Boosting Predictability: Towards Rapid Estimation of Organic Molecule Solubility

Arsalan Hashemi,[1,*] Pekka Peljo,[2,3] Tapio Ala-Nissila,[1,4] and Kari Laasonen[3,†]

[1]*Quantum Technology Finland Center of Excellence, Department of Applied Physics, Aalto University, P.O. Box 15600, FI-00076 Aalto, Finland*

[2]*Research Group of Battery Materials and Technologies, Department of Mechanical and Materials Engineering, Faculty of Technology, University of Turku, 20014 Turun Yliopisto, Finland*

[3]*Department of Chemistry and Material Science, School of Chemical Engineering, Aalto University, FI-00076 Aalto, Finland*

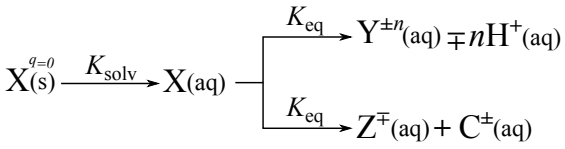
[4]*Interdisciplinary Centre for Mathematical Modelling and Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, United Kingdom*

The water solubility of organic molecules is critical for optimizing the performance and stability of aqueous flow batteries, as well as for various other applications. Although relatively straightforward to measure in some cases, the theoretical prediction of the solubility remains a considerable challenge. To this end, machine learning algorithms have become increasingly important tools in the past decade. High-quality data and effective descriptors are essential for constructing reliable data-driven estimation models. We systematically investigate the effectiveness of enhanced structure-based descriptors and an outlier detection procedure for improving aqueous solubility predictability. We train and evaluate random forest regression models using various descriptors to predict experimental solubility. Outliers are identified through an iterative maximum-error deletion procedure. We discover that descriptors derived from hydration free energy and weighted fingerprints, along with other established features, are effective. Notably, solvation energy, octanol-water partition coefficient, atomic charge polarizability interactions, and the presence of a full-carbon aromatic ring are critical for solubility prediction. Furthermore, the effectiveness of the outlier detection protocol is validated by improving the performance of the model and detailed analysis of the dataset. This study significantly improves the predictive capacity of supervised machine learning for molecular properties, enabling advancements in various technological applications.

I. Introduction

In many disciplines, such as synthetic, medicinal, and environmental chemistry, knowing the aqueous solubility of organic molecules is critical [1–5]. This importance is exemplified by the following applications: (i) Flow batteries (FBs) require highly soluble compounds [6–9], as higher solubility allows the solution to store more electrons, thereby enhancing the energy capacity [10]. (ii) In drug discovery, soluble compounds are essential to minimize side effects [11]. When working with poorly soluble drugs, advanced delivery techniques are necessary to improve therapeutic efficacy, which can complicate the development process [12, 13].

Aqueous solubility refers to the maximum amount of a compound that dissolves in water through *dissolution process* [14]:



Scheme1: Dissolution process mechanisms.

In this context, X(s) and X(aq) represent compounds in the solid and aqueous phases, respectively, which have an equilibrium constant of K_{solv} (solvation process). Furthermore, solvated species can increase their solubility by undergoing either an acid-base reaction or salt dissociation/formation, ultimately reaching equilibrium defined by K_{eq}. X(s) is charge-neutral ($q = 0$). If the molecule of interest is charged (e.g., $Z^{\mp}$), a counterion ($C^{\pm}$) is needed to stabilize the crystal. $Y^{\pm n}$ denotes the (de)protonated form of X.

Measuring solubility requires time-consuming experiments, often taking tens of hours to perform accurately, and may present significant challenges [15–18]. Consequently, estimating solubility at the early discovery stage without using physical samples is highly beneficial. This estimation allows for the prediction of whether a proposed molecule falls outside the desired solubility range before it undergoes experimental testing or production. Should this be the case, the compound might be rejected or alternatively, it could be structurally modified to enhance its solubility. A critical question then arises: what is the most effective method for estimating the aqueous solubility of organic molecules?

Estimating solubility is a challenging task [19, 20] because it requires theoretical models that capture interactions within the solid (molecule-molecule) and solution (molecule-solvent) phases, as well as the transitions between them. This difficulty becomes more evident when accounting for factors such as solvent reorganization, electrostatic interactions, and environmental variables such as temperature, pressure, supporting electrolytes, pH, and polymorphism in a *single* model to assess the dissolution process. The quantum chemistry approach is accurate but highly time-consuming, especially when dynamics are included. Therefore, this method is not preferred for high-throughput screening. Empirical methods and molecular simulations are used to study longer time scales, but their accuracy and reliability remain inadequate for multi-phase studies [21–24].

* arsalan.hashemi@aalto.fi
† kari.laasonen@aalto.fi

To mitigate these limitations, data-driven approaches [25, 26] can be used to leverage statistical models and link molecular descriptors with relevant data. These descriptors, derived from molecular modeling and structural properties, are paired with solubility data collected from experiments. As a result, the target values inherently covers all realistic effects, enabling the application of various data-driven models such as machine learning (ML) methods to efficiently predict solubility [27–35].

The search for ML solubility models has received considerable attention in recent years, particularly with regard to drugs and drug-like compounds [36–45]. Notably, Zhang et al. [38] utilized data-driven predictions and 3D molecular representations to bridge the gap between computational models and experimental validation for quinone-like molecules. Chaka et al. [39] enhanced the predictability of solubility by employing a graph neural network (GNN) named MolGAT [46]. Because various ML algorithms were used across both descriptor- and graph-based models for predicting solubility, systematic research into the descriptors is called for to improve the accuracy of predictions, especially in medium-sized databases.

Data quality is critical to effective deployment of ML models [18, 47]. However, experimental measurements are prone to errors, which raises concerns regarding data reliability. To address this concern, outliers (molecules whose solubility cannot be reliably predicted) can typically be identified using descriptor dissimilarity [48] or clustering [49] methods. In recent work [50] it was proposed that outliers of solubility datasets could be detected based on the *prediction error distribution* [51]. The data were divided into k training and testing datasets. As a result, the samples in each testing set are predicted k times by the models trained on distinct data. Samples that frequently fall beyond a predetermined error range are considered outliers. Nevertheless, this approach may be inefficient for smaller datasets, high-dimensional feature spaces, or complex ML models.

In this study, we propose a workflow to improve aqueous solubility prediction. We select a database that should be suitable for organic FB applications. Our methodology comprises four steps: (i) cleaning the database, (ii) developing structure-based descriptors, (iii) implementing an outlier detection process to refine the data, and (iv) conducting a thorough discussion of the results to suggest potential directions for future research.

II. Data Analysis

The Solubility of Organic Molecules in Aqueous Solutions (SOMAS) database, which was previously evaluated [52], consists of critically validated records. Each record includes the molecule's name, Chemical Abstracts Service (CAS) number, a reference, and its experimental aqueous solubility, the latter expressed as a floating-point number. Additionally, the database incorporates eight quantum descriptors, calculated using density functional theory [53, 54] (DFT), and employs SMILES (Simplified Molecular Input Line Entry System) chemical notation to detail the molecular structures. It also records the temperature (in K) at which each measurement was made. In total, the database contains 11696 records, with solubility reported in milligrams per liter (mg/L) and designated as the target variable.

