Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00456j

# A case study on hybrid machine learning and quantum-informed modelling for solubility prediction of drug compounds in organic solvents

Weiling Wang, [a] Isabel Cooley,[b] Morgan R. Alexander, [c] Ricky D. Wildman,[d] Anna K. Croft [b] and Blair F. Johnston *[a]

Received 10th October 2025
Accepted 19th December 2025

DOI: 10.1039/d5dd00456j

rsc.li/digitaldiscovery

Solubility is a physicochemical property that plays a critical role in pharmaceutical formulation and processing. While COSMO-RS offers physics-based solubility estimates, its computational cost limits large-scale application. Building on earlier attempts to incorporate COSMO-RS-derived solubilities into Machine Learning (ML) models, we present a substantially expanded and systematic hybrid QSAR framework that advances the field in several novel ways. The direct comparison between COSMOtherm and openCOSMO revealed consistent hybrid augmentation across COSMO engines and enhanced reproducibility. Three widely used ML algorithms, eXtreme Gradient Boosting, Random Forest, and Support Vector Machine, were benchmarked under both 10-fold and leave-one-solute-out cross-validation. The comparison between four major descriptor sets, including MOE, Mordred, RDKit descriptors, and Morgan Fingerprints, offering the first descriptor-level assessment of how COSMO-RS calculated solubility augmentation interacts with diverse chemical feature space. The statistical Y-scrambling was conducted to confirm that the hybrid improvements are genuine and not artefacts of dimensionality. SHAP-based feature analysis further revealed substructural patterns linked to solubility, providing interpretability and mechanistic insight. This study demonstrates that combining physics-informed features with robust, interpretable ML algorithms enables scalable and generalisable solubility prediction, supporting data-driven pharmaceutical design.

## 1 Introduction

Solubility is a key determinant of pharmaceutical formulation and compound screening,[1,2] affecting solute–solvent compatibility, mixture stability, and bioavailability.[3] Reliable solubility models can accelerate early-stage development by narrowing down viable solvent–solute combinations before costly experimental testing. Among available tools, the COnductor-like Screening Model for Real Solvents (COSMO-RS)[4–7] has been widely adopted for its quantum chemistry-based accuracy, offering reasonable solubility predictions across diverse solvent systems. However, the requirement for computationally costly geometry optimisations and COSMO energy calculations restricts its use in high-throughput or exploratory workflows.

Machine Learning (ML) offers a complementary route by learning structure–property patterns from data. Once trained on an appropriate feature set, ML models can deliver predictions at lower computational cost. When the feature set is augmented with COSMO-RS-derived descriptors, the resulting models not only improve predictive accuracy but also preserve a degree of physics-based interpretability. More broadly, ML has emerged as a powerful tool for property prediction in chemistry and materials science.[8,9] It can effectively capture complex links between structure, composition, and properties, allowing the design of targeted compounds and even the generation of novel materials.[10–13] Within this domain, the Quantitative–Structure–Property Relationship (QSPR) and Quantitative–Structure–Activity Relationship (QSAR) approaches serve as a powerful framework that correlate molecular structure with experimentally measured physicochemical or material properties through linear or nonlinear modelling based on various descriptors.[14–18] Applications span the prediction of solubility,[19–21] boiling point,[16,22,23] polarisability,[24,25] and viscosity,[26–28] while offering insights into the contributions of specific functional groups and structural motifs.[29–31] It thereby supports the rational design and optimisation of molecules and materials across chemical, pharmaceutical, and materials sciences.

[a] CMAC, University of Strathclyde, Glasgow, G1 1RD, UK. E-mail: weiling.wang@strath.ac.uk; blair.johnston@strath.ac.uk

[b] Department of Chemical Engineering, Loughborough University, Loughborough, LE11 3TU, UK

[c] School of Pharmacy, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

[d] Centre for Additive Manufacturing, University of Nottingham, Nottingham, NG7 2RD, UK

Given the high dimensionality introduced by molecular descriptors and fingerprints, modelling efforts often face the "curse of dimensionality", where available data sparsely samples the chemical space.[32] To address this challenge, we benchmarked three widely used, non-linear algorithms well suited to sparse, high-dimensional problems: eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM).

Although the aqueous solubility of drug compounds has been widely investigated, studies on solubility in organic solvents remain comparatively limited.[33–37] In this study, we presented a QSAR-based solubility prediction framework tailored for drug-like compounds in organic solvents. Extending previous work that used RF with Molecular Operating Environment (MOE)[33] descriptors, we expanded the modelling framework to include XGBoost and SVM. Multiple descriptor types are systematically compared: MOE,[38] Mordred descriptors,[39] RDKit descriptors,[40] Morgan Fingerprints,[41] and COSMO-RS predictions from COSMOtherm and openCOSMO-RS. Mordred descriptors were computed using the Mordred cheminformatics package; RDKit descriptors were calculated using the RDKit cheminformatics toolkit; and Morgan Fingerprints were generated with the RDKit package. COSMO-RS simulated solubility is incorporated as an auxiliary feature rather than used standalone, following evidence that hybrid models outperform both descriptor-only ML and COSMO-RS alone.

Beyond accuracy, interpretability is emphasised through SHapley Additive exPlanations (SHAP),[42] which decompose predictions into feature contributions and highlight substructural motifs influencing solubility. This not only elucidates the contribution of individual input features to model predictions but also highlights the factors most critical in determining solubilities under the QSAR framework. Among the selected algorithms, XGBoost,[43] a regularised boosting algorithm, sequentially optimises decision trees to correct residual errors and has demonstrated strong performance in various molecular prediction tasks,[44–46] including feature reconstruction tasks such as Raman spectra,[47] Near-Infrared (NIR) spectra,[48] and Infrared (IR) spectra prediction.[49] In contrast, RF aggregates independently built decision trees using bootstrap resampling, providing robustness and stability but lacking the iterative refinement of boosting. SVM uses kernel functions to project input data into higher dimensions, offering strong performance for high-dimensional and small-sample problems. However, their performance can be sensitive to parameter selection, and scalability may be limited for extensive datasets. These three algorithms are frequently selected and compared in cheminformatics studies for their robust predictive capabilities.

Several recent works highlight their utility. Kim *et al.*[50] predicted the antioxidant activity of 2,2-diphenyl-1-picrylhydrazyl (DPPH) using XGBoost, RF, and SVM with RDKit descriptors. Qu *et al.*[51] compared XGBoost, RF, SVM, and K-Nearest Neighbour (KNN) for retention time prediction of proteolysis-targeting chimeras, using fingerprints, physicochemical descriptors, along with chromatographic-condition features. Ghuriani *et al.*[52] developed an XGBoost-driven biomarker identification pipeline that fed selected features into RF, SVM, and logistic regression for cancer prediction from gene sequencing. Danishuddin *et al.*[53] benchmarked XGBoost, RF, SVM, and Multi-Layer Perceptron (MLP) to model HIV-1 integrase inhibitor activity, comparing PaDEL and RDKit descriptors with ECFP4 fingerprints (Morgan Fingerprints). Collectively, these studies underscore the prominence of XGBoost, RF, and SVM as standard benchmarks in cheminformatics, where they are routinely employed in parallel to assess predictive performance under varying descriptor types and molecular contexts, supporting their selection here for solubility modelling.

Recent advances have emphasised hybrid strategies that integrate physics-based knowledge into ML frameworks, aiming to preserve mechanistic interpretability while enhancing predictive scalability. Beyond descriptors and fingerprints, the output of physics-based models is increasingly incorporated as auxiliary features to enrich the input space. For example, Vassileiou *et al.*[33] showed that including COSMOtherm solubility predictions as features improved the RF drug solubility models. Xiong *et al.*[54] combined first-principles descriptors generated *via* Multiwfn with conventional descriptors to predict flotation behaviour, while Lu *et al.*[55] embedded quantum-derived electronic features into deep learning for the prediction of drug–drug interaction.

These studies collectively demonstrate that augmenting data-driven QSAR/QSPR with physics-based descriptors yields models that are both more predictive and mechanistically interpretable. Although the trends identified align with established chemical intuition, the strength of this framework lies in its ability to validate and generalise these relationships systematically across hundreds of solute–solvent pairs. Building on this foundation, we designed a workflow that systematically integrates diverse descriptor types with COSMO-RS features to benchmark ML algorithms for solubility prediction, as shown in Fig. 1.

The dataset used in this study is based on the solubility collection compiled by Vassileiou *et al.*,[33] which combines experimental measurements extracted from the literature with their in-house data. The dataset contains 714 solubility measurements at room temperature, covering 75 organic solutes and 49 solvents. The solutes span a chemically diverse set of drug-like small molecules, making the dataset representative of pharmaceutical formulation challenges.

For comparison, Sodaei *et al.*[56] integrated MD-derived properties with ML and reported a gradient boosting model that achieved an $R^2$ of 0.87 and an RMSE of 0.537 for aqueous drug solubility prediction under 10-fold CV. Alqarni *et al.*[57] employed one-hot encoded solvents, temperature, and mass fraction as inputs to predict rivaroxaban solubility, where the light gradient boosting model achieved the best performance with an $R^2$ of 0.988 and an RMSE of $9.13 \times 10^{-5}$ under 5-fold CV. Jiang *et al.*[58] utilised temperature and pressure to model the solubility of nonsteroidal anti-inflammatory drugs in green solvents, achieving an $R^2$ of 0.987 and an RMSE of 13.7 under 10-fold CV using AdaBoost with Gaussian Process Regression (ADA–GPR).

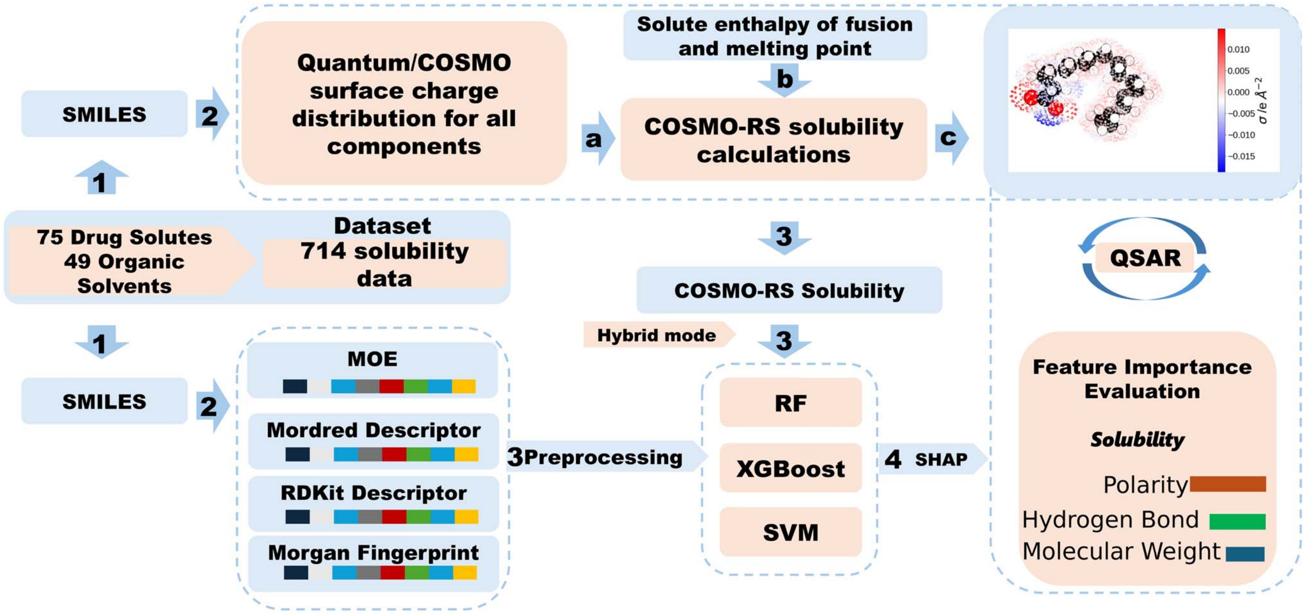

Fig. 1 Workflow for solubility modelling and interpretation. (1) A dataset of 714 binary solute–solvent systems is encoded using SMILES. (2) These SMILES serve as inputs for: openCOSMO solubility prediction utilising (a) surface charge distributions obtained from BP86/def2TZVPD and COSMO calculations (b) in cases where the solute is a solid, solute enthalpies of fusion and melting points, and (c) a representative COSMO surface charge density visualisation shown for illustration as part of the COSMO solubility output; and for the generation of MOE, RDKit, and Mordred descriptors as well as Morgan Fingerprints. (3) The resulting COSMO-RS solubility estimates and preprocessed descriptor sets are combined as input under a hybrid mode. Machine learning models (RF, XGBoost, SVM) are trained to predict solubility. (4) SHAP-based heatmaps then decompose model outputs into descriptor and fingerprint contributions, translating predictions into QSAR insights.

