

Improved Solubility Predictions in scCO₂ Using Thermodynamics-Informed Machine Learning Models

Dmitriy M. Makarov,^{*,†} Nikolai N. Kalikin,[†] Yury A. Budkov,^{*,†,‡} Pavel Gurikov,[¶]
Sergey E. Kruchinin,[§] Abolghasem Jouyban,^{||} and Michael G. Kiselev[⊥]

[†]*Laboratory of Multiscale Modeling of Molecular Systems, G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Akademicheskaya Street, Ivanovo 153045, Russia*

[‡]*Laboratory of Computational Physics, National Research University Higher School of Economics (HSE University), Tallinskaya Street, 34, Moscow 123458, Russia*

[¶]*Laboratory for Development and Modelling of Novel Nanoporous Materials, Hamburg University of Technology, Eißendorfer Straße 38, Hamburg, Germany*

[§]*G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Akademicheskaya Street, Ivanovo 153045, Russia*

^{||}*Pharmaceutical Analysis Research Center and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz 5165665811, Iran*

[⊥]*Laboratory of NMR-spectroscopy and numerical methods of investigation of liquid systems, G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Akademicheskaya Street, Ivanovo 153045, Russia*

E-mail: dmm@isc-ras.ru; ybudkov@hse.ru

Abstract

Accurate solubility prediction in supercritical carbon dioxide (scCO₂) is crucial for optimizing experimental design by eliminating unnecessary and costly trials at an early stage, thereby streamlining the workflow. A comprehensive solubility database containing 31,975 records has been compiled, providing a foundation for developing predictive models applicable to a diverse class of chemical compounds, with a particular focus on drug-like substances. In this study, we propose a domain-aware machine learning approach that incorporates thermodynamic properties governing phase transitions to solubility predictions in scCO₂. Predictive models were developed using the CatBoost algorithm and a graph-based architecture employing directed message passing to identify the most effective approach. Furthermore, auxiliary properties of the solute, including melting point, critical parameters, enthalpy of vaporization, and Gibbs free energy of solvation, were predicted as part of this work. The findings underscore the efficacy of incorporating domain-specific thermodynamic features to enhance the predictive accuracy of scCO₂ solubility modeling. The interpretation and the applicability domain assessment have confirmed the qualitative selection of the employed descriptors, demonstrating their ability to generalize to unique compounds that fall outside the defined domain.

Keywords: Solubility; Supercritical carbon dioxide; Machine learning; Drug-like compounds.

Introduction

Supercritical fluids are an environmentally friendly and efficient alternative to traditional chemical solvents used in various industrial processes.¹ Among the most widely utilized supercritical fluids is carbon dioxide (scCO₂), which is characterized by its safety, inertness, non-flammability, ease of recycling, and high solvation potential for non-polar compounds. With a critical temperature of 304 K and a pressure of 73.8 bar, scCO₂ is a cost-effective fluid and particularly well-suited for processing thermally sensitive drugs/solutes, food-related,

and biological materials.

The application of scCO₂ spans a wide range of fields, including the extraction of compounds,^{2,3} biomass conversion,⁴ the production of micro- and nanoparticles,⁵⁻⁷ control of polymorphic forms,⁸⁻¹⁰ impregnation,^{11,12} and polymer synthesis,¹³ among others. The supercritical fluid extraction method is thus comprehensively applied across various industries, including pharmaceuticals, food, cosmetics, and beverages.^{14,15} Understanding solubility is crucial for optimizing extraction parameters (temperature, pressure, and time) to maximize the yield of target compounds. Furthermore, knowledge of solubility helps to predict the behavior of substances during extraction, which is particularly important for developing cost-effective and environmentally sustainable technologies. Another critical area where solubility in scCO₂ plays a pivotal role is in the modification of pharmaceutical compounds, which is essential for enhancing their bioavailability, dosing efficiency, and therapeutic applicability. Nanonization technologies leveraging scCO₂ have emerged as a promising approach for producing drug nanoparticles, offering improved solubility, enhanced permeability of hydrophobic drugs, and greater stability.¹⁶ In addition, providing microparticles (usually with the size range of 1–5 µm) of the inhalation drugs to be used in treatment of respiratory and/or systemic diseases as inhaled pharmaceutical formulations¹⁷ is another important industrial application for scCO₂ technologies.

However, the experimental investigation of solute solubility in scCO₂ presents significant challenges, including substantial time requirements, high equipment costs, and variability in results.¹⁸ Consequently, the development of computational modeling methods for predicting solubility has become a critical task, as it can reduce the need for extensive laboratory experiments and accelerate research and development processes.¹⁹

Currently, there are no universal methods for predicting the solubility of substances in scCO₂. A variety of computational approaches have been developed for predicting solubility in scCO₂.²⁰⁻²⁴ These methods encompass density-based correlation models,²⁴ equation-of-state (EoS) approaches,²⁵ quantitative structure–property relationship models,²⁶ classical

density functional theory (cDFT),^{22,27–29} molecular dynamics (MD) simulations^{20,21,30,31} and others. Among these, density-based correlations and EoS methods are the most widely employed due to their relative simplicity and ease of use. However, these approaches rely on experimental solubility data to calibrate computational parameters through fitting procedures. In contrast, more complex methodologies, such as cDFT or full-atomic MD simulation, focus primarily on solvation contributions and require experimental sublimation data for the compounds under investigation.

Recently, machine learning (ML)-based approaches have emerged as a particularly promising direction for solubility prediction, as evidenced by a growing body of literature dedicated to this topic.¹⁹ This study does not consider literature ML approaches based on limited datasets, which typically include only a few compounds. Such models are generally unsuitable for predicting the solubility of novel solutes. However, there have been notable exceptions. For instance, studies^{26,32} have employed substantially larger datasets, encompassing about a hundred compounds, to develop predictive models. Nevertheless, these studies have employed random data splitting, resulting in an overly optimistic estimation of prediction error. In contrast, our recent study³³ employed a dataset comprising 187 substances, where a rigorous data split revealed a substantially higher prediction error compared to random splitting. However, this dataset remains insufficient for achieving robust generalization.

In this study, we developed a global model based on the most extensive dataset, capable of predicting solubility at specified pressures and temperatures solely from the molecular structure of the solute. The proposed approach simultaneously addresses closely related tasks, including the prediction of solubility in scCO₂, melting point, enthalpy of vaporization, critical pressures and temperatures, as well as Gibbs free energy of solvation. By advancing toward more sustainable solutions, the methodology presented in this study serves as a valuable tool for researchers aiming to optimize supercritical fluid processes. We anticipate that the integration of large-scale processes utilizing supercritical carbon dioxide as a medium will accelerate the adoption of green chemistry methods and pave the way for a more sustainable

future, offering efficient alternatives to traditional organic solvents.

Methodology

Datasets and Data curation

A comprehensive dataset on the solubility of various solutes in subcritical and supercritical carbon dioxide was compiled from 771 original literature sources. The resulting dataset comprises a total of 31,975 data points encompassing 1,065 distinct compounds (Table 1). Each data point includes corresponding experimental conditions, specifically temperature and pressure, along with the solubility of the solute expressed as a mole fraction (y_2). The structures of all solutes were standardized and converted into computer-readable identifiers using the Simplified Molecular Input Line Entry System (SMILES).³⁴

The primary challenge associated with the compiled dataset was the presence of inter-laboratory duplicates, i.e. solubility data for the same dissolved compounds were reported by multiple sources under identical pressure and temperature conditions. In cases where multiple experimental values were available for a given compound at the same pressure and temperature, the median solubility value was selected. For identical isotherms, a systematic data reduction procedure was implemented. This procedure involved merging data for each isotherm (with a tolerance of ± 1 K) and subsequently approximating the natural logarithm of solubility values using a linear function dependent on CO₂ density. It is obvious that the quality of data and their validity is a major parameter in evaluating the models' performances. There are some differences in the reported data in the literature³⁵ and large variability is expected in repeating the measurement of the solubility in scCO₂ even at the same experimental condition.

Prior to this step, CO₂ density for all experimental state parameters was calculated using the Span and Wagner EoS.³⁶ Given that a sharp increase in solubility can occur during the transition from the subcritical to the supercritical state of CO₂,¹⁸ the data cleaning procedure

was applied exclusively to measurements conducted in the supercritical region. Solubility values were averaged over three or more data points as a function of CO₂ density. A relative deviation of 10% between the approximated solubility value and the experimental data was considered acceptable. Data points exceeding this deviation threshold were classified as outliers and subsequently excluded from the purified dataset.

The application of this data purification procedure, illustrated using naphthalene as an example (Figure S1, Supplementary Materials), resulted in a final dataset comprising 30,771 data points across 1,065 compounds. The adherence to these preprocessing criteria aimed to optimize the dataset for enhanced reliability in predictive modeling and meaningful analytical insights. The complete dataset included only the solubility data of pure compounds in subcritical and supercritical CO₂, excluding any systems with cosolvents. For further analysis, two distinct subsets were derived. First, data corresponding to solubilities measured under subcritical CO₂ conditions specifically at temperatures below 304.3 K and pressures below 73.9 bar were excluded. The resulting subset was designated as the “scCO₂ dataset.” The second subset, referred to as the “Drug-like dataset”, comprised only compounds that satisfied Lipinski’s rule of five.³⁷ This subset excluded salts and complex compounds in which metals were coordinated to the organic framework via coordination bonds.

The refined datasets were utilized to develop auxiliary predictive models. The melting point dataset was compiled from two sources: one containing molecular compounds³⁸ with 47,427 substances, and another covering salts³⁹ with 3,064 substances. Data for critical temperature (T_c), critical pressure (P_c), and enthalpy of vaporization (ΔH_v) were obtained from Yaws’ publication,⁴⁰ which provides both experimental values and estimated data derived from group contribution methods. The dataset included 4,031 compounds for T_c , 4,109 for P_c , and 3,522 for ΔH_v .

Table 1: Characteristics of datasets

Dataset	Data points	Chemicals	Temperature ranges, K	Pressure ranges, bar
Full	31975	1065	285.9 - 523	1 - 3494
Full (curated)	30771	1065	285.9 - 523	1 - 3494
scCO ₂	29029	1065	304.9 - 523	73.9 - 3494
Drug-like	19831	656	305 - 510.9	73.9 - 2700

Chemical Structure Representation and Machine Learning Algorithms

The choice of molecular descriptors and feature representations is crucial in machine learning tasks related to physicochemical properties, as they significantly impact prediction accuracy, model generalization, and interpretability.⁴¹ The SMILES notation was employed to generate two distinct molecular representation modalities: molecular descriptors and graph-based representations. Building on insights from previous studies,³³ we selected only the descriptor set that demonstrated the highest predictive performance for scCO₂ solubility in a dataset comprising 187 solutes.

The Chemistry Development Kit (CDK, version 2.8)⁴² was utilized to calculate 14 different types of molecular fingerprints and 287 molecular descriptors. These descriptors encompass quantitative properties, structural fragments, graph invariants, and three-dimensional (3D) molecular characteristics. However, due to numerous errors and the loss of crucial data resulting from the structural complexity of certain compounds during optimization, the calculation of 3D descriptors was omitted. For salt descriptor calculations, no specialized preprocessing was applied to remove small counterions; instead, all ions were considered. The final salt descriptor values were obtained by averaging the characteristics of all ionic components.

Subsequently, a standard descriptor filtering procedure⁴³ was conducted. In the initial stage, descriptors with zero variance, as well as those with absolute values exceeding 999,999, were excluded. To minimize data redundancy, highly correlated descriptors (Pearson correla-

tion coefficient > 0.95) were grouped, and only a single representative (the first in sequence) was retained from each group. The resulting filtered set of descriptors was used for further model development.

Two machine learning approaches were employed for model development:

CatBoost⁴⁴ is an advanced gradient boosting algorithm based on decision trees, specifically designed to handle categorical data efficiently without the need for prior encoding. Gradient boosting constructs models iteratively by combining weak learners, with each subsequent model correcting the residual errors of the previous iterations. A distinguishing feature of CatBoost is its ability to process categorical features directly, improving both model accuracy and training speed. Previous studies have demonstrated the effectiveness of CatBoost-based models in predicting molecular properties for small datasets, including molecular systems,³³ mixtures,⁴⁵ and chemical reactions.⁴⁶

The Graph Convolutional Neural Network (GCNN) model implemented in Chemprop⁴⁷ extracts molecular features by first computing simple atomic and bond-level descriptors, followed by multiple rounds of message passing through neural network layers to aggregate structural information across the entire molecule. This process enables the generation of a comprehensive molecular representation, which is subsequently processed by a fully connected feedforward neural network (FNN).