A. Data Cleaning and Preprocessing

To prepare the data for a thorough analysis, we first cleaned it systematically. The criteria we used align with FB applications:

(i) *Aromaticity* [55, 56] is a critical factor that enables the dispersion of π-electrons across the conjugated ring. This dispersion lowers the ground-state energy and enhances molecular stability [57–59]. Cyclic molecular structures also offer a platform for facile molecular engineering, applicable either to the core structure or to the functional group decoration. Consequently, a total of 6845 records were excluded from the database due to their non-aromatic character.

(ii) Water solubility is significantly affected by the *net charge* of organic molecules. A prototypical example is methyl viologen dication [60], which exhibits a solubility of 1.5 M in aqueous solutions. As electrons are transferred through reduction reactions, the solubility decreases, until the neutral molecule becomes practically insoluble. It can therefore be argued that the solubility of neutral molecules involved in the redox processes may represent a critical bottleneck. Only three records were excluded after filtering out charged molecules.

(iii) The organic molecules used in FBs are composed of *common elements*, primarily carbon (C), nitrogen (N), oxygen (O), sulfur (S), and hydrogen (H). To a lesser extent, they also contain phosphorus (P), fluorine (F), chlorine (Cl), bromine (Br), and iodine (I), indicating an abundant availability of raw materials. These elements can combine to form various functional groups, including hydroxyl ($-$OH), carbonyl (C=O), carboxyl ($-$COOH), amino ($-NH_2$), phosphate ($-PO_4$), and halides ($-$F, $-$Cl, $-$Br, $-$I), each possessing unique properties. Consequently, additional 20 records were excluded because the molecules in these records contained boron (B), arsenic (As), silicon (Si), selenium (Se), or deuterium (D) atoms.

(iv) An aqueous *solubility of* 10^6 mg/L (equivalent to 1 kg/L) for a substance is exceptionally high. This observation could indicate either an error or a special case occurring under specific conditions that are not typical for most substances at room temperature and atmospheric pressure. Most neutral organic compounds exhibit low solubility in water. We excluded 41 records reporting solubilities of $\geq 10^6$ mg/L from our analysis.

(v) The *molecular weight* of the redox species should also be considered, as smaller molecules are more likely to achieve higher gravimetric energy density and solubility [61] and lower cost [62]. Accordingly, we established a cutoff of 500 Da for molecular weight, in accordance with Lipinski's Rule [63]. Using this criterion, we removed a total of 122 records.

(vi) We also identified 36 complex compounds featuring multiple stereocenters by analyzing their molecular structures.

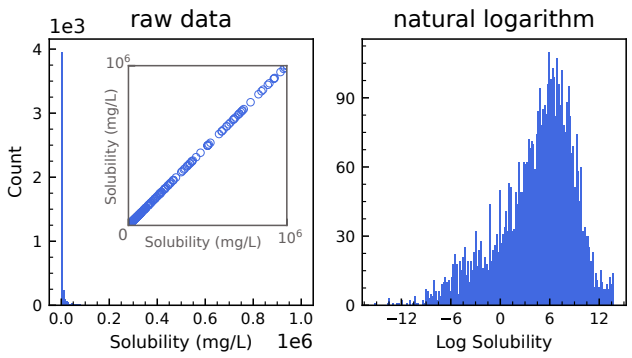


Figure 1. Distribution of (left panel) raw, (right panel) logarithmically scaled solubility. To demonstrate the presence of higher solubility values, an inset scatter plot displays the raw solubility data.

These molecules possess a tetrahedral configuration that includes "[C@@]" and/or "[C@]" as chiral centers in their SMILES representations. Typically, such complex molecules exhibit low symmetry, and their physical properties, including solubility, are influenced by chirality. To preserve the integrity of our processed dataset, we decided to exclude these *stereoisomeric* records.

The final dataset includes cyclic molecules composed of rings containing 3, 4, 5, 6, or 7 carbon atoms, which may be fused with nitrogen (N), sulfur (S), or oxygen (O) atoms. To ensure accuracy, the database underwent a comprehensive check for duplicates using both SMILES representations and compound names. This process confirmed the absence of duplicates or missing values. Our data cleansing procedure yielded a total of 4635 samples.

B. Target Data Transformation

The distribution of the solubility data is illustrated in Figure 1. The raw data follow an exponential distribution. Approximately 95% of the collected records were below 0.05×10^6 mg/L. It is necessary to transform the target data before training the model to ensure efficient predictions.

By applying the natural logarithm, we linearized our data to approximate a quasi-normal distribution. Logarithmic scaling yields real numerical values since all the experimental solubility values are nonzero. The scaled values range from -15.29 to 13.82, with a mean of 4.39, a standard deviation of 4.61, and a median of 5.26. The median being slightly greater than the mean suggests a left-skewed distribution.

III. Descriptors

To introduce a molecule into the ML model, we initially utilized property-based descriptors in their original form as stored in the SOMAS database. These descriptors include various chemical properties: molecular mass, solvation energy, dipole and quadrupole moments, molecular volume, surface area, highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, and the HOMO–LUMO energy gap. The temperature was also recorded. Electronic structures were estimated using DFT calculations, implemented in NWChem software [64], and incorporated the implicit COSMO solvation model [65].

However, these descriptors necessitate additional computational resources and an in-depth theoretical understanding of electronic structure calculations. Consequently, there is growing interest in employing SMILES as input for ML applications [66–68]. SMILES is a single-line molecular notation system that does not require any further (pre-processing) DFT calculations. Since ML models cannot process categorical variables in string format, these must be converted into numerical values. To achieve this, we explored three distinct methods, which are outlined below.

(*1*) A list of numeric values representing 208 physicochemical properties, such as the octanol-water partition coefficient (log P), van der Waals surface area (LabuteASA), and number of rings, was estimated using the RDKit open-source cheminformatics software [69]. Initially, the SMILES strings were converted into RDKit *mol* objects. If an error occurred during this conversion, the input SMILES string was deemed invalid. All records successfully passed this stage. Following the generation of the descriptors, we implemented an additional filtering protocol to eliminate columns with zero variance or missing values from the descriptor space. Consequently, 204 features were retained for each sample. The names of these attributes are listed in Sec. I of the Supporting Information (SI). Additionally, we concatenated the temperature data within the feature space. This feature space comprises 205 dimensions.

(*2*) The Gibbs free energy of hydration, expressed in kJ/mol and denoted by `dgtot`, along with its polar and apolar components (`dgp`/`dga`), and the hydrogen-bond strength of donor and acceptor atoms, can be estimated using a Python library named Jazzy [70]. In this context, `sdc`, `sdx`, and `sa` represent the strengths of the C−H donor, X−H donor (where X denotes non-carbon atoms), and acceptor, respectively. Jazzy also provides atomic features such as partial charge (q) and charge-dependent dynamic polarizability (α). These features are post-processed to describe interactions within a molecule as follows:

$$V_q = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} q_i q_j e^{-\beta |\mathbf{r}_i - \mathbf{r}_j|}; \tag{1}$$

$$V_\alpha = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \alpha_i \alpha_j e^{-\beta |\mathbf{r}_i - \mathbf{r}_j|}. \tag{2}$$

These formulas mimic the Yukawa potential [71] and yield real numerical values. Here, $\mathbf{r}_i$, q_i, and α_i denote the coordinates, partial charge, and charge-dependent dynamic polarizability of atom i, respectively. N is the total number of atoms, and $\beta = 0.01$ Å^{-1} is used as an arbitrary tuning parameter. The 3D coordinates, including hydrogens, were generated from SMILES strings and optimized using the universal force

field [72] (UFF) implemented in the RDkit package. The temperatures were added in this set of descriptors, too.