It is important to note that these studies were trained on different datasets, often restricted to a single solvent system or a narrow chemical domain. As a result, direct numerical comparisons to our results can be misleading, since dataset composition and chemical diversity strongly influence apparent model accuracy. Moreover, most of these works rely on *n*-fold CV, which typically yields higher apparent performance than LOSO while providing a weaker measure of generalisability.

By contrast, the focus of our work lies in integrating diverse cheminformatics descriptors with multiple ML frameworks and providing physicochemical interpretation across a broader range of drug-like molecules and organic solvents, rather than solely pursuing the highest apparent predictive accuracy.

Our integrated framework benchmarked descriptor sets, algorithms, and hybrid strategies, while linking substructure-level contributions to solubility behaviour. The remainder of this paper presents the dataset and descriptors (Section 2.1), COSMO-RS method (Section 2.2), ML modelling and evaluation procedures (Section 2.3), predictive performance and interpretability outcomes (Section 3.1), and concludes with broader implications for data-driven pharmaceutical design (Section 3.2).

## 2 Model and method

In the original study, Vassileiou *et al.*[33] trained hybrid machine learning models on this solubility dataset, achieving the coefficient of determination, $R^2$, of 0.56 and 0.78, Mean Absolute Errors (MAE) of 0.36 and 0.59, and Root Mean Square Errors (RMSE) of 0.79 and 0.55 under Leave One Solute Out (LOSO) and 10-fold Cross Validation (CV), respectively. Building on these results, this dataset provides the foundation for our benchmarking and model development, supporting comparative evaluation of hybrid feature sets and learning algorithms for solubility prediction.

Solubility arises from the interactions between solute and solvent, and cannot be explained solely by the properties of either in isolation. A solute that dissolves readily in a polar solvent can be expected to display very low solubility in a non-polar organic solvent with contrasting chemical characteristics. It is therefore important, when examining the dataset, to define the chemical space spanned by the solvents, so that any conclusions about solute features contributing to solubility are interpreted within the context of solvent type. The dataset used in this work was originally designed to include organic solvents,[33] which generally display high lipophilicity and low hydrophilicity, participating in favourable interactions with non-polar organic solutes. The octanol/water partition coefficient, $\log P$(octanol/water), is a widely used measure of lipophilicity/hydrophilicity. In this work, it was calculated using the MOE $\log P$(o/w) descriptor, a fragment-based method trained on experimental data. Higher $\log P$(o/w) values indicate greater lipophilicity and lower polarity, whereas lower values reflect greater hydrophilicity. A compound with a $\log P$(o/w) of 0 would partition equally between an octanol and a water phase, with positive $\log P$(o/w) values favouring octanol (hydrophobic) and negative $\log P$(o/w) values favouring water (hydrophilic). In our dataset, solvent $\log P$(o/w) values range from −1.1 to 7.8. In total, 8 solvents have $\log P$(o/w) below 0, while the other 41 have log *P*(o/w) above 0. The solvents are predominantly lipophilic, although the presence of several low-log *P* solvents introduces chemical diversity that complicates simple trends in solute–solvent solubility relationships. As introduced at the end of Section 1, the 714 solute–solvent pairs consist of 75 organic solutes and 49 solvents. The 12 most commonly occurring solvent molecules in this dataset each appear in at least 20 solvent/solute mixtures and together account for 381 total mixtures, over half of the dataset. These are ethanol, methanol, ethyl acetate, 2-propanol, acetone, 1-butanol, acetonitrile, chloroform, water, 1-propanol, 1-octanol and tetrahydrofuran. Thus, the range includes water and a number of relatively small organic molecules, all containing some functionality which can contribute to dipolar or hydrogen bonds. Some of the longer chain molecules among this group have lipophilic characteristics, particularly 1-octanol in which polar compounds are only sparingly soluble and which is used to define log *P*(o/w) and has a log *P*(o/w) of 2.8. However, there is enough polar functionality among the set of solvents that solvent polarity must also be considered when concluding trends in the structure of solutes.

### 2.1 Molecular descriptors

To numerically represent the chemical features of each solute and solvent, we computed three complementary sets of molecular descriptors, in addition to retaining the original MOE descriptors. Mordred descriptors (2D only) were generated using the Mordred Python package,[39] yielding over 1600 features spanning physicochemical, topological, and constitutional properties. Morgan Fingerprints (extended-connectivity fingerprints, ECFP)[41,59] were generated using RDKit,[40] with a radius of 2 and a 2048 bit length. The associated *bitInfo* metadata was preserved, enabling active bits to be mapped back to atomic substructures. To reduce redundancy, low-information bits were filtered using a Shannon entropy[60] threshold of 0.001, based on binary activation profiles across all molecules. RDKit descriptors were computed using RDKit's built-in set,[40] encompassing diverse physicochemical and topological properties. For all descriptor types, features with >10 per cent missing values were discarded. The remaining features were zero-filled and filtered to remove zero-variance columns prior to analysis.

After applying all filtering steps described above, the final dimensionalities reported here refer to the combined solute–solvent descriptor space used as input to the ML models. In total, the retained descriptors comprised 357 MOE features, 322 RDKit features, 2439 Mordred features, and 875 informative Morgan Fingerprint bits. These dimensionalities therefore reflect the sum of all filtered solute and solvent descriptors entering the final models. Although the Mordred representation remains larger than the other descriptor families, its effective dimensionality was substantially reduced from the initial raw feature set, and all retained descriptors passed the missingness, variance, and stability criteria. Moreover, all models were trained exclusively under rigorous CV schemes, which provide a strong safeguard against overfitting. The differences in predictive performance between descriptor families therefore cannot be attributed solely to the number of retained features.

It is also worth noting that all selected modelling approaches: XGBoost, RF, and SVM, are widely recognised for their robustness in high-dimensional settings owing to intrinsic regularisation, non-linear feature compression, and resistance to overfitting. Together with the descriptor filtering steps and the use of strict CV, this ensures that the observed differences in predictive performance cannot be attributed solely to the size of the feature space. To further validate that the models learn genuine chemical signal rather than spurious correlations arising from high dimensionality, we performed Y-scrambling with $B = 200$ permutations for every descriptor family and modelling approach. As shown in S2-SI, in all cases, the scrambled models collapsed to chance level, and the probability that the original model performance could be reproduced under randomly permuted targets was $p = 0.005$. These results confirm that the input features contain statistically significant information, and that the hybrid ML-physics models do not rely on spurious correlations introduced by descriptor dimensionality.

### 2.2 Physics-based solubility estimates

Mechanistic solubility estimates were generated for each solute–solvent pair using the COSMO-RS method.[4–7] The predicted log solubility served as an additional numerical input, providing quantum chemistry-derived insights into solution behaviour. Importantly, COSMO-RS outputs were incorporated solely as input descriptors rather than as synthetic training targets, ensuring that all models were trained exclusively on experimental solubilities. The hybrid approach integrates COSMO-RS predictions with structure-encoded descriptors, allowing the model to learn residual structure–property relationships beyond the physics-based baseline.

In the interest of reproducibility, open science and the FAIR (Findable, Accessible, Interoperable and Reusable) data principles, we investigated the use of the open source openCOSMO-RS software[61,62] to generate the physics-based features. Our openCOSMO-RS workflow involved accounting for the influence of multiple conformers as described by Klamt[63] and has been previously used by Schindl *et al.*[64] We calculated solubility iteratively from COSMO-RS activity coefficients using infinite dilution as an initial guess, as described elsewhere.[64,65] Solubility calculations were performed at 298.15 K, and for molecules whose melting point was above working temperature, the free energy of fusion was calculated from literature experimental melting point and enthalpy of fusion values taken from the original dataset of Vassileiou *et al.*[4,33] Newly generated openCOSMO-RS results were benchmarked against the original COSMOtherm data[33] before being used as inputs for each QSAR model.

Given the nature of this work, it is important to note that our current solute–solvent dataset is already highly challenging, containing substantial chemical diversity and many intrinsically difficult cases. Expanding to external datasets such as AqSolDB[66] or BigSolDB[67] would require COSMO-RS calculations for every associated solute–solvent pair, which is computationally prohibitive at present. As our models rely on COSMO-RS-derived solubility inputs, such benchmarking is deferred to future work once large-scale COSMO-RS data become available.

Instead, a more efficient strategy is to perform targeted validation on a representative subset of compounds drawn from these larger datasets. This allows us to benchmark model accuracy externally without undertaking full COSMO-RS enumeration. The results will be discussed in detail in Section 3.1.5.

### 2.3 Machine learning model

We benchmarked three supervised regression algorithms, eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM), for predicting the solubility of drug-like compounds in organic solvents. XGBoost models were implemented using *XGBRegressor* for regularised gradient boosting, while RF and SVM models employed *RandomForestRegressor* and *SVR*, respectively, *via* the scikit-learn framework.

To minimise hardware-induced variability, XGBoost was run in CPU-only mode (*CUDA_VISIBLE_DEVICES* = −1, *device* = '*cpu*') using the histogram tree method (*tree_method* = '*hist*'). OpenMP parallelism was restricted to eight threads (*OMP_NUM_THREADS* = *8*), and the regressor was configured with eight worker threads (*n_jobs* = *8*) while CV fits were run sequentially (*RandomizedSearchCV(n_jobs* = *1)*). These settings control the degree of parallelism and help stabilise floating-point round-off behaviour, so that repeated runs on different machines produce numerically comparable CV metrics.

Model inputs included molecular descriptors derived from one of four descriptor sets: MOE descriptors, RDKit descriptors, Mordred descriptors, or Morgan Fingerprints. COSMO-RS-predicted solubility was incorporated as an additional input feature in hybrid models. Model performance was evaluated using 10-fold CV and Leave-One-Solute-Out (LOSO) CV, with results reported using $R^2$, RMSE, and MAE, averaged across folds. LOSO ensures that no solute appears in both training and test folds, preventing solute memorisation and reducing leakage from solute recurrence.

**2.3.1 Hyperparameter optimisation.** Each model was subjected to hyperparameter optimisation using *RandomizedSearchCV*[68] with three-fold internal CV, a search budget of $n_{\text{iter}} = 10$, employing negative root mean squared error as the scoring criterion. This strategy enabled efficient sampling of the hyperparameter space while reducing the risk of overfitting to specific folds. Full search spaces are provided in SI Section 1.

To quantify descriptor importance and generate heatmaps, SHAP values were computed. For XGBoost regressors, we used *TreeExplainer* in *interventional* mode with a fixed background of 100 samples. For SVM regressors, SHAP values were obtained with *KernelExplainer* using a *K-means* background (≤ 10 clusters), with perturbation sampling limited for efficiency and *l1_reg (num_features(10))* applied to stabilise attributions. Random Forest models were not analysed with SHAP due to weaker predictive performance and limited interpretability within this framework.

*2.3.1.1 Y-scrambling.* To assess the risk of chance correlations, we performed Y-scrambling (response permutation) under both CV schemes (10-fold and LOSO), preserving the frozen splits of the observed models. For 10-fold CV, the same fixed random seed and fixed fold assignment were reused for every permutation, and no replicated resampling was performed. Only the training targets were permuted within each split, and models were re-fit using the same fixed hyperparameters. One-sided empirical permutation *p*-values were computed following the standard formulation,[69] with $B = 200$ random permutations per model. Complete methodological details and permutation histograms are provided in SI-S2.

