RESEARCH

Open Access

Towards the prediction of drug solubility in binary solvent mixtures at various temperatures using machine learning

Zeqing Bao[1], Gary Tom[2,3,4], Austin Cheng[2,3,4], Jeffrey Watchorn[5], Alán Aspuru-Guzik[2,3,4,5,6,7,8,9] and Christine Allen[1,5,7*]

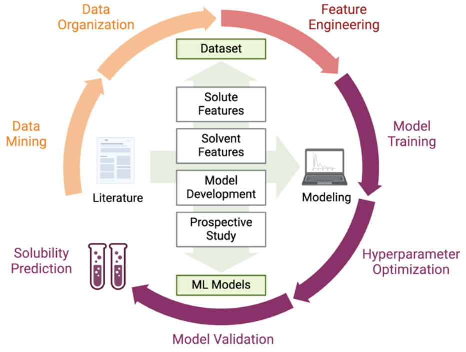
Abstract Drug solubility is an important parameter in the drug development process, yet it is often tedious and challenging to measure, especially for expensive drugs or those available in small quantities. To alleviate these challenges, machine learning (ML) has been applied to predict drug solubility as an alternative approach. However, the majority of existing ML research has focused on the predictions of aqueous solubility and/or solubility at specific temperatures, which restricts the model applicability in pharmaceutical development. To bridge this gap, we compiled a dataset of 27,000 solubility datapoints, including solubility of small molecules measured in a range of binary solvent mixtures under various temperatures. Next, a panel of ML models were trained on this dataset with their hyperparameters tuned using Bayesian optimization. The resulting top-performing models, both gradient boosted decision trees (light gradient boosting machine and extreme gradient boosting), achieved mean absolute errors (MAE) of 0.33 for LogS (S in g/100 g) on the holdout set. These models were further validated through a prospective study, wherein the solubility of four drug molecules were predicted by the models and then validated with in-house solubility experiments. This prospective study demonstrated that the models accurately predicted the solubility of solutes in specific binary solvent mixtures under different temperatures, especially for drugs whose features closely align within the solutes in the dataset (MAE < 0.5 for LogS). To support future research and facilitate advancements in the field, we have made the dataset and code openly available.

Scientific contribution

Our research advances the state-of-the-art in predicting solubility for small molecules by leveraging ML and a uniquely comprehensive dataset. Unlike existing ML studies that predominantly focus on solubility in aqueous solvents at fixed temperatures, our work enables prediction of drug solubility in a variety of binary solvent mixtures over a broad temperature range, providing practical insights on the modeling of solubility for realistic pharmaceutical applications. These advancements along with the open access dataset and code support significant steps in the drug development process including new molecule discovery, drug analysis and formulation.

*Correspondence:
Christine Allen
cj.allen@utoronto.ca
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

In pharmaceutical development, the solubility of drugs is a critical factor that influences various stages, including drug discovery [1–3], drug analysis [2, 4, 5], and formulation design [6–8]. Drug solubility measurements primarily include thermodynamic and kinetic methods [9, 10]. Thermodynamic methods assess the solute concentration in a solution at equilibrium between dissolved and undissolved compounds, while kinetic methods identify the solute concentration at which precipitation first occurs [11]. Although both methods are effective and efforts have been made to enhance their throughput [1, 4, 12], they remain tedious and can pose challenges for costly drugs or when the available quantity of drug is limited. To address these challenges, researchers have begun to explore machine learning (ML) as an alternative method [13, 14]. As a data-driven approach, ML models can be trained on available solubility datasets to learn solute–solvent interactions for prediction of unmeasured solubility. To date, most of these ML models were developed to model the solute solubility in a specific solvent [15–25], with a primary focus on aqueous solubility due to availability of data for model development. Compared to models designed for predicting solubility across multiple solvents, these solvent-specific models generally deliver improved accuracy because their design does not require them to account for variations between different solvents. However, for the same reason, the applications of these solvent-specific models are limited to the specific solvents for which they were developed. To extend the applicability of these models, researchers have started to include a range of solvents into their datasets, allowing ML models to learn and adapt to different solvents [26–28]. A notable example is the work of Vassileiou et al., who compiled a dataset primarily from the literature, supplemented with in-house and industrial partner data [26]. This dataset includes 714 solubility data points measured at room temperature, covering 75 solutes and 49 solvents [26]. Using this dataset, ML models were trained to predict solubility across solvents, with the resulting model mean absolute error (MAE) ranging from 0.39 to 0.58 (LogS, S in g/100 g) [26]. Another significant contribution is from Ye et al., who collected a larger dataset of 5081 data points from the literature [27]. This dataset includes the solubility of 266 compounds in 123 solvents measured at various temperatures, and the optimal model resulted in predictions for solubility with an MAE of 0.47 (LogS, S in mol/L) [27].

Despite these significant advancements in predicting solubility within single solvents, the complicated nature of pharmaceutical research often necessitates an understanding of drug solubility in solvent mixtures [29–32]. These mixtures allow greater flexibility through adjusting solvent combinations and ratios, enabling solubility to be tailored to meet specific needs and to co-dissolve other necessary materials (e.g., excipients [33, 34]). However, the wide variety of possible solvent combinations/ratios and the influence of temperature significantly complicate the experimental measurements. This complexity further extends beyond experimental methods to ML approaches, which require extensive datasets covering a broad spectrum of these conditions for effective model training. The need for such an extensive dataset potentially contributes to the scarcity of research on modeling solubility in solvent mixtures with ML approaches. One of the very few examples in this field is work by Chinta et al. [35], who trained ML models to predict solubility in binary solvent mixtures. In that study, models were trained and evaluated using a relatively small dataset, containing around 600 solubility results for 27 solutes predominantly measured at a single temperature [35]. While that work makes for important progress, the limited size and scope of the dataset restrict the broader application of these models.

To bridge the existing knowledge gap, we compiled an extensive dataset from the literature, consisting of 27,000 solubility data points measured in binary solvent mixtures at various temperatures (Fig. 1). The collected solubility studies were characterized by the conditions under which the measurements were made, as well as the experimental and computational features of the solutes and solvents. These features were then refined, with the goal of reducing the dimensionality of the dataset in terms of feature variance and correlation. Next, Bayesian hyperparameter optimization was applied to a selection of ML models to identify the most effective models and their optimal hyperparameter configurations. The identified models were further validated through a prospective study, demonstrating their potential for predicting drug solubility in specific binary solvent mixtures and at various temperatures. These models were particularly effective for drugs with properties well-represented within the dataset's scope. The collected dataset and Python code are published for open access to facilitate future research in solubility modeling.

Results

Dataset overview

The dataset for this study was compiled through a literature review using the Web of Science database, with search keywords "solubility" and "binary system". This literature search resulted in a dataset of 27,000 solubility data points, including 123 small-molecule solutes, 44 solvents, 110 binary solvent mixtures, and 373 unique solute-binary solvent systems. About 30% of the solutes are FDA-approved drugs [36]. The remaining compounds are mainly pharmaceutical intermediates, drug metabolites, and compounds with therapeutic potential. To the best of the authors' knowledge, this is the most extensive

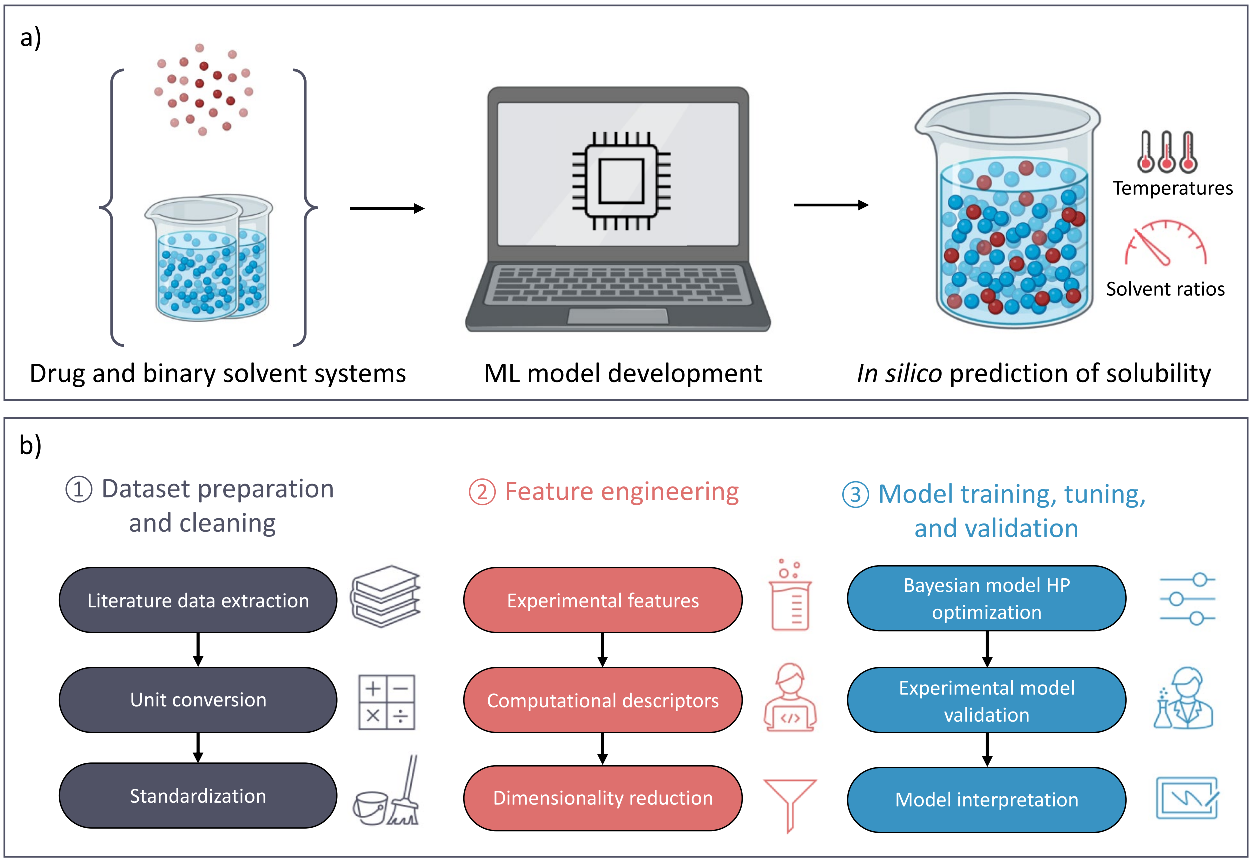

a)
Drug and binary solvent systems
ML model development
Temperatures
Solvent ratios
In silico prediction of solubility
b)
① Dataset preparation and cleaning
Literature data extraction
Unit conversion
Standardization
② Feature engineering
Experimental features
Computational descriptors
Dimensionality reduction
③ Model training, tuning, and validation
Bayesian model HP optimization
Experimental model validation
Model interpretation

Fig. 1 **a** Schematic outlining the study's objective of utilizing ML approaches for predicting drug solubility in binary solvent mixtures. **b** Outline of the methodology deployed in this study. First, a comprehensive dataset was compiled through a literature review, followed by cleaning and standardization of the data. Next, both experimental and computational descriptors were incorporated into the dataset as input features, which were then refined to mitigate the risks associated with high-dimensional datasets. Lastly, this refined dataset was used to optimize the models' hyperparameters (HPs), leading to identification of the top-performing models. Following its identification, these models were experimentally validated and then interpreted open-access dataset on solubility in binary solvent systems available to date. Figure 2 provides an overview of the dataset. In Fig. 2a, the data points are categorized into quantiles based on their solubilities, which are reported in LogS (S in g/100 g) in this study. The color gradient of the bars represents specific temperature ranges, illustrating the effects of temperature on the solubility. As expected, higher temperatures are usually associated with increased solubility [37–39]. For example, in the lowest solubility quantile ([− 5.48, − 0.43]), only 13.3% of samples were recorded at higher temperatures (above 313.15 K). This proportion nearly doubles to 27.6% in the fourth quantile, which corresponds to higher solubility levels ([1.00, 1.99]).