(*3*) Extended-connectivity fingerprints (ECFPs), as implemented in the RDkit package, were utilized [73]. These fingerprints are based on the distinctiveness of the environments surrounding individual atoms, pairs of atoms, and trios of atoms. The types of atoms, bonds, and their connectivity were assessed and encoded as a binary vector with a length of 1024, where a '1' indicates the presence of a specific subfragment, and a '0' denotes its absence. We further adapted the ECFPs by assigning a weight to each bit corresponding to the number of occurrences, n, of that bit. As a result, the vectors are encoded using '0's and n_i values, where n_i represents the number of times subfragment i appears in a molecule (sample). We refer to this modified descriptor as w-ECFP. Additionally, temperature data were normalized to the maximum observed value (469.15 K) and incorporated into the fingerprint as an additional column. The feature space comprises 1025 dimensions.

All scripts and files related to this study are available in A.H.'s **GitHub repository**. The repository contains all the codes and data used in the present work.

IV. Machine Learning Models

In our study, we utilized the Random Forest Regressor (RFR) in the Scikit-learn package [74] for high-throughput screening. To determine the optimal hyperparameters, we employed cross-validation, calculating scores across a predefined grid of hyperparameter spaces. Specifically, the training dataset was divided into ten subsets; nine were used for training the model, while the remaining subset served to evaluate its performance. We chose the mean squared error (MSE) as the metric for optimizing hyperparameters. To prepare our models, we first randomized the dataset, allocating 85% of the data for training and the remaining 15% for validation. The performance of each trained model was evaluated using three metrics: the mean absolute error (MAE), the root mean squared error (RMSE), and the coefficient of determination (R^2) calculated from the predictions on the test set as described in SI (Sect. II). To report the performance metrics, we calculated each parameter 20 times and then computed the average value along with the standard deviation.

The models were trained using different feature spaces (see Table I). For simplicity, we used $D_1 - D_5$ to denote the descriptors. Among these, D_2, comprising structure-derived chemical attributes, exhibited the best performance. D_1, D_3, and D_4 performed comparably well. Additionally, when comparing ECFP with w-ECFP, we observed that the modification to the fingerprint significantly enhanced model performance.

The larger the difference between RMSE and MAE, the greater the likelihood of predicting records with substantial errors, as the RMSE overemphasizes outliers. To enhance the predictability of the model, it is crucial to identify and eliminate (i) noise within the feature space (irrelevant features) and (ii) noise in the records (outliers in particular). We explored the former using Gini importance [75] and addressed the latter by removing outliers from the datasets through an iterative protocol, which we will discuss in detail later.

The six most important features for each descriptor are illustrated in Figure 2(a)-(d). A higher feature score indicates a greater influence on the prediction of the target variable. Within the DFT feature space, the volume of the DFT-optimized structure and the solvation energy exert the most significant effects on model predictions. Compared to the other 204 chemical attributes calculated by the RDKit package, log P shows a high correlation with the target variable. In D_3, the post-evaluated V_α exhibits the most substantial impact on the descriptors.

In the w-ECFPs feature space, the most significant molecular sub-fragments identified are: (i) an aromatic carbon (C) atom with three bonds within a ring, (ii) an aromatic C atom with a single hydrogen (H) atom, forming part of a ring structure with two connections, (iii) a structural pattern where a chlorine (Cl) atom, acting as a functional group, is single-bonded to a C atom that is part of a ring and connected to three other atoms, (iv) a molecular fragment containing a Cl atom that is not incorporated into a ring structure, (v) a configuration of C atoms in a ring where two C atoms, each bonded to one H atom, have two connections and are linked through a C atom that is part of an aromatic ring with three connections, and (vi) an aromatic C atom in a ring connected to three other atoms, which is single-bonded to a Cl atom (not part of a ring and connected only to one atom), and is also double-bonded to another aromatic C (which is part of the same ring and has three connections); this second aromatic C is further double-bonded to another aromatic C, which is also part of the ring, bears one H atom, and is connected to two other atoms. Figure 2(d) illustrates the structures of these sub-fragments in a representative molecule.

There was no indication that temperature was an important feature in any feature-importance analyses. This is likely due to the narrow range of temperatures, centered around room temperature, in which all measurements were taken. The dataset includes temperature readings with a mean of 296.6 K and standard deviation of 6.1 K, indicating moderate variability. More statistical information can be found in Figure S1.

To investigate the correlation between the target variable and the most significant features within each D_1-D_4 space, we illustrated these relationships in Figures S2-S5. We observed an inverse linear correlation with log P and, to a lesser degree, with molecular volume.

To investigate potential noise in the feature space, we gathered key features to reduce dimensionality, resulting in a new dataset comprising six-key-features (6KFs) descriptors. Consequently, the 6KFs space encompasses 24 features. Performance metrics indicate that the predictability of the final model remained almost unchanged and performed similarly to other models. This finding encouraged us to establish a systematic protocol for detecting noise in the dataset, specifically targeting outliers, i.e., measurements that are incorrect or scarcely represented in the dataset.

MAE (log)	RMSE (log)	R^2	N^{est}	$f^{\max}$	NOF	Descriptor	Package
1.77 ± 0.06	2.44 ± 0.09	0.72 ± 0.02	1880	0.42	12	D_1: chemical attributes	NWChem (DFT)
1.50 ± 0.05	2.17 ± 0.10	0.78 ± 0.02	600	0.23	205	D_2: chemical attributes	RDkit (SMILES)
1.78 ± 0.06	2.53 ± 0.12	0.72 ± 0.03	2250	0.45	9	D_3: chemical attributes	Jazzy (SMILES)
1.73 ± 0.07	2.41 ± 0.10	0.73 ± 0.02	950	0.25	1025	D_4: $w-$ECFP	RDkit (SMILES)
2.10 ± 0.07	2.83 ± 0.11	0.62 ± 0.02	425	0.40	1025	D_5: ECFP	RDkit (SMILES)
1.55 ± 0.06	2.24 ± 0.10	0.77 ± 0.03	1500	0.31	24	Six Key Features (6KFs)	

Table I. Detailed information on the RFR models and descriptors: test set RMSE, MAE, and R^2 were calculated for the model performance assessment. N^{est} and $f^{\max}$ denote the number of decision trees that will be running in a model and the maximum number of features a model considers when determining a split, respectively. $f^{\max}$ defines the ratio of features to consider when looking for the best split while training each tree in the forest. NOF is the number of features in each descriptor space.