**2.3.2 Model explainability and substructure analysis.** To interpret model predictions and identify key molecular characterisation driving solubility, we employed SHapley Additive exPlanations (SHAP),[42,70,71] which attributes changes in predicted output to individual input features based on cooperative game theory. SHAP values were computed using the best-performing model refitted on the full dataset after CV, ensuring maximum exposure to training data. For descriptor-based models, global feature importance was quantified by ranking descriptors according to their mean absolute SHAP values. Inputs comprised per-observation solute–solvent feature vectors. Preprocessing steps preserved sample-level granularity. SHAP therefore reports local contributions for each sample.

The top 20 most influential descriptors were visualised as a heatmap and exported for further analysis, highlighting the physicochemical properties that predominantly impact predicted solubility. For fingerprint-based models using Morgan Fingerprints, we extended the analysis by mapping top-ranked fingerprint bits to substructures using RDKit.[40] Substructures corresponding to high-ranking bits were extracted and visualised across all activating solutes and solvents, enabling interpretation of fragment-level contributions. Additionally, bit-wise SHAP heatmaps were generated, offering a chemically intuitive view of substructure importance.

## 3 Results and discussion

### 3.1 Model performance

**3.1.1 Re-implementation of baseline RF model.** To establish a consistent and reproducible baseline, we first re-implemented a previously developed RF solubility prediction model. The original version was written in R with the *randomForest* package and MOE descriptors, which we translated into Python using scikit-learn. All original settings were preserved, including descriptor input, original COSMOtherm solubility predictions as hybrid features, model parameters (default in both R and Python), and CV strategy. In Table 1, minor performance differences arise from inherent implementation-level distinctions between R and Python libraries (*e.g.*, bootstrap sampling routines, random number generator algorithms, and default tree-splitting criteria). These differences are purely technical, numerically insignificant, and do not affect the scientific conclusions. The equations for these performance metrics are provided in SI-S4.

**3.1.2 Integration with openCOSMO.** To evaluate the utility of open-source COSMO methods, we compared openCOSMO against COSMOtherm using the same Python machine learning

Table 1 Performance comparison of Random Forest solubility models re-implemented from R to Python using identical MOE descriptors and CV settings. Minor differences reflect implementation-level variations in random sampling and tree construction

| Version | CV | $R^2$ | RMSE | MAE |
|---|---|---|---|---|
| Original R | 10-Fold | 0.783 | 0.553 | 0.363 |
| Rewritten Python | 10-Fold | 0.785 | 0.551 | 0.364 |
| Original R | LOSO | 0.562 | 0.786 | 0.59 |
| Rewritten Python | LOSO | 0.556 | 0.791 | 0.581 |

pipeline, RF model with MOE descriptor. One of the solutes in the dataset, gsk-Q, is an ion with no counterion specified. It was therefore omitted from openCOSMO calculations. Meanwhile, ORCA COSMO calculations for one of the solutes, 3-((6-*O*-(6-deoxy-α-L-mannopyranosyl)-β-D-glucopyranosyl)oxy)2-(3,4-dihydroxyphenol)-5,7-dihydroxy-4*H*-1-benzopyran4-one, failed to converge with the openCOSMO-RS default functional and basis set, BP86/def2-TZVPD, so the smaller basis set def2-TZVP was used instead. ORCA calculations for a second solute, iodopropynyl butylcarbamate, failed to converge using either basis set, and so this solute was omitted. For the gsk-S systems, the openCOSMO-RS iterative solubility calculations did not converge reliably, leading to unphysical solubility values; these systems were therefore excluded from the dataset. Meanwhile, the RDKit software was unable to correctly parse the SMILES string of a final solute, 311-03-5, during the openCOSMO-RS conformer workflow, so this solute was also omitted from the openCOSMO-RS dataset. In total, four solute molecules were removed. All of their solute–solvent combinations were omitted from the openCOSMO-RS calculations, which resulted in the loss of 27 data points compared to the COSMOtherm dataset. To ensure a fair comparison, we restricted the analysis to the subset of entries with successful openCOSMO predictions, and filtered the COSMOtherm results to the same subset (Table 2).

The standalone performance of openCOSMO was poor ($R^2 = -0.098$, RMSE > 1), indicating large deviations from the experimental solubility values. In contrast, COSMOtherm showed stronger predictive accuracy in the same dataset ($R^2 = 0.314$), although this still reflects limited standalone reliability. The openCOSMO-RS software is newer than the established COSMOtherm and is still under development. Although the latest parameterisation[72] provides a less accurate performance than COSMOtherm, its performance is improving. The solute molecules in this dataset for which openCOSMO-RS displays the poorest agreement with experiment tend to be larger molecules featuring more complex ring systems than those for which the openCOSMO-RS agreement is the best. They also contain a wider variety of different polar and hydrogen-bonding groups, and are more likely to contain less common atoms, like S or P, which featured less heavily in the latest openCOSMO-RS parameterisation set than C, O, and N. When combined with machine learning models, both COSMO sources contributed to improved hybrid predictions. In particular, models using openCOSMO features still benefit from error correction *via* ML, narrowing the performance gap with COSMOtherm-based hybrids. Default settings are used for the RF model. The COSMO-RS–predicted log solubility was incorporated as an auxiliary numerical feature, providing a physically grounded, quantum-chemistry-derived summary of the underlying solvation thermodynamics for each solute–solvent pair. By combining this physics-based estimate with detailed structure-encoded descriptors, the hybrid models are able to learn the remaining structure–property relationships that COSMO-RS does not capture on its own, thereby systematically improving upon the purely physics-based baseline. These findings highlight the potential of openCOSMO as a lower-cost, open-access alternative for hybrid solubility prediction, especially when high-fidelity quantum chemical tools are not available. While its standalone accuracy remains limited, integration with ML enables correction of systematic errors, particularly for polar or strongly hydrogen-bonding systems where COSMO-based models might struggle, or areas of chemical space where COSMO-based models are less reliable. Accordingly, the relative improvement obtained by including openCOSMO–RS features is more important than its absolute standalone performance. Given its stronger performance and broader data coverage, COSMOtherm is selected as the hybrid COSMO source for subsequent analyses and discussion.

Table 2 Standalone performance of openCOSMO and COSMOtherm solubility predictions (log(g/100 g)) against experimental values, evaluated on a reduced dataset where openCOSMO results were available

| COSMO model | $R^2$ | RMSE | MAE |
|---|---|---|---|
| OpenCOSMO | −0.098 | 1.243 | 0.940 |
| COSMOtherm | 0.314 | 0.983 | 0.707 |

**3.1.3 Benchmarking with hyperparameter optimisation.** We applied hyperparameter optimisation using RandomizedSearchCV (see Section 2.3.1) to assess its impact on hybrid model performance. Three algorithms, RF, XGB, and SVM were trained using MOE descriptors together with COSMOtherm solubilities as hybrid features, and evaluated under both 10-fold and LOSO CV (Fig. 2 and 3). For 10-fold CV, all models were trained and tested using the fixed random seed to ensure that the fold partitions were identical across (i) the tuned models, (ii) the untuned baselines, and (iii) all 10-fold Y-scrambling experiments. This guarantees that the performance differences reported here arise solely from model behaviour and not from variations in fold composition. The LOSO protocol, by construction, is deterministic and therefore required no seed control.

For RF, tuning produced only minor and mixed effects. Under LOSO, performance shifted marginally in the favourable direction (*e.g.*, $R^2$ increased from 0.5578 to 0.5582, although MAE rose from 0.581 to 0.593), whereas under 10-fold CV the tuned model performed slightly worse (*e.g.*, $R^2$ decreased from 0.777 to 0.769). These changes are numerically small and consistent with the well-known stability of RF, which typically operates close to its default optimum. Because tuning alters tree depth and feature-sampling behaviour, modest fluctuations in performance across validation schemes are expected. For

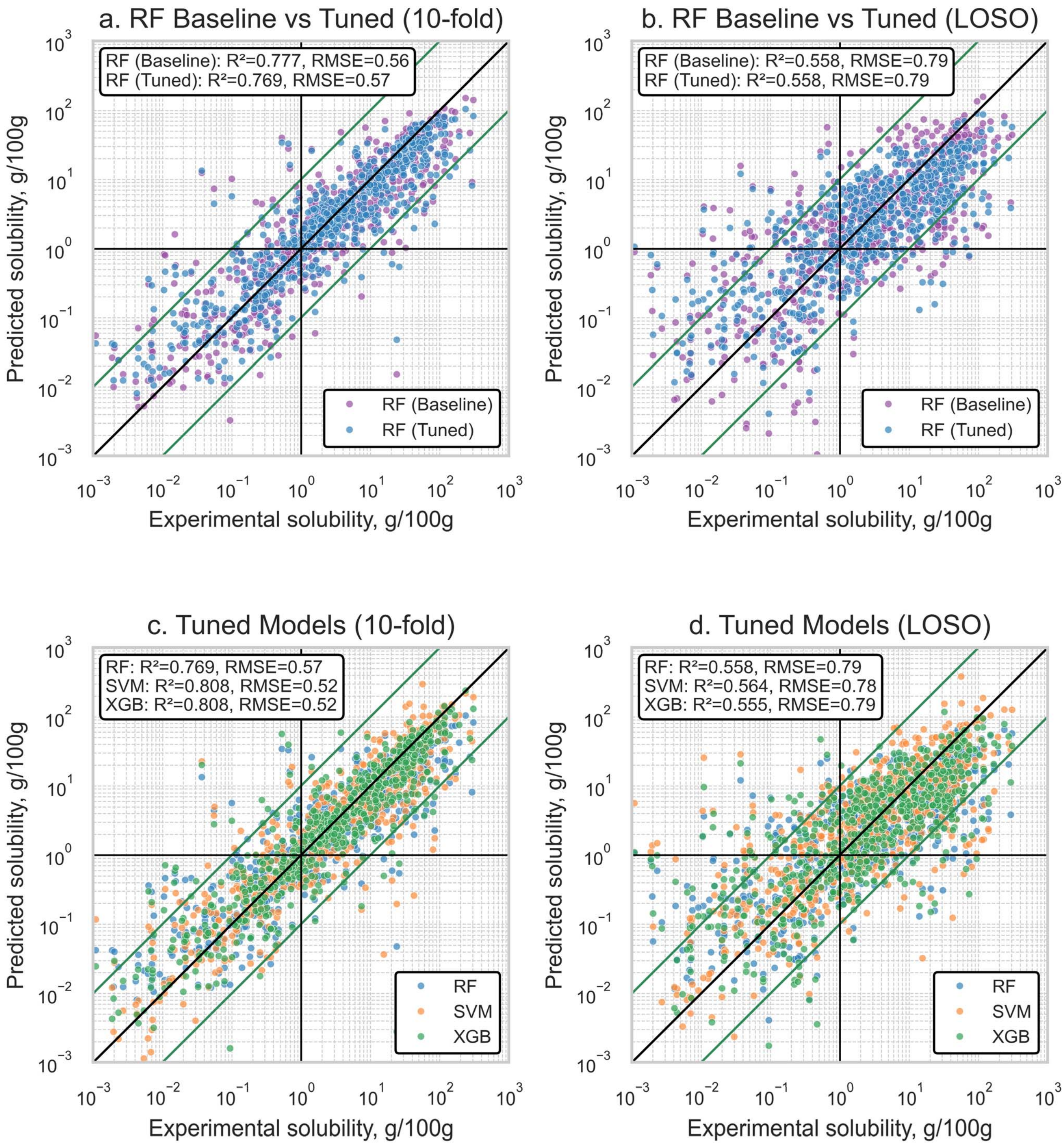

Fig. 2 Model performance comparison for solubility prediction using RF, XGBoost, and SVM trained on pre-processed MOE descriptors, evaluated under two CV schemes: 10-fold (left panels) and LOSO (right panels). (a–d) Parity plots comparing predicted and experimental solubilities (log scale), with diagonal lines indicating ±1 log unit error bounds.

consistency with the tuned XGB and SVM baselines, we therefore retain the tuned RF configuration in subsequent comparisons.