The dataset was then divided into two distinct parts: 75% for the training set and 25% for the test set as described in the method section. The training set was used for feature engineering and model hyperparameter optimization, while the test set for model evaluation to identify the optimal models. Figure 2b to e show the distributions of four key attributes of solute–solvent systems including investigation temperature, solute molecular weight, solvent melting temperature, and reported solubility, across both subsets. These plots demonstrate that the test set generally aligns with the training set, evidenced by their comparable quantile values, ensuring that the model evaluation is representative and thorough.

Feature engineering

Feature engineering is a critical step in the development of ML models, as it involves identifying the most informative features that can greatly improve accuracy by making it easier for models to learn the relationship between the inputs and the target [40–43]. For this study, the feature set initially collected from the literature was

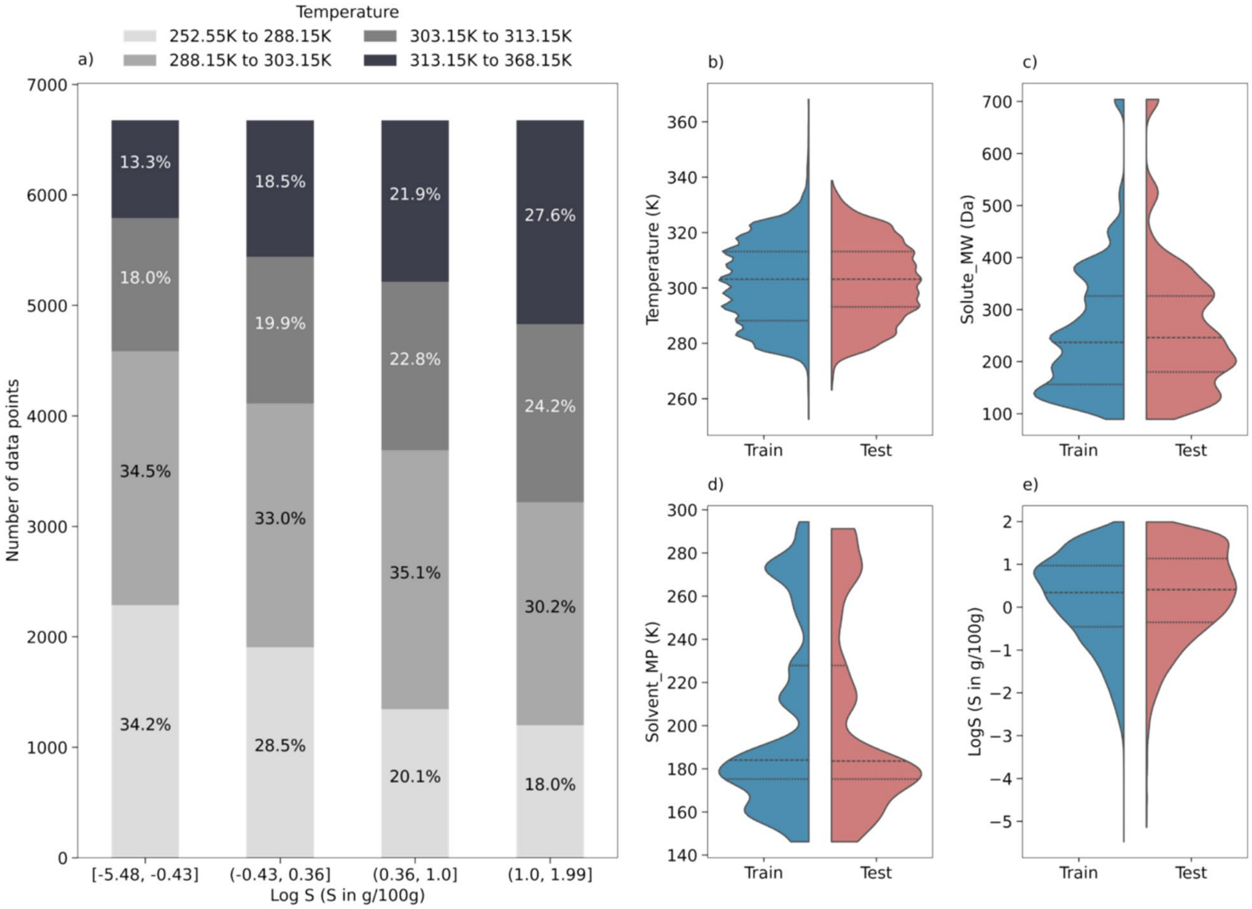


Fig. 2 **a** Shows an overview of the dataset, illustrating the distribution of solubility values across the whole dataset. The bars are color-coded with a gradient to signify the range of investigation temperatures: lighter shades represent lower temperatures, while darker shades indicate higher temperatures. This dataset is divided into training and test subsets, with their investigation temperature, solute molecular weight (MW), solvent melting temperature (MP), and solubility distributions shown in violin plots (**b**–**e**). The dashed lines in the violin plots indicate the first, second, and third quartiles of the feature distribution

limited to solubility measurement parameters only, including solvent ratios and temperature of investigation. However, to predict the solubility in solvent mixtures, it was necessary to also include descriptors of the solute and the two solvents.

The selection of compound melting temperatures as a crucial experimental attribute was informed by the well-established effect of melting temperature on solubility, as noted in the literature [44–46], and its widespread availability from open access sources such as the Chemical Abstracts Service (CAS) database [47] and supplier websites. To ensure accuracy, each melting temperature was obtained from three different sources, and the average was used for modeling in this work. The melting temperatures from these various sources demonstrated high consistency, with a mean coefficient of variation (CV) of 1% and a maximum CV of 5%. In addition to collecting solute melting temperatures from public sources, separately we generated predicted solute melting temperatures using an open-access model [48] and included in the dataset. Our comparison of model performance, as detailed in Sect. "The effects of melting temperature on model performance", when trained on published versus predicted solute melting temperatures aims to expand utility of the model to applications wherein molecules are not readily available for experimental determinations. Beyond melting temperatures, the dataset was enriched with computational descriptors for both solutes and solvents, incorporating MACCS molecular fingerprints and RDKit features to provide a comprehensive characterization. These features were generated by the RDKit package [49], a toolset for cheminformatics that processes molecular structures encoded in SMILES notation. A full list of these computational descriptors is included in Table S1.

However, the inclusion of the RDkit descriptors and MACCS fingerprints led to a substantial increase in dataset dimensionality, introducing over 1000 features for each solute-binary solvent system. High-dimensional data can pose challenges, such as overfitting, increased computational demands for model development, and difficulties in model interpretability [50–52]. To mitigate these challenges while maintaining dataset information, a standard feature refinement process based on feature variance and correlation was implemented. Initially, features demonstrating zero variance were removed, as their constant values across all observations provide no discriminative capability, thereby not enhancing the models' predictive performance. Next, features exhibiting high correlation (Pearson correlation coefficient > 0.8 [53–55]) with others in the dataset were also removed to reduce redundancy, which can burden the model and diminish its ability to generalize. Through these steps, the dataset was condensed to 362 features, effectively reducing it to approximately one-third of its original volume. This refined dataset was utilized in subsequent studies.

Beyond these measures, principal component analysis (PCA) was also employed as an additional step for further feature reduction. PCA is another widely employed technique to manage high-dimensional data by transforming a large set of variables into a smaller number of principal components [56–59]. For instance, as illustrated in Figure S1, PCA was performed on the refined dataset (containing 362 features) and 85 principal component features were identified as capable of explaining the majority of the dataset's variance (> 95%). However, employing PCA for dimensionality reduction comes with certain limitations, including challenges in model interpretability [60, 61] and the potential for dataset information loss during the transformation process [62, 63]. As shown in Figure S2, applying PCA for further feature reduction resulted in decreased model performance. Therefore, the dataset processed by PCA was not utilized further in this study.

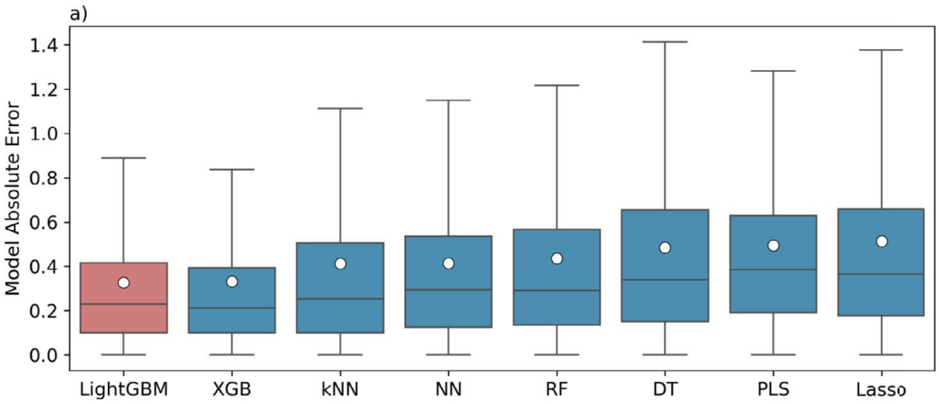
ML model development and evaluation

Using the refined dataset, a selection of ML models was trained and finetuned on the training set before their performance was assessed on the test set. The training process employed a ten-fold group cross-validation technique, a strategy that iterates the training and evaluation process across all folds. This approach ensures that the available data is utilized comprehensively, with the model accuracy from each fold being averaged to determine the model performance. To enhance the model performance, the model hyperparameters were tuned to identify the optimal ML configuration via Bayesian optimization [64]. Bayesian optimization has been found to be more efficient than traditional random and full grid search methods as it develops a probabilistic model that maps hyperparameters to a performance metric, facilitating a strategic search compared to brute-force approaches [65–67]. In addition, the literature also highlights Bayesian optimization's capability to manage continuous parameter spaces provided a finer search resolution, unlike grid-based approaches constrained by discrete parameters [68, 69]. This Bayesian optimization proceeded for 100 iterations to search for the optimal hyperparameters within the search space (Table S2). As shown in Figure S3, this Bayesian hyperparameter optimization improved the performance of all models investigated, but different models showed varying sensitivity to this process. For example, light gradient boosting machine (LightGBM) showed a marginal improvement in MAE (LogS) from 0.31 to 0.29, while the random forest (RF) model showed a more significant improvement in MAE (LogS) from 0.63 to 0.42. The hyperparameters identified for each model are summarized in Table S3.