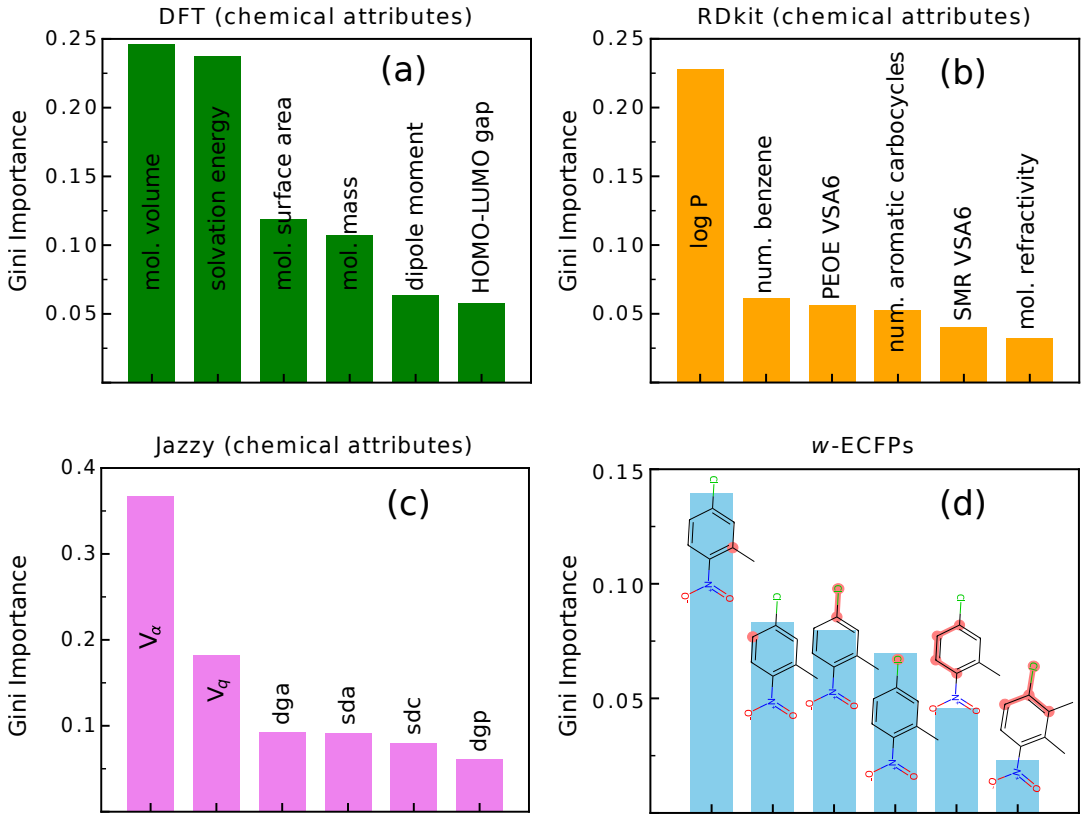


Figure 2. Feature Importance Plot: The top six important features are sorted based on the importance for different descriptors. Each bar is labeled with the feature name, except for (d) the highlighted sub-fragments are displayed for an optional molecule.

V. Outlier Detection

To enhance the integrity of our dataset, we utilize an outlier detection algorithm that operates by selectively removing data points exhibiting maximal prediction errors. This iterative process continues until the overall prediction error converges within an acceptable range. The pseudocode for this algorithm is illustrated in Figure 3.

To elucidate the computational steps, we begin with an explanation of the nested loop, which operates as follows:
Train model: In accordance with standard procedures, the ML model is trained on 85% of the data (a randomly generated training set), while the main hyperparameters are maintained as specified in Table I.
Identify maximum error: The trained model predicts log solubility for the test set and calculates the absolute prediction error for each molecule to identify the maximum absolute error.
Update dataset: If a molecule's error exceeds an acceptable threshold, it is removed from the dataset. The data indices are then reset to prepare for another iteration, if necessary.
Iterate: This process is repeated, involving the recalculation of the maximum error and reassessment for outliers. The iter-

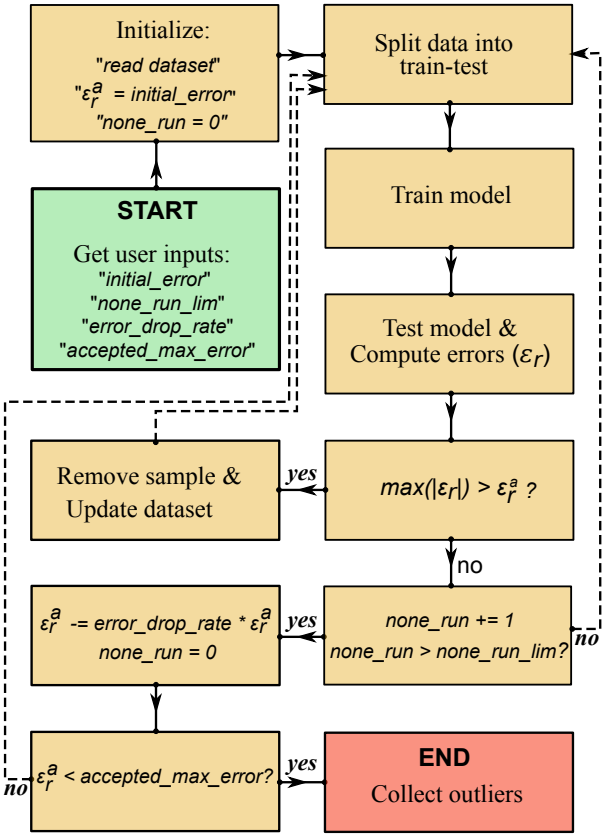


Figure 3. Pseudocode diagram for outlier detection.

ative cycle continues until no further outliers are detected or until a predetermined number of iterations or deletions have been achieved.

To assess the final set of outliers, we implemented the procedure described above using various descriptors (D_1, D_2, D_3, and D_4). The overlap of the outputs from these descriptors yielded the final set of outliers, defined mathematically as:

$$\mathcal{S}_{\text{outlier}} = \mathcal{S}_{\text{out}}^{\text{DFT}} \cap \mathcal{S}_{\text{out}}^{\text{RDkit}} \cap \mathcal{S}_{\text{out}}^{\text{Jazzy}} \cap \mathcal{S}_{\text{out}}^{w-\text{ECFP}}. \quad (3)$$

In this expression, $\mathcal{S}_{\text{outlier}}$ denotes the intersection of all sets, containing elements common across all specified sets. The individual sets $\mathcal{S}_{\text{out}}^{\text{DFT}}$, $\mathcal{S}_{\text{out}}^{\text{RDkit}}$, $\mathcal{S}_{\text{out}}^{\text{Jazzy}}$, and $\mathcal{S}_{\text{out}}^{w-\text{ECFP}}$ represent the outliers identified in the feature spaces of D_1, D_2, D_3, and D_4, respectively. Given that each descriptor is tailored to a specific application domain, samples that consistently fail to be modeled across these domains are classified as outliers.

To identify outliers, we utilized the following parameters: "*initial_error*" (15), "*none_run_lim*" (20), "*error_drop_rate*" (0.1), and "*accepted_max_error*" (3). The detection process required approximately 5000 cycles to complete. As a result, 750 records were excluded from the dataset. The outliers demonstrated a wide range of solubility values, from low to high, as depicted in Figure 4(a).

Re-optimization of the models using cleaned data significantly improved performance, with predictability increasing by up to 10%. Changes in model complexity varied depending on the specific case. The descriptor D, composed of descriptors D_1 to D_4, is the most effective, as shown in Table II. A scatter plot comparing experimental log solubility with predicted log solubility (Figure 4 (b)) underscores the model's good predictive accuracy for the test set data.