In contrast, both XGB and SVM exhibited clear improvements following tuning. For XGB, all metrics improved under both CV strategies (10-fold $R^2$ increased from 0.781 to 0.808; LOSO $R^2$ from 0.506 to 0.555). SVM showed the largest gains overall, particularly under LOSO, where $R^2$ increased from 0.492 to 0.564 and RMSE decreased from 0.846 to 0.784, yielding the strongest out-of-distribution performance among all tuned models. A notable observation is that SVM outperforms XGB under the LOSO protocol, even though their 10-fold CV performance is nearly identical: both models achieve comparable $R^2$ and RMSE values, with XGB showing only a slightly lower MAE.

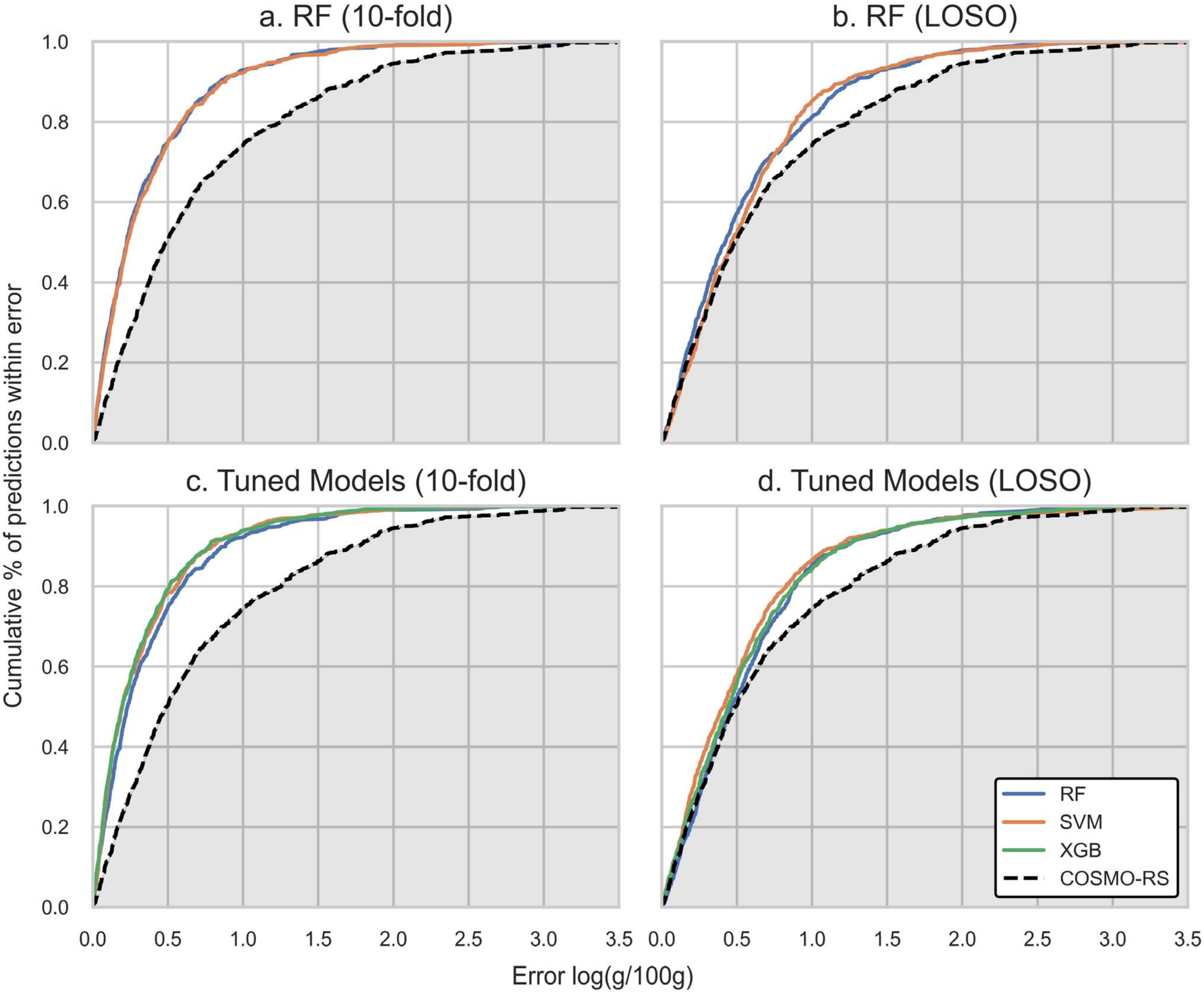

Fig. 3 Model performance comparison for solubility prediction using RF, XGBoost, and SVM across two CV schemes: 10-fold (left panels) and LOSO (right panels). (a) RF model performance under 10-fold CV. (b) RF model performance under LOSO CV. (c) Performance of RF, XGBoost, and SVM models under 10-fold CV. (d) Performance of RF, XGBoost, and SVM models under LOSO. All panels show cumulative error plots, reporting the proportion of predictions within a given absolute error. COSMO-RS predictions are included as a baseline for comparison.

This behaviour is well understood in high-dimensional chemical datasets where the number of available samples is limited relative to the descriptor space. XGB relies on iterative, boosted tree expansions that exploit correlations present within the training folds; when the held-out solute is structurally dissimilar to the training set, these correlations can become unreliable, leading to reduced stability and higher variance in the predictions. In contrast, the SVM with an RBF kernel imposes a stronger inductive bias and heavier regularisation, and the model capacity is controlled primarily by the kernel bandwidth and the regularisation parameter. As a result, SVMs tend to produce smoother decision boundaries and are less sensitive to small fold-to-fold fluctuations in descriptor–property correlations. Under LOSO, where the task explicitly tests out-of-distribution generalisation across different solutes, this regularisation leads to improved robustness, yielding lower RMSE and higher $R^2$ than XGB. Thus, the superior LOSO performance of the tuned SVM reflects its more conservative extrapolative behaviour in hybrid descriptor spaces.

Overall, these trends demonstrate that XGB and SVM benefit substantially from hyperparameter optimisation due to their

Table 3 Performance of RF models trained with different COSMO feature inputs under 10-fold and LOSO CV, without hyperparameter tuning

| COSMO feature | CV | $R^2$ | RMSE | MAE |
|---|---|---|---|---|
| None | 10-Fold | 0.653 | 0.699 | 0.450 |
| | LOSO | 0.349 | 0.958 | 0.699 |
| OpenCOSMO | 10-Fold | 0.747 | 0.598 | 0.387 |
| | LOSO | 0.510 | 0.831 | 0.600 |
| COSMOtherm | 10-Fold | 0.777 | 0.561 | 0.364 |
| | LOSO | 0.549 | 0.797 | 0.581 |

Table 4 Model performance ($R^2$, RMSE, MAE) for RF, XGB, and SVM using MOE descriptors, with and without hyperparameter tuning under LOSO and 10-fold CV. (Hyper P: hyperparameters)

| | | $R^2$ | | RMSE | | MAE | |
|---|---|---|---|---|---|---|---|
| Model | Hyper P | LOSO | 10-Fold | LOSO | 10-Fold | LOSO | 10-Fold |
| RF | Default | 0.558 | 0.777 | 0.789 | 0.561 | 0.581 | 0.369 |
| | Tuned | 0.558 | 0.769 | 0.789 | 0.571 | 0.593 | 0.378 |
| XGB | Default | 0.506 | 0.781 | 0.835 | 0.556 | 0.627 | 0.358 |
| | Tuned | 0.555 | 0.808 | 0.792 | 0.520 | 0.575 | 0.332 |
| SVM | Default | 0.492 | 0.707 | 0.846 | 0.643 | 0.615 | 0.432 |
| | Tuned | 0.564 | 0.808 | 0.784 | 0.520 | 0.550 | 0.339 |

sensitivity to regularisation strength and margin or learning-rate parameters, whereas RF is already strongly regularised by design, remains close to its default optimum. Full results are provided in Table 4.

As mentioned in Section 3.1.2, four solutes were removed, eliminating 27 associated solute–solvent pairs from the openCOSMO-RS calculations. For a fair comparison, both openCOSMO and COSMOtherm predictions were restricted to the subset of entries for which openCOSMO results were available. Therefore, the hybrid RF results without tuning in Fig. 2 and Table 4 will be different from what has been shown in Table 3.

### 3.1.4 Descriptor comparison across models.

To compare model generalisation across descriptor types, we evaluated tuned RF, XGB, and SVM models using MOE, RDKit descriptors, Mordred descriptors, and Morgan Fingerprints. All setups included hybrid COSMO-RS predictions and were evaluated using both 10-fold and LOSO validation.

Performance values reported in Table 5 correspond to the mean ± standard deviation obtained from $N = 20$ replicated 10-fold CV runs, each performed using a distinct random seed. This replicated protocol was applied to every model–descriptor combination to provide a fair and statistically robust comparison. In contrast, LOSO-CV is deterministic with respect to data partitioning and therefore yields a single, seed-independent result. The method details are illustrated in S3-SI, and the observed variability across the 20 seeds was consistently small (RMSE, MAE, and $R^2$ differing only in the second decimal place), confirming that hybrid improvements are stable and not driven by stochastic variation in training or fold assignment.

Notably, the XGBoost model using MOE descriptors achieved the highest overall predictive performance under 10-fold CV ($R^2 = 0.801 \pm 0.010$, RMSE $= 0.517 \pm 0.012$, MAE $= 0.337 \pm 0.005$; Table 5), with the SVM–MOE model performing comparably ($R^2 = 0.801 \pm 0.007$, RMSE $= 0.517 \pm 0.010$, MAE $= 0.339 \pm 0.005$). These results highlight the strong representational power of curated MOE descriptors, particularly when combined with regularised algorithms such as XGBoost and SVM.

Across all models and descriptor types, performance declined under LOSO compared to 10-fold CV, consistent with the increased difficulty of predicting solubility for unseen solutes. The largest reductions in $R^2$ were observed for SVM using RDKit descriptors (from 0.782 ± 0.011 to 0.466; $\Delta R^2 = 0.316$) and SVM using Mordred descriptors (from 0.726 ± 0.012 to 0.415; $\Delta R^2 = 0.311$), indicating that kernel methods are particularly sensitive to descriptor redundancy and high-dimensional noise when forced to extrapolate to new solute structures. RF performed most robustly under LOSO, achieving $R^2 = 0.568$ with RDKit and $R^2 = 0.543$ with Morgan Fingerprints. This suggests that RF's ensemble averaging helps stabilise predictions when solute diversity increases. By contrast, RF was less effective with lower-dimensional inputs such as MOE and RDKit under 10-fold CV, where XGBoost and SVM achieved markedly higher accuracy.

XGBoost showed the strongest resilience with Mordred descriptors ($R^2 = 0.613$ under LOSO; 0.791 ± 0.007 under 10-fold), demonstrating that gradient boosting can effectively extract signal from large, heterogeneous descriptor sets. Under 10-fold CV, XGBoost also achieved the highest overall accuracy with all types of descriptors and fingerprints, with particularly strong performance for MOE ($R^2 = 0.801 \pm 0.010$).