The message passing mechanism in GCNN is highly effective for capturing local atomic environments, however, it may struggle to capture global molecular features, particularly in large molecules where the longest molecular path exceeds the number of message passing iterations.⁴⁸ To enhance the predictive power of the GCNN model and incorporate global molecular characteristics, 12 additional descriptors from RDKit⁴⁹ were included: molecular weight (**MolWt**), octanol-water partition coefficient (**MolLogP**), topological polar surface area (**TPSA**), number of rotatable bonds (**NumRotatableBonds**), Balaban index (**BalabanJ**), number of hydrogen bond acceptors (**NumHAcceptors**) and donors (**NumHDonors**), total ring count (**RingCount**), number of aliphatic and aromatic rings (**NumAliphaticRings**, **NumAromaticRings**),

fraction of sp³ hybridized carbons (**FractionCSP3**), and heavy atom count (**HeavyAtomCount**). These descriptors provide high-level structural and physicochemical insights into the molecules, improving the model’s ability to capture global molecular properties.

Distributed hyperparameter optimization for GCNN models was conducted using Ray Tune.⁵⁰ For CatBoost models, the default hyperparameters were employed.

Model Validation

A critical aspect in evaluating the predictive performance of a model is the division of experimental data into a training set, used for model calibration, and a validation set, used to assess the model’s generalization ability and prediction uncertainty. In this study, all models underwent a 5-fold cross-validation (5CV) procedure. In 5CV, the dataset was randomly partitioned into five subsets, with the model trained and validated five times. In each iteration, one subset served as the validation set, while the remaining four subsets were used for training. This approach enables predictions to be generated for all data points, in contrast to the conventional single-split method, which randomly (or stratified) divides the data into separate training and test sets.

For datasets where properties depend on state variables, different cross-validation strategies can be employed.⁵¹ In this study, three commonly used partitioning strategies were applied: random split, strict splitting based on unique SMILES, and strict splitting based on molecular scaffolds.

Molecular scaffolds represent the core structures of molecules, obtained by removing side chains and replacing specific functional groups. For scaffold-based partitioning, the **MurckoScaffold** function from the RDKit library was utilized to extract the structural backbone of each molecule. In strict partitioning schemes within the 5CV framework, data were split such that molecules with the same SMILES or scaffold structures were not present in both training and validation sets. To eliminate stereochemical redundancy in the unique SMILES-based partitioning, canonical SMILES representations were used. This ensured that

stereoisomers- compounds with similar solubility values were consistently placed in the same fold, preventing potential data leakage across subsets.

Model performance was evaluated using three statistical metrics:

$$R^2 = 1 - \frac{\sum_{i=1}^n (\lg(y_{\text{pred},i}) - \lg(y_{\text{exp},i}))^2}{\sum_{i=1}^n (\lg(y_{\text{exp},i}) - \lg(\bar{y}_{\text{exp}}))^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\lg(y_{\text{pred},i}) - \lg(y_{\text{exp},i}))^2} \quad (2)$$

$$AARD = \frac{1}{n} \sum_{i=1}^n \left| \frac{\lg(y_{\text{pred},i}) - \lg(y_{\text{exp},i})}{\lg(y_{\text{exp},i})} \right| 100 \quad (3)$$

where n is the number of data points; y_{exp} is the experimental solubility value, and y_{pred} is the predicted solubility value.

Model Interpretation and Applicability Domain Analysis

Interpreting the developed models is a crucial step in gaining insights into the obtained results and understanding the underlying relationships within the data. The contribution of individual features to model predictions was analyzed using the SHapley Additive ex-Planations (SHAP) method.⁵² SHAP provides a comprehensive framework for interpreting machine learning model outputs, leveraging principles derived from cooperative game theory to fairly attribute the contribution of each feature to the final prediction. By applying SHAP values, it is possible to quantify the impact of each molecular descriptor on solubility predictions, thereby enhancing the interpretability and transparency of the developed models.

Training models based on experimental data is inherently susceptible to the presence of outliers, which can adversely affect predictive accuracy and limit model generalization. To mitigate these risks, it is crucial to establish a well-defined applicability domain (AD)

for the model.⁵³ One of the most commonly used approaches for defining AD is the leverage method,⁵⁴ which is widely applied in machine learning to determine the boundaries of applicability. This method relies on two key components: leverage (h), which quantifies the influence of a given sample on the regression model, and standardized residuals, which measure the deviation between predicted and experimental values.

However, for nonlinear models, the conventional leverage calculation may be overly optimistic, leading to inaccurate assessments of the AD. To address this limitation, the leverage method was combined with a SHAP-based approach, which provides an estimate of the contribution of each descriptor to individual predictions. By summing the absolute SHAP values for each descriptor, the degree of influence of the solute on the solubility prediction was estimated. This measure served as an analogue to leverage and was calculated using the following equation:

$$h_i = \sum_{j=1}^d |SHAP_{ij}| \quad (4)$$

where $SHAP_{ij}$ represents the contribution of the j -th descriptor for the i -th sample, and d denotes the total number of descriptors.

Unlike conventional leverage, SHAP leverage is not normalized and its magnitude is directly dependent on the SHAP values themselves. Therefore, instead of using a classical leverage threshold, the 95th percentile of the leverage distribution was employed to identify influential points.

Standardized residuals (SDR) were computed in accordance with the standard leverage approach, calculated as the difference between the experimental values y and the predicted values y_{pred} , followed by standardization. Predictions with residuals falling outside the range $[-3, 3]$ were classified as outliers based on the residuals criterion.

This combined methodology ensures a more robust definition of the model's applicability domain by integrating both descriptor contributions and residual analysis, enhancing the reliability of the developed predictive models.

Results and discussion

Dataset Analysis

The second phase of this study involved a detailed examination of the dataset to gain a comprehensive understanding of its content and to identify any potential limitations that could affect the modeling process. The distribution of decimal logarithmic solubility values in the subCO₂ and scCO₂ datasets is shown in Figure 1a. Given that this distribution approximates a normal distribution, it is anticipated that there will be no systematic bias in the model predictions. The solubility values in the total dataset range from -9.1 to -0.19 on a decimal logarithmic scale, with a mean value of -3.96 .

The experimental state parameter values range from 285.9 K to 523.0 K, in temperature, with a mean 327.3 K, and from 1 to 3494 bar in pressure, with a mean of 198 bar (Figures 1b and 1c). Solubility measurements at temperatures above 400 K were sparse, and data for pressures exceeding 500 bar were isolated. The dataset includes compounds exhibiting a wide chemical diversity, featuring 42 unique atom types. Molecular sizes range from 1 to 273 atoms, and molecular weights from 16 to 1819 g mol⁻¹. The molecular weight distribution closely approximates a normal distribution, with an average value of 325 g mol⁻¹ (Figure S2). This suggests that most compounds in the dataset are of medium to high molecular weight.

The total number of compounds in the dataset includes 1039 organic compounds and 26 inorganic compounds. The compounds were analyzed based on the most common functional groups they contain. The quantitative distribution of these functional groups in the full dataset is presented in Figure 1d. The histogram shows that structures with aromatic, carbonyl, halogen, and hydroxyl groups are the most common, while those containing phosphate, sulfo-, or nitro- groups are significantly less common. The chemical space of the full and drug-like subsets was analyzed using t-distributed stochastic neighbor embedding (t-SNE)⁵⁵ (Figure S3). The first two projections generated by t-SNE captured most of the

variation present in the dataset, which was originally represented in a high-dimensional space derived from the Extended Connectivity Fingerprint (ECFP) with a circular neighborhood radius of 2, encoded in a 2048-bit vector using RDKit. The resulting scatter plot reveals that the compounds in the drug-like subset are clustered into a more compact region, indicating a higher degree of molecular homogeneity within this subset. In contrast, the full set displays a broader distribution, reflecting a greater diversity of chemical structures.

Development of Auxiliary Models

Representation-based neural network models, increasingly utilized to describe complex physicochemical properties, require a substantial amount of training data. One approach to mitigate this need and enhance predictive performance is to incorporate chemical and physical knowledge into the model architecture.

The thermodynamic cycle of dissolution can be conceptualized as a process involving the transition of a substance from solid to liquid (melting), from liquid to gas (evaporation), and then from the gas phase to the solvent (solvation), ultimately forming a solution (see scheme in Figure 2). Building on this understanding, we can correlate each step of the cycle with a physicochemical property that can be estimated or predicted to improve solubility models. Thus, we integrated the following physicochemical properties of the solute: melting point (T_m), standard enthalpy of vaporization (ΔH_v), and standard Gibbs free energy of solvation in an infinitely dilute CO₂ solution (ΔG_{solv}). Their physical meanings are as follows: the melting point relates to the enthalpy of melting via the Clausius-Clapeyron equation, with substances having higher melting points generally exhibiting lower solubility, as more energy is required to break their crystal lattice; enthalpy of vaporization characterizes the energy required for the phase transition from liquid to gas; and Gibbs free energy of solvation reflects the interactions between solvent and solute molecules, with higher absolute values typically indicating a greater tendency for the compound to dissolve in CO₂.

Thus, we developed auxiliary models to predict key physicochemical properties, including

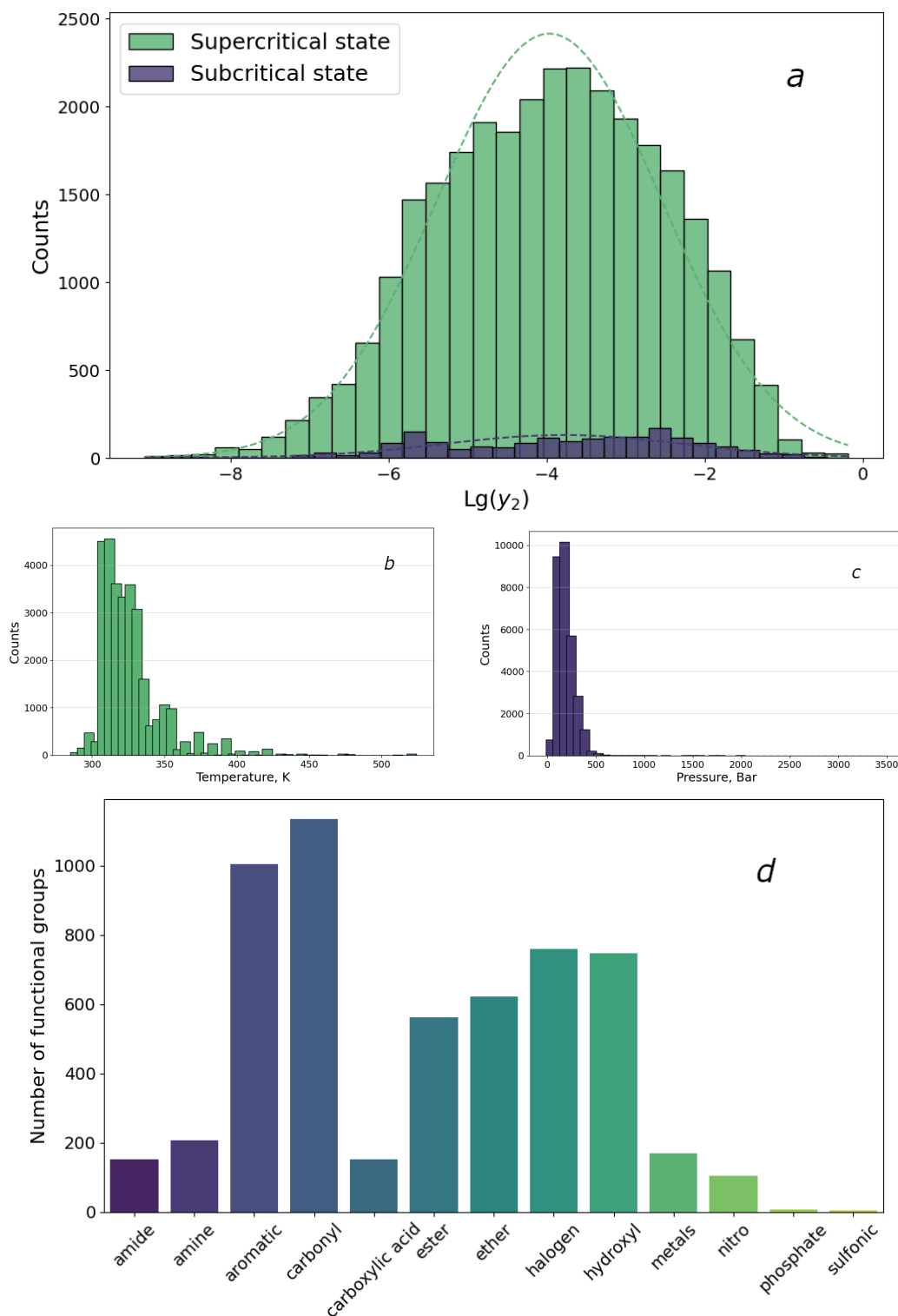


Figure 1: Dataset Visualization. (a) Quantitative distribution of the logarithmic solubility values for compounds in the subcritical and supercritical CO_2 regions; Distribution of (b) temperatures and (c) pressures values in the dataset; (d) Distribution of functional groups in the dataset.

melting point, enthalpy of evaporation, critical temperature, and critical pressure. The latter two are necessary for calculating the standard Gibbs free energy of solvation via the approach based on classical density functional theory (cDFT) as will be discussed further.