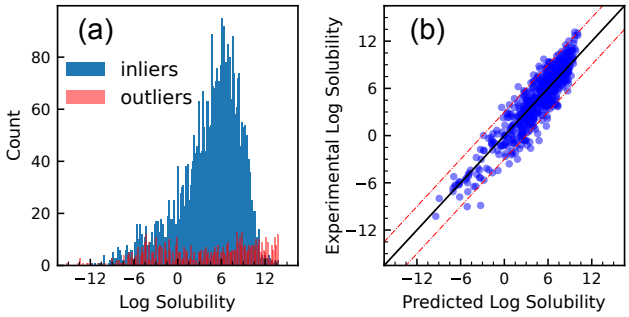


Figure 4. (a) Distribution of inlier (blue) and outlier (red) log solubility data. (b) Scatter plot of the experimental log solubility versus predicted log solubility. The red dashed line shows a deviation of $\varepsilon_r = \pm 3$ from the ideal prediction line.

MAE (log)	RMSE (log)	R^2	N^{est}	$f^{\max}$	Descriptor
1.36 ± 0.03	1.76 ± 0.04	0.81 ± 0.01	850	0.52	D_1
1.24 ± 0.03	1.62 ± 0.04	0.85 ± 0.02	950	0.49	D_2
1.25 ± 0.04	1.63 ± 0.05	0.84 ± 0.01	1500	0.46	D_3
1.29 ± 0.03	1.64 ± 0.04	0.83 ± 0.02	950	0.49	D_4
1.07 ± 0.03	1.38 ± 0.05	0.90 ± 0.01	800	0.32	D

Table II. Performance of models applied on the clean dataset: test set RMSE, MAE, and R^2 were calculated for model performance assessment. N^{est} and $f^{\max}$ denote the number of decision trees that will be running in a model and the maximum number of features a model considers when determining a split, respectively. The descriptors labeled as D_1 to D_4 follow the same conventions as presented in Table I. Descriptor D is composed of descriptors D_1 to D_4.

VI. Challenges and Cases of Applications

In our analysis of the refined data, we found that solubility is enhanced under two conditions: (i) disruption of the central symmetry in core structures, exemplified by the substitution of an N/O atom with a carbon atom within the benzene ring, which increases polarity; and (ii) the incorporation of functional groups such as –OH, –COOH, $–NH_2$, and $–SO_3H$, which enhances hydrogen bonding and solute-solvent interactions. These concepts have been extensively discussed in chemistry textbooks.

Returning to Scheme 1, our models were designed to predict the solubility of neutral molecules (i.e., related to the solvation process), without considering the equilibration process. This is referred to as *intrinsic solubility* and is generally expected to have low molarity in the aqueous solutions.

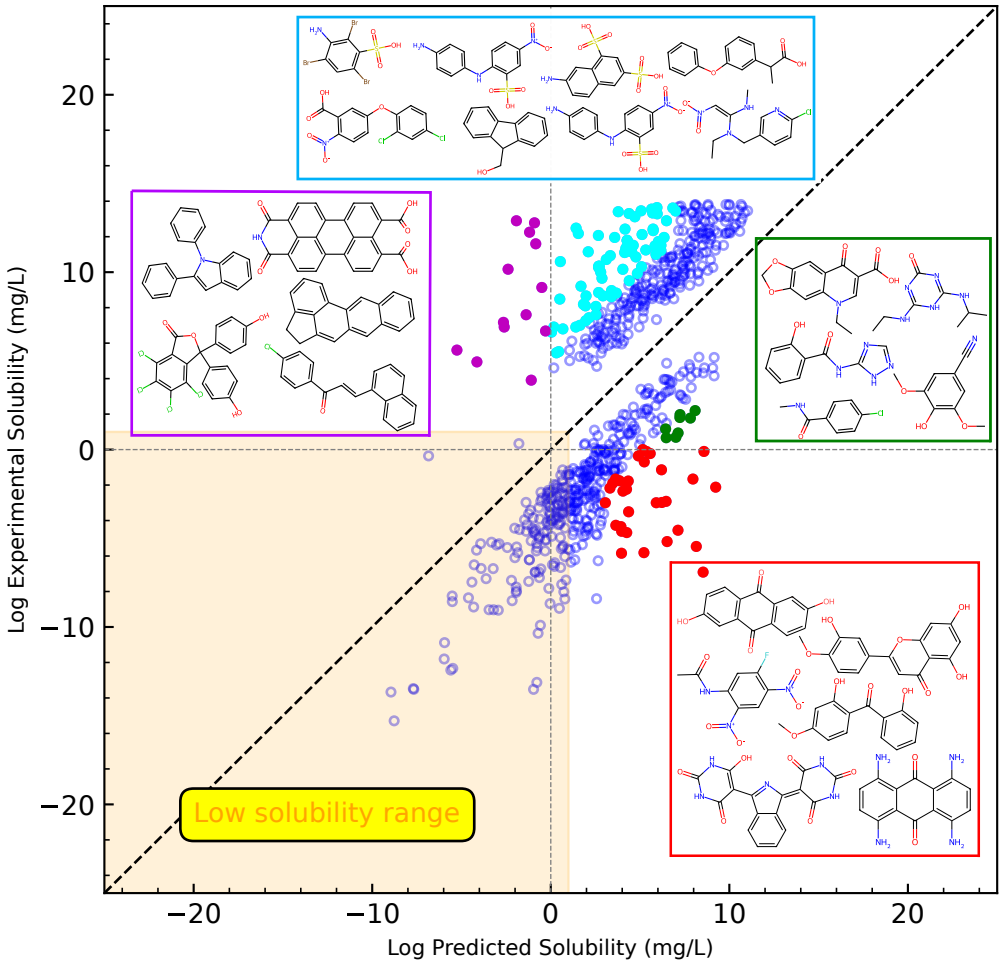


Figure 5. Scatter plot of experimental versus predicted log solubility for the outliers, identified by the algorithms shown in Figure 3, is presented. The predictions were obtained using the model trained on descriptor D. Molecular structures corresponding to the most highly deviating data points, colorized in magenta, cyan, green, and red, are illustrated.

As shown in Figure 5, the outlier detection process identified two distinct groups: those with overestimated solubility and those with underestimated solubility. Among the data, there is a low solubility range where, logically, even if the absolute error $|\varepsilon_r|$ is larger than 3, both the predicted and actual target values are negative (log solubility < 0). This means that the deviation is negligible on a linear scale and that these data can be returned to the dataset. However, to keep the algorithm universal, general, and simple, we avoid applying this constraint at this stage, although this should be considered in future refinements.

The majority of the data identified during the outlier detection procedure contained acidic or basic sub-fragments. This influences the formation of ionic species, which depends on the $\mathrm{p}K_\mathrm{a}$ value of the solute and the pH of the solution. If a compound exhibits an acidic character in its neutral form, it will deprotonate at pH values greater than its $\mathrm{p}K_\mathrm{a}$, resulting in increased solubility. Similarly, molecules with basic characteristics exhibit higher solubility as the pH decreases below $\mathrm{p}K_\mathrm{a}$, leading to protonation. This highlights a significant limitation of many current databases, which often lack $\mathrm{p}K_\mathrm{a}$ (or $\mathrm{p}K_\mathrm{b}$) and pH information for each species.