Following hyperparameter tuning, all the models configured with their optimized hyperparameters were evaluated using the test set, with their performance shown in Fig. 3. Figure 3a shows the absolute error between the model predictions and the actual targets, arranging the models in ascending order based on their MAE (LogS) from the lowest to the highest. Notably, the LightGBM and XGB models exhibited the lowest MAE, signifying superior performance relative to the other models analyzed. The performance of these models was further measured by additional model metrics as shown in Fig. 3b, where the LightGBM and XGB models consistently exceeded the performance of the other models across these metrics. This effectiveness of gradient-boosted tree models (e.g., LightGBM and XGB) aligns with literature findings on chemical tabular data within similar data regimes [23, 34, 70–72]. In addition, the accuracy associated with these models are comparable, if not better, to benchmark ML studies in the literature, where MAE values (LogS) for new solute–solvent combinations are typically between 0.4 and 0.5 [26, 27]. This performance is especially notable given that most prior studies on solubility focus on less complicated systems, such as single solvents or constant temperature. Conversely, our model was trained to predict solubility in binary solvent systems measured over a variety of temperatures.

Prospective study

The performance of the developed LightGBM and XGB models was further assessed through an experimental prospective solubility study on four small-molecule



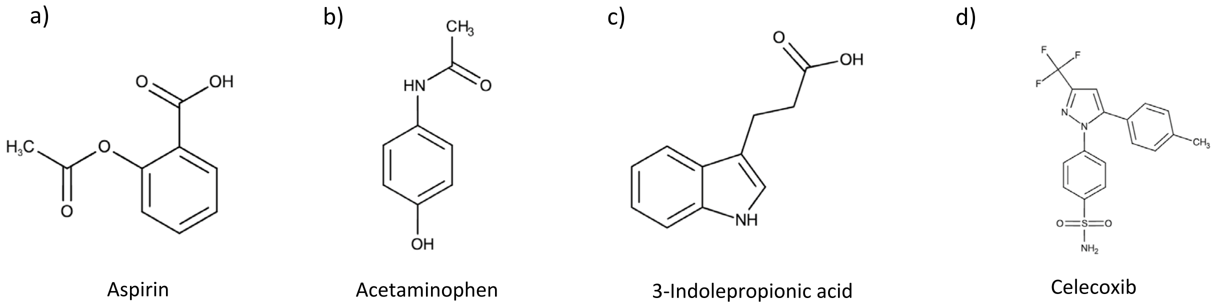
b)

	LightGBM	XGB	kNN	NN	RF	DT	PLS	Lasso
MAE	**0.33**	**0.33**	0.41	0.41	0.44	0.49	0.50	0.51
MedAE	0.23	**0.21**	0.25	0.30	0.29	0.34	0.39	0.36
RMSE	**0.48**	0.50	0.65	0.61	0.62	0.70	0.68	0.72
MSE	**0.23**	0.25	0.42	0.37	0.39	0.49	0.46	0.51
PCC	**0.90**	0.89	0.82	0.83	0.82	0.77	0.77	0.77
SCC	**0.88**	0.87	0.79	0.80	0.80	0.75	0.76	0.73

Fig. 3 **a** Illustrates the distribution of absolute error between experimental values and predictions for eight models, based on evaluations with the test set. Each boxplot highlights the mean absolute error (MAE) and median absolute error (MedAE) using white circles and black lines, respectively. Outliers are not included for clarity, and the detailed plots including outliers are available in Figure S4. **b** Summarizes the performance of these models using six metrics: MAE, MedAE, root mean square error (RMSE), mean square error (MSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC). Within the evaluated models, the two gradient-boosted tree models, namely LightGBM and XGB, showed superior performance compared to the rest, standing out across all considered metrics

compounds: aspirin (ASA), acetaminophen (ACM), 3-indolepropionic acid (IPA), and celecoxib (CXB). Among these compounds, ASA, ACM, and CXB are FDA-approved drugs, and IPA has been evaluated for it neuroprotective and anti-inflammatory properties for use in indications such as brain injury [73–75], stroke [76–78], and neurodegenerative diseases [79–81]. As shown in Fig. 4, these molecules were purposefully selected for their varied structural features, molecular weights, and LogP values. In addition, ACM and CXB were previously included in our dataset, while ASA and IPA represent novel solutes. The solubility measurements for these molecules were performed in ethanol/water mixtures, presenting the models with solute-binary solvent systems they had not previously encountered in the training set. To examine the effects of measurement parameters on solubility and model predictions, these studies were performed at two temperatures (298.15 and 313.15 K) and three solvent ratios (ethanol content = 0.2, 0.6, and 0.8; weight ratio). In total, 24 solubility studies were performed in triplicate as detailed in Fig. 4e.

The solubility measurement results are shown in Fig. 5a to d, where slashed and non-slashed bars represent results at higher (313.15 K) and lower (298.15 K) temperatures. The solubility values measured spanned from − 3 to 1.5 LogS (S in g/100 g), which is consistent with the range observed in the dataset. As expected, the results show that solubility measurement parameters significantly influence solute solubility, with an increase in solubility resulting from a rise in temperature or increase in ethanol content in the solvent mixture. Following the solubility measurements, the models' ability to predict solubility for these solutes was evaluated, using six metrics to assess accuracy for each compound. For the



e)

		Aspirin	Acetaminophen	3-Indolepropionic acid	Celecoxib
CAS No.		50-78-2	103-90-2	830-96-6	169590-42-5
LogP		1.31	1.35	2.19	3.51
Molecular weight (Da)		180	151	189	381
Drug in the dataset		No	Yes	No	Yes
Drug-solvent system in the dataset		None	DMSO/Water 4-Formylmorpholine/Water DMF/Water	None	NMP/1-Propanol NMP/Ethanol
Prospective study	Solvent mixture	Ethanol (w1)/water			
	Temperature	[298.15K, 313.15K]			

Note: w1 represents the mass ratios of ethanol [0.2, 0.5, 0.8] in the ethanol-water solvent systems.
NMP, N-methyl-2-pyrrolidone; DMF, Dimethylformamide; DMSO, Dimethyl sulfoxide.

Fig. 4 Overview of the prospective study design including four compounds: aspirin, acetaminophen, 3-indolepropionic acid, and celecoxib. **a**–**d** Show the structures of each solute. **e** Provides a summary of the compound properties, their inclusion in the dataset, and the conditions under which the solubility studies were performed

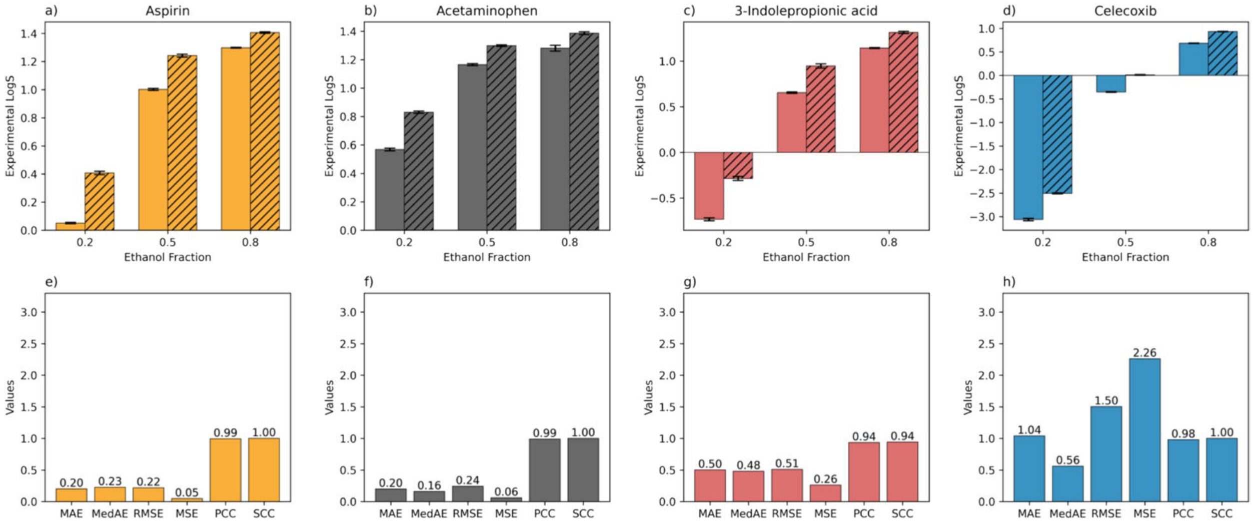


Fig. 5 a–**d** The experimental solubility data for aspirin, acetaminophen, 3-indolepropinoic acid, and celecoxib in three ethanol/water mixtures (i.e., weight ratio of ethanol = 0.2, 0.5, 0.8) at temperatures of 298.15 K (non-slashed bars) and 313.15 K (slashed bars). **e**–**h** The predictive accuracy of the LightGBM model for these solubility measurements, evaluated through various metrics including mean absolute error (MAE), median absolute error (MedAE), root mean square error (RMSE), mean square error (MSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC)

LightGBM model, Fig. 5e to g indicate that the LightGBM model is effective at predicting solubility for ASA, ACM, and IPA, with MAE of 0.20, 0.20, and 0.50, respectively. The model's ability to accurately predict the solubility of ACM may be due to its presence in the training dataset. Although ACM's solubility in ethanol/water mixtures was not included in the dataset, the data on ACM's solubility in other solvents likely contributed to the model's proficiency in handling new combinations of solute and solvent. Despite ASA and IPA being new additions to the model, their predicted solubilities were notably accurate and on par with the top-performing models described in existing research, which have focused on solubility in single solvents under both constant and variable temperature conditions [26, 27]. This accuracy in prediction could stem from the fact that most of the key features for ASA and IPA fall within the scope of the dataset (to be detailed in the following paragraph). Such alignment enables the model to effectively generalize to these two new compounds, leveraging the diverse and extensive range of features present in the existing dataset which encompasses more than 100 distinct solutes.