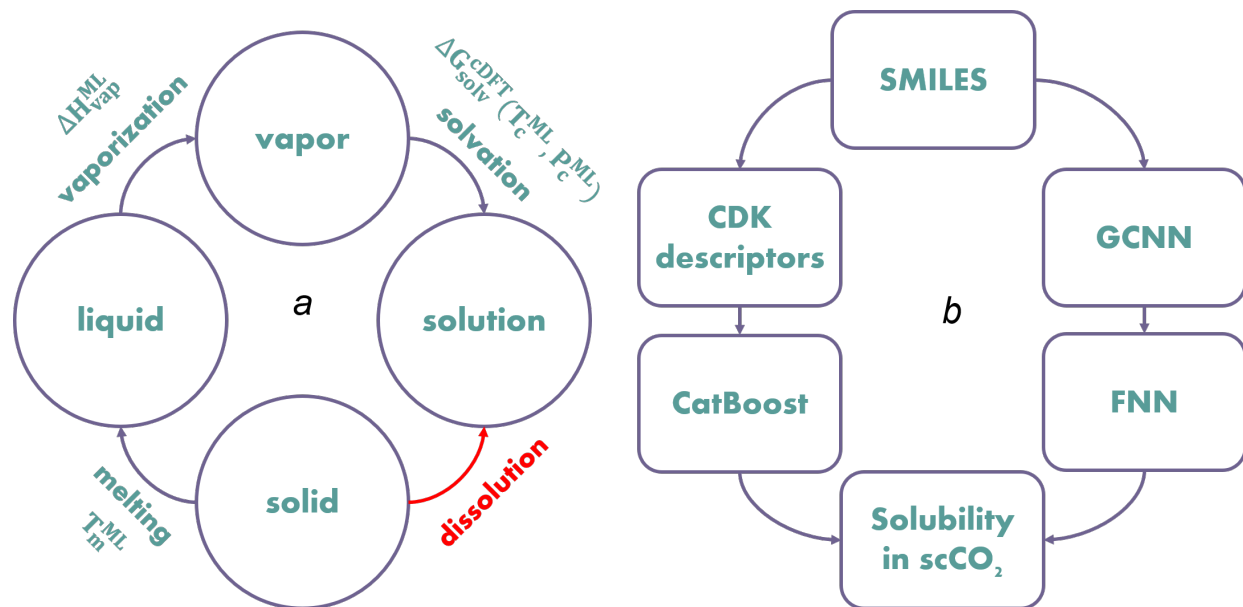


Figure 2: Visualization of the working process. (a) The thermodynamic cycle of dissolution; (b) Framework for developing models to predict solubility in scCO_2 .

Melting Point

Prediction models for melting point were developed using the methodologies outlined in Section methods. Comparative statistics for the models are shown in Figure S4. Our results were also benchmarked against previous studies.^{38,56} The first study³⁸ utilized a molecular dataset of 47,427 compounds and a consensus of 11 individual models with different descriptor sets, achieving an *RMSE* of 37.1 K during 5-fold cross-validation. Another study,⁵⁶ built on the same dataset, and employing Gaussian Process Regression with bioinformatics descriptors from RDKit, as well as structural and quantum chemical features, yielded a mean absolute error (*MAE*) of 28.85 K. Our dataset, in addition to the same molecular compounds, included 3,054 salts. The best-performing model, developed using a GCNN with

twelve additional RDKit descriptors, demonstrated superior performance with an *RMSE* of 35.3 K and an *MAE* of 25.5 K in 5-fold cross-validation.

Gibbs Free Energy of Solvation and Vaporization Enthalpy

The solvation free energy of a solute in scCO₂, which directly correlates with its solubility in the fluid, was estimated using a previously developed approach based on cDFT.^{27–29} In this framework, the grand thermodynamic potential of the system is expressed as a functional of the fluid density, perturbed by a spherically symmetric effective potential generated by a single site representing the solute molecule. The solvation free energy in the infinite dilution approximation is then defined as the difference between the grand thermodynamic potential of the perturbed by the solute fluid density and the potential of the unperturbed fluid density at fixed temperature and pressure. The parameters of the effective Lennard-Jones interaction potential between the fluid particles were determined by fitting the solvent critical point and reproducing the coexistence curve using the model equation of state. The solute parameters were obtained in the same manner.

Beyond inherent model simplifications, such as coarsening the internal chemical structure of the solute and solvent molecules, a major challenge is determining the critical parameters of the solutes, as most compounds degrade thermally before reaching the temperatures of interest. To predict critical temperatures, pressures, and enthalpies of vaporization, a multi-task learning (MTL) GCNN model was additionally employed, augmented with molecular descriptors accounting for intermolecular interactions. The MTL model outperformed single-task GCNN:RDKit models but was less accurate than the CatBoost:CDK models (Table S1). The training dataset included both experimental and calculated from group contribution methods values, increasing the dataset size fourfold.

Models for Predicting Solubility in scCO₂

Models for predicting the solubility of compounds in sub- and supercritical carbon dioxide were developed using the curated full dataset, both with and without the additional generated thermodynamic features. An important but often overlooked aspect in the literature – evaluation of the model’s performance on out-of-training set data – was addressed in this study. The ability of a model to extrapolate to chemical structures that were not encountered during training is rarely assessed in studies where property values depend on external state parameters. This oversight may lead to an inaccurate perception of the model’s predictive accuracy. To assess this capability, we employed various validation protocols that tested both interpolation (prediction of solubility for chemicals within the dataset) and extrapolation (prediction for new molecular structures). The reliability and stability of all models presented in this section were evaluated using statistical parameters, including the mean and standard deviation, based on the results of 5CV.

Random Data Splitting

In the baseline model, which combines the CatBoost method with filtered CDK descriptors, additional features providing specific information about the phase transition process and fluid density were incorporated. A stepwise analysis was conducted to evaluate the impact of each added feature on the solubility prediction performance. Changes in the *AARD* of these models, evaluated on randomly split test sets during cross-validation, are shown in Figure 3a. All other metrics are presented in Table S2.

The results indicate that incorporating individual thermodynamic parameters – such as T_m , ΔH_v , ΔG_{solv} and ρ_{CO_2} – into molecular CDK descriptors, as well as their various combinations, consistently reduced errors compared to the baseline model. For graph-based models, these thermodynamic characteristics further improved the internal graph representations, leading to even lower prediction errors. The GCNN:RDKit models outperformed traditional CatBoost:CDK models, reducing prediction error by an average of 12%. A scat-

ter plot illustrating the performance of the GCNN model with all descriptors is provided in Figure 3b.

Overall, for both approaches, the inclusion of external features improved model performance, reducing *AARD* by 9% compared to the baseline model. This highlights the importance of considering thermodynamic parameters when training models for solubility prediction.

However, the strong performance of the models presented in this section should be interpreted with caution. In a random data split, the same molecular structure under different state conditions could appear in different folds during cross-validation. As a result, the model might have recognized specific compounds rather than developing a true generalization capability.

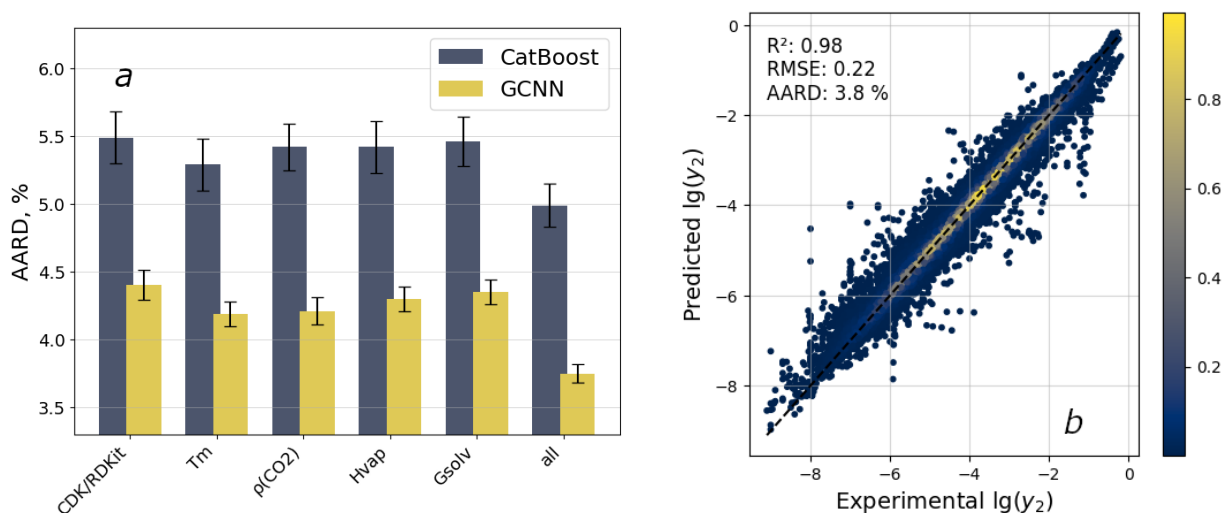


Figure 3: (a) Change in *AARD* of model predictions with the inclusion of thermodynamic features in the base models CatBoost:CDK and GCNN:RDKit, where the uncertainty represents the standard deviation obtained from 5 predictions during cross-validation. (b) The dependence of predicted versus experimental solubility values for the GCNN:RDKit model with the inclusion of all thermodynamic parameters.

Strict Validation

Under strict data splitting, where the model is required to predict entirely new and unique chemical structures — thereby demonstrating its true generalization capability — baseline models exhibited significantly higher errors compared to random splitting. Additionally, model performance showed greater variability, as reflected in an increase in standard deviation. This can be attributed to the varying complexity of the test set, which depends on the degree of similarity between test compounds and those in the training set.

Notably, the baseline GCNN:RDKit model demonstrated higher errors than the traditional CatBoost model, particularly when predicting the solubility of two small molecules: water and iodine (Figure S5). In contrast, the descriptor-based model did not show such significant deviations in its predictions for these compounds. To investigate this further, we excluded these two compounds from the dataset and retrained the models.

In the extrapolative validation protocol, incorporating fluid density and Gibbs free energy of solvation had little impact on model performance (Figure 4a). However, adding other thermodynamic properties individually led to slight improvements in predictive accuracy. When all descriptors related to solute phase changes and solvent behavior were included, predictions improved by an average of 9% in terms of *RMSE*. A t-test comparing the prediction results for models with and without thermodynamic parameters revealed *p*-values below 0.05, indicating statistically significant differences. This finding reinforces the importance of taking into account thermodynamic parameters not only for random data splits but also when predicting the solubility of entirely new molecular structures that the model has not encountered before.

Unlike SMILES-based splitting, which organizes samples to ensure unique SMILES representations do not appear in different folds, scaffold-based splitting groups molecules by their core structural frameworks and partitions the data accordingly. This approach allows the model to be tested on chemical structures that differ significantly from those in the training set. Such a method increases the complexity of the task and better simulates real-

world conditions, where the model encounters novel molecular types and must generalize its knowledge to new chemical classes.

Baseline models trained using scaffold-based splitting exhibited the highest prediction errors compared to other data-splitting strategies. Specifically, the GCNN:RDKit model achieved an *RMSE* of 1.02, while the CatBoost:CDK model reached 0.97 $\lg(y_2)$. This outcome is expected, as clustered splitting is intentionally designed to challenge the model with more complex test examples.

Consistent with previous validation protocols, incorporating additional physicochemical and thermodynamic descriptors into the models improved predictive performance, reducing *RMSE* by an average of 12%. The final *RMSE* values for the enhanced models with account of thermodynamic parameters (TP) were 0.90 for GCNN:RDKit and 0.85 $\lg(y_2)$ for CatBoost:CDK.