Our interpretation is further reinforced by a comprehensive high-throughput study of the solubility of quinone derivatives [76]. Under strongly acidic conditions, various quinone derivatives were nearly insoluble because of their $\mathrm{p}K_\mathrm{a} >$ pH. These derivatives remain unaffected by chemical reactions and equilibrate with the solvent in their unionized forms. However, their solubility increases dramatically in a strong basic medium, where the $-\mathrm{OH}$ and $-\mathrm{SO_3H}$ functional groups undergo deprotonation reactions, resulting in the formation of $-\mathrm{O}^-$ and $-\mathrm{SO}_3^-$, respectively.

In addition to pH and $\mathrm{p}K_\mathrm{a}$, information regarding supporting electrolytes is another critical property that is currently missing from existing data banks. The presence of additional ions from the supporting electrolyte alters the concentration of free ions with opposite charges and shields interactions between solute ions, which can lead to modifications in solubil-

ity [77]. For example, the solubility of *2,2,6,6-tetramethyl-1-piperidinyloxy* [78] has increased from 0.8 M to 4.8 M [79] through the use of different salts.

Although laboratory errors in reported data are always possible, recent publications [80, 81] provide good examples to shed light on another existing reason for the discrepancies in data. A wide range of solubility has been reported for compounds from the same family, despite identical pH conditions and supporting electrolytes. Our computed pK_{a} values (see SI Sect. V) suggest that in each study, these molecules have very similar acidic or basic characteristics and exist in ionic form in the solutions studied, which would typically result in high solubility. This discrepancy can be attributed to crystallization processes. The low-soluble compounds will likely form more stable crystals even if they are chemically similar. In other words, the molecular packing in the crystal is very sensitive to the molecular structure and this can lead to large changes in crystal binding energy [82] (*polymorphism impact*).

In summary, the solubility data quality can be reliably assessed by obtaining information on the structural variations, acidity of the molecules, pH of the solution, and supporting electrolytes. While pK_{a} is relatively straightforward to evaluate nowadays, predicting structural variations presents more challenges and is still an open question [83]. Nevertheless, progress in ML-based potentials and sophisticated structure search methodologies offer substantial potential for accurately predicting the energy landscapes of polymorphic structures.

VII. Conclusions

The aim of this study has been to develop machine learning (ML)-based solubility models as a viable alternative to quantum computational methods. To achieve this, we undertook a systematic process that included data cleaning and analysis, descriptor development, and outlier detection. We proposed a workflow to refine experimental data, focusing on producing useful models through effective outlier detection.

Our efforts facilitate the development of models for predicting intrinsic solubility. The results indicate that both hydration free energy and subfragment-frequency-weighted fingerprints perform comparably to established descriptors. Notably, solvation energy, log P, V_{α}, and the number of ringed aromatic carbon atoms forming three bonds emerged as the most significant features for predicting solubility.

The effectiveness of our proposed outlier detection approach was validated through improved model performance. This method also contributed to data quality assessment and addressed missing information to mitigate discrepancies. We believe that further advancements in solubility prediction could be achieved by integrating relevant data from the literature on flow batteries.

The insights gained from this research have significant implications for both industrial applications and educational initiatives aimed at advancing solubility models.

VIII. Author Contributions

Arsalan Hashemi: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing original draft. Pekka Peljo: writing - review & editing. Tapio Ala-Nissila: resources, funding acquisition, writing – review & editing. Kari Laasonen: conceptualization, resources, funding acquisition, project administration, supervision, writing – review & editing.

IX. Conflicts of Interest

There are no conflicts to declare.

X. Acknowledgments

We acknowledge the Digipower project, supported by Teknologiateollisuuden 100v Säätiö and Jane ja Aatos Erkon Säätiö. A.H. and T.A-N. have been supported by the Academy of Finland through QTF Center of Excellence program (Project No. 312298) and the European Union – NextGenerationEU instrument grant 353298. We also express our gratitude to CSC–IT Center for Science Ltd for generous allocation of computing time.

References

[1] A. Chanda and V. V. Fokin, Organic synthesis "on water", Chemical Reviews **109**, 725 (2009), pMID: 19209944.

[2] M.-O. Simon and C.-J. Li, Green chemistry oriented organic synthesis in water, Chem. Soc. Rev. **41**, 1415 (2012).

[3] A.-R. Coltescu, M. Butnariu, and I. Sarac, The importance of solubility for new drug molecules, Biomedical and Pharmacology Journal **13**, 577 (2020).

[4] V. Singh and H. R. Byon, Solubility and stability of redox-active organic molecules in redox flow batteries, ACS Applied Energy Materials **7**, 7562 (2024).

[5] T. A. Welsh and E. R. Draper, Water soluble organic electrochromic materials, RSC Adv. **11**, 5245 (2021).

[6] M. E. Carrington, K. Sokołowski, E. Jónsson, E. W. Zhao, A. M. Graf, I. Temprano, J. A. McCune, C. P. Grey, and O. A. Scherman, Associative pyridinium electrolytes for air-tolerant redox flow batteries, Nature **623**, 949 (2023).

[7] M. Pan, M. Shao, and Z. Jin, Development of organic redox-active materials in aqueous flow batteries: Current strategies and future perspectives, SmartMat **4**, e1198 (2023).

[8] L. Zhang, R. Feng, W. Wang, and G. Yu, Emerging chemistries and molecular designs for flow batteries, Nature Reviews Chemistry **6**, 524 (2022), iD: Zhang2022.

[9] Y. Ding, C. Zhang, L. Zhang, Y. Zhou, and G. Yu, Molecular engineering of organic electroactive materials for redox flow