However, the LightGBM model's predictions for CXB were less accurate despite CXB being present in the dataset. To explore the underlying reasons for this discrepancy, we examined the top 15 solute features deemed most influential to the LightGBM model's decisions (Fig. 6a). The values for these features for the four compounds (ASA, ACM, IPA, and CXB) were then plotted compared with overall distribution of all the solutes within the dataset. As shown in Figure S6, it was observed that CXB deviates significantly from the dataset's distribution in over half of (8 out of 15) of these crucial features, either falling below the first quantile, exceeding the third quantile, or being considered outliers. We attribute the worse predictive performance of the model on CXB due to higher levels of deviation from the training set in the feature space than ASA (6 of out 15), ACM (1 out of 15), and IPA (5 out of 15). The deviation of the four compounds from the dataset was further quantified by calculating their Mahalanobis distances, as discussed in Sect. "Model interpretation to understand prediction results and model limitations".

Considering the comparable performance of the LightGBM and XGB models during the development stage, a parallel analysis was performed for the XGB model. This included both validation and interpretation with respect to critical solute attributes. During model validation, the XGB model's performance mirrored that of the LightGBM model, demonstrating improved accuracy in predicting the solubilities of ASA, ACM, and IPA compared to CXB, with a similar accuracy level to LightGBM (difference in MAE for LogS within 0.15). Further examination of the feature distribution highlighted by XGB revealed a pattern consistent with the analysis of the LightGBM model, indicating higher deviation in CXB for most of the critical features relative to ASA, ACM, and IPA. These results are summarized in Figures S7, S8, S9, and S10.

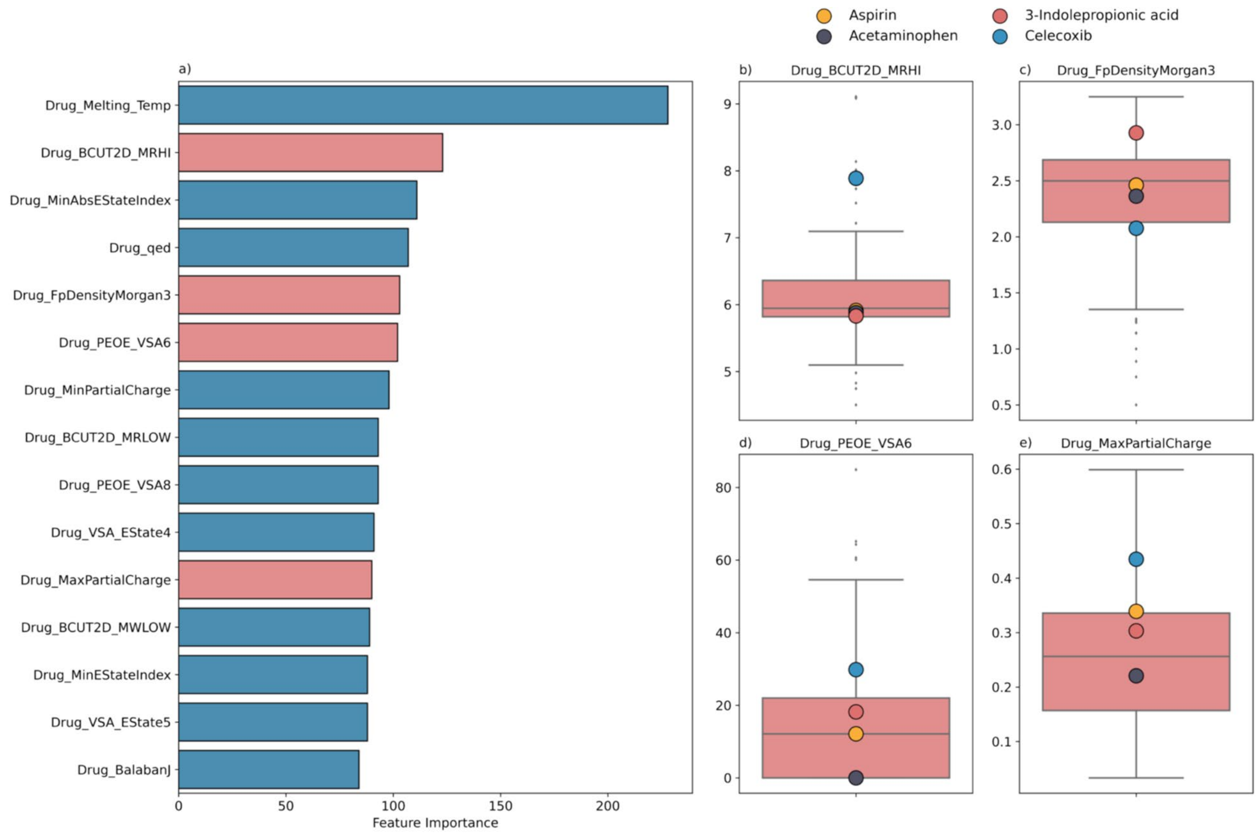


Fig. 6 **a** The ranking of the top 15 most important solute features identified by the LightGBM model, ordered by their importance from top to bottom, with representative features for further analysis highlighted in red. **b**–**e** Comparison of the values of these features for the four compounds evaluated in the prospective study, juxtaposed against the distribution of these features across all solutes in the dataset, providing a visual depiction of their alignment or deviation from the dataset range

Discussion

The effects of melting temperature on model performance

The melting temperature of a compound is crucial for understanding and predicting its solubility. This has been well-established in the literature with many of solubility equations [82, 83] and ML models [84, 85] developed based on the melting temperature. Similarly, in this study, melting temperature was also found to be one of the most significant features as indicated by the feature importance analysis (Fig. 6). However, the inclusion of melting temperature can also limit the application of the models as the melting temperature of a compound is not always available, which can be a common issue in fields where new compound synthesis is a primary focus.

To address this limitation, in addition to the models developed using solute melting temperatures collected from public sources, we also evaluated the performance of models developed using predicted solute melting temperatures as well as models developed omitting solute melting temperature as an input feature. Given the nature of our dataset, the predicted solute melting temperatures were generated using an open-access model [86] specifically designed for drug-like compound melting temperature prediction. These predicted melting temperatures were compared to those reported by public sources, and the chosen melting temperature model achieved a MAE of approximately 30 K, which is aligned with the accuracy of this type of model, as reported in the literature [87–90]. As shown in Table S4, models trained with both collected and predicted melting temperatures performed comparably and outperformed models that did not include solute melting temperatures. For example, the LightGBM model trained with collected and predicted melting temperatures showed an MAE for ASA LogS within 0.2, while those trained without melting points had a significantly higher MAE at 0.66 (Table S4). These results show that although melting temperature is crucial for predicting solubility accurately, using predicted melting temperatures can achieve comparable predictive accuracy. For applications where using predicted melting temperatures is necessary, it is recommended that users employ a model specifically developed for their compound type to achieve more accurate melting point predictions, and then use those predicted melting points as an input feature to generate solubility predictions using the solubility models developed in this study.

Featurizing solutes and solvents with descriptors

Feature engineering aims to create and identify meaningful features that effectively represent the dataset, thereby improving the performance of models. This step is crucial in ML as it directly impacts the model's ability to make accurate predictions [91, 92]. One commonly used feature engineering approach for molecular structures is to convert them into numerical features to facilitate model training [93–95]. In this study, this was done by generating RDKit descriptors and MACCS fingerprints for solutes and solvents based on their structures. These features enable the ML models to differentiate between various solutes and solvents effectively. The adoption of these computational descriptors has seen broad application in various fields, including drug discovery [96–99] and materials science [25, 100–106]. In addition to their prevailing usage, the rationale for selecting MACCS molecular fingerprints and RDKit features was also attributed to their interpretability. These descriptors offer either direct associations with specific molecular structures or yield computational molecular properties.

To evaluate whether additional features could improve model accuracy, more solute/solvent descriptors, primarily including 3D descriptors and quantum mechanics (QM)-based descriptors as listed in Table S1, were incorporated into the dataset to develop the models using the same workflow. However, as shown in Table S4, these additional features generally reduced model performance rather than improving accuracy. For example, when using the collected melting temperatures to train the models, the resulting XGB model showed a 0.2 to 0.3 increase in MAE for LogS, indicating a decrease in model performance with these additional features compared to models trained without them (Table S1). This observation aligns with existing literature, indicating that these more complex descriptors do not always enhance model performance [107–110]. For example, it was reported that 3D descriptors usually do not perform as well as 2D descriptors in quantitative structure–activity relationship (QSAR) and ML applications for molecular representations [107]. Moreover, these descriptors generally require greater computational power for feature generation and can complicate model interpretation. Consequently, these features were not further explored in this research. Nonetheless, recognizing that the impact of these features can vary depending on the problem, the methods for generating them have been included in the method section and published code to support their future application.

Model interpretation to understand prediction results and model limitations

ML models are often seen as black boxes because the process they use to transform input data into output predictions is typically complex and not easily interpretable [111, 112]. However, interpreting ML models is important for understanding their performance and building trust in the model's accuracy. In this study, feature importance analysis was performed as an interpretative tool to understand the deviation in the model accuracy for CXB compared to ASA, ACM, and IPA. This analysis demonstrated that the high-importance features which describe CXB deviated significantly from those of most solutes in the dataset. For instance, Fig. 6c shows that most solutes in the dataset have a BCUT2D_MRHI value ranging from 5.5 to 6.5, with ASA, ACM, and IPA fitting well within this range, whereas CXB stands out as an outlier. The BCUT2D represents a group of 2D molecular topology descriptors that have been employed in the literature as input features for solubility modeling [113–116]. Upon identifying this deviation in CXB's featurization, where it presents as a molecule near the bounds of the high importance descriptors, we performed clustering analysis to further quantify the distribution of feature values in the dataset. All the solutes in the dataset were clustered using k-means clustering based on the solute descriptors used for model development. As shown in Table S5, eight clusters were identified, with more than half of the solutes falling into the two largest clusters. Next, the Mahalanobis distance was computed between the four tested compounds (ASA, ACM, IPA, and CXB) and all the eight clusters. The results showed that CXB significantly deviates from the two main clusters, with distances ranging from 8 to 10, almost double compared to the other three drugs, which ranged from 4 to 6. This indicates that CXB is relatively distant from the solutes currently in the dataset compared to the other three compounds, which are predicted with better accuracy. Thus, the applicability domain of the developed models corresponds to the predominant type of solutes in the dataset, which are primarily of lower molecular weight and relatively high aqueous solubility.

However, the dominance of these types of solutes also highlights one of the limitations of our dataset. Drugs that are approved and currently under development have a wide range of physico-chemical properties and thus are not fully represented by the molecules in our dataset. In response to this limitation and to further expand model applicability, we are openly sharing the dataset and source code. This is designed to foster additional research and innovation, with the hope that other research groups will build upon our dataset, thereby broadening its scope to encompass a more extensive range of drug properties. Such expansion is anticipated to refine the models' predictive accuracy across a more diverse array of compounds. By making these resources available [117], we aim to catalyze advancements in the field, encouraging collaborative efforts to overcome existing challenges and push the boundaries of what is currently achievable in predictive modeling.