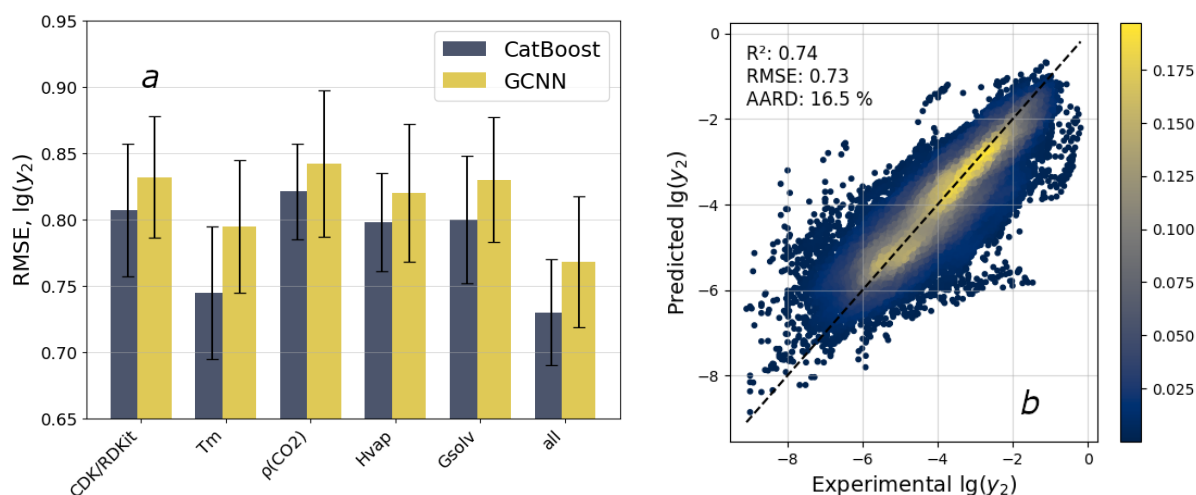


Figure 4: Changes in *RMSE* of model predictions during strict 5-fold cross-validation with the inclusion of various thermodynamic features in the baseline models CatBoost:CDK and GCNN:RDKit (a). Uncertainty bars represent variations between predictions during cross-validation. The dependence of predicted versus experimental solubility values for the CatBoost:CDK model with all thermodynamic parameters included (b)

Subset Models

When modeling with filtered datasets (Table 1), we used the same model architectures as for the complete dataset, incorporating all calculated features that had previously demonstrated the best performance. A comparison of model performance on the two subsets was conducted to evaluate the impact of rare and structurally distinct data on prediction quality. This approach helps determine whether filtering improves overall modeling outcomes by removing noisy data or, conversely, leads to the loss of important information needed for more accurate predictions.

The first subset included data exclusively related to the supercritical state of CO₂. In this state, carbon dioxide exhibits a density comparable to that of liquids while retaining the high mobility and diffusion properties characteristic of gases. This makes CO₂ a universal solvent capable of effectively dissolving substances that remain poorly soluble in the subcritical state. Data from the subcritical state of CO₂ were excluded. Models trained solely on supercritical condition data showed no changes in their statistical metrics compared to general models, indicating that joint training on both subcritical and supercritical data is feasible.

The second subset was formed based on Lipinski's rule. All rare examples, including inorganic compounds, organic salts, and substances containing metals, were excluded to focus solely on organic molecules.

Training the model on the drug-like subset using the SMILES protocol reduced *RMSE* for CatBoost:CDK from 0.74 to 0.69 $\lg(y_2)$ compared to the full dataset, and for GCNN:RDKit, from 0.77 to 0.71 $\lg(y_2)$. Using the scaffold protocol, the *RMSE* for CatBoost:CDK decreased from 0.85 to 0.82 $\lg(y_2)$, and for GCNN:RDKit, from 0.91 to 0.83 $\lg(y_2)$. Separation of drug-like molecules increased the homogeneity of the sample, improving prediction accuracy within this subgroup. The GCNN:RDKit method demonstrated a particularly noticeable improvement, approaching the performance of CatBoost:CDK. This higher performance may be due to the graph-based model's ability to better handle organic compounds with similar topology and atomic types when trained predominantly on such structures. At the same

time, it may struggle to account for unique atoms and interactions characteristic of salts and large molecules present in the full dataset. However, it should be noted that such filtering limits the model’s applicability to a broader range of substances. Table 2 presents the final results for models based on the full and drug-like datasets, validated using three different protocols.

A consensus approach, averaging the predictions of the CatBoost and GCNN models, achieved the lowest *RMSE* values for data splitting by SMILES (0.70 for the full dataset and $0.67 \lg(y_2)$ for the drug-like subset). This approach leverages the strengths of each model, contributing to a reduction in variance and systematic bias in the predictions. According to the literature,⁵⁷ optimal results in consensus modeling are achieved by integrating heterogeneous methods or diverse molecular representations, further validating the effectiveness of the proposed approach.

It is important to acknowledge the generally high prediction error for new compounds; however, such a result is expected, as achieving precise predictive accuracy remains challenging even for large datasets, including those involving aqueous solubility.⁵⁸ Experimental methods for determining solubility in scCO₂ are known to be time- and resource-intensive,⁵⁹ and the accuracy of such measurements is influenced by numerous factors. These include the precision of state parameter measurements (e.g., pressure and temperature), control of phase equilibrium, the choice of measurement method (static, dynamic, or gravimetric), and the purity of the solute.⁶⁰ Additionally, phenomena such as polymorphism⁶¹ and chemical interactions between CO₂ and the solute^{62,63} can significantly affect solubility, which must be carefully considered when developing predictive models. These factors collectively contribute to high experimental uncertainty, with substantial inter-laboratory variability.^{18,64} This variability may cast doubt on the reliability of reported solubility data in scCO₂. Since the inherent error is embedded in the data itself, an aleatoric limit is established. Attempts to improve predictive performance on such data may lead to overfitting, compromising the model’s generalizability.

Table 2: Statistical parameters of the model for the full and drug-like datasets

	Random split		SMILES split		Scaffold split	
	CatBoost	GCNN	CatBoost	GCNN	CatBoost	GCNN
Full						
R^2	0.967 ± 0.002	0.983 ± 0.002	0.73 ± 0.02	0.70 ± 0.05	0.59 ± 0.03	0.52 ± 0.07
$RMSE$	0.261 ± 0.009	0.223 ± 0.008	0.74 ± 0.05	0.77 ± 0.04	0.85 ± 0.09	0.91 ± 0.08
$AARD$, %	4.99 ± 0.16	3.75 ± 0.14	16.6 ± 1.4	16.9 ± 1.1	18.4 ± 5.2	18.9 ± 4.8
Drug-like						
R^2	0.977 ± 0.001	0.988 ± 0.002	0.76 ± 0.04	0.74 ± 0.04	0.64 ± 0.08	0.63 ± 0.06
$RMSE$	0.213 ± 0.005	0.200 ± 0.004	0.69 ± 0.05	0.71 ± 0.05	0.82 ± 0.09	0.83 ± 0.1
$AARD$, %	4.14 ± 0.08	3.24 ± 0.14	16.3 ± 2.3	16.8 ± 1.9	20.6 ± 5.6	20.5 ± 4.8

Model Interpretation

The transformation of input data into output predictions is often complex and challenging to interpret, particularly for vector representations derived from natural language processing methods and graph-based approaches. In this study, we conducted a feature importance analysis for the CatBoost:CDK with TP model using the complete dataset. Figure 5 illustrates the contributions of ten most influential features on the prediction of solubility in scCO₂. The analysis was performed using 5CV with a data splitting protocol based on unique SMILES identifiers. The bee swarm plot depicts the impact of each feature on the model’s predictions, where each point represents an individual compound with its corresponding solubility, and the SHAP value is plotted on the abscissa axis. The interpretation of the results (Figure 5) reveals that the added physicochemical descriptors play a pivotal role in the model’s performance. Solubility is governed by the balance between the solute-solvent interaction energy and the energy required to disrupt the crystal lattice or intermolecular bonds within the solute. For compounds with high melting points and enthalpies of vaporization, the energy needed to break these bonds significantly exceeds the interaction energy with scCO₂, which, as shown in the figure, reduces solubility. Notably, the Gibbs free energy of solvation ranked only in the thirtieth tier of importance. This may be due to high uncertainty associated with its theoretical estimation, where the input parameters (predicted critical parameters) could contain systematic errors. Additionally, its contribution may be

implicitly captured by other descriptors in the model.

An increase in the values of certain descriptors related to molecular branching (e.g., WTPT-5 and WPATH) and the number of hydrogen bond donors (nHBDon) leads to decrease in predicted solubility, as they determine the type and strength of intermolecular interactions. Conversely, an increase in fluid density and state parameters, as expected, enhances solubility. Qualitative analysis indicates that the model successfully captures the fundamental structure-property relationships, which align well with established physical and chemical principles of solubility modeling, thereby validating the its correct operation.

The descriptors are grouped into clusters based on their correlations, indicating similarities in their influence on solubility (Figure 5b). For instance, such descriptors as density of CO₂ and pressure in the system, as well as melting point and nAtomP (a constitutional descriptor representing the size of the longest conjugated system in the molecule), show the highest positive correlations with each other. This clustering highlights the interconnected nature of these features and their collective impact on solubility predictions.

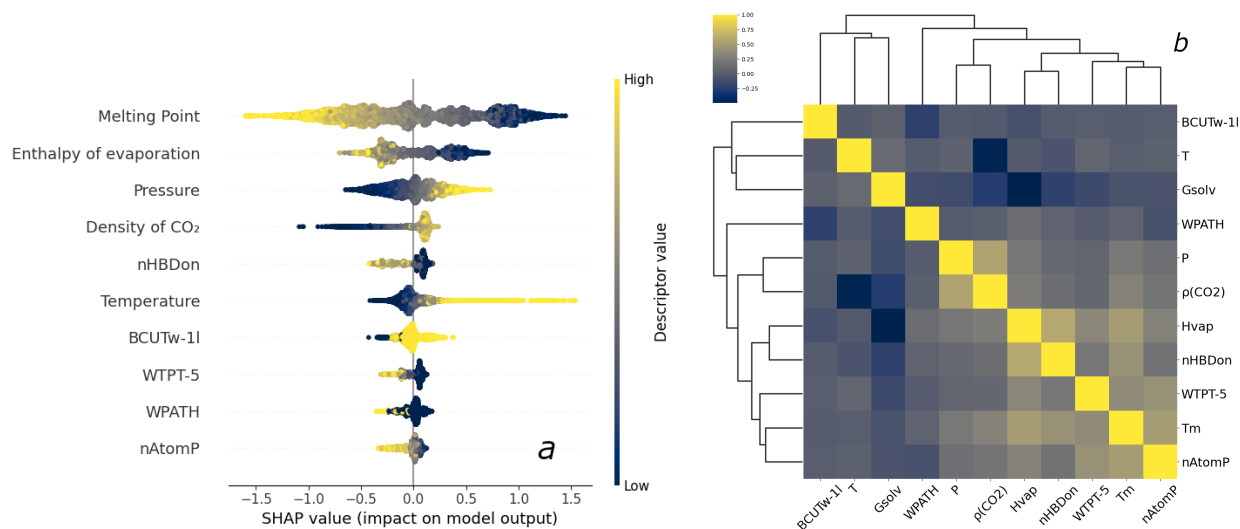


Figure 5: (a) Results from the SHAP approach illustrate the contributions of ten most significant features for predicting solubility in scCO₂. (b) Clustering of correlated features.

Applicability Domain and Error Analysis

To identify outliers in the datasets and establish the range of compounds for reliable predictions, the applicability domain was assessed using SHAP leverage values for the CatBoost:CDK model. For more stable evaluation of the applicability domain, the analysis was performed with 5CV. To account for model variability during cross-validation, SHAP leverage values from each test set were aggregated, and the results were presented in a William's plot (Figure 6). The delineation of the reliable zone (shaded area) helps identify data points that align with model expectations and fall within statistical reliability, facilitating accurate dataset evaluation and interpretation.

For the complete dataset and the drug-like subset, the critical SHAP leverage values (h^*) were 3.86 and 3.58, respectively. In the complete dataset, 1539 data points corresponding to 163 compounds exceeded the h^* threshold, while in the drug-like subset, 969 data points corresponding to 101 compounds were beyond this limit, indicating that predictions for these compounds may be less reliable. In the complete dataset, these 163 compounds included a significant number of salts, dyes, and metal-containing compounds, which often exhibit unique structural features (e.g., specific cations or anions, large number of functional groups) that complicate predictions. These features may not be sufficiently represented in the training set, leading to anomalous SHAP values and leverage. However, it is worth noting that the prediction errors (standardized residuals) for these structurally distinct compounds were not excessively large. This may be attributed to the fact that compounds with high leverage values can represent unusual yet valid samples that the model interprets correctly, suggesting that the model is well-trained on the primary dataset and demonstrates robustness and strong generalization capabilities.

Notably, the drug-like subset contained fewer compounds beyond the h^* threshold with high prediction errors. For compounds with high leverage, certain descriptors may have a pronounced influence, but the model's predictions remain stable due to their consistency with the trained model. This highlights the thoughtful selection of descriptors, which effec-

tively describe both the main distribution and rare data points. Thus, such outliers can be considered "suitable" for the model, even if they lie outside the primary distribution.

Most data points fall within the range of standardized residuals, indicating that the models perform correctly for the majority of the dataset. In the complete dataset, 458 data points from 67 compounds exceeded the residuals boundaries, suggesting potential outliers or errors in the experimental data. In the drug-like subset, this number was smaller, with 259 data points from 37 compounds.

The distributions of absolute relative deviations (*ARD*) for the CatBoost:CDK model across different functional groups are presented in Figure 6c. The highest errors are observed for compounds containing amide and metal-associated functional groups, which exhibit higher median errors compared to most other groups. Additionally, these groups display substantial variability in prediction errors, indicating the increased complexity of modeling such compounds.