The higher performance of XGBoost is also consistent with the structure of the dataset. The descriptor sets used here are high-dimensional and contain correlated features, and XGBoost's sequential boosting, residual fitting, and built-in regularisation allow it to exploit such feature spaces more effectively than RF. This is particularly advantageous in the hybrid setting, where the COSMO-RS solubility provides

Table 5 Cross-descriptor benchmarking of tuned RF, XGB, and SVM models under LOSO and replicated 10-fold CV (20 seeds; mean ± std)

| | | $R^2$ | | RMSE | | MAE | |
|---|---|---|---|---|---|---|---|
| Model | Descriptors | LOSO | 10-Fold | LOSO | 10-Fold | LOSO | 10-Fold |
| RF | MOE | 0.558 | 0.765 ± 0.006 | 0.789 | 0.565 ± 0.007 | 0.593 | 0.381 ± 0.003 |
| | RDKit | 0.568 | 0.764 ± 0.006 | 0.780 | 0.567 ± 0.007 | 0.577 | 0.384 ± 0.002 |
| | Mordred | 0.567 | 0.756 ± 0.006 | 0.781 | 0.576 ± 0.007 | 0.584 | 0.388 ± 0.003 |
| | Morgan | 0.543 | 0.757 ± 0.004 | 0.803 | 0.574 ± 0.005 | 0.576 | 0.394 ± 0.003 |
| XGBoost | MOE | 0.555 | 0.801 ± 0.010 | 0.792 | 0.517 ± 0.012 | 0.575 | 0.337 ± 0.005 |
| | RDKit | 0.567 | 0.794 ± 0.011 | 0.781 | 0.527 ± 0.014 | 0.567 | 0.343 ± 0.006 |
| | Mordred | 0.613 | 0.791 ± 0.007 | 0.738 | 0.529 ± 0.009 | 0.538 | 0.350 ± 0.005 |
| | Morgan | 0.511 | 0.775 ± 0.009 | 0.831 | 0.547 ± 0.010 | 0.595 | 0.365 ± 0.005 |
| SVM | MOE | 0.564 | 0.801 ± 0.007 | 0.784 | 0.517 ± 0.010 | 0.550 | 0.339 ± 0.005 |
| | RDKit | 0.466 | 0.782 ± 0.011 | 0.868 | 0.541 ± 0.014 | 0.628 | 0.348 ± 0.007 |
| | Mordred | 0.415 | 0.726 ± 0.012 | 0.908 | 0.607 ± 0.014 | 0.661 | 0.406 ± 0.007 |
| | Morgan | 0.504 | 0.605 ± 0.017 | 0.836 | 0.729 ± 0.015 | 0.595 | 0.481 ± 0.007 |

a physics-based baseline and the ML component must learn the remaining structure-dependent deviations. RF, which does not explicitly model residuals, is less able to capture these fine-grained corrections.

Among descriptor sets, MOE provided consistently strong performance for both XGBoost ($R^2 = 0.801 \pm 0.010$ in 10-fold; 0.555 in LOSO) and SVM ($R^2 = 0.801 \pm 0.007$ in 10-fold; 0.564 in LOSO), reflecting the curated, property-focused nature of MOE features. It is noteworthy that, even with the same descriptor set, the tuned RF–MOE model reached a slightly lower 10-fold performance ($R^2 = 0.765 \pm 0.006$), whereas both XGB–MOE and SVM–MOE achieved the highest accuracies observed in this study. This improvement does not arise from differences in descriptor quality, which is held fixed across models, but from differences in model capacity and how each algorithm extracts structure–property relationships from MOE features. RF relies on axis-aligned, shallow tree partitions, which stabilise predictions but limit its ability to model smooth nonlinear interactions among correlated MOE descriptors. In contrast, XGBoost leverages sequential residual fitting to capture higher-order nonlinearities, while the RBF–SVM constructs a continuous similarity landscape that can more fully exploit the physicochemical signal encoded in the MOE feature space. As a result, both XGB and SVM are able to extract more predictive information from MOE features than RF, explaining the systematic improvement from $R^2 = 0.765 \pm 0.006$ (RF–MOE, 10-fold CV) to $\approx$ 0.801 (XGB–MOE and SVM–MOE) under 10-fold CV.

Under LOSO, SVM–MOE markedly higher than SVM trained on RDKit ($R^2 = 0.466$), Mordred descriptors ($R^2 = 0.415$), or Morgan Fingerprints ($R^2 = 0.504$). This pattern reflects differences in descriptor design. MOE descriptors form a compact, property-focused feature set with limited redundancy and well-defined chemical meaning, enabling the RBF kernel to construct smooth similarity functions without overfitting to spurious dimensions. In contrast, the high-dimensional, highly correlated Mordred and Morgan representations create a much more irregular kernel landscape, making the SVM sensitive to descriptor noise and leading to pronounced performance degradation under LOSO. RDKit descriptors lie between these extremes but still lack the curated, physicochemical structure encoded in MOE. Consequently, the SVM benefits most from the balanced dimensionality and targeted chemical relevance of MOE features, which support stable extrapolation to unseen solutes. RDKit descriptors occupied an intermediate position, performing competitively with RF under LOSO and outperforming Morgan and Mordred descriptors under 10-fold CV. Morgan Fingerprints performed best with XGBoost ($R^2 = 0.775 \pm 0.009$ in 10-fold) but lagged behind MOE and RDKit overall, while Mordred descriptors exhibited the significant LOSO degradation across all models due to their high redundancy.

Fig. 4 compares prediction performance ($R^2$) across all descriptor-model-CV combinations, using mean $R^2$ over replicated 10-fold CV runs and single-run $R^2$ for deterministic LOSO. For both RDKit descriptor and MOE, their relatively low redundancy and high interpretability make them well-suited to both tree-based and kernel-based models.

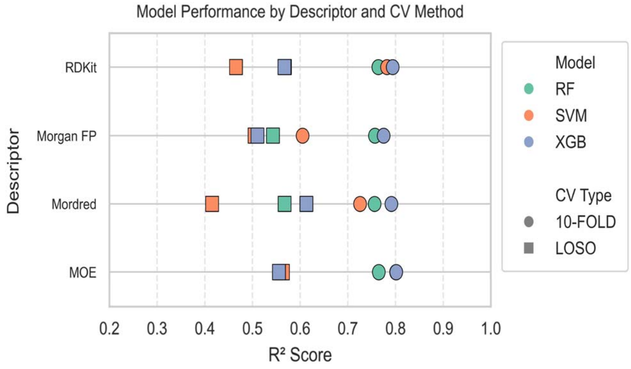

Fig. 4 Prediction performance ($R^2$) of RF, XGB, and SVM models across RDKit, Morgan Fingerprint, Mordred and MOE descriptors. Marker colour indicates model type; marker shape distinguishes CV scheme (10-fold points show the mean over replicated runs, while LOSO points correspond to a single deterministic split).

A portion of the residual unexplained variance may arise from experimental heterogeneity in the solid-state form of solutes, including differences in polymorphic form, amorphous content, sample history (*e.g.* cooling or quenching rates), or impurities. Such factors influence the Gibbs free energy of fusion but are seldom reported in the literature sources from which the dataset is constructed. As these effects cannot be represented by structural descriptors or COSMO-RS features, they set an intrinsic upper bound on achievable predictive accuracy.

Under LOSO, RF outperforms XGB and SVM for most descriptor sets, whereas under 10-fold CV it is often surpassed by XGB and SVM. This divergence reflects the validation objective: LOSO enforces generalisation to "unseen solutes" (grouped CV), while 10-fold permits the same solute to appear in both training and test folds. RF's bagging and feature subsampling reduce variance, yielding stable predictions that transfer better across solute identity, whereas the higher capacity of XGB and SVM captures solute-specific interactions that improve 10-fold scores but do not necessarily translate to LOSO. Therefore LOSO should be prioritised for model selection when the deployment target is new solutes.

**3.1.5 External validation.** To further assess model generalisability beyond the training chemical space, external validation was performed using solubility data for organic compounds collected from BigSolDB 2.0,[67] an open-source solubility database for organic compounds across diverse solvents and temperatures. Six solutes absent from the original training set were selected, yielding 63 solute–solvent pairs with solubility measurements reported at 298.15 K, consistent with the conditions used in the present study.

The ML models, trained exclusively on the original dataset, were frozen and directly applied to the external dataset. For external prediction, we used the ensemble of models obtained from the 10-fold CV: each of the ten fold-specific models was applied to the external systems, and their predictions were averaged to obtain a single estimate for each solute–solvent pair. External predictions were generated using models trained

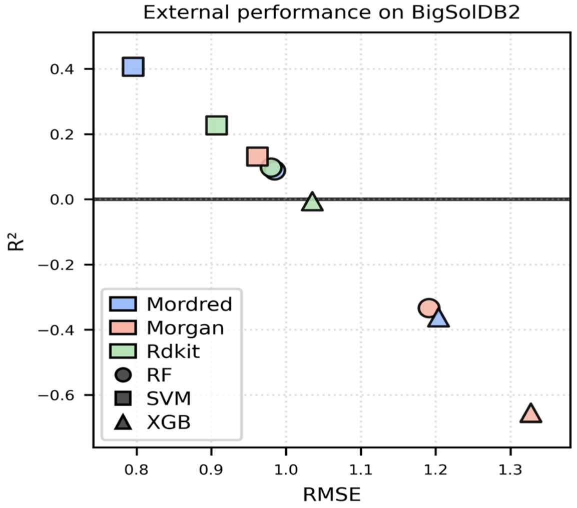

Fig. 5 External performance of tuned RF, XGB, and SVM models on the BigSolDB 2.0 subset. Points show RMSE *versus* $R^2$ for each combination of descriptor set (colour) and model type (marker).

with a fixed random seed to ensure a deterministic and reproducible evaluation. This approach makes full use of the available training data while providing a statistically stable prediction for unseen compounds. LOSO was used solely as an internal validation scheme and was not considered for external prediction. Descriptor calculation and preprocessing steps were defined entirely from the training data and subsequently fixed, so that external compounds were projected into the same descriptor space. COSMOtherm-predicted solubilities for the external systems were generated independently and used solely as input features in the hybrid models.

For the external validation, we focused on descriptor sets that are openly implementable or broadly accessible (Morgan Fingerprints, Mordred, and RDKit descriptors). Although MOE descriptors demonstrated competitive performance under internal CV, their use relies on commercial software that cannot be readily redistributed, and they were therefore omitted from the external benchmark in favour of more reproducible options.

Fig. 5 summarises the external performance of tuned RF, XGB, and SVM models with Morgan Fingerprints, Mordred, and RDKit descriptors. Across all descriptor spaces, SVM models show the most robust generalisation (highest $R^2$ and lowest RMSE), followed by RF. In contrast, XGBoost exhibits a pronounced degradation in external performance, yielding negative $R^2$. This behaviour is consistent with the higher sensitivity of boosted tree ensembles to covariate shift, arising from changes in the distribution of input descriptors between the training and external datasets, as well as to sparse descriptor activation, whereas SVM and RF models show improved robustness when extrapolating beyond the training domain.

## 3.2 Interpreting molecular features and fingerprints

To elucidate the physicochemical determinants of solubility, we analysed the relative importance of both molecular descriptors (RDKit, MOE, Mordred) and structural fingerprints across models. Feature contributions were quantified using SHAP values, allowing us to connect model predictions back to interpretable chemical features. This dual perspective captures both engineered descriptors that summarise known molecular properties and fingerprint bits that encode specific substructural motifs. By comparing feature rankings across high-performance models, we aim to identify common drivers of solubility, highlight algorithm-specific sensitivities, and assess whether machine learning rediscovered known heuristics.

**3.2.1 Descriptor–property relationships and solvent–solute space alignment.** Principal Component Analysis (PCA) was used to assess how well each descriptor set (RDKit, MOE, Mordred) encodes solubility-relevant chemical variation. The first two principal components (PC1 and PC2) of each descriptor matrix were analysed, and coloured projections by experimental solubility, hydrophobicity ($\log P$-like), molecular weight ($\log_{10}$(MW)), and polarity (TPSA) are provided in the SI (Section 4). These confirm that PC1 consistently encodes a multivariate gradient of increasing molecular weight, lipophilicity, and polarity. In all cases, solubility increases toward the negative end of PC1, aligning with smaller, less lipophilic, and less polar solutes. While the relatively high solubility of smaller solutes is expected, the apparent favourability of both less lipophilic and less polar solutes may at first appear somewhat counterintuitive. Typically, a more polar solute is less lipophilic and therefore soluble in different solvents than a lipophilic solute. However here, as discussed in Section 1, this set of solvents contains a range of different chemical features. In line with this, the PCA observations suggest that solutes with intermediate properties are most likely to display high solubility across the chemically diverse solvent set.

Importantly, the apparent dominance of solute features compared to solvent features in PCA cannot be fully understood without considering the corresponding solvent space. If the solvents are averaged or biased toward one end of the polarity or lipophilicity spectrum, the apparent dominance of solute features in PCA becomes expected, consistent with the classical principle of “like-dissolves-like”. To assess this directly, we compared solute and solvent rankings for polarity (TPSA) and lipophilicity ($\log P$) generated from MOE descriptors. Spearman's rank correlation coefficient ($\rho$) was calculated to assess the correspondence between solute and solvent properties. Unlike Pearson's linear correlation coefficient ($r$), which is sensitive to absolute scales and assumes linearity, Spearman's $\rho$ evaluates the degree of monotonic alignment between ranked variables. This rank-based approach mitigates the impact of differing property ranges and provides a more appropriate measure of whether the solute and solvent spaces are systematically aligned.