Conclusion

In conclusion, this study reports the development, optimization, and validation of ML models for predicting the solubility of drugs in binary solvent mixtures at different temperatures. To construct this model, a comprehensive dataset containing 27,000 solubility entries was sourced from the literature. This dataset was further enriched with the melting temperatures of solutes and solvents, as well as computational descriptors. A subset of this dataset was utilized to train and optimize several ML models through Bayesian hyperparameter optimization. Among the investigated models, those demonstrating superior performance on the test set were further validated via prospective solubility studies using FDA-approved drugs and a compound that has been shown to have therapeutic potential. The validation results highlight the potential of the developed models for predicting drug solubility in given binary solvent mixtures and under varying temperature conditions. To advance research in the area and encourage enhancements to the models, the dataset, models, and source code used in this study have been made available. This initiative seeks to inspire future work aimed at expanding the dataset's comprehensiveness and improving the model's utility for a broader spectrum of compounds.

Methods

Data collection

The dataset for this study was collected via a literature review conducted using the Web of Science database and keywords "solubility" and "binary system". Search results were limited to research articles published since 2019. Each paper was manually reviewed to ensure relevance and clarity in reporting necessary experimental parameters for solubility measurement, resulting in a total of 125 relevant papers published by different research groups. Information was extracted from these papers including the names of solutes and solvents, solvent ratios, temperatures, and solubility values. In this dataset, the two solvents in the binary mixtures were designated as "Solvent 1" and "Solvent 2". For a specific drug-binary solvent system, "Solvent 1" refers to the solvent in a binary mixture that has a higher Pearson correlation with the solubility of the solute compared to "Solvent 2". Solvent ratios, reported in the literature as either molar or mass ratios, were interconverted based on the reported values and the molecular weights of the solvents. Solubility values, originally reported in mole percent or weight percent, were converted to standardized LogS (S in g/100 g) for consistency with previous research. SMILES strings and melting temperatures of the solutes and solvents were obtained from the open-access databases (Chemical Abstracts Service [47], PubChem [118], Wikipedia [119], DrugBank [120], ChemSpider [121], ChemBK [122], Pesticide Properties Database [123], and ChemSrc [124]) as well as supplier websites (LKT Labs [125], Chemical Book [126], Sigma [127], Santa Cruz Biotechnology [128], Fisher Scientific [129], AK Scientific [130], Moltus Research Laboratories [131], TCI America [132], GuideChem [133], ECHEMI [134], and EBCLink [135]). To ensure accuracy of the collected melting temperature, each melting temperature was collected from three different sources, and their averages were used for modeling. In addition, for evaluating model performance when solute melting temperatures are unavailable, predicted melting temperatures were generated using the OCHEM Predictor [48] based on the models reported by Tetko et al. [86].

Data preprocessing and splitting

The collected dataset was then preprocessed to check for duplicates, conflicting values, and ensure compound structural standardization. After examining the entire dataset with Python, around 1300 data points (less than 5% of the total dataset) were identified as duplicates; in these cases, the first data entry was retained, and subsequent duplicates were removed. These duplicates occurred when the same solubility data was reported in different tables within the same paper for comparison purposes, so no conflicting results were found. For structural standardization, tautomers and charge states were standardized using the rdMolStandardize module in RDKit. Additionally, possible diastereomers were reviewed to ensure that the collected structures accurately represented the specific diastereomers reported in the literature.

This preprocessed dataset was split into training and test sets using a group-based dataset splitting method via GroupShuffleSplit. The parameter 'group' was set as the solute-binary solvent systems to ensure that the same solute-binary solvent systems measured at different solvent ratios or temperatures were not repeated across the training and test sets. The training set (75%) was used for feature engineering and model development while the test set (25%) was used for model evaluation.

Feature engineering

Solutes and solvents in the preprocessed dataset were featurized using MACCS molecular fingerprints and RDKit features for the solute and solvents, which were calculated using the RDKit package [49]. To evaluate the effects of an expanded feature set, additional features including 3D descriptors, molecular volume, dipole moment, dielectric constant, and QM-based descriptors were also generated for comparison. Specifically, molecular volume and dipole moment were computed using the RDKit package [49], solvent dielectric constant was collected from public databases [136–138], 3D descriptors were generated using rdkit. Chem.Descriptors3D module [139], and QM-based descriptors were generated using the Morfeus package [140]. A full list of these descriptors is included in Table S1. The features were then refined by removing those with no variability (variance = 0) and those that were highly correlated with others (Pearson correlation coefficient > 0.8).

ML model training and hyperparameter optimization

A panel of eight ML models was investigated, including the light gradient boosting machine (LightGBM), extreme gradient boosting (XGB), k-nearest neighbors (kNN), neural network (NN), random forest (RF), decision tree (DT), linear regression with least absolute shrinkage and selection operator (Lasso) regularization, and partial least squares (PLS). The implementation of all models was carried out using the Scikit-learn library [141], except for LightGBM and XGB, which were developed using the LightGBM [142] and XGB [143] packages, respectively. The optimization of the models' hyperparameters was conducted using Bayesian optimization, facilitated by the BayesSearchCV function from the Scikit-Optimize library [64]. This hyperparameter optimization process entailed a ten-fold group cross-validation strategy, executed for 100 iterations to search for the optimal set of model hyperparameters. Same as the dataset splitting method (Sect. "Data preprocessing and splitting"), the parameter 'group' was set as the solute-binary solvent systems. Model development was performed on a Mac mini with an 8-core Apple M1 chip and 8 GB of RAM. With this computer, the majority of models (kNN, PLS, Lasso, NN, DT) take less than one hour to complete each hyperparameter optimization, while other models (RF, LightGBM, and XGB) take three to five hours each.

ML model evaluation

Following the optimization of hyperparameters, the models were evaluated to identify the top performers. This involved generating solubility predictions for the test set, where the predictive accuracy of each model was determined by comparing the predictions to experimental values collected from the literature. For this assessment, six ML metrics were implemented: mean absolute error (MAE), median absolute error (MedAE), root mean square error (RMSE), mean square error (MSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC). The formulas for these metrics are as follows.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{1}$$

$$\mathrm{MedAE} = \mathrm{median}(|y_i - \hat{y}_i|) \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \tag{3}$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N} \tag{4}$$

$$\mathrm{PCC} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{5}$$

$$\mathrm{SCC} = \frac{\mathrm{cov}(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{6}$$

ML model feature importance analysis

Following the model evaluation using the test set, feature importance of the top-performing models was computed to understand the underlying factors contributing to its predictive capability. Specifically, the identified models (i.e., LightGBM and XGB) with their optimized hyperparameters were trained on the entire dataset. The importance of each feature was computed from the trained model using the built-in feature importance function in the LightGBM [144] and XGB [143] packages.

Cluster analysis

To better understand the dataset and the model prediction results, solutes in the dataset were clustered. This process involved standardizing the solute features used for modeling with sklearn.preprocessing. StandardScaler [145], followed by clustering using sklearn.cluster.KMeans [146]. Subsequently, the

Mahalanobis distance between these identified clusters and the four compounds (ASA, ACM, IPA, and CXB) in the prospective study was computed using scipy.spatial.distance.mahalanobis [147].

Prospective study

Materials

Aspirin (acetylsalicylic acid, ASA, ≥ 99.0%), acetaminophen (ACM, ≥ 99.0%), 3-indolepropionic acid (IPA, 99%), acetic acid (≥ 99.0%), and trifluoroacetic acid (TFA, 99%) were purchased from Sigma Aldrich (ON, CA). Celecoxib (CXB, > 98.0%) was purchased from TCI Chemicals (Japan). Ethanol (anhydrous) was purchased from Commercial Alcohols (ON, CA). Acetonitrile (ACN, HPLC grade) was purchased from Fisher Chemical (ON, CA). Methanol (HPLC grade) was purchased from Caledon Laboratories (ON, CA). Ammonium acetate (≥ 97.0%) was purchased from BioShop (ON, CA).

Solubility measurement

To test the model performance prospectively, the solubility of four drug or drug-like compounds (ASA, ACM, IPA, and CXB) was measured in ethanol–water mixtures with varying solvent ratios and temperatures. Ethanol–water binary mixtures were prepared in three different ratios, with ethanol present as 20%, 50%, or 80% by weight of the mixture (measured by PB303-S balance, Mettler Toledo). For each compound, an excess of solute was added to each solvent mixture, sealed, and stored in darkness overnight. Each solubility experiment was performed in triplicates (n = 3) to ensure reproducibility and accuracy of the results. Two temperatures, 298.15 K and 313.15 K, were maintained using an incubator (INCU-Line®, VWR) to examine their effects on solubility. After the incubation period, the entire mixture (containing dissolved and undissolved solute) was then filtered using a syringe filter (0.22 μm pore size, polyvinylidene fluoride, Millex®, Sigma Aldrich) to separate the solution from the undissolved. The filtrates (i.e., saturated solution) were appropriately diluted with ethanol for high-performance liquid chromatography (HPLC) analysis to determine their concentrations.

HPLC analysis

The concentrations of ASA, ACM, IPA, and CXB were measured via HPLC using the Agilent Technologies 1260 Infinity II system. ACM was measured using a Restek Raptor ARC-18 column (150 mm × 4.6 mm, 2.7 μm). ASA, IPA, and CXB were measured using an Eclipse XDB-C18 column (150 mm × 4.6 mm, 5 μm). The detailed HPLC parameters are summarized in Table S6.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00911-3.

Additional file 1.

Acknowledgements
The graphical abstract was created using BioRender.

Author contributions

Z.B. co-led the conceptualization of this work, led dataset construction, led the computational studies, led the data analysis, performed the experimental studies, and wrote the first draft of the paper. G.T., A.C., and J.W. assisted with the computational studies, assisted with data analysis, and assisted with the review/editing of the paper. A.A.G. assisted with supervising the work, acquired research funding, and assisted with review/editing the paper. C.A. co-led conceptualization of the work, led the supervision of the work, acquired research funding, and led the review/editing of the paper. All authors read and approved the final manuscript.

Funding
This research was undertaken thanks in part to funding provided by an NSERC Discovery grant (RGPIN-2022-04910) to C.A. and to the University of Toronto's Acceleration Consortium from the Canada First Research Excellence Fund CFREF-2022-00042. A.A.G. acknowledges support from the Defense Advanced Research Projects Agency under the Accelerated Molecular Discovery Program under Cooperative Agreement No. HR00111920027 dated August 1, 2019. The content of the information presented in this work does not necessarily reflect the position or the policy of the U.S. government. Z.B. acknowledges a CPO scholarship from the Centre for Pharmaceutical Oncology, Leslie Dan Faculty of Pharmacy, University of Toronto. G.T. acknowledges financial support from NSERC and the Vector Institute. A.A.G. would like to thank Dr. Anders Frøseth for his support.