The sulfonic group, was least represented in the dataset, demonstrated the smallest median errors and a narrower range of *ARD* values, suggesting high prediction accuracy for compounds containing this functional group. Most functional groups had similar median *ARD* values; however, certain groups, such as aromatic and carboxylic acid, exhibited a wider spread of error values (Figure S6).

Recently, experimental data on solubility in scCO₂ for two drugs, Teriflunomide⁶⁵ and Rifampin,⁶⁶ were published, covering a temperature range of 308 to 338 K and pressures from 120 to 300 bar. These data were selected for external model validation. Notably, Rifampin was absent from the training dataset, while Teriflunomide was included but only at pressures up to 270 bar within the same temperature range. As a result, data points at all temperatures at 300 bar were chosen for model validation.

Predictions were made using two CatBoost:CDK with TP models—one trained on the full dataset and another on the drug-like subset (Figure 7). Their respective *RMSE* values were 0.3 and 0.2 lg(y_2). As expected, this shows that the model specialized in pharmaceuticals

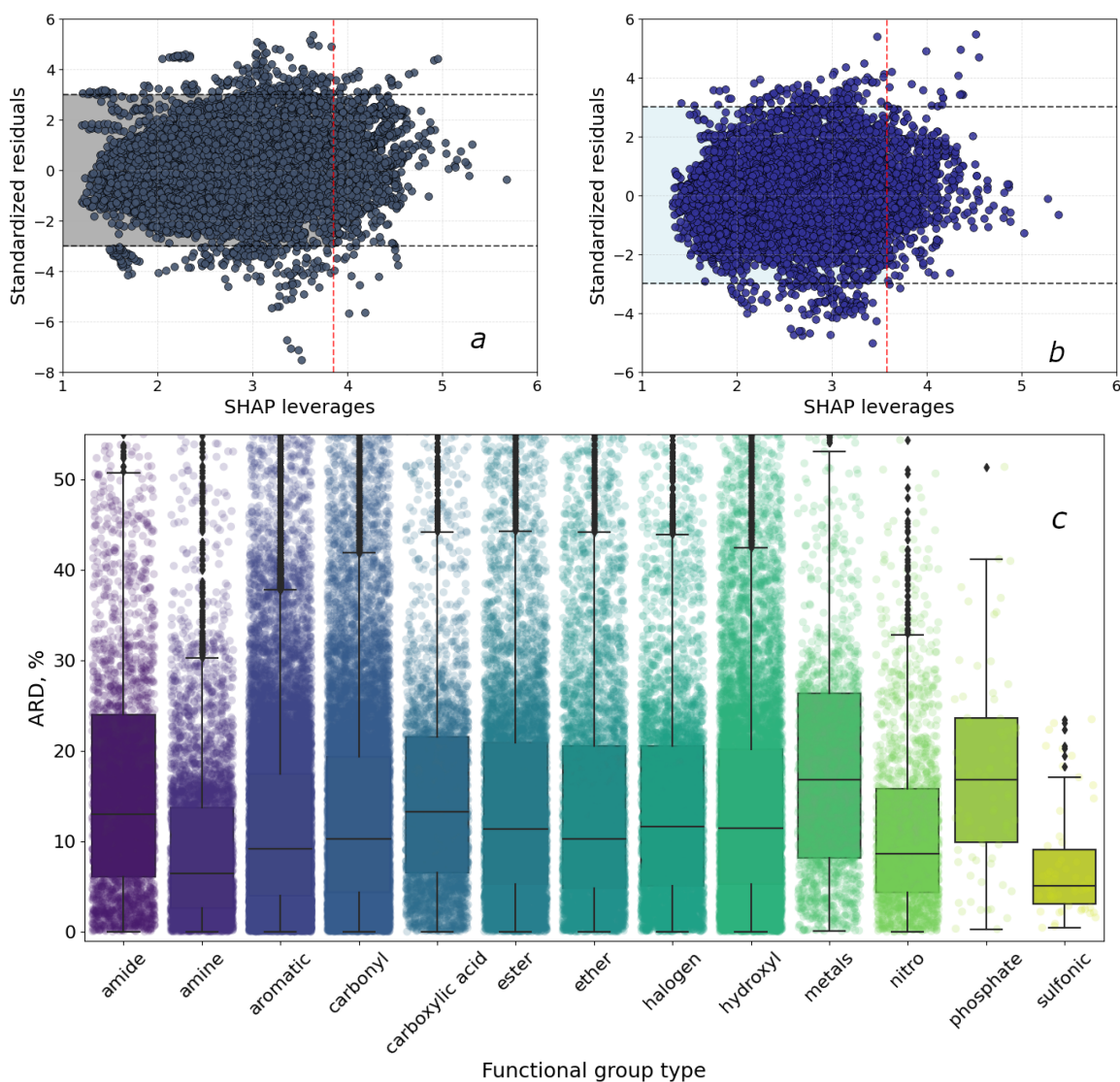


Figure 6: Applicability domain assessment using William's plot: (a) Full dataset; (b) Drug-like dataset. Black dashed lines represent the boundaries of the standardized residuals, while the red dashed line indicates the critical SHAP leverage value. (c) Distributions of absolute relative deviations of the CatBoost:CDK model across different functional groups.

outperformed the one trained on larger, more diverse dataset. The latter model had to generalize across a broad spectrum of data, which may have reduced its accuracy for drug-related predictions.

Notably, for Rifampicin, a chemical structure previously unseen by the model (Figure S7), it successfully reproduced observed in the experiment solubility crossover at a pressure of 180 bar. Overall predictions for this compound were excellent at lower pressures, but less accurate at higher ones. The structure of Teriflunomide was familiar to the model, but predictions were made for new conditions (300 bar across several temperatures), leading to slightly underestimated solubility values when comparing to the experimental ones.

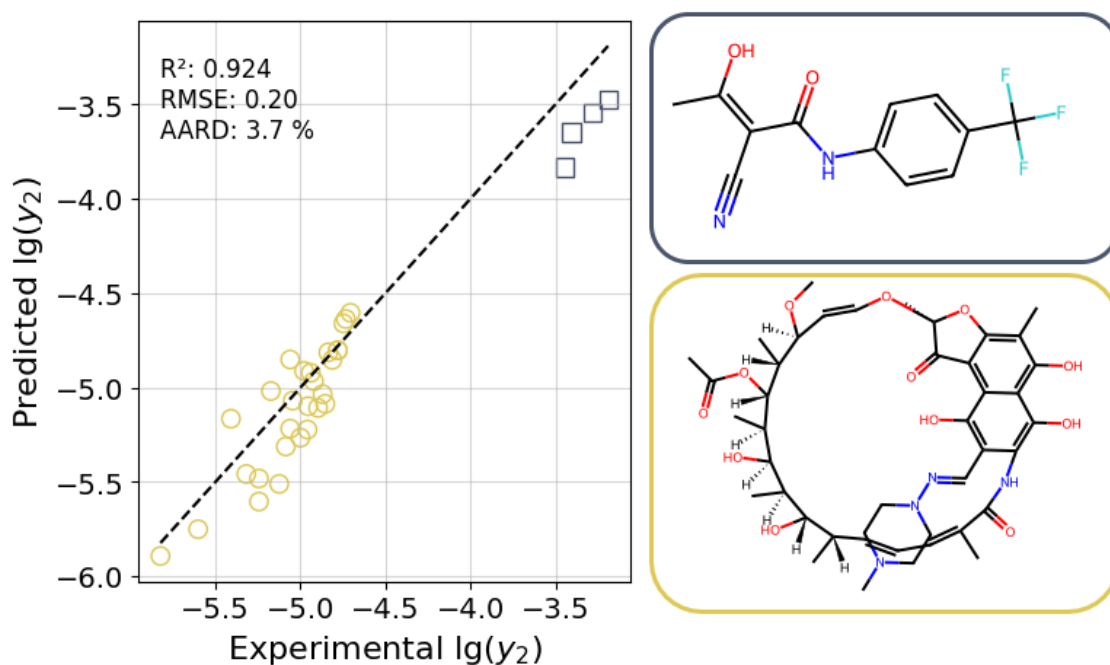


Figure 7: The dependence of predicted versus experimental solubility values for the CatBoost:CDK model with TP. Teriflunomide (dark blue) and Rifampicin (yellow).

Conclusion

The largest database on the solubility of chemical compounds in supercritical carbon dioxide has been gathered, comprising 31,975 records. A CatBoost-based model with CDK

molecular descriptors and phase transition-related parameters was developed. This model outperformed the GCNN model and demonstrated superior generalization ability for prediction of solubility in supercritical fluid. Additionally, a consensus model, derived by averaging the predictions of the CatBoost and GCNN models, achieved the lowest *RMSE* of 0.70.

The modeling in this study demonstrated that strong predictive performance on a randomly split dataset does not guarantee superior predictions under a strict evaluation protocol. This highlights the importance of thoroughly validating a model's generalization capability during its development.

The effectiveness of combining thermodynamic characteristics related to phase transitions with fundamental CDK molecular descriptors and graph-based representations for solubility prediction in scCO₂ was demonstrated. This finding opens new possibilities for enhancing predictive models of solubility in water and other solvents.

The SHAP-based approach was integrated with the leverage method to assess the models' applicability domain, confirming the robustness and generalizability of selected descriptors, even for rare compounds.

Models were deployed based on both the complete dataset and a subset restricted to drug-like molecules. These models have been integrated into a production environment, capable of providing real-time predictions for generating new compound-related data, including melting points, critical parameters, enthalpy of vaporization, and solubility in scCO₂. The models are accessible at `chem-predictor.isc-ras.ru`. To obtain a prediction, users need to input the SMILES representation of the solute (or draw its structure), along with the temperature and pressure of the supercritical CO₂.

ASSOCIATED CONTENT

Data Availability Statement

The dataset and the final machine learning models are available at: https://github.com/MDMISC/Solubility_scCO2.

Supporting Information

The Supporting Information is available free of charge at

Figure S1. Dependence of naphthalene solubility on CO₂ density at 308 K.

Figure S2. Distribution of the molecular weight of dissolved substances in scCO₂.

Figure S3. t-SNE visualization of the chemical space for compounds in the full and drug-like datasets.

Figure S4. The dependence of the melting points values predicted on the values experimentally for the CatBoost:CDK and GCNN:RDkit models.

Table S1. Comparison of the RMSE for critical properties and enthalpies of vaporization across models developed using different methods.

Table S2. Statistical parameters of the models for solubility prediction in scCO₂ with random data splitting.

Figure S5. The dependence of predicted versus experimental solubility values for the GCNN:RDKit model and tree of dissolved substances.

Table S3. Statistical parameters of the models for solubility prediction in scCO₂ with SMILES data splitting.


Figure S6. Distributions of ARD of the CatBoost:CDK model across different functional groups.

AUTHOR INFORMATION

Corresponding Authors


Dmitriy M. Makarov - G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo 153045, Russia; 


E-mail: dmm@isc-ras.ru


Yury A. Budkov - Laboratory of Computational Physics, National Research University Higher School of Economics (HSE University), Tallinskaya Street, 34, Moscow 123458, Russia; 


E-mail: ybudkov@hse.ru


Authors

Nikolai N. Kalikin - G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo 153045, Russia; 

Pavel Gurikov - Laboratory for Development and Modelling of Novel Nanoporous Materials, Hamburg University of Technology, Eißendorfer Straße 38, Hamburg, Germany; 

Sergey E. Kruchinin - G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo 153045, Russia; 

Abolghasem Jouyban - Pharmaceutical Analysis Research Center and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz 5165665931, Iran; 

Michael G. Kiselev - G. A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Ivanovo 153045, Russia; 

Author contributions

The manuscript was written with contributions from all authors. All authors have approved the final version of the manuscript.

Acknowledgement

Hyperparameter optimizations were performed on the supercomputer facilities provided by NRU HSE. The article was prepared within the framework of the project 'Mirror Laboratories' HSE University.

Notes

The authors declare no competing financial interest.