The resulting rank–rank plots (Fig. 6 and 7) showed only weak correlations (Spearman $\rho$ = 0.10, 0.14) with horizontal banding that reflects a narrower property range for solvents than for solutes. To further explore the solvent contribution, PCA projections coloured by solvent properties ($\log P$, MW, and TPSA generated from MOE) are shown in Fig. 8. These reveal that the solvent space occupies a narrower range of lipophilicity

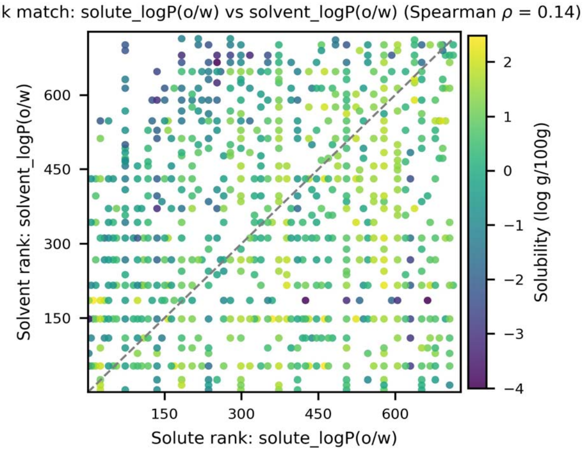

Fig. 6 Rank–rank scatter comparing solute and solvent lipophilicity ($\log P$(o/w)). The negligible correlation (Spearman $\rho = 0.14$) indicates that solute and solvent lipophilicity are only loosely aligned across the dataset. The horizontal banding reflects the limited diversity of solvent $\log P$ values relative to solutes, confirming that solute lipophilicity spans a broader chemical space than solvent lipophilicity.

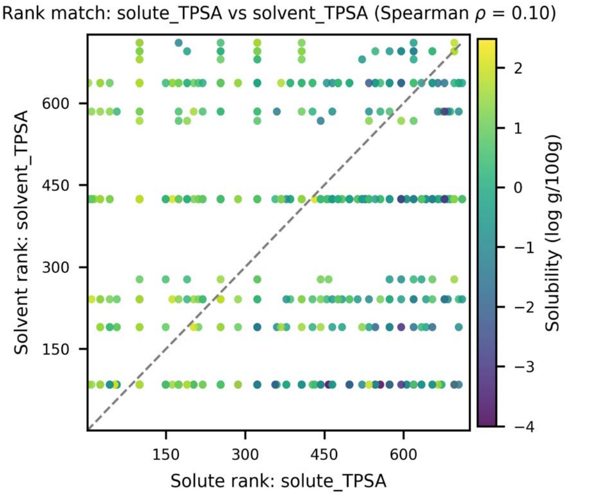

Fig. 7 Rank–rank scatter comparing solute and solvent polarity (TPSA). The negligible correlation (Spearman $\rho = 0.10$) and strong horizontal banding highlight that solvents occupy a compressed polarity range compared with solutes. This explains why PCA trends are dominated by solute polarity, as solvent variation provides less discriminative power.

and polarity compared with the corresponding solute properties. This compression explains the horizontal banding observed in rank–rank comparisons and why solute descriptors dominate the PCA trends. In other words, while solvents introduce some modulation, the dataset is primarily defined by solute chemical diversity.

As illustrated in SI-Section 4, although PCA revealed broadly similar chemical organisation across RDKit, MOE, and Mordred descriptor spaces, predictive performance varied between models. XGBoost performed best with RDKit (under 10-fold) and Mordred, while SVM performed better with MOE (under LOSO). These differences likely arise from how each algorithm interacts with descriptor characteristics such as dimensionality, collinearity, and feature scaling, rather than differences in the underlying chemical information. For instance, the higher dimensionality and redundancy of Mordred may favour ensemble methods like XGBoost, which can down-weight irrelevant features, whereas the more compact and curated MOE descriptors align better with kernel-based methods such as SVM. This highlights that comparable PCA structures do not necessarily translate into uniform model performance, underscoring the importance of jointly optimising both descriptor representation and modelling approach.

### 3.2.2 Complementary descriptor coverage by XGBoost and SVM

*3.2.2.1 Descriptor-level comparison between XGBoost and SVM.* To elucidate the physicochemical features governing solubility, we analysed the top 20 MOE descriptors ranked by SHAP values from tuned XGBoost and SVM models under 10-fold CV. These two models showed consistently strong and g-eneralisable performance across both CV strategies, making them representative models for interpreting descriptor importance. Although both models captured overlapping chemical themes, differences in descriptor prioritisation revealed their distinct inductive biases and sensitivity to specific structural patterns.

As shown in SI-S10, Fig. S6 and S7, descriptor names are colour-coded by molecular role: solute-derived descriptors (blue) and solvent-derived descriptors (green). Each row represents a solute, ordered by decreasing averaged experimental solubility across available solvents (top to bottom). Columns indicate individual MOE descriptors, selected based on their global SHAP importance. Cell colours represent signed SHAP values, with red indicating a positive influence on predicted solubility and blue indicating a negative influence. Grey cells correspond to near-zero contributions ($|\mathrm{SHAP}| < 1\times10^{-30}$). While the top 20 includes both solute and solvent descriptors along with the COSMOtherm feature, the majority originate from the solute. This is not unexpected, since solubility depends on the interactions between the solvent and the solute, and the dataset contains more different solute molecules than solvent molecules. Although there is variation within the solvent set, including in hydrophilicity and lipophilicity, the solvents may be considered to broadly fall within the same category of organic solvents with heteroatom functionality. Meanwhile, the solutes in the dataset cover a wider range of chemical space, as seen in Section 3.2.1. In particular, they include more molecules that are more hydrophilic, with the lowest $\log P$(o/w) value for a solute being −3.5, although the highest is 7.0, close to the highest value among the solvents. Variations in solvent chemical environment in the dataset are likely to have a smaller impact on solubility than the larger variations between solutes. Descriptors are ranked by their mean absolute SHAP values across all solutes, reflecting their overall contribution magnitude to model predictions regardless of direction. The details of each top-ranked MOE descriptor are listed in SI Sections S6 and S7.

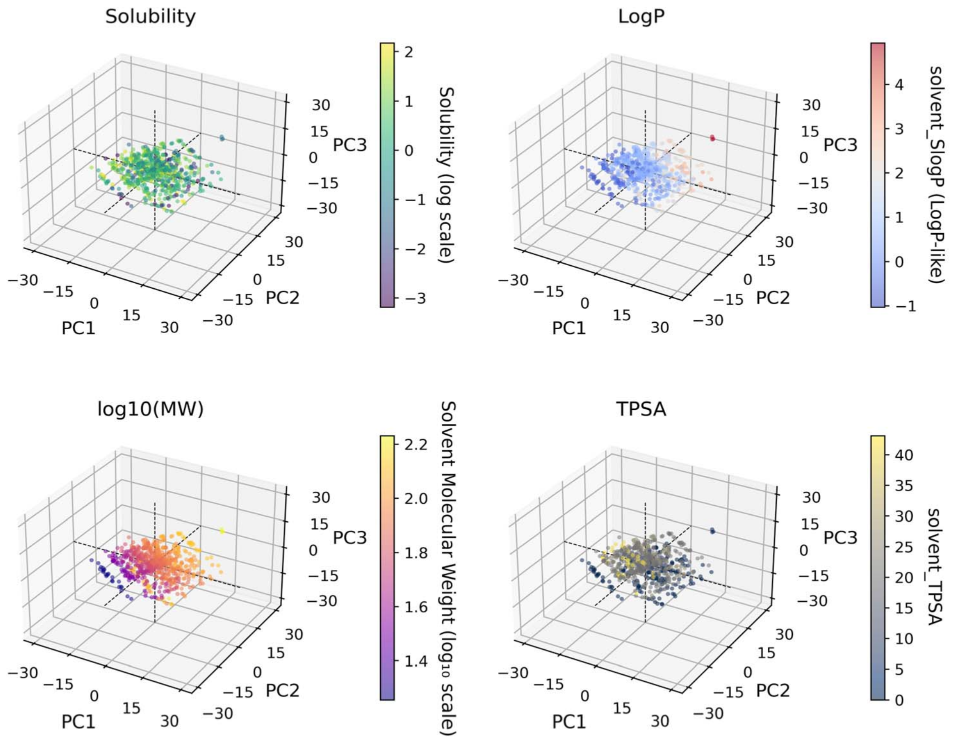

Fig. 8 PCA projections of the dataset coloured by experimental solubility, log scale, solvent hydrophobicity, log *P*(o/w), solvent molecular weight, $\log_{10}$(MW), and solvent polarity, TPSA. These plots illustrate the distribution of solvent chemical space relative to solubility. The narrow ranges of solvent log *P*(o/w) and TPSA values produce compressed colour gradients and horizontal banding, confirming that solvent variation is less extensive than solute variation. In contrast, solubility and molecular weight span broader ranges, indicating that solubility prediction is primarily modulated by solute properties, with only limited modulation by solvent descriptors.

*3.2.2.2 XGBoost: polarity, hydrogen bonding, and solvation geometry.* XGBoost highlighted descriptors spanning molecular geometry, charge distribution, and polar surface exposure. For example, solute_a_ICM and solvent_a_ICM quantify atom information content and internal coordinate moments, reflecting molecular compactness and topological structure. Charge-related descriptors such as solute_PEOE_RPC-(relative negative partial charge) and solvent_PEOE_VSA_FNEG (fractional negative van der Waals surface area) capture the strength and extent of electron-rich regions, while solute_TPSA and solute_vsa_pol measure accessible polar surface area. Additional contributors, including solute_h_emd (EHT donor strength, sum) and solvent_h_emd_C (donor strength restricted to carbon atoms), highlight electrotopological states that encode the electronic environment of atoms in their bonding context, emphasising the role of electron delocalisation and substitution patterns in solubility.

SHAP analysis further identified classic drug-likeness descriptors log *P* (solute_h_log_pbo), H-bond donor and acceptor counts (solute_a_donscc), H-bond acceptor counts (Lipinski-style acceptor count: solute_lip_acc and general acceptor count: solute_a_acc), and topological polar surface area (TPSA, solute_TPSA) among the top contributors. This convergence demonstrates that the XGBoost model effectively rediscovered the same physicochemical drivers underlying Lipinski's Rule of 5 (ref. 73) and its related Veber criteria.[74] While these rules were originally formulated in the context of aqueous environments, their recurrence here reflects the general importance of polarity, hydrogen bonding, and size-related features across solvent systems. The specific impact of each descriptor is modulated by solvent polarity and lipophilicity, with our analysis focusing on general organic solvent systems.

Beyond these primary descriptors, several related features also appeared among the top-ranked contributors, including additional lipophilicity measures (solute_$S\log P$_VSA1, solute_GCUT_$S$ LOG $P$_1), estimated molecular aqueous solubility (solute_h_ema), polar surface metrics (vsa_pol, solute_PEOE_RPC-, solute_GCUT_PEOE_1, and alternative formulations of hydrogen-bonding capacity (total count of donor and acceptor atoms together: solute_a_donacc). The recurrence of multiple log *P*-, TPSA-, and H-bond–related descriptors underscores the robustness of hydrophobicity, polarity, and hydrogen bonding as central solubility determinants across descriptor classes.

Finally, although solute descriptors dominate overall, XGBoost also incorporates more solvent-specific features than SVM. This suggests a greater sensitivity of the tree-based model to subtle solvent differences, reinforcing that solubility is governed not only by intrinsic solute properties but also by the solute–solvent match within the studied chemical space.