Data availability

The dataset, codes, and models that support the findings of this study are available at the GitHub (https://github.com/Christine-Allen-Lab/Solubility_ML).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

C.A. is a cofounder and CEO of Intrepid Labs Inc., A.A.G. is a cofounder of Intrepid Labs Inc., Kebotix Inc., and Zapata AI.

Author details
[1]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON M5S 3M2, Canada. [2]Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada. [3]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada. [4]Vector Institute for Artificial Intelligence, Toronto, ON M5S 1M1, Canada. [5]Acceleration Consortium, Toronto, ON M5S 3H6, Canada. [6]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, ON M5S 1M1, Canada. [7]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S 3E5, Canada. [8]Department of Materials Science and Engineering, University of Toronto, Toronto, ON M5S 3E4, Canada. [9]CIFAR Artificial Intelligence Research Chair, Vector Institute, Toronto, ON M5S 1M1, Canada.

Received: 26 March 2024 Accepted: 28 September 2024
Published online: 28 October 2024

References

1. Alsenz J, Kansy M (2007) High throughput solubility measurement in drug discovery and development. Adv Drug Deliv Rev 59:546–567. https://doi.org/10.1016/j.addr.2007.05.007
2. Salo-Ahen OMH, Alanko I, Bhadane R, Bonvin AMJJ, Honorato RV, Hossain S, Juffer AH, Kabedev A, Lahtela-Kakkonen M, Larsen AS, Lescrinier E, Marimuthu P, Mirza MU, Mustafa G, Nunes-Alves A, Pantsar T, Saadabadi A, Singaravelu K, Vanmeert M (2021) Molecular dynamics simulations in drug discovery and pharmaceutical development. Processes 9:71. https://doi.org/10.3390/pr9010071
3. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK (2021) Artificial intelligence in drug discovery and development. Drug Discov Today 26:80–93. https://doi.org/10.1016/j.drudis.2020.10.010
4. Veseli A, Žakelj S, Kristl A (2019) A review of methods for solubility determination in biopharmaceutical drug characterization. Drug Dev Ind Pharm 45:1717–1724. https://doi.org/10.1080/03639045.2019.1665062
5. Pedersen-Bjergaard S, Rasmussen KE, Brekke A, Ho TS, Grønhaug Halvorsen T (2005) Liquid-phase microextraction of basic drugs—selection of extraction mode based on computer calculated solubility data. J Sep Sci 28:1195–1203. https://doi.org/10.1002/jssc.200401935
6. Salunke S, O'Brien F, Cheng Thiam Tan D, Harris D, Math M-C, Ariën T, Klein S, Timpe C (2022) Oral drug delivery strategies for development of poorly water soluble drugs in paediatric patient population. Adv Drug Delivery Rev 190:114507. https://doi.org/10.1016/j.addr.2022.114507
7. Khan KU, Minhas MU, Badshah SF, Suhail M, Ahmad A, Ijaz S (2022) Overview of nanoparticulate strategies for solubility enhancement of poorly soluble drugs. Life Sci 291:120301. https://doi.org/10.1016/j.lfs.2022.120301
8. Ainurofiq A, Putro DS, Ramadhani DA, Putra GM, Do Espirito Santo LDC (2021) A review on solubility enhancement methods for poorly water-soluble drugs. J Reports Pharm Sci 10:137. https://doi.org/10.4103/jrptps.JRPTPS_134_19
9. Saal C, Petereit AC (2012) Optimizing solubility: kinetic versus thermodynamic solubility temptations and risks. Eur J Pharm Sci 47:589–595. https://doi.org/10.1016/j.ejps.2012.07.019
10. Barrett JA, Yang W, Skolnik SM, Belliveau LM, Patros KM (2022) Discovery solubility measurement and assessment of small molecules with drug development in mind. Drug Discovery Today 27:1315–1325. https://doi.org/10.1016/j.drudis.2022.01.017
11. Csicsák D, Borbás E, Kádár S, Tőzsér P, Bagi P, Pataki H, Sinkó B, Takács-Novák K, Völgyi G (2021) Towards more accurate solubility measurements with real time monitoring: a carvedilol case study. New J Chem 45:11618–11625. https://doi.org/10.1039/D1NJ01349A
12. Sou T, Bergström CAS (2018) Automated assays for thermodynamic (equilibrium) solubility determination. Drug Discov Today Technol 27:11–19. https://doi.org/10.1016/j.ddtec.2018.04.004
13. Huang G, Guo Y, Chen Y, Nie Z (2023) Application of machine learning in material synthesis and property prediction. Materials 16:5977. https://doi.org/10.3390/ma16175977
14. Mitchell JBO (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481. https://doi.org/10.1002/wcms.1183
15. Stienstra CMK, Ieritano C, Haack A, Hopkins WS (2023) Bridging the Gap between differential mobility, Log S, and Log P using machine learning and SHAP analysis. Anal Chem 95:10309–10321. https://doi.org/10.1021/acs.analchem.3c00921
16. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN (2020) Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun 11:5753. https://doi.org/10.1038/s41467-020-19594-z
17. Li M, Chen H, Zhang H, Zeng M, Chen B, Guan L (2022) Prediction of the aqueous solubility of compounds based on light gradient boosting machines with molecular fingerprints and the cuckoo search algorithm. ACS Omega 7:42027–42035. https://doi.org/10.1021/acsomega.2c03885
18. Tosca EM, Bartolucci R, Magni P (2021) Application of artificial neural networks to predict the intrinsic solubility of drug-like molecules. Pharmaceutics 13:1101. https://doi.org/10.3390/pharmaceutics13071101
19. Ahmad W, Tayara H, Chong KT (2023) Attention-Based graph neural network for molecular solubility prediction. ACS Omega 8:3236–3244. https://doi.org/10.1021/acsomega.2c06702
20. Cui Q, Lu S, Ni B, Zeng X, Tan Y, Chen YD, Zhao H (2020) Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. Front Oncol. https://doi.org/10.3389/fonc.2020.00121
21. Lovrić M, Pavlović K, Žuvela P, Spataru A, Lučić B, Kern R, Wong MW (2021) Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: generalization, complexity, or predictive ability? J Chemom 35:e3349. https://doi.org/10.1002/cem.3349
22. Delaney JS (2004) ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 44:1000–1005. https://doi.org/10.1021/ci034243x
23. Tom G, Hickman RJ, Zinzuwadia A, Mohajeri A, Sanchez-Lengeling B, Aspuru-Guzik A (2023) Calibration and generalizability of probabilistic models on low-data chemical datasets with DIONYSUS. Digital Discovery 2:759–774. https://doi.org/10.1039/D2DD00146B
24. Griffiths RR, Klarner L, Moss H, Ravuri A, Truong S, Du Y, Stanton S, Tom G, Rankovic B, Jamasb A, Deshwal A, Schwartz J, Tripp A, Kell G, Frieder S, Bourached A, Chan A, Moss J, Guo C, Durholt J, Chaurasia S, Strieth-Kalthoff F, Lee AA, Cheng B, Aspuru-Guzik A, Schwaller P, Tang J (2023) GAUCHE: a library for gaussian processes in chemistry. https://doi.org/10.48550/arXiv.2212.04450
25. Kim S, Jinich A, Aspuru-Guzik A (2017) MultiDK: a multiple descriptor multiple kernel approach for molecular discovery and its application to organic flow battery electrolytes. J Chem Inf Model 57:657–668. https://doi.org/10.1021/acs.jcim.6b00332
26. Vassileiou AD, Robertson MN, Wareham BG, Soundaranathan M, Ottoboni S, Florence AJ, Hartwig T, Johnston BF (2023) A unified ML framework for solubility prediction across organic solvents. Digital Discovery 2:356–367. https://doi.org/10.1039/D2DD00024E
27. Ye Z, Ouyang D (2021) Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. J Cheminform 13:98. https://doi.org/10.1186/s13321-021-00575-3
28. Vermeire FH, Chung Y, Green WH (2022) Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. J Am Chem Soc 144:10785–10797. https://doi.org/10.1021/jacs.2c01768
29. Osorio IP, Martínez F, Peña MÁ, Jouyban A, Acree WE Jr (2021) Solubility of sulphadiazine in some Carbitol® (1) + water (2) mixtures: determination, correlation, and preferential solvation. Phys Chem Liq 59:890–906. https://doi.org/10.1080/00319104.2020.1858420
30. Rahimpour E, Azarmir O, Hassanzadeh D, Nokhodchi A, Jouyban A (2021) Solubility of paracetamol in the ternary solvent mixtures of water + ethanol + glycerol at 298.2 and 303.2 K. Phys Chem Liq 59:827–834. https://doi.org/10.1080/00319104.2020.1849208
31. Maheri A, Ghanbarpour P, Rahimpour E, Acree WE Jr, Jouyban A, Azarbayjani AF, Kouhkan M (2021) Solubilisation of dexamethasone: experimental data, co-solvency and Polarised Continuum Modelling. Phys Chem Liq 59:817–826. https://doi.org/10.1080/00319104.2020.1836640
32. Jagdale SK, Nawale RB (2020) Estimation and correlation of solubility of practically insoluble drug itraconazole in 1,4-butanediol + water mixtures using extended hildebrand solubility approach. J Pharm Innov 15:344–356. https://doi.org/10.1007/s12247-019-09384-6
33. Gasmi H, Siepmann F, Hamoudi MC, Danede F, Verin J, Willart J-F, Siepmann J (2016) Towards a better understanding of the different release phases from PLGA microparticles: dexamethasone-loaded systems. Int J Pharm 514:189–199. https://doi.org/10.1016/j.ijpharm.2016.08.032
34. Bannigan P, Bao Z, Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A, Allen C (2023) Machine learning models to accelerate the design of polymeric long-acting injectables. Nat Commun 14:35. https://doi.org/10.1038/s41467-022-35343-w
35. Chinta S, Rengaswamy R (2019) Machine learning derived quantitative structure property relationship (QSPR) to predict drug solubility in binary solvent systems. Ind Eng Chem Res 58:3082–3092. https://doi.org/10.1021/acs.iecr.8b04584
36. Drugs@FDA: FDA-Approved Drugs (n.d.) https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm (accessed March 13, 2024)
37. Zheng B, McClements DJ (2020) Formulation of more efficacious curcumin delivery systems using colloid science: enhanced solubility,

stability, and bioavailability. Molecules 25:2791. https://doi.org/10.3390/molecules25122791