References

- (1) Veiga, G. C. d.; Mafaldo, Í. M.; Barão, C. E.; Baú, T. R.; Magnani, M.; Pimentel, T. C. Supercritical carbon dioxide technology in food processing: Insightful comprehension of the mechanisms of microbial inactivation and impacts on quality and safety aspects. *Comprehensive Reviews in Food Science and Food Safety* **2024**, *23*, e13345.
- (2) Herzyk, F.; Piłakowska-Pietras, D.; Korzeniowska, M. Supercritical Extraction Techniques for Obtaining Biologically Active Substances from a Variety of Plant Byproducts. *Foods* **2024**, *13*, 1713.
- (3) Moreira, R. C.; de Melo, R. P. F.; Martínez, J.; Marostica Junior, M. R.; Pastore, G. M.; Zorn, H.; Bicas, J. L. Supercritical CO₂ as a valuable tool for aroma technology. *Journal of Agricultural and Food Chemistry* **2023**, *71*, 9201–9212.
- (4) Yan, B.; Hu, Y.; Wang, J.; Tao, J.; Xia, S.; Yang, W.; Zhang, Y.; Chen, G.; Zhou, W.; Chen, G. State-of-the-art conceptual design of supercritical carbon dioxide as a green technology involved in bioresource conversion processes. *Chemical Engineering Journal* **2024**, 150166.
- (5) Sodeifian, G.; Sajadian, S. A.; Derakhsheshpour, R. CO₂ utilization as a supercritical

- solvent and supercritical antisolvent in production of sertraline hydrochloride nanoparticles. *Journal of CO2 Utilization* **2022**, *55*, 101799.
- (6) Sodeifian, G.; Usefi, M. M. B. Solubility, Extraction, and nanoparticles production in supercritical carbon dioxide: A mini-review. *ChemBioEng Reviews* **2023**, *10*, 133–166.
- (7) Franco, P.; De Marco, I. Nanoparticles and nanocrystals by supercritical CO₂-assisted techniques for pharmaceutical applications: a review. *Applied Sciences* **2021**, *11*, 1476.
- (8) Oparin, R. D.; Vaksler, Y. A.; Krestyaninov, M. A.; Idrissi, A.; Shishkina, S. V.; Kiselev, M. G. Polymorphism and conformations of mefenamic acid in supercritical carbon dioxide. *The Journal of Supercritical Fluids* **2019**, *152*, 104547.
- (9) Oparin, R. D.; Vaksler, Y. A.; Krestyaninov, M. A.; Idrissi, A.; Kiselev, M. G. High temperature polymorphic conversion of carbamazepine in supercritical CO₂: A way to obtain pure polymorph I. *Journal of Molecular Liquids* **2021**, *323*, 114630.
- (10) Khodov, I. A.; Belov, K. V.; Sobornova, V. V.; Dyshin, A. A.; Kiselev, M. G. Exploring the temperature-dependent proportions of lidocaine conformers equilibria in supercritical carbon dioxide via NOESY. *Journal of Molecular Liquids* **2023**, 122620.
- (11) Oparin, R. D.; Vaksler, Y. A.; Krestyaninov, M. A.; Idrissi, A.; Kiselev, M. G. Possibility of dopant morphology control in the process of polymer impregnation with pharmaceuticals in a supercritical CO₂ medium. *Journal of Molecular Liquids* **2021**, *330*, 115657.
- (12) Fathi, M.; Sodeifian, G.; Sajadian, S. A. Experimental study of ketoconazole impregnation into polyvinyl pyrrolidone and hydroxyl propyl methyl cellulose using supercritical carbon dioxide: Process optimization. *The Journal of Supercritical Fluids* **2022**, *188*, 105674.

- (13) Daneshyan, S.; Sodeifian, G. Synthesis of cyclic polystyrene in supercritical carbon dioxide green solvent. *The Journal of Supercritical Fluids* **2022**, *188*, 105679.
- (14) Campalani, C.; Amadio, E.; Zanini, S.; Dall'Acqua, S.; Panozzo, M.; Ferrari, S.; De Nadai, G.; Francescato, S.; Selva, M.; Perosa, A. Supercritical CO₂ as a green solvent for the circular economy: Extraction of fatty acids from fruit pomace. *Journal of CO₂ Utilization* **2020**, *41*, 101259.
- (15) Shekunov, B. Y.; York, P. Crystallization processes in pharmaceutical technology and drug delivery design. *Journal of crystal growth* **2000**, *211*, 122–136.
- (16) Liu, H.; Liang, X.; Peng, Y.; Liu, G.; Cheng, H. Supercritical Fluids: An Innovative Strategy for Drug Development. *Bioengineering* **2024**, *11*, 788.
- (17) Kole, E.; Jadhav, K.; Shirsath, N.; Dudhe, P.; Verma, R. K.; Chatterjee, A.; Naik, J. Nanotherapeutics for pulmonary drug delivery: An emerging approach to overcome respiratory diseases. *Journal of Drug Delivery Science and Technology* **2023**, *81*, 104261.
- (18) Knez, Z.; Cor, D.; Knez Hrnčić, M. Solubility of solids in sub-and supercritical fluids: a review 2010–2017. *Journal of Chemical & Engineering Data* **2017**, *63*, 860–884.
- (19) Alamoudi, J. A. Recent advancements toward the increment of drug solubility using environmentally-friendly supercritical CO₂: a machine learning perspective. *Frontiers in Medicine* **2024**, *11*, 1467289.
- (20) Noroozi, J.; Paluch, A. S. Microscopic structure and solubility predictions of multifunctional solids in supercritical carbon dioxide: a molecular simulation study. *The Journal of Physical Chemistry B* **2017**, *121*, 1660–1674.
- (21) Noroozi, J.; Ghotbi, C.; Sardroodi, J. J.; Karimi-Sabet, J.; Robert, M. A. Solvation free energy and solubility of acetaminophen and ibuprofen in supercritical carbon dioxide: Impact of the solvent model. *The Journal of Supercritical Fluids* **2016**, *109*, 166–176.

- (22) Sang, J.; Jin, J.; Mi, J. Solubility prediction of naphthalene in carbon dioxide from crystal microstructure. *The Journal of Chemical Physics* **2018**, *148*.
- (23) Roach, L.; Rignanese, G.-M.; Erriguible, A.; Aymonier, C. Applications of machine learning in supercritical fluids research. *The Journal of Supercritical Fluids* **2023**, *106051*.
- (24) Bouhallas, C.; Ammi, Y.; Belghait, A.; others Assessment of new semi-empirical density based model for prediction the solubility of pharmaceutical components in supercritical carbon dioxide. *The Journal of Supercritical Fluids* **2024**, *213*, 106351.
- (25) Azim, M. M.; Ushiki, I.; Miyajima, A.; Takishima, S. Modeling the solubility of non-steroidal anti-inflammatory drugs (ibuprofen and ketoprofen) in supercritical CO₂ using PC-SAFT. *The Journal of Supercritical Fluids* **2022**, *186*, 105626.
- (26) Euldji, I.; Si-Moussa, C.; Hamadache, M.; Benkortbi, O. QSPR Modelling of the Solubility of Drug and Drug-like Compounds in Supercritical Carbon Dioxide. *Molecular Informatics* **2022**, *41*, 2200026.
- (27) Budkov, Y.; Kolesnikov, A.; Ivlev, D.; Kalikin, N.; Kiselev, M. Possibility of pressure crossover prediction by classical dft for sparingly dissolved compounds in scco₂. *Journal of Molecular Liquids* **2019**, *276*, 801–805.
- (28) Kalikin, N.; Budkov, Y.; Kolesnikov, A.; Ivlev, D.; Krestyaninov, M.; Kiselev, M. Computation of drug solvation free energy in supercritical CO₂: Alternatives to all-atom computer simulations. *Fluid Phase Equilibria* **2021**, *544*, 113096.
- (29) Kalikin, N.; Oparin, R.; Kolesnikov, A.; Budkov, Y.; Kiselev, M. A crossover of the solid substances solubility in supercritical fluids: What is it in fact? *Journal of Molecular Liquids* **2021**, *334*, 115997.

- (30) Frolov, A. I.; Kiselev, M. G. Prediction of cosolvent effect on solvation free energies and solubilities of organic compounds in supercritical carbon dioxide based on fully atomistic molecular simulations. *The Journal of Physical Chemistry B* **2014**, *118*, 11769–11780.
- (31) Su, Z.; Maroncelli, M. Simulations of solvation free energies and solubilities in supercritical solvents. *The Journal of chemical physics* **2006**, *124*.
- (32) Euldji, I.; Belghait, A.; Si-Moussa, C.; Benkortbi, O.; Amrane, A. A new hybrid quantitative structure property relationships-support vector regression (QSPR-SVR) approach for predicting the solubility of drug compounds in supercritical carbon dioxide. *AIChE Journal* **2023**, *69*, e18115.
- (33) Makarov, D. M.; Kalikin, N. N.; Budkov, Y. A. Prediction of Drug-like Compounds Solubility in Supercritical Carbon Dioxide: A Comparative Study between Classical Density Functional Theory and Machine Learning Approaches. *Industrial & Engineering Chemistry Research* **2024**, *63*, 1589–1603.
- (34) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36, Publisher: American Chemical Society.
- (35) Jouyban, A. Comments on “Artificial intelligence aided pharmaceutical engineering: Development of hybrid machine learning models for prediction of nanomedicine solubility in supercritical solvent”. *Journal of Molecular Liquids* **2025**, 126979.
- (36) Span, R.; Wagner, W. A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. *Journal of physical and chemical reference data* **1996**, *25*, 1509–1596.
- (37) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **1997**, *23*, 3–25.

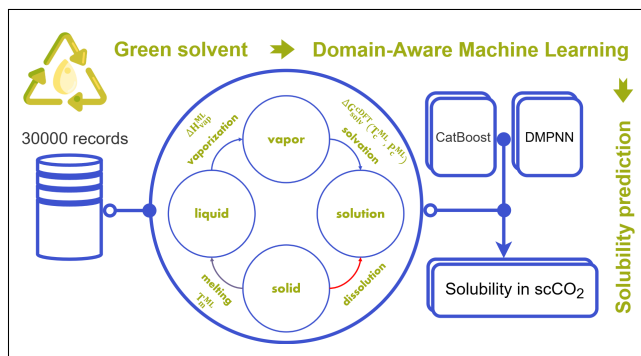
- (38) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How accurately can we predict the melting points of drug-like compounds? *Journal of chemical information and modeling* **2014**, *54*, 3320–3329.
- (39) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V. Machine learning models for phase transition and decomposition temperature of ionic liquids. *Journal of Molecular Liquids* **2022**, *366*, 120247.
- (40) Yaws, C. L. *Thermophysical properties of chemicals and hydrocarbons*; William Andrew, 2008.
- (41) Wang, R.; Chen, J.; Song, Z.; Qi, Z. Bridging Machine Learning and Redlich–Kister Theory for Solid–Liquid Equilibria Prediction of Binary Eutectic Solvent Systems. *Industrial & Engineering Chemistry Research* **2023**, *62*, 5382–5393.
- (42) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; others The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics* **2017**, *9*, 1–19.
- (43) Makarov, D.; Fadeeva, Y. A.; Shmukler, L.; Tetko, I. Beware of proper validation of models for ionic Liquids! *Journal of Molecular Liquids* **2021**, *344*, 117722.
- (44) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* **2018**, *31*.
- (45) Makarov, D. M.; Kolker, A. M. Viscosity of deep eutectic solvents: Predictive modeling with experimental validation. *Fluid Phase Equilibria* **2025**, *587*, 114217.

- (46) Makarov, D. M.; Lukanov, M. M.; Rusanov, A. I.; Mamardashvili, N. Z.; Ksenofontov, A. A. Machine learning approach for predicting the yield of pyrroles and dipyrromethanes condensation reactions with aldehydes. *Journal of Computational Science* **2023**, *74*, 102173.
- (47) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling* **2023**, *64*, 9–17.
- (48) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; others A deep learning approach to antibiotic discovery. *Cell* **2020**, *180*, 688–702.
- (49) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **2020**, *12*, 1–16.
- (50) Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J. E.; Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* **2018**,
- (51) Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E. Predictive modeling of physicochemical properties and ionicity of ionic liquids for virtual screening of novel electrolytes. *Journal of Molecular Liquids* **2023**, *391*, 123323.
- (52) Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* **2017**,
- (53) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of chemometrics* **2010**, *24*, 202–208.

- (54) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR & combinatorial science* **2007**, *26*, 694–701.
- (55) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
- (56) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vázquez-Mayagoitia, Á.; Aspuru-Guzik, A.; Vishwanath, V.; De Pablo, J. J. A machine learning workflow for molecular analysis: application to melting points. *Machine Learning: Science and Technology* **2020**, *1*, 025015.
- (57) Hunklinger, A.; Hartog, P.; Šícho, M.; Godin, G.; Tetko, I. V. The openOCHEM consensus model is the best-performing open-source predictive model in the First EUOS/SLAS joint compound solubility challenge. *SLAS Discovery* **2024**, *29*, 100144.
- (58) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. Will we ever be able to accurately predict solubility? *Scientific Data* **2024**, *11*, 303.
- (59) Gupta, R. B.; Shim, J.-J. *Solubility in supercritical carbon dioxide*; CRC press, 2006.
- (60) Balbinot Filho, C. A.; Dias, J. L.; Rebelatto, E. A.; Lanza, M. Solubility of Food-Relevant Substances in Pure and Modified Supercritical Carbon Dioxide: Experimental Data (2011–Present), Modeling, and Related Applications. *Food Engineering Reviews* **2023**, *15*, 466–490.
- (61) Khodov, I.; Belov, K.; Dyshin, A.; Krestyaninov, M.; Kiselev, M. Pressure effect on lidocaine conformational equilibria in scCO₂: A study by 2D NOESY. *Journal of Molecular Liquids* **2022**, *367*, 120525.
- (62) Oparin, R.; Krestyaninov, M.; Vorobyev, E.; Pokrovskiy, O.; Parenago, O.; Kiselev, M. An insight into possibility of chemical reaction between dense carbon dioxide and methanol. *Journal of Molecular Liquids* **2017**, *239*, 83–91.