batteries, Chem. Soc. Rev. **47**, 69 (2018).
[10] V. Singh, S. Kim, J. Kang, and H. R. Byon, Aqueous organic redox flow batteries, Nano Research **12**, 1988 (2019), iD: Singh2019.
[11] E. H. Kerns, L. Di, and G. T. Carter, In vitro solubility assays in drug discovery, Current Drug Metabolism **9**, 879 (2008).
[12] Y. S. Krishnaiah, Pharmaceutical technologies for enhancing oral bioavailability of poorly soluble drugs, J Bioequiv Availab **2**, 28 (2010).
[13] T. C. Ezike, U. S. Okpala, U. L. Onoja, C. P. Nwike, E. C. Ezeako, O. J. Okpara, C. C. Okoroafor, S. C. Eze, O. L. Kalu, E. C. Odoh, U. G. Nwadike, J. O. Ogbodo, B. U. Umeh, and E. Ossai, Advances in drug delivery systems, challenges and future directions, Journal of Drug Delivery Science and Technology **56**, 123 (2023).
[14] S. Kalepu and V. Nekkanti, Insoluble drug delivery strategies: review of recent advances and business prospects, Acta Pharmaceutica Sinica B **5**, 442 (2015).
[15] M. Hewitt, M. T. D. Cronin, S. J. Enoch, J. C. Madden, D. W. Roberts, and J. C. Dearden, In silico prediction of aqueous solubility: The solubility challenge, Journal of Chemical Information and Modeling **49**, 2572 (2009), pMID: 19877720.
[16] D. S. Palmer and J. B. O. Mitchell, Is experimental data quality the limiting factor in predicting the aqueous solubility of drug-like molecules?, Molecular Pharmaceutics **11**, 2962 (2014), pMID: 24919008.
[17] A. Llinas, I. Oprisiu, and A. Avdeef, Findings of the second challenge to predict aqueous solubility, Journal of Chemical Information and Modeling **60**, 4791 (2020), pMID: 32794744.
[18] P. Llompart, C. Minoletti, S. Baybekov, D. Horvath, G. Marcou, and A. Varnek, Will we ever be able to accurately predict solubility?, Scientific Data **11**, 303 (2024).
[19] F. Silva, F. Veiga, S. P. J. Rodrigues, C. Cardoso, and A. C. Paiva-Santos, Cosmo models for the pharmaceutical development of parenteral drug formulations, European Journal of Pharmaceutics and Biopharmaceutics **187**, 156 (2023).
[20] A. J. Hopfinger, E. X. Esposito, A. Llinàs, R. C. Glen, and J. M. Goodman, Findings of the challenge to predict aqueous solubility, J. Chem. Inf. Model. **49**, 1 (2009).
[21] M. Murase and D. Nakamura, Hansen solubility parameters for directly dealing with surface and interfacial phenomena, Langmuir **39**, 10475 (2023), pMID: 37463335.
[22] Z. Bjelobrk, D. Mendels, T. Karmakar, M. Parrinello, and M. Mazzotti, Solubility prediction of organic molecules with molecular dynamics simulations, Crystal Growth & Design **21**, 5198 (2021).
[23] J. Ali, P. Camilleri, M. B. Brown, A. J. Hutt, and S. B. Kirton, Revisiting the general solubility equation: in silico prediction of aqueous solubility incorporating the effect of topographical polar surface area, Journal of Chemical Information and Modeling **52**, 420 (2012).
[24] S. Boothroyd, A. Kerridge, A. Broo, D. Buttar, and J. Anwar, Solubility prediction from first principles: a density of states approach, Phys. Chem. Chem. Phys. **20**, 20981 (2018).
[25] L.-S. Berg, J. Hamaekers, and A. Maass, Machine learning for fb electrolyte screening, in *Flow Batteries* (John Wiley Sons, Ltd, 2023) Chap. 21, pp. 487–506.
[26] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018), iD: Butler2018.
[27] G. Klopman and H. Zhu, Estimation of the aqueous solubility of organic molecules by the group contribution approach, Journal of Chemical Information and Computer Sciences **41**, 439 (2001), pMID: 11277734.
[28] J. S. Delaney, Esol: Estimating aqueous solubility directly from molecular structure, Journal of Chemical Information and Computer Sciences **44**, 1000 (2004), pMID: 15154768.
[29] S. H. Hilal, S. W. Karickhoff, and L. A. Carreira, Prediction of the solubility, activity coefficient and liquid/liquid partition coefficient of organic compounds, QSAR & Combinatorial Science **23**, 709 (2004).
[30] B. Sanchez-Lengeling, L. M. Roch, J. D. Perea, S. Langner, C. J. Brabec, and A. Aspuru-Guzik, A bayesian approach to predict solubility parameters, Advanced Theory and Simulations **2**, 1800069 (2019).
[31] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, Machine learning for materials scientists: An introductory guide toward best practices, Chemistry of Materials **32**, 4954 (2020).
[32] M. C. Sorkun, J. V. A. Koelman, and S. Er, Pushing the limits of solubility prediction via quality-oriented data selection, iScience **24**, 101961 (2021).
[33] T. Li, C. Zhang, and X. Li, Machine learning for flow batteries: opportunities and challenges, Chem. Sci. **13**, 4740 (2022).
[34] A. Khetan, High-throughput virtual screening of quinones for aqueous redox flow batteries: Status and perspectives, Batteries **9**, 10.3390/batteries9010024 (2023).
[35] M. R. Tuttle, E. M. Brackman, F. Sorourifar, J. Paulson, and S. Zhang, Predicting the solubility of organic energy storage materials based on functional group identity and substitution pattern, The Journal of Physical Chemistry Letters **14**, 1318 (2023), pMID: 36724735.
[36] F. Wang, J. Li, Z. Liu, T. Qiu, J. Wu, and D. Lu, Computational design of quinone electrolytes for redox flow batteries using high-throughput machine learning and theoretical calculations, Frontiers in Chemical Engineering **4**, 10.3389/fceng.2022.1086412 (2023).
[37] M. Su and E. Herrero, Creation and interpretation of machine learning models for aqueous solubility prediction, Exploration of Drug Science **1**, 388 (2023).
[38] Q. Zhang, A. Khetan, E. Sorkun, F. Niu, A. Loss, I. Pucher, and S. Er, Data-driven discovery of small electroactive molecules for energy storage in aqueous redox flow batteries, Energy Storage Materials **47**, 167 (2022).
[39] M. D. Chaka, Y. S. Mekonnen, Q. Wu, and C. A. Geffe, Advancing energy storage through solubility prediction: leveraging the potential of deep learning, Phys. Chem. Chem. Phys. **25**, 31836 (2023).
[40] A. D. Vassileiou, M. N. Robertson, B. G. Wareham, M. Soundaranathan, S. Ottoboni, A. J. Florence, T. Hartwig, and B. F. Johnston, A unified ml framework for solubility prediction across organic solvents, Digital Discovery **2**, 356 (2023).
[41] S. Lee, H. Park, C. Choi, W. Kim, K. K. Kim, Y.-K. Han, J. Kang, C.-J. Kang, and Y. Son, Multi-order graph attention network for water solubility prediction and interpretation, Scientific Reports **13**, 957 (2023).
[42] J. Yu, C. Zhang, Y. Cheng, Y.-F. Yang, Y.-B. She, F. Liu, W. Su, and A. Su, Solvbert for solvation free energy and solubility prediction: a demonstration of an nlp model for predicting the properties of molecular complexes, Digital Discovery **2**, 409 (2023).
[43] M. Lovrić, K. Pavlović, P. Žuvela, A. Spataru, B. Lučić, R. Kern, and M. W. Wong, Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability?, Journal of Chemometrics **35**, e3349 (2021).

[44] S. Boobier, D. R. J. Hose, A. J. Blacker, and B. N. Nguyen, Machine learning with physicochemical relationships: solubility prediction in organic solvents and water, Nature Communications **11**, 5753 (2020).

[45] D. S. Palmer, N. M. O'Boyle, R. C. Glen, and J. B. O. Mitchell, Random forest models to predict aqueous solubility, Journal of Chemical Information and Modeling **47**, 150 (2007), pMID: 17238260.

[46] M. D. Chaka, C. A. Geffe, A. Rodriguez, N. Seriani, Q. Wu, and Y. S. Mekonnen, High-throughput screening of promising redox-active molecules with molgat, ACS Omega **8**, 24268 (2023).

[47] L. Tang, P. Leung, Q. Xu, and C. Flox, Machine learning orchestrating the materials discovery and performance optimization of redox flow battery, ChemElectroChem **11**, e202400024 (2024).

[48] C. Trinh, S. Lasala, O. Herbinet, and D. Meimaroglou, On the development of descriptor-based machine learning models for thermodynamic properties: Part 2mdash;applicability domain and outliers, Algorithms **16**, 10.3390/a16120573 (2023).

[49] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, Comparison of different approaches to define the applicability domain of qsar models, Molecules **17**, 4791 (2012).

