An Improved LogP Predictor Trained Using Predicted Data. *Aust. J. Chem.* **2018**, *10*, 61.

(19) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177−1185.

(20) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing K-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412−2422.

(21) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How Accurately Can We Predict the Melting Points of Drug-like Compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320−3329.

(22) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.

(23) ENAMINE Ltd. https://enamine.net.

(24) Bradley, J. C.; Lang, A.; Williams, A. Jean-Claude Bradley Double Plus Good (Highly Curated and Validated) Melting Point Dataset. *figshare* **2014**, *10*, m9.

(25) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146−1158.

(26) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vázquez-Mayagoitia, Á.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. J. A Machine Learning Workflow for Molecular Analysis: Application to Melting Points. *Mach. Learn. Sci. Technol.* **2020**, *1*, No. 025015.

(27) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. B* **2016**, *72*, 171−179.

(28) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(29) Hill, A. P.; Young, R. J. Getting Physical in Drug Discovery: A Contemporary Perspective on Solubility and Hydrophobicity. *Drug Discovery Today* **2010**, *15*, 648−655.

(30) RDKit: Open-Source Cheminformatics. Http://Www.Rdkit.Org.

(31) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(32) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077−2088.

(33) Martin, E. J.; Polyakov, V. R.; Zhu, X.-W.; Tian, L.; Mukherjee, P.; Liu, X. All-Assay-Max2 PQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, *59*, 4450−4459.

(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(35) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016, pp. 785−794.
(36) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.
(37) Avdeef, A.; Kansy, M. Predicting Solubility of Newly-Approved Drugs (2016−2020) with a Simple ABSOLV and GSE(Flexible-Acceptor) Consensus Model Outperforming Random Forest Regression. *J. Solution Chem.* **2022**, *51*, 1020.
(38) Llinas, A.; Oprisiu, I.; Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2020**, *60*, 4791−4803.
(39) García Jiménez, D.; Rossi Sebastiano, M.; Vallaro, M.; Mileo, V.; Pizzirani, D.; Moretti, E.; Ermondi, G.; Caron, G. Designing Soluble PROTACs: Strategies and Preliminary Guidelines. *J. Med. Chem.* **2022**, *65*, 12639.
(40) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(41) Panapitiya, G.; Girard, M.; Hollas, A.; Sepulveda, J.; Murugesan, V.; Wang, W.; Saldanha, E. Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **2022**, *7*, 15695−15710.
(42) Wang, X.; Li, Z.; Jiang, M.; Wang, S.; Zhang, S.; Wei, Z. Molecule Property Prediction Based on Spatial Graph Embedding. *J. Chem. Inf. Model.* **2019**, *59*, 3817−3828.
(43) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(44) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin. Drug Discov.* **2010**, *5*, 235−248.
(45) Gimeno, A.; Ojeda-Montes, M. J.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20*, 1375.

https://doi.org/10.1021/acs.jcim.3c00308
J. Chem. Inf. Model. 2023, 63, 2948−2959