- (63) Makarov, D. M.; Krestyaninov, M. A.; Dyshin, A. A.; Golubev, V. A.; Kolker, A. M. CO₂ capture using choline chloride-based eutectic solvents. An experimental and theoretical investigation. *Journal of Molecular Liquids* **2024**, *413*, 125910.
- (64) Skerget, M.; Knez, Z.; Knez-Hrncic, M. Solubility of solids in sub-and supercritical fluids: a review. *Journal of Chemical & Engineering Data* **2011**, *56*, 694–719.
- (65) Askarizadeh, M.; Esfandiari, N.; Honarvar, B.; Sajadian, S. A.; Azdarpour, A. Solubility of teriflunomide in supercritical carbon dioxide and co-solvent investigation. *Fluid Phase Equilibria* **2025**, *590*, 114284.
- (66) Peyrovedin, H.; Sajadian, S. A.; Bahmanzade, S.; Zomorodian, K.; Khorram, M. Studying the rifampin solubility in supercritical CO₂ with/without co-solvent: Experimental data, modeling and machine learning approach. *The Journal of Supercritical Fluids* **2025**, *218*, 106510.

TOC Graphic



Supporting Information

Improved Solubility Predictions in scCO₂ Using Thermodynamics-Informed Machine Learning Models

Dmitriy M. Makarov,^{*,†} Nikolai N. Kalikin,[†] Yury A. Budkov,^{*,†,‡} Pavel Gurikov,[¶] Sergey Kruchinin,[†] Abolghasem Jouyban,[§] and Michael G. Kiselev[†]

[†]G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Akademicheskaya Street, Ivanovo 153045, Russia

[‡]Laboratory of Computational Physics, National Research University Higher School of Economics (HSE University), Tallinskaya Street, 34, Moscow 123458, Russia

[¶]Laboratory for Development and Modelling of Novel Nanoporous Materials, Hamburg University of Technology, Eißendorfer Straße 38, Hamburg, Germany

[§]Pharmaceutical Analysis Research Center and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz 5165665931, Iran

*E-mail: dmm@isc-ras.ru; ybudkov@hse.ru

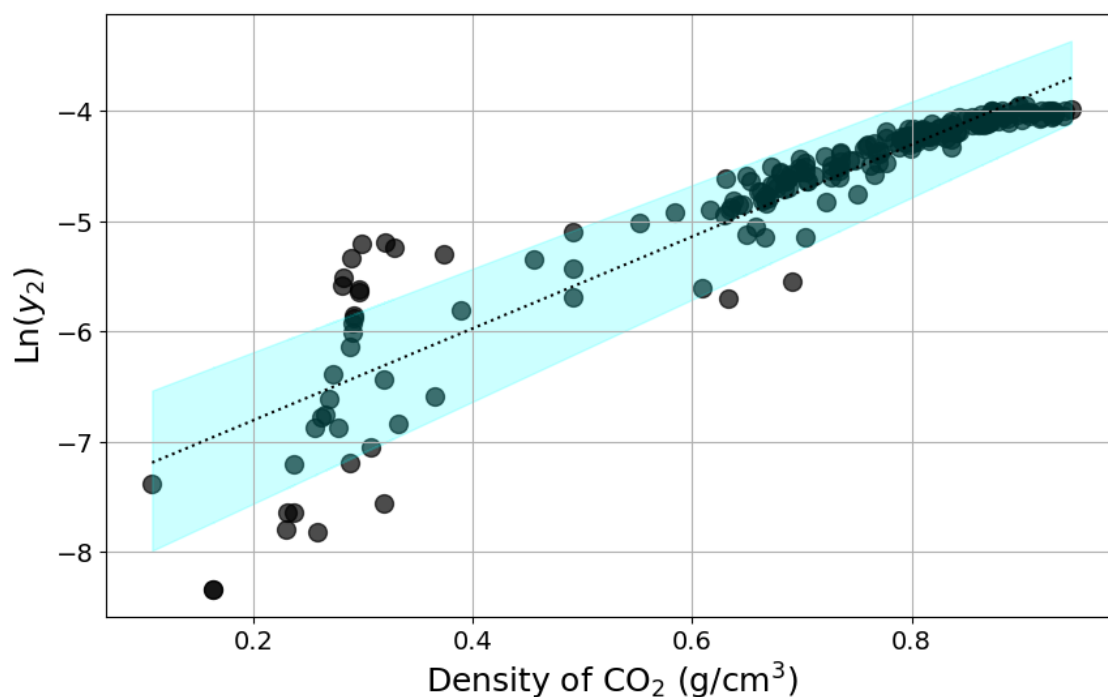


Figure S1. Dependence of naphthalene solubility on CO₂ density at 308 K. The dashed line represents the linear approximation $\ln(y_2) = f(\rho_{\text{CO}_2})$. The blue shaded area indicates the acceptable deviation range ($\pm 10\%$). All experimental points out-side this range were excluded.

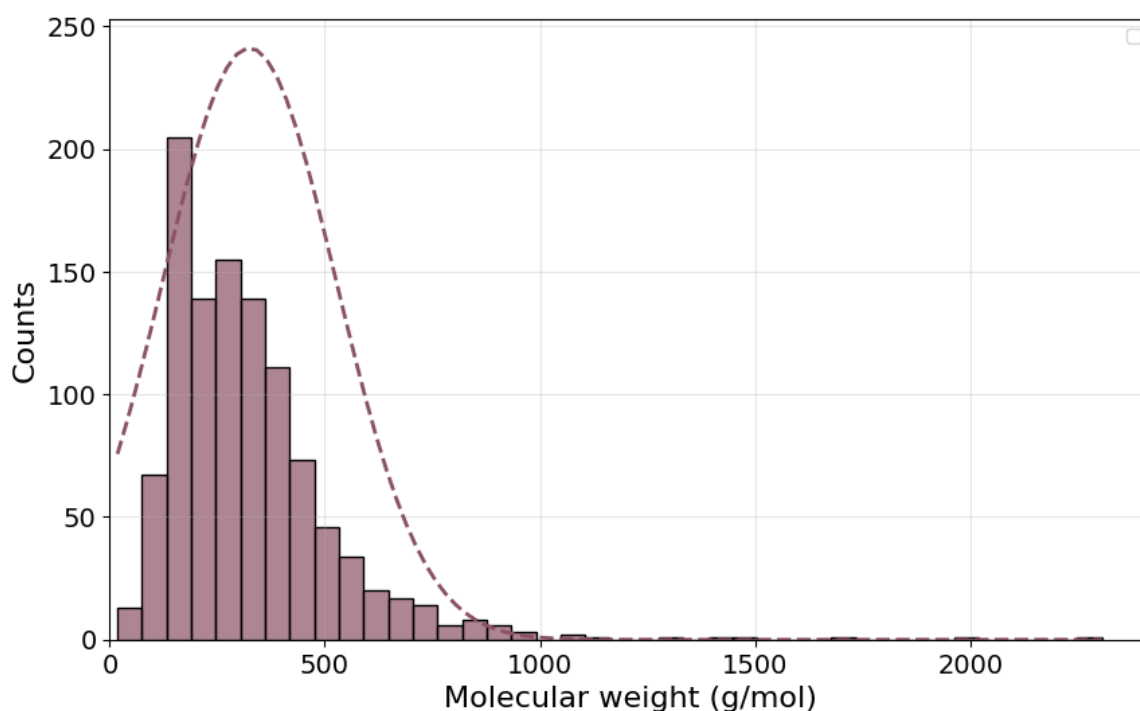


Figure S2. Distribution of the molecular weight of dissolved substances in scCO₂ in the full dataset.

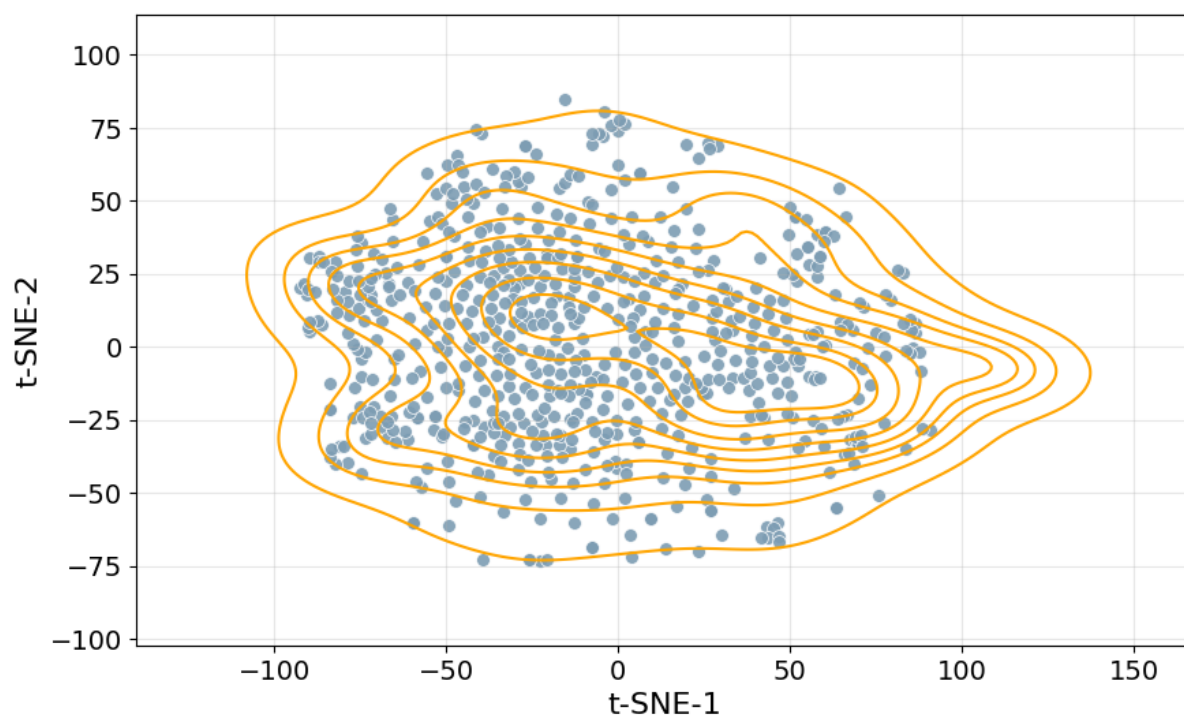
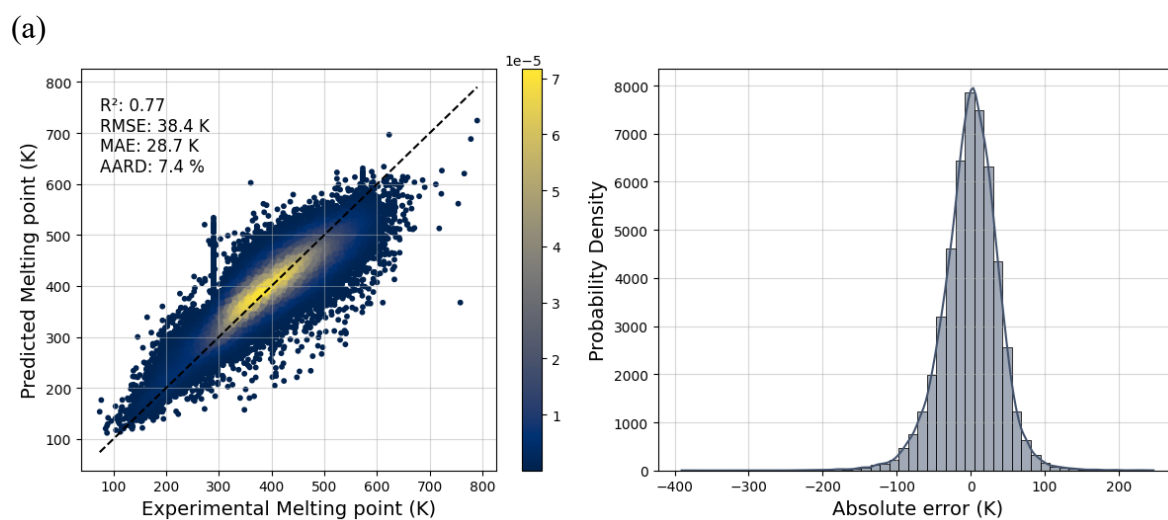


Figure S3. t-SNE visualization of the chemical space for compounds in the Full dataset (orange line) and Drug-like dataset (blue dots)