*3.2.2.3 SVM: broader spectrum including reactivity and accessibility.* In contrast to XGBoost, the SVM model distributed importance more broadly across descriptors spanning hydrophobicity, polarity, electronic structure, and synthetic accessibility. Lipophilicity-related features were again prominent, including solute_$S\log P$_VSA6 and solute_$S\log P$_VSA1 (surface area contributions partitioned by $S\log P$ values), together with the classical octanol–water partition coefficient $\log P$, solute_log $P$(o/w), highlighting the contribution of both global hydrophobicity and localised lipophilic surface exposure. Charge-based descriptors such as solute_PEOE_RPC- and solute_PEOE_RPC+ (relative negative and positive partial charges), solute_PEOE_VSA+4 and solute_PEOE_VSA-4, solute_PEOE_VSA_FPPOS (fractional positive polar surface area), and solute_GCUT_PEOE_2 (graph-cut from PEOE charges) reflect detailed electron distribution patterns. Similarly, solute_SMR_VSA4 (van der Waals surface area weighted by molar refractivity) integrates both surface area and polarizability effects.

Topological and geometric measures also featured, including solute_a_ICM (atom information content), solute_balabanJ (Balaban connectivity index), solute_radius (molecular size), and solute_chiral_u (number of unconstrained chiral centres), together indicating that SVM is sensitive to molecular complexity, stereochemistry, and global shape. Beyond these physicochemical determinants, the model selected descriptors linked to reactivity and synthetic tractability, including solute_reactive, solute_rsynth, and solvent_rsynth. While not mechanistic solubility drivers *per se*, these properties correlate with size, functional group composition, and overall polarity, indirectly shaping solubility across solvents.

The broader distribution of influential descriptors suggests that SVM captures a more diffuse representation of solubility, spanning lipophilicity, electronic structure, geometry, and accessibility. This complements the sharper physicochemical emphasis of XGBoost and underscores the multifactorial nature of solubility in organic solvents.

*3.2.2.4 Comparative interpretation of model biases.* Several descriptors top-ranked by SVM overlapped with those emphasised by XGBoost, notably solute_PEOE_RPC-, solute_a_ICM, and solute_$S\log P$_VSA1, indicating consistent relevance of partial charges, 3D shape, and surface lipophilicity across models.

Taken together, the results and patterns in Fig. S6 and S7 suggest that XGBoost concentrated on a tighter set of polarity and H-bonding metrics, while SVM captured a broader, more diffuse spectrum of structural, electronic, and accessibility-related features. SHAP heatmaps revealed clear differences in how the two models attributed importance across solutes, consistent with their distinct learning paradigms. XGBoost, which leverages decision tree ensembles, produced smooth, continuous SHAP gradients, particularly for global descriptors such as solute_TPSA, solute_PEOE_RPC-, and solute/solvent_a_ICM. This reflects its ability to split the feature space using threshold-based decisions recursively and to aggregate weak learners across multiple feature partitions. The resulting heatmaps suggest that XGBoost captures multiscale structure–property relationships, where continuous features contribute to solubility predictions across a broad chemical space.

In contrast, SVM, trained with a Radial Basis Function (RBF) kernel, yielded discrete and clustered SHAP patterns, with contributions sharply concentrated on specific solutes for descriptors such as rsynth and reactive. These descriptors often encode binary or threshold-like properties, which align with the SVM's reliance on support vectors to define decision boundaries in a projected feature space. Because SHAP values for kernel models are computed *via* background sampling (*KernelExplainer*), they tend to reflect abrupt shifts in predicted output due to localised descriptor influence. The SVM heatmaps, therefore, highlight class-like behaviour, solutes with distinct physicochemical filters or synthetic characteristics, rather than broad, continuous trends.

Together, these findings suggest that XGBoost identifies solubility determinants driven by continuous physicochemical gradients, particularly polarity, charge distribution, and hydrogen-bonding capacity, while SVM highlights discrete structural or accessibility filters that partition solutes into distinct subgroups. This divergence reflects their underlying learning biases: XGBoost leverages recursive thresholding to capture fine-grained physicochemical variation, whereas SVM relies on support vectors to enforce boundary-driven classification in descriptor space. The convergence on common features such as lipophilicity and 3D shape, coupled with their complementary sensitivities, reinforces the robustness of our conclusions and underscores the value of combining tree-based and kernel methods to map a richer landscape of solubility-relevant features.

**3.2.3 Interpreting substructural patterns from fingerprints.** To complement descriptor-based insights, we analysed SHAP-ranked Morgan Fingerprints derived from the XGBoost model, which achieved the highest overall performance. Unlike predefined descriptors, these fingerprint bits encode algorithmically-derived molecular substructures based on atom connectivity, rather than predefined chemically-informed functional groups. By mapping key fingerprint bits back to chemical fragments, we identified interpretable substructures associated with solubility trends across the dataset. Fig. S8 and S9 in SI-S10 show the top-ranked fingerprint bits for solutes and solvents across SHAP values of solutes ordered by solubility. Rows represent solutes ordered by decreasing averaged experimental solubility across available solvents (top to bottom); columns show fingerprint bits, with the importance ranking from left to right. Cell values indicate SHAP contributions, with positive values promoting solubility and negative values reducing it.

Due to hashing and bit folding, the same Morgan Fingerprint bit can correspond to multiple distinct substructures across different molecules. In cases where RDKit could not recover a mapped atom environment for an active bit, we return to highlighting the single atom. S8 and S9 in SI illustrated the

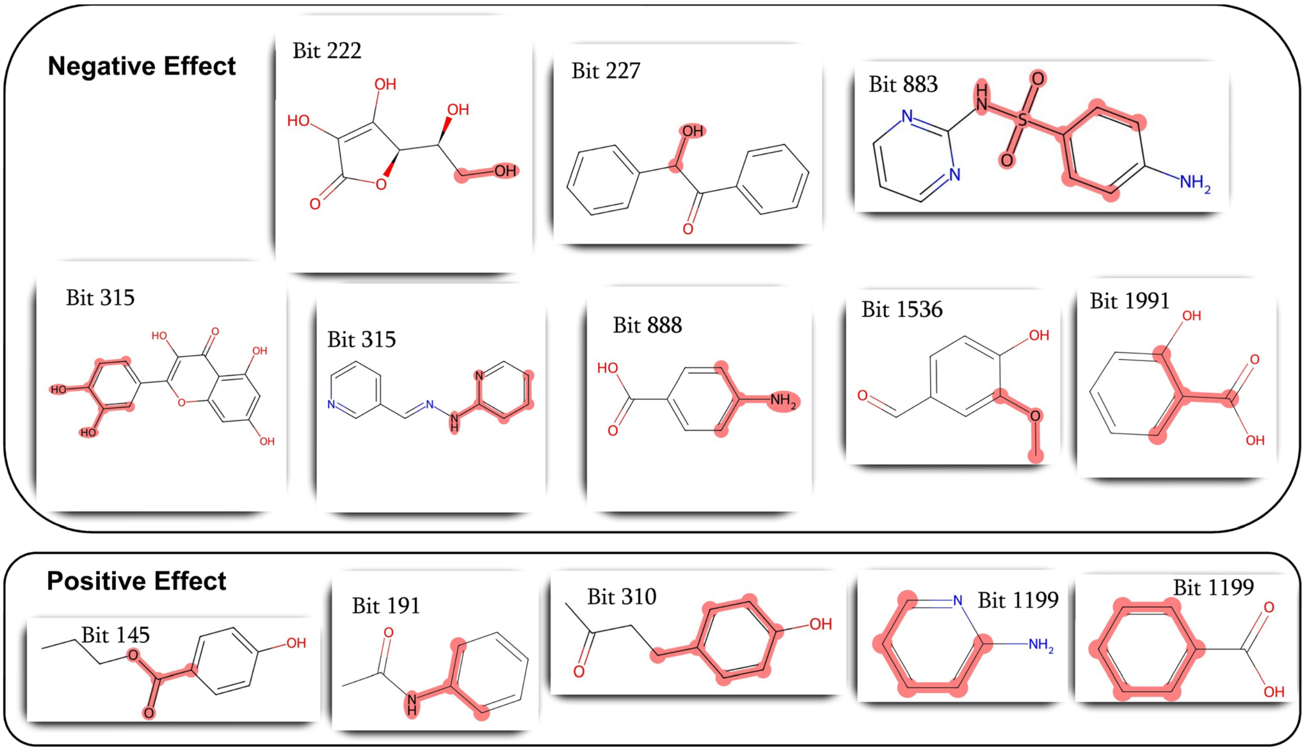

Fig. 9 Representative solute substructures corresponding to the top 20 SHAP-ranked Morgan Fingerprint bits that appeared in more than two different solutes. Substructures are highlighted in red and classified according to their negative or positive effects on solubility.

top 20 Morgan Fingerprint bits that are visualisable from the solutes and solvents individually.

Fig. 9 highlights recurring solute substructures among the top 20 SHAP-ranked bits, restricted to those appearing in more than two different solutes, thereby emphasising motifs with consistent solubility effects across the dataset. Among the top 20 SHAP-ranked Morgan Fingerprint bits for solutes, several showed consistent directional contributions across multiple molecules. Bits such as 1876, 114, 145, 191, 310, and 1199 are predominantly associated with increased predicted solubility, while the rest of the bits contribute negatively. Both the bits with positive contributions and those with negative contributions include a combination of organic sections and heteroatom-based functional groups. The positive bits feature more contribution from larger parts of aromatic rings than the negative bits. Aromatic groups without heteroatoms can contribute to solubility in solvents with significant organic character, suggesting the extended aromatic sections contribute to solubility in the less polar solvents among this dataset. Meanwhile, aromatic groups containing heteroatoms can contribute to solubility in polar solvents or solvents with polar groups. This environment-dependent behaviour may explain why aromatic groups feature heavily in general over the whole dataset.

For a subset of high-ranking bits (*e.g.* 378, 935, 114, 656, 1683), a valid atom substructure could not be recovered, likely due to bit collisions or unresolved hashing. In these cases, visualisation defaulted to the central atom. This highlights a key limitation of hashed fingerprints: while they capture predictive patterns, interpretability is constrained by the one-to-many mapping between bits and substructures. Despite this, SHAP analysis at the bit level offers a useful route for prioritising solubility-relevant features, particularly when combined with substructure visualisation.

Among the top SHAP-ranked solvent fingerprint bits, directional contributions varied across solutes, highlighting the context-dependent nature of solvation. Although both solute and solvent substructures are fixed for a given molecule, the influence of solvent fingerprints on solubility is inherently relational; their effect emerges from compatibility with specific solute features, rather than from intrinsic properties alone. For instance, the same solvent bit encoding an ether or halogen may enhance solubility for a polar solute by enabling favourable dipolar or hydrogen bonding interactions, but reduce solubility for a non-polar solute that cannot experience these interactions. Meanwhile, a substantial carbon chain or non-decorated aromatic group may enhance solubility for a non-polar solvent, but reduce it for a solvent which relies on the presence of polar bonds. This variability reflects the principle that solubility emerges from solute–solvent complementarity, rather than intrinsic solvent features in isolation. Although this limits direct attribution of solvent bits to fixed solubility effects, the overall SHAP patterns reveal solvent environments that are more or less compatible with the range of diverse solute classes present in the current dataset.

## 4 Conclusions

Predicting solubility remains a central challenge in chemical and pharmaceutical research because of the complex interplay of molecular structure, intermolecular interactions, and solvent environment. Classical thermodynamic approaches, such as COSMO-RS, provide a strong theoretical foundation by linking molecular structure to solubility behaviour through quantum chemical calculations and thermodynamic modelling. In contrast, descriptor-based QSPR/QSAR frameworks capture these structure–property relationships using molecular descriptors through machine learning methods, offering a more flexible way to learn patterns directly from data.

In this work, we develop an integrated pipeline that unifies descriptor generation, flexible model selection (RF, SVM, XGBoost), and hybridisation with COSMO-RS outputs, together with systematic CV strategies. By incorporating interpretable ML methods, notably SHAP-based feature attribution, the framework not only achieves competitive predictive performance but also reveals the molecular features most strongly governing solubility. The pipeline further supports multiple descriptor types (Morgan Fingerprints, Mordred, MOE, and RDKit descriptors), enabling a comprehensive and modular evaluation across data representations.

In particular, SHAP analysis revealed that the most influential features across the models corresponded to well-established physicochemical determinants of solubility. This alignment with Lipinski's Rule of Five demonstrates that the ML models effectively rediscover classical medicinal chemistry heuristics, while extending them into a broader solubility context. Crucially, the contribution of the present work is in showing that such relationships can be captured, quantified, and generalised at scale across hundreds of solute–solvent combinations. This systematic and data-driven validation provides confidence that intuitive chemical principles hold in diverse contexts, while highlighting where deviations may occur. Such convergence across descriptor types underscores the reliability of our framework and highlights the value of interpretable ML in providing chemically meaningful insights.