38. An M, Yi D, Qiu J, Liu H, Hu S, Han J, Guo Y, Huang H, He H, Wang P (2020) Measurement and correlation for solubility of moroxydine hydrochloride in pure and binary solvents. J Chem Eng Data 65:2611–2618. https://doi.org/10.1021/acs.jced.0c00015
39. Moradi M, Rahimpour E, Hemmati S, Martinez F, Barzegar-Jalali M, Jouyban A (2020) Solubility of mesalazine in polyethylene glycol 400 + water mixtures at different temperatures. J Mol Liq 314:113546. https://doi.org/10.1016/j.molliq.2020.113546
40. Verdonck T, Baesens B, Óskarsdóttir M, van den Broucke S (2021) Special issue on feature engineering editorial. Mach Learn. https://doi.org/10.1007/s10994-021-06042-2
41. Zheng A, Casari A (2018) Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly Media, Inc.
42. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C (2021) Machine learning directed drug formulation development. Adv Drug Deliv Rev 175:113806. https://doi.org/10.1016/j.addr.2021.05.016
43. Bao Z, Bufton J, Hickman RJ, Aspuru-Guzik A, Bannigan P, Allen C (2023) Revolutionizing drug formulation development: the increasing impact of machine learning. Adv Drug Deliv Rev 202:115108. https://doi.org/10.1016/j.addr.2023.115108
44. Nordström FL, Rasmuson ÅC (2009) Prediction of solubility curves and melting properties of organic and pharmaceutical compounds. Eur J Pharm Sci 36:330–344. https://doi.org/10.1016/j.ejps.2008.10.009
45. Wyttenbach N, Niederquell A, Kuentz M (2020) Machine estimation of drug melting properties and influence on solubility prediction. Mol Pharmaceutics 17:2660–2671. https://doi.org/10.1021/acs.molpharmaceut.0c00355
46. Tam Do H, Zen Chua Y, Kumar A, Pabsch D, Hallermann M, Zaitsau D, Schick C, Held C (2020) Melting properties of amino acids and their solubility in water. RSC Adv 10:44205–44215. https://doi.org/10.1039/D0RA08947H
47. Empowering Innovation & Scientific Discoveries | CAS (n.d.) https://www.cas.org/ (accessed February 7, 2024)
48. Online Chemical Modeling Environment (n.d.) https://ochem.eu/predictor/show.do (accessed July 19, 2024)
49. RDKit (n.d.) https://www.rdkit.org/ (accessed February 7, 2024)
50. Jeon H, Oh S (2020) Hybrid-recursive feature elimination for efficient feature selection. Appl Sci 10:3211. https://doi.org/10.3390/app10093211
51. Singh D, Climente-Gonzalez H, Petrovich M, Kawakami E, Yamada M (2023) FsNet: Feature Selection Network on High-dimensional Biological Data, in: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. https://doi.org/10.1109/IJCNN54540.2023.10191985
52. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 143:106839. https://doi.org/10.1016/j.csda.2019.106839
53. Meng H, Yu R, Tang Z, Wen Z, Yu H, Chu Y (2023) Formation ability descriptors for high-entropy diborides established through high-throughput experiments and machine learning. Acta Mater 256:119132. https://doi.org/10.1016/j.actamat.2023.119132
54. Shrestha N (2020) Detecting multicollinearity in regression analysis. Am J Appl Math Stat 8:39–42
55. Zhang W, Fang M, Dong D, Wang X, Ke X, Zhang L, Hu C, Guo L, Guan X, Zhou J, Shan X, Tian J (2020) Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. Radiother Oncol 145:13–20. https://doi.org/10.1016/j.radonc.2019.11.023
56. Zhao B, Dong X, Guo Y, Jia X, Huang Y (2022) PCA dimensionality reduction method for image classification. Neural Process Lett 54:347–368. https://doi.org/10.1007/s11063-021-10632-5
57. Brauner N, Shacham M (2000) Considering precision of data in reduction of dimensionality and PCA. Comput Chem Eng 24:2603–2611. https://doi.org/10.1016/S0098-1354(00)00616-5
58. van der Maaten L, Postma E, Herik (2007) Dimensionality reduction: a comparative review. J Mach Learn Res JMLR 10
59. Stuart S, Watchorn J, Gu FX (2023) An interpretable machine learning framework for modelling macromolecular interaction mechanisms with nuclear magnetic resonance. Digital Discovery 2:1697–1709. https://doi.org/10.1039/D3DD00009E
60. Gibson EA, Goldsmith J, Kioumourtzoglou M-A (2019) Complex mixtures complex analyses: an emphasis on interpretable results. Curr Envir Health Rpt 6:53–61. https://doi.org/10.1007/s40572-019-00229-5
61. Monti RP, Gibberd A, Roy S, Nunes M, Lorenz R, Leech R, Ogawa T, Kawanabe M, Hyvärinen A (2020) Interpretable brain age prediction using linear latent variable models of functional connectivity. PLoS ONE 15:e0232296. https://doi.org/10.1371/journal.pone.0232296
62. Trinh C, Meimaroglou D, Hoppe S (2021) Machine learning in chemical product engineering: the state of the art and a guide for newcomers. Processes 9:1456. https://doi.org/10.3390/pr9081456
63. Kim S, Yoon H-K (2023) Application of classification coupled with PCA and SMOTE, for obtaining safety factor of landslide based on HRA. Bull Eng Geol Environ 82:381. https://doi.org/10.1007/s10064-023-03403-0
64. scikit-optimize: sequential model-based optimization in Python—scikit-optimize 0.8.1 documentation, (n.d.). https://scikit-optimize.github.io/stable/ (accessed February 7, 2024)
65. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian Optimization of Machine Learning Algorithms, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html (accessed February 8, 2024)
66. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H (2019) Hyperparameter optimization for machine learning models based on bayesian optimizationb. J Electron Sci Technol 17:26–40
67. Ban T, Ohue M, Akiyama Y (2017) Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction, in: 2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp. 1–6. https://doi.org/10.1109/ICCABS.2017.8114299
68. Shekhar S, Bansode A, Salim A (2021) A Comparative study of Hyper-Parameter Optimization Tools, in: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. https://doi.org/10.1109/CSDE53843.2021.9718485
69. Stuke A, Rinke P, Todorović M (2021) Efficient hyperparameter tuning for kernel ridge regression with Bayesian optimization. Mach Learn Sci Technol 2:035022. https://doi.org/10.1088/2632-2153/abee59
70. Shwartz-Ziv R, Armon A (2022) Tabular data: deep learning is not all you need. Inf Fusion 81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011
71. Boldini D, Grisoni F, Kuhn D, Friedrich L, Sieber SA (2023) Practical guidelines for the use of gradient boosting for molecular property prediction. J Cheminform 15:73. https://doi.org/10.1186/s13321-023-00743-7
72. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54:1937–1967. https://doi.org/10.1007/s10462-020-09896-5
73. Xie Y, Zou X, Han J, Zhang Z, Feng Z, Ouyang Q, Hua S, Liu Z, Li C, Cai Y, Zou Y, Tang Y, Jiang X (2022) Indole-3-propionic acid alleviates ischemic brain injury in a mouse middle cerebral artery occlusion model. Exp Neurol 353:114081. https://doi.org/10.1016/j.expneurol.2022.114081
74. Zhao Q, Chen T, Ni C, Hu Y, Nan Y, Lin W, Liu Y, Zheng F, Shi X, Lin Z, Zhu J, Lin Z (2022) Indole-3-propionic acid attenuates HI-related blood-brain barrier injury in neonatal rats by modulating the PXR signaling pathway. ACS Chem Neurosci 13:2897–2912. https://doi.org/10.1021/acschemneuro.2c00418
75. Zheng Z, Wang S, Wu C, Cao Y, Gu Q, Zhu Y, Zhang W, Hu W (2022) Gut Microbiota dysbiosis after traumatic brain injury contributes to persistent microglial activation associated with upregulated Lyz2 and shifted tryptophan metabolic phenotype. Nutrients 14:3467. https://doi.org/10.3390/nu14173467
76. Zhou Y, Chen Y, He H, Peng M, Zeng M, Sun H (2023) The role of the indoles in microbiota-gut-brain axis and potential therapeutic targets: a focus on human neurological and neuropsychiatric diseases. Neuropharmacology 239:109690. https://doi.org/10.1016/j.neuropharm.2023.109690
77. Bhave VM, Ament Z, Patki A, Gao Y, Kijpaisalratana N, Guo B, Chaudhary NS, Guarniz A-LG, Gerszten R, Correa A, Cushman M, Judd S, Irvin MR, Kimberly WT (2023) Plasma metabolites link dietary patterns to stroke risk. Ann Neurol 93:500–510. https://doi.org/10.1002/ana.26552