[50] G. Panapitiya and E. Saldanha, Outlier-based domain of applicability identification for materials property prediction models (2023), arXiv:2302.06454 [cond-mat.mtrl-sci].

[51] D.-S. Cao, Z.-K. Deng, M.-F. Zhu, Z.-J. Yao, J. Dong, and R.-G. Zhao, Ensemble partial least squares regression for descriptor selection, outlier detection, applicability domain assessment, and ensemble modeling in qsar/qspr modeling, Journal of Chemometrics **31**, e2922 (2017), e2922 CEM-17-0029.R1, https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.2922.

[52] P. Gao, A. Andersen, J. Sepulveda, G. U. Panapitiya, A. Hollas, E. G. Saldanha, V. Murugesan, and W. Wang, Somas: a platform for data-driven material discovery in redox flow battery development, Scientific Data **9**, 740 (2022).

[53] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2004).

[54] What is density functional theory?, in *Density Functional Theory* (John Wiley Sons, Ltd, 2009) Chap. 1, pp. 1–33.

[55] D. Chen, T. Shen, K. An, and J. Zhu, Adaptive aromaticity in s0 and t1 states of pentalene incorporating 16 valence electron osmium, Communications Chemistry **1**, 18 (2018).

[56] C. Liu, M. E. Sandoval-Salinas, Y. Hong, T. Y. Gopalakrishna, H. Phan, N. Aratani, T. S. Herng, J. Ding, H. Yamada, D. Kim, D. Casanova, and J. Wu, Macrocyclic polyradicaloids with unusual super-ring structure and global aromaticity, Chem **4**, 1586 (2018).

[57] Z. Niu, H. Wu, Y. Lu, S. Xiong, X. Zhu, Y. Zhao, and X. Zhang, Orbital-dependent redox potential regulation of quinone derivatives for electrical energy storage, RSC Adv. **9**, 5164 (2019).

[58] Q. Wang, T. Y. Gopalakrishna, H. Phan, T. S. Herng, S. Dong, J. Ding, and C. Chi, Cyclopenta ring fused bisanthene and its charged species with open-shell singlet diradical character and global aromaticity/ anti-aromaticity, Angewandte Chemie International Edition **56**, 11415 (2017).

[59] T. M. Krygowski and M. K. Cyrański, Structural aspects of aromaticity, Chemical Reviews **101**, 1385 (2001), pMID: 11710226.

[60] C. DeBruler, B. Hu, J. Moss, X. Liu, J. Luo, Y. Sun, and T. L. Liu, Designer two-electron storage viologen anolyte materials for neutral aqueous organic redox flow batteries, Chem **3**, 961 (2017).

[61] E. I. Romadina and K. J. Stevenson, Small-molecule organics for redox flow batteries – creation of highly-soluble and stable compounds, Electrochimica Acta **461**, 142670 (2023).

[62] D. Cremoncini, G. Di Lorenzo, G. F. Frate, A. Bischi, A. Baccioli, and L. Ferrari, Techno-economic analysis of aqueous organic redox flow batteries: Stochastic investigation of capital cost and levelized cost of storage, Applied Energy **360**, 122738 (2024).

[63] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Advanced Drug Delivery Reviews **23**, 3 (1997), in Vitro Models for Selection of Development Candidates.

[64] M. Valiev, E. Bylaska, N. Govind, K. Kowalski, T. Straatsma, H. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. Windus, and W. de Jong, Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations, Computer Physics Communications **181**, 1477 (2010).

[65] A. Klamt and G. Schüürmann, Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, J. Chem. Soc., Perkin Trans. 2 , 799 (1993).

[66] A. Hashemi, R. Khakpour, A. Mahdian, M. Busch, P. Peljo, and K. Laasonen, Density functional theory and machine learning for electrochemical square-scheme prediction: an application to quinone-type molecules relevant to redox flow batteries, Digital Discovery **2**, 1565 (2023).

[67] K. Farshadfar, A. Hashemi, R. Khakpour, and K. Laasonen, Kinetics of n2 release from diazo compounds: A combined machine learning-density functional theory study, ACS Omega **0**, null (2023).

[68] A. Tayyebi, A. S. Alshami, Z. Rabiei, X. Yu, N. Ismail, M. J. Talhuk, and J. Power, Prediction of organic compound aqueous solubility using machine learning: A comparison study of descriptor-based and fingerprints-based models, Journal of Cheminformatics **15**, 99 (2023).

[69] G. Landrum, Rdkit: Open-source cheminformatics software, (2016).

[70] G. M. Ghiandoni and E. Caldeweyher, Fast calculation of hydrogen-bond strengths and free energy of hydration of small molecules, Scientific reports **13**, 4143 (2023).

[71] H. YUKAWA, On the interaction of elementary particles. i, Proceedings of the Physico-Mathematical Society of Japan. 3rd Series **17**, 48 (1935).

[72] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. I. Goddard, and W. M. Skiff, Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations, Journal of the American Chemical Society **114**, 10024 (1992).

[73] D. Rogers and M. Hahn, Extended-connectivity fingerprints, Journal of Chemical Information and Modeling **50**, 742 (2010), pMID: 20426451.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).

[75] L. Breiman, Random forests, Machine learning **45**, 5 (2001).

[76] K. Wedege, E. Dražević, D. Konya, and A. Bentien, Organic redox species in aqueous flow batteries: Redox potentials, chemical stability and solubility, Scientific Reports **6**, 39101 (2016).

[77] S. Creager, 3 - solvents and supporting electrolytes, in *Handbook of Electrochemistry*, edited by C. G. Zoski (Elsevier, Amsterdam, 2007) pp. 57–72.

[78] N. Prakash, R. Rajeev, A. John, A. Vijayan, L. George, and A. Varghese, 2,2,6,6-tetramethylpiperidinyloxyl (tempo) radical mediated electro-oxidation reactions: A review, ChemistrySelect **6**, 7691 (2021).

[79] W. Zhou, W. Liu, M. Qin, Z. Chen, J. Xu, J. Cao, and J. Li, Fundamental properties of tempo-based catholytes for aqueous redox flow batteries: effects of substituent groups and electrolytes on electrochemical properties, solubilities and battery performance, RSC Adv. **10**, 10.1039/D0RA03424J (2020).

[80] E. F. Kerr, Z. Tang, T. Y. George, S. Jin, E. M. Fell, K. Amini, Y. Jing, M. Wu, R. G. Gordon, and M. J. Aziz, High energy density aqueous flow battery utilizing extremely stable, branching-induced high-solubility anthraquinone near neutral ph, ACS Energy Letters **8**, 600 (2023).

[81] X. Liu, T. Li, C. Zhang, and X. Li, Benzidine derivatives: A class of high redox potential molecules for aqueous organic flow batteries, Angewandte Chemie International Edition **62**, e202307796 (2023).

[82] S. N. Garcia, X. Yang, L. Bereczki, and D. Kónya, Aqueous solubility of organic compounds for flow battery applications: Symmetry and counter ion design to avoid low-solubility polymorphs, Molecules **26**, 10.3390/molecules26051203 (2021).

[83] A. J. Cruz-Cabeza, N. Feeder, and R. J. Davey, Open questions in organic crystal polymorphism, Communications Chemistry **3**, 142 (2020).