Looking ahead, this pipeline demonstrates how mechanistic insight and ML can be synergistically combined into a reproducible and extensible workflow. By coupling COSMO-RS physical descriptors with modern ML architectures and interpretable feature analysis, the framework establishes a robust template for the prediction of solubility with improved accuracy and transparency. Most importantly, the workflow enables scalable, reproducible screening of the solubility across large chemical spaces, supporting practical applications where solubility is a critical determinant of molecular performance. Such advances can accelerate applications in drug discovery, sustainable chemistry, and pharmaceutical design.

## Author contributions

W. Wang: conceptualisation, methodology, software, validation, formal analysis, visualisation, writing – original draft, writing – review & editing. I. Cooley: methodology – openCOSMO simulation, chemical analysis, writing – review & editing. M. R. Alexander: funding acquisition, project administration, writing – review & editing. R. D. Wildman: funding acquisition, resources, project administration, writing – review & editing. A. K. Croft: supervision, funding acquisition, project administration, writing – review & editing. B. F. Johnston: supervision, funding acquisition, project administration, writing – review & editing.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

The source code used for model training and analysis, together with the SI and a processed, non-confidential version of the dataset suitable for direct use, are publicly available *via* Zenodo at https://doi.org/10.5281/zenodo.1794948.

## Acknowledgements

All authors acknowledge support from the EPSRC "Dialling up performance for on-demand manufacturing" (EP/W017032/1). AKC and IC acknowledge additional support from the EPSRC's Physical Sciences Data Infrastructure Phase 1b (EP/X032701/1).

## Notes and references

1 A. Sitovs and V. Mohylyuk, *Drug Discov. Today*, 2024, **29**, 104214.

2 W. Ge, R. De Silva, Y. Fan, S. A. Sisson and M. H. Stenzel, *Macromol. Rapid Commun.*, 2025, e00251.

3 C. Lipinski, *Am. Pharm. Rev*, 2002, **5**, 82–85.

4 A. Klamt, F. Eckert and W. Arlt, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 101–122.

5 A. Klamt, *J. Phys. Chem.*, 1995, **99**, 2224–2235.

6 F. Eckert and A. Klamt, *AIChE J.*, 2002, **48**, 369–385.

7 A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074–5085.

8 A. K. Nangia, *Cryst. Growth Des.*, 2024, **24**, 6888–6910.

9 K. A. Ali, S. K. Mohin, P. Mondal, S. G. Susmita, S. Choudhuri and S. Ghosh, *J. Pharm. Sci.*, 2024, **10**, 53.

10 M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

11 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

12 Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, *et al.*, *Nat. Mach. Intell.*, 2021, **3**, 76–86.

13 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.

14 M. Elkabous, A. Karzazi and Y. Karzazi, *Comput. Mater. Sci.*, 2024, **243**, 113146.

15 A. Sadeghi, M. Shariatmadar, S. Amoozadeh, A. M. Nahavandi and M. Mahdavian, *J. Taiwan Inst. Chem. Eng.*, 2025, **169**, 105998.

16 S. Kour and J. R. Sankar, *Contemp. Math.*, 2024, **5**, 6515–6526.

17 A. Yang, S. Sun, L. Qi, Z. Y. Kong, J. Sunarso and W. Shen, *Green Chem. Eng.*, 2025, **6**, 193–199.

18 N. Zand, A. E. Gorji, S. Riahi and M. MohammadiKhanaposhtani, *Fuel*, 2026, **403**, 136001.

19 Z. Chen, J. Chen, Y. Qiu, J. Cheng, L. Chen, Z. Qi and Z. Song, *ACS Sustain. Chem. Eng.*, 2024, **12**, 6648–6658.

20 A. E. Gorji and V. Alopaeus, *Int. J. Hydrogen Energy*, 2024, **90**, 803–816.

21 P. Mikulskis, M. R. Alexander and D. A. Winkler, *Adv. Intell. Syst.*, 2019, **1**, 1900045.

22 Y. Beghour and Y. Lahiouel, *Chem. Eng. Sci.*, 2025, **309**, 121228.

23 L. Cheng, H. Liao, Y. Zhu, Z. Tang, L. Lv and H. Ren, *Ind. Eng. Chem. Res.*, 2025, **64**, 4669–4684.

24 O. Ejima, M. S. Abubakar, S. S. Sarkin Pawa, A. H. Ibrahim and K. O. Aremu, *Phys. Scr.*, 2024, **99**, 106009.

25 H. Qin, M. Rehman, M. F. Hanif, M. Y. Bhatti, M. K. Siddiqui and M. A. Fiidow, *Sci. Rep.*, 2025, **15**, 1742.

26 H. Cho, J. Kim, K. T. No and H. Lim, *EPJ Quantum Technol.*, 2025, **12**, 79.

27 A. K. Chew, M. Sender, Z. Kaplan, A. Chandrasekaran, J. C. Elk, A. R. Browning, H. S. Kwak, M. D. Halls and M. A. F. Afzal, *J. Cheminf.*, 2024, **16**, 31.

28 M. Mohan, K. D. Jetti, S. Guggilam, M. D. Smith, M. K. Kidder and J. C. Smith, *ACS Sustain. Chem. Eng.*, 2024, **12**, 7040–7054.

29 V. H. Masand, S. Al-Hussain, G. S. Masand, A. Samad, R. Gawali, S. Jadhav and M. E. A. Zaki, *Comput. Biol. Chem.*, 2025, **115**, 108324.

30 X. Li, X. Li, X. Ma, Z. Cao, H. Zhong and S. Wang, *Sep. Purif. Technol.*, 2025, **373**, 133550.

31 W. Zhang, J. Ralston, R. Zheng, W. Sun, S. Xu, J. Cao, X. Jin, Z. Feng and Z. Gao, *Sep. Purif. Technol.*, 2024, **332**, 125855.

32 R. E. Bellman, *A Guided Tour*, Princeton University Press, 1961.

33 A. D. Vassileiou, M. N. Robertson, B. G. Wareham, M. Soundaranathan, S. Ottoboni, A. J. Florence, T. Hartwig and B. F. Johnston, *Digit. Discov.*, 2023, **2**, 356–367.

34 P. Cysewski, T. Jelinski, M. Przybyłek, W. Nowak and M. Olczak, *Pharmaceutics*, 2022, **14**, 2828.

35 Z. Ye and D. Ouyang, *J. Cheminf.*, 2021, **13**, 98.

36 S. Boobier, D. R. Hose, A. J. Blacker and B. N. Nguyen, *Nat. Commun.*, 2020, **11**, 5753.

37 F. Cenci, S. Diab, P. Ferrini, C. Harabajiu, M. Barolo, F. Bezzo and P. Facco, *Int. J. Pharm.*, 2024, **660**, 124233.

38 Q. C. Montreal, Molecular Operating Environment (MOE), *Chemical Computing Group ULC*, 2022.

39 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.

40 G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, S. Riniker, P. Gedeck, G. Jones, E. Kawashima, N. Schneider, D. Nealschneider, A. Dalke, T. Hurst, M. Swain, B. Cole, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, H. Faara, R. Walker, V. F. Scalfani, D. Probst, K. Ujihara, N. Maeder, A. Pahl, G. Godin and J. Lehtivarjo, *Q1 2025 Release*, 2025, DOI: 10.5281/zenodo.16996017.

41 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.

42 L. S. Shapley, *17. A Value for n-Person Games*, Princeton University Press, Princeton, 1953, pp. 307–318.

43 T. Chen and C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

44 R. Satheeskumar, *Intell. Pharm.*, 2025, **3**, 127–140.

45 Q. Yang, L. Fan, E. Hao, X. Hou, J. Deng, Z. Xia and Z. Du, *J. Pharm. Sci.*, 2024, **113**, 1155–1167.

46 S. Wang and Y. Ji, *AIChE J.*, 2024, **70**, e18359.

47 Q. Wang, X. Zou, Y. Chen, Z. Zhu, C. Yan, P. Shan, S. Wang and Y. Fu, *Spectrochim. Acta, Part A*, 2024, **323**, 124917.

48 X. Zou, Q. Wang, Y. Chen, J. Wang, S. Xu, Z. Zhu, C. Yan, P. Shan, S. Wang and Y. Fu, *Food Chem.*, 2025, **463**, 141053.

49 B. G. Beglaryan, A. S. Zakuskin, V. A. Nemchenko and T. A. Labutin, *J. Chem. Inf. Model.*, 2025, **65**, 4854–4865.

50 H. C. Kim, S. Y. Ha and J.-K. Yang, *J. King Saud Univ., Sci.*, 2025, **37**, 5552024.

51 X. Qu, C. Jiang, M. Shan, W. Ke, J. Chen, Q. Zhao, Y. Hu, J. Liu, L.-P. Qin and G. Cheng, *J. Chem. Inf. Model.*, 2025, **65**, 613–625.

52 V. Ghuriani, J. T. Wassan, P. Tripathi and A. Chauhan, *Int. J. Mol. Sci.*, 2025, **26**, 5590.

53 M. A. H. Danishuddin, G. Madhukar, Q. M. S. Jamal, J.-J. Kim and K. Ahmad, *Pharmaceuticals*, 2025, **18**, 714.

54 W. Xiong, J. Tan, H. Zhong, Z. Cao, H. Cai, X. Ma and S. Wang, *Sep. Purif. Technol.*, 2025, **376**, 133879.

55 S. Lu, N. J. Huls, K. Basu and T. Li, *J. Chem. Inf. Model.*, 2025, **65**, 1188–1197.

56 Z. Sodaei, S. Ekrami and S. M. Hashemianzadeh, *Sci. Rep.*, 2025, **15**, 26955.

57 M. Alqarni and A. Alqarni, *Sci. Rep.*, 2025, **15**, 2241.

58 L. Jiang, Q. Li, H. Liao, H. Liu and B. Tan, *Sci. Rep.*, 2025, **15**, 19456.

59 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

60 C. E. Shannon, *Bell Syst. Tech. J.*, 1948, **27**, 379–423.

61 S. Müller, T. Nevolianis, M. Garcia-Ratés, C. Riplinger, K. Leonhard and I. Smirnova, *Fluid Phase Equilib.*, 2025, **589**, 114250.

62 T. Gerlach, S. Müller, A. G. de Castilla and I. Smirnova, *Fluid Phase Equilib.*, 2022, **560**, 113472.

63 A. Klamt, *COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design*, Elsevier, 2005.

64 A. Schindl, M. L. Hagen, I. Cooley, C. M. Jäger, A. C. Warden, M. Zelzer, T. Allers and A. K. Croft, *RSC Sustain.*, 2024, **2**, 2559–2580.

65 R. P. Schwarzenbach, P. M. Gschwend and D. M. Imboden, Activity Coefficient and Solubility in Water, in *Environmental Organic Chemistry*, ed. R. P. Schwarzenbach, P. M. Gschwend and D. M. Imboden, Wiley, Hoboken, 2003, p. 2.

66 M. C. Sorkun, A. Khetan and S. Er, *Sci. Data*, 2019, **6**, 143.

67 L. Krasnov, D. Malikov, M. Kiseleva, S. Tatarin, S. Sosnin and S. Bezzubov, *Sci. Data*, 2025, **12**, 1236.

68 J. Bergstra and Y. Bengio, *J. Mach. Learn. Res.*, 2012, **13**, 281–305.

69 P. Good, *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Science & Business Media, 2013.

70 S. M. Lundberg and S.-I. Lee, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, pp. 4768–4777.

71 H. Chen, I. C. Covert, S. M. Lundberg and S.-I. Lee, *Nat. Mach. Intell.*, 2023, **5**, 590–601.

72 S. Müller, T. Nevolianis, M. Garcia-Ratés, C. Riplinger, K. Leonhard and I. Smirnova, *Fluid Phase Equilib.*, 2025, **589**, 114250.

73 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.

74 D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, 2002, **45**, 2615–2623.