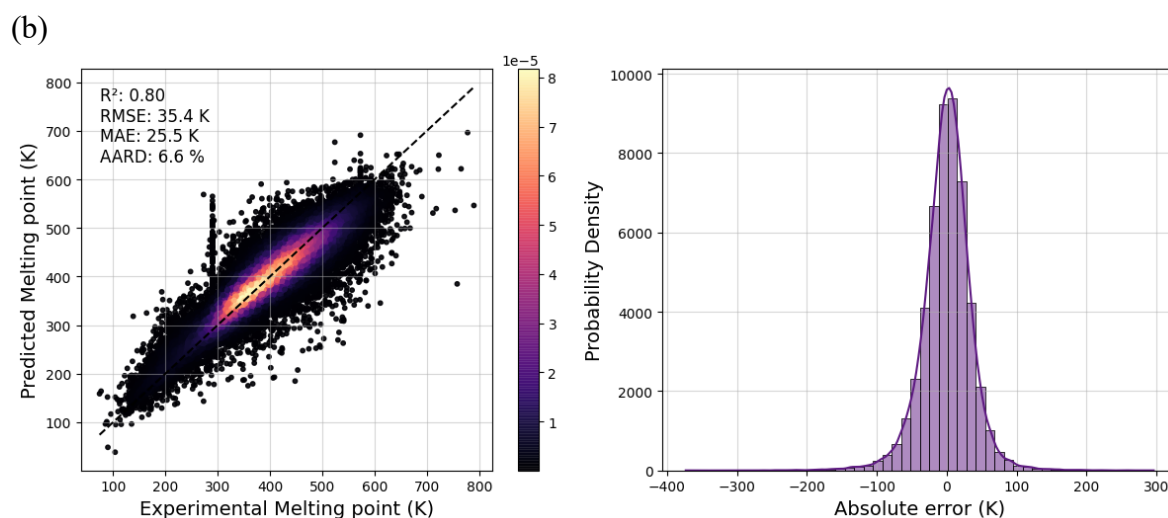


Figure S4. The dependence of the melting points values predicted on the values experimentally and histograms of absolute errors: catBoost:CDK (a); GCNN:RDkit (b).

Table S1. Comparison of the *RMSE* for critical properties and enthalpies of vaporization across models developed using different methods.*

	STL:GCNN:RDKit	MTL:GCNN:RDKit	CatBoost:CDK
T_c / K	22.8±1.9	20.9±1.4	19.2±1.9
P_c / MPa	2.5±0.6	2.4±0.3	2.2±0.4
ΔH_v / kJ/mol	4.2±0.6	4.0±0.5	3.9±0.5

*STL - single-task model; MTL - multi-task model

Table S2. Statistical parameters of the models for solubility prediction in scCO₂ with random data splitting.*

CatBoost	CDK	T_m	ΔH_v	ΔG_{solv}	ρ_{CO_2}	all
R^2	0.962 ± 0.002	0.964 ± 0.002	0.963 ± 0.002	0.963 ± 0.002	0.964 ± 0.002	0.967 ± 0.002
$RMSE$	0.278 ± 0.008	0.272 ± 0.008	0.276 ± 0.009	0.277 ± 0.009	0.271 ± 0.009	0.261 ± 0.009
$AARD$, %	5.49 ± 0.19	5.29 ± 0.19	5.42 ± 0.19	5.46 ± 0.18	5.26 ± 0.17	4.99 ± 0.16
GCNN	RDKit	T_m	ΔH_v	ΔG_{solv}	ρ_{CO_2}	all
R^2	0.970± 0.002	0.976± 0.002	0.974± 0.002	0.973± 0.002	0.979± 0.002	0.983 ± 0.002
$RMSE$	0.247± 0.009	0.236± 0.011	0.242± 0.010	0.245± 0.009	0.230± 0.007	0.223 ± 0.008
$AARD$, %	4.36 ± 0.09	4.21± 0.08	4.28± 0.10	4.31± 0.11	4.10± 0.08	3.75 ± 0.14

*The thermodynamic parameters were added to the baseline models, both individually and collectively (all).

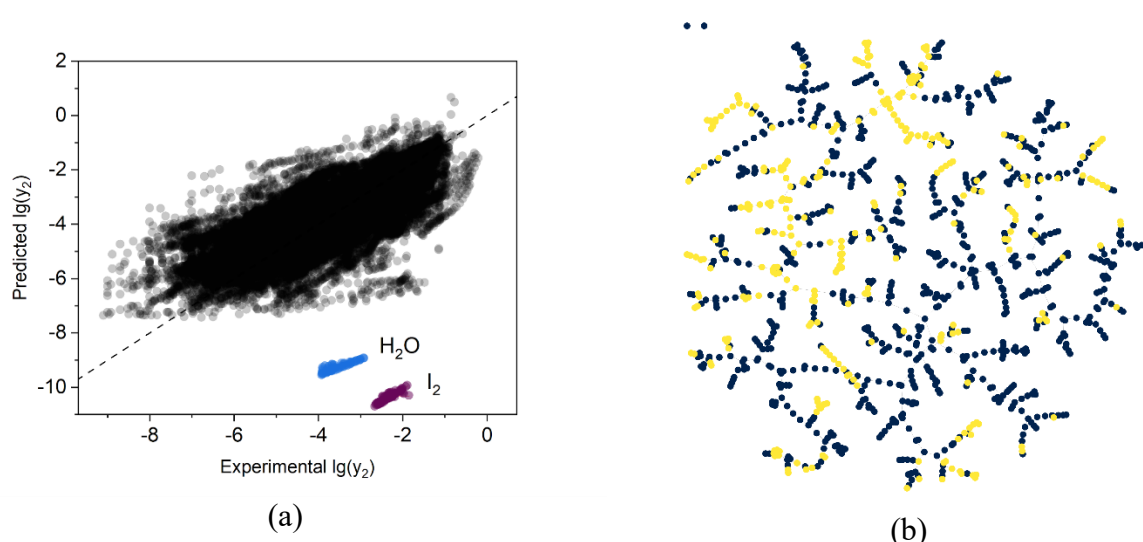


Figure S5. The dependence of predicted versus experimental solubility values for the GCNN:RDKit model (a). Tree of dissolved substances (b).

A tree of dissolved substances was constructed by visualizing the full dataset and highlighting molecular drug-like compounds based on chemical structure similarity using the TMAP algorithm. Two substances, water and iodine, were excluded from the main connected tree.

Table S3. Statistical parameters of the models for solubility prediction in scCO₂ with SMILES data splitting.*

CatBoost	CDK	T_m	ΔH_v	ΔG_{solv}	ρ_{CO_2}	all
R^2	0.68 ± 0.06	0.72 ± 0.02	0.68 ± 0.05	0.68 ± 0.06	0.67 ± 0.08	0.73 ± 0.02
$RMSE$	0.80 ± 0.08	0.75 ± 0.04	0.80 ± 0.08	0.80 ± 0.07	0.81 ± 0.09	0.74 ± 0.05
$AARD, \%$	18.5 ± 2.2	16.8 ± 1.4	17.8 ± 1.8	18.4 ± 2.0	18.5 ± 2.4	16.6 ± 1.4
GCNN	RDKit	T_m	ΔH_v	ΔG_{solv}	ρ_{CO_2}	all
R^2	0.64 ± 0.04	0.67 ± 0.03	0.65 ± 0.04	0.64 ± 0.05	0.63 ± 0.07	0.70 ± 0.02
$RMSE$	0.83 ± 0.07	0.80 ± 0.05	0.82 ± 0.05	0.83 ± 0.05	0.85 ± 0.06	0.77 ± 0.04
$AARD, \%$	19.1 ± 2.1	18.4 ± 1.3	18.9 ± 1.4	18.9 ± 1.8	19.3 ± 1.9	17.8 ± 1.4

*The thermodynamic parameters were added to the baseline models, both individually and collectively (all).

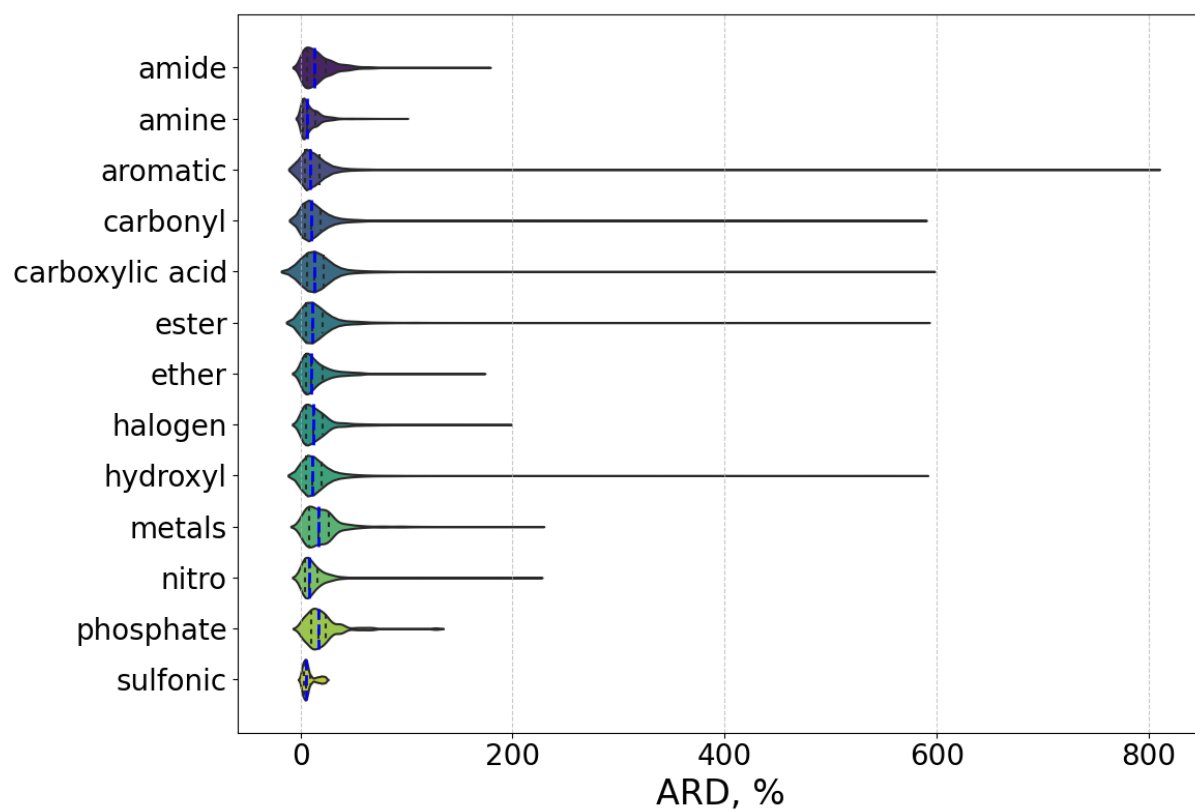


Figure S6. Distributions of ARD of the CatBoost:CDK model across different functional groups

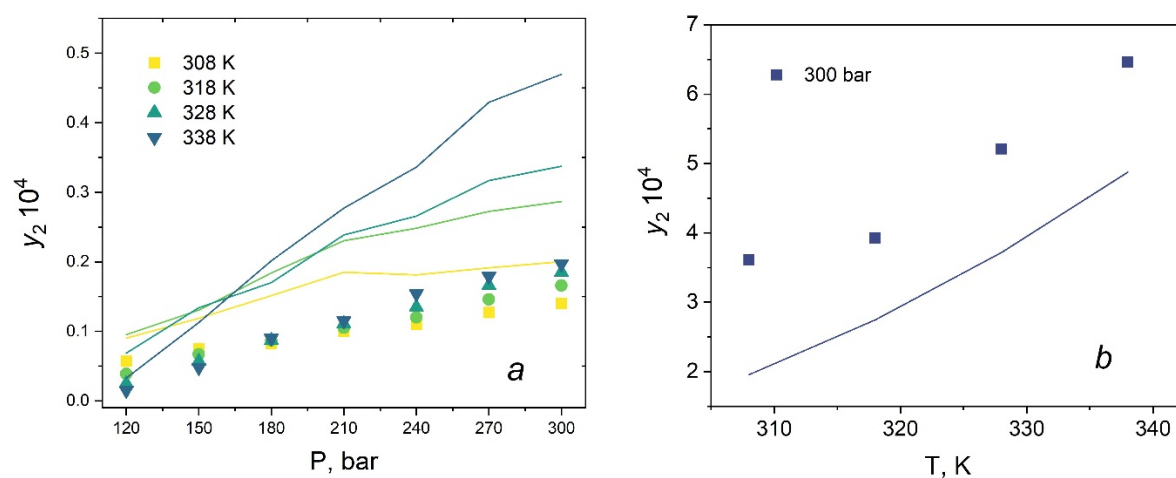


Figure S7. Temperature dependence of solubility for Rifampin (a) and Teriflunomide (b). The dots represent experimental values, while the lines indicate predictions.