78. Zhang S, Jin M, Ren J, Sun X, Zhang Z, Luo Y, Sun X (2023) New insight into gut microbiota and their metabolites in ischemic stroke: a promising therapeutic target. Biomed Pharmacother 162:114559. https://doi.org/10.1016/j.biopha.2023.114559
79. Zhang B, Jiang M, Zhao J, Song Y, Du W, Shi J (2022) The mechanism underlying the influence of indole-3-propionic acid: a relevance to metabolic disorders. Front Endocrinol. https://doi.org/10.3389/fendo.2022.841703
80. Jiang H, Chen C, Gao J (2023) Extensive summary of the important roles of indole propionic acid, a gut microbial metabolite in host health and disease. Nutrients 15:151. https://doi.org/10.3390/nu15010151
81. Garcez ML, Tan VX, Heng B, Guillemin GJ (2020) Sodium butyrate and indole-3-propionic acid prevent the increase of cytokines and kynurenine levels in LPS-induced human primary astrocytes. Int J Tryptophan Res. https://doi.org/10.1177/1178646920978404
82. Ran Y, He Y, Yang G, Johnson JLH, Yalkowsky SH (2002) Estimation of aqueous solubility of organic compounds by using the general solubility equation. Chemosphere 48:487–509. https://doi.org/10.1016/S0045-6535(02)00118-2
83. Ran Y, Yalkowsky SH (2001) Prediction of drug solubility by the general solubility equation (GSE). J Chem Inf Comput Sci 41:354–357. https://doi.org/10.1021/ci000338c
84. Ge K, Ji Y (2021) Novel computational approach by combining machine learning with molecular thermodynamics for predicting drug solubility in solvents. Ind Eng Chem Res 60:9259–9268. https://doi.org/10.1021/acs.iecr.1c00998
85. Ma Y, Gao Z, Shi P, Chen M, Wu S, Yang C, Wang J, Cheng J, Gong J (2022) Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. Front Chem Sci Eng 16:523–535. https://doi.org/10.1007/s11705-021-2083-5
86. Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, Charochkina L, Asiri AM (2014) How accurately can we predict the melting points of drug-like compounds? J Chem Inf Model 54:3320–3329. https://doi.org/10.1021/ci5005288
87. Sivaraman G, Jackson NE, Sanchez-Lengeling B, Vázquez-Mayagoitia Á, Aspuru-Guzik A, Vishwanath V, de Pablo JJ (2020) A machine learning workflow for molecular analysis: application to melting points. Mach Learn Sci Technol 1:025015. https://doi.org/10.1088/2632-2153/ab8aa3
88. Galeazzo T, Shiraiwa M (2022) Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings. Environ Sci Atmos 2:362–374. https://doi.org/10.1039/D1EA00090J
89. Venkatraman V, Evjen S, Knuutila HK, Fiksdahl A, Alsberg BK (2018) Predicting ionic liquid melting points using machine learning. J Mol Liq 264:318–326. https://doi.org/10.1016/j.molliq.2018.03.090
90. Zhu X, Polyakov VR, Bajjuri K, Hu H, Maderna A, Tovee CA, Ward SC (2023) Building machine learning small molecule melting points and solubility models using CCDC melting points dataset. J Chem Inf Model 63:2948–2959. https://doi.org/10.1021/acs.jcim.3c00308
91. Uddin MF, Lee J, Rizvi S, Hamada S (2018) Proposing enhanced feature engineering and a selection model for machine learning processes. Appl Sci 8:646. https://doi.org/10.3390/app8040646
92. Li Z, Ma X, Xin H (2017) Feature engineering of machine-learning chemisorption models for catalyst design. Catal Today 280:232–238. https://doi.org/10.1016/j.cattod.2016.04.013
93. Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T (2019) Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform 20:1878–1912. https://doi.org/10.1093/bib/bby061
94. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25:1315–1360. https://doi.org/10.1007/s11030-021-10217-3
95. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, Coley CW, Xiao C, Sun J, Zitnik M (2021) Therapeutics data commons: machine learning datasets and tasks for drug discovery and development. https://doi.org/10.48550/arXiv.2102.09548
96. Capecchi A, Probst D, Reymond J-L (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminf 12:43. https://doi.org/10.1186/s13321-020-00445-4
97. Hutter MC (2022) Differential multimolecule fingerprint for similarity search—making use of active and inactive compound sets in virtual screening. J Chem Inf Model 62:2726–2736. https://doi.org/10.1021/acs.jcim.2c00242
98. Xie L, Xu L, Kong R, Chang S, Xu X (2020) Improvement of prediction performance with conjoint molecular fingerprint in deep learning. Front Pharmacol. https://doi.org/10.3389/fphar.2020.606668
99. Breslin W, Pham D (2023) Machine learning and drug discovery for neglected tropical diseases. BMC Bioinformatics 24:165. https://doi.org/10.1186/s12859-022-05076-0
100. Nguyen P, Loveland D, Kim JT, Karande P, Hiszpanski AM, Han TY-J (2021) Predicting energetics materials' crystalline density from chemical structure by machine learning. J Chem Inf Model 61:2147–2158. https://doi.org/10.1021/acs.jcim.0c01318
101. Katubi KM, Saqib M, Mubashir T, Tahir MH, Halawa MI, Akbar A, Basha B, Sulaman M, Alrowaili ZA, Al-Buriahi MS (2023) Predicting the multiple parameters of organic acceptors through machine learning using RDkit descriptors: an easy and fast pipeline. Int J Quantum Chem 123:e27230. https://doi.org/10.1002/qua.27230
102. Packwood D, Nguyen LTH, Cesana P, Zhang G, Staykov A, Fukumoto Y, Nguyen DH (2022) Machine learning in materials chemistry: an invitation. Mach Learn Appl 8:100265. https://doi.org/10.1016/j.mlwa.2022.100265
103. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, Metni H, van Hoesel C, Schopmans H, Sommer T, Friederich P (2022) Graph neural networks for materials science and chemistry. Commun Mater 3:1–18. https://doi.org/10.1038/s43246-022-00315-6
104. Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, Seress LR, Román-Salgado C, Trepte K, Atahan-Evrenk S, Er S, Shrestha S, Mondal R, Sokolov A, Bao Z, Aspuru-Guzik A (2014) Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry-the Harvard Clean Energy Project. Energy Environ Sci 7:698–704. https://doi.org/10.1039/C3EE42756K
105. Pyzer-Knapp EO, Simm GN, Guzik AA (2016) A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. Mater Horiz 3:226–233. https://doi.org/10.1039/C5MH00282F
106. Stuart S, Watchorn J, Gu FX (2023) Sizing up feature descriptors for macromolecular machine learning with polymeric biomaterials. Npj Comput Mater 9:1–10. https://doi.org/10.1038/s41524-023-01040-5
107. Chuang KV, Gunsalus LM, Keiser MJ (2020) Learning molecular representations for medicinal chemistry. J Med Chem 63:8705–8722. https://doi.org/10.1021/acs.jmedchem.0c00385
108. Brown RD, Martin YC (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Comput Sci 37:1–9. https://doi.org/10.1021/ci960373c
109. Sato A, Miyao T, Jasial S, Funatsu K (2021) Comparing predictive ability of QSAR/QSPR models using 2D and 3D molecular representations. J Comput Aided Mol Des 35:179–193. https://doi.org/10.1007/s10822-020-00361-7
110. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M (2006) Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. J Med Chem 49:6802–6810. https://doi.org/10.1021/jm060902w
111. Zhang Y, Zhang X, Razbek J, Li D, Xia W, Bao L, Mao H, Daken M, Cao M (2022) Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome. BMC Endocr Disord 22:214. https://doi.org/10.1186/s12902-022-01121-4
112. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x
113. Manallack DT, Tehan BG, Gancia E, Hudson BD, Ford MG, Livingstone DJ, Whitley DC, Pitt WR (2003) A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. J Chem Inf Comput Sci 43:674–679. https://doi.org/10.1021/ci0202741
114. Gozalbes R, Pineda-Lucena A (2010) QSAR-based solubility model for drug-like compounds. Bioorg Med Chem 18:7078–7084. https://doi.org/10.1016/j.bmc.2010.08.003
115. Gao H, Shanmugasundaram V, Lee P (2002) Estimation of aqueous solubility of organic compounds with QSPR approach. Pharm Res 19:497–503. https://doi.org/10.1023/A:1015103914543

116. Xue N, Zhang Y, Liu S (2024) Evaluation of Machine Learning Models for Aqueous Solubility Prediction in Drug Discovery, https://doi.org/10.1101/2024.06.10.598383
117. Christine-Allen-Lab/Solubility_ML, GitHub (n.d.). https://github.com/Christine-Allen-Lab/Solubility_ML (accessed March 26, 2024)
118. PubChem, PubChem, (n.d.). https://pubchem.ncbi.nlm.nih.gov/ (accessed March 21, 2024)
119. Main Page, Wikipedia, the Free Encyclopedia (2024). https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=1212457119 (accessed March 21, 2024)
120. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074–D1082. https://doi.org/10.1093/nar/gkx1037
121. ChemSpider | Search and share chemistry, (n.d.). https://www.chemspider.com/ (accessed March 21, 2024)
122. Chemical Database Online, (n.d.). https://www.chembk.com/en (accessed July 19, 2024).
123. Pesticide Properties Database, (n.d.). https://sitem.herts.ac.uk/aeru/ppdb/en/ (accessed July 19, 2024)
124. CAS Number Search—Chemsrc, (n.d.). https://www.chemsrc.com/en/ (accessed July 19, 2024)
125. LKT Labs—Biochemicals for Life Science Research, (n.d.). https://lktlabs.com/ (accessed March 21, 2024)
126. ChemicalBook, (n.d.). https://www.chemicalbook.com/ProductIndex_EN.aspx (accessed March 21, 2024)
127. MilliporeSigma | Life Science Products & Service Solutions, (n.d.). https://www.sigmaaldrich.com/CA/en (accessed March 21, 2024)
128. Antibodies, Gene Editors, Chemicals & Lab Supplies For Research | Santa Cruz Biotechnology, (n.d.). https://www.scbt.com/home (accessed March 21, 2024)
129. Lab Equipment and Lab Supplies | Fisher Scientific, (n.d.). https://www.fishersci.com/us/en/home.html (accessed March 21, 2024)
130. Home—AK Scientific (n.d.) https://aksci.com/ (accessed July 19, 2024)
131. Aziridine, Benzyl Isothiocyanate & Benzoyl Isothiocyanate Manufacturers, MOLTUS RESEARCH LABORATORIES PRIVATE LIMITED (n.d.) https://www.moltuslab.com/ (accessed July 19, 2024)
132. TCI AMERICA | Homepage (n.d.) https://www.tcichemicals.com/CA/en/ (accessed July 19, 2024)
133. Guidechem chemical B2B network provides information on china and global chemical market quotation and relative chemical Information. Guidechem Chemical Network providing the most complete information of the chemical industry., GuideChem (n.d.). https://www.guidechem.com (accessed July 19, 2024)
134. ECHEMI: Online Chemical Company to Buy Chemical Products, ECHEMI (n.d.) https://www.echemi.com (accessed July 19, 2024)
135. EBCLink, Drug Delivery (2024). http://www.ebclink.com/ (accessed July 19, 2024).
136. Dielectric Constant (n.d.) https://macro.lsu.edu/HowTo/solvents/Dielectric%20Constant%20.htm (accessed July 19, 2024)
137. Solvent Physical Properties (n.d.) https://people.chem.umass.edu/xray/solvent.html (accessed July 19, 2024)
138. Dielectric constant (n.d.) https://depts.washington.edu/eooptic/linkfiles/ (accessed July 19, 2024)
139. rdkit.Chem.Descriptors3D (n.d.) https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors3D.html# (accessed July 19, 2024)
140. Jacot-Descombes L, Turcani L, Jorner K, morfeus (2024) https://github.com/digital-chemistry-laboratory/morfeus (accessed July 19, 2024)
141. scikit-learn: machine learning in Python—scikit-learn 1.4.0 documentation (n.d.) https://scikit-learn.org/stable/ (accessed February 7, 2024)
142. Welcome to LightGBM's documentation!—LightGBM 4.3.0.99 documentation (n.d.) https://lightgbm.readthedocs.io/en/latest/ (accessed February 7, 2024)
143. XGBoost Python Package—xgboost 2.1.0-dev documentation (n.d.) https://xgboost.readthedocs.io/en/latest/python/index.html (accessed February 7, 2024)
144. lightgbm.plot_importance—LightGBM 4.3.0.99 documentation (n.d.) https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.plot_importance.html (accessed March 12, 2024)
145. StandardScaler, Scikit-Learn (n.d.) https://www.scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed July 19, 2024)
146. KMeans, Scikit-Learn (n.d.) https://www.scikit-learn/stable/modules/generated/sklearn.cluster.KMeans.html (accessed July 19, 2024)
147. mahalanobis—SciPy v1.14.0 Manual (n.d.) https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.mahalanobis.html (accessed July 19, 2024)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.