Advancing Solubility Prediction by Integrating Computational Models, Multi-Solvent Systems, and Sourced PubChem Data

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften

des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M. Sc. Mushtaq Ali

aus New Delhi, India

Dekan: Prof. Dr. Martin Bastmeyer

Referent: Prof. Dr. Stefan Bräse

Korreferent: Prof. Dr. Pascal Friederich

Tag der mündlichen Prüfung : 05.05.2025

Table of Contents

List of figures

List of Tables

Honesty Declaration

This work was carried out from July 1st, 2021, until April 30th, 2025, at the Institute of Biological and Chemical Systems – Functional Molecular Systems (IBCS–FMS), submitted to the Faculty of Chemistry and Biosciences at the Karlsruhe Institute of Technology (KIT) Campus North under the supervision of Prof. Dr. Stefan Bräse.

Die vorliegende Arbeit wurde im Zeitraum vom 1. Juli 2021 bis zum 30. April 2025 am Institut für Biologische und Chemische Systeme – Funktionelle Molekulare Systeme (IBCS–FMS) der Fakultät für Chemie und Biowissenschaften am Karlsruher Institut für Technologie (KIT) Campus Nord unter der Leitung von Prof. Dr. Stefan Bräse angefertigt.

Hiermit versichere ich, Mushtaq Ali, die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet sowie Zitate kenntlich gemacht zu haben. Die Dissertation wurde bisher an keiner anderen Hochschule oder Universität eingereicht. Die „Regeln zur Sicherung guter wissenschaftlicher Praxis am Karlsruher Institut für Technologie (KIT)“ wurden beachtet.

I, Mushtaq Ali, hereby declare that I completed the work independently, without any improper assistance, and that all material published by others is properly cited. This thesis has not been submitted to any other university before this submission.

Abstract

Aqueous solubility is a key property of chemical compounds that determines their potential applications, from drug development to materials science. To further improve the current solubility prediction methods, we followed three complementary approaches: (1) added more data and representation, (2) refined model development with added features, and (3) expanded to a multi-solvent system.

In the first project, a high-throughput pipeline was developed to extract, clean, and validate solubility data from the PubChem database using SMILES (Simplified Molecular Input Line Entry System) representations. A total of 50 million SMILES were processed using the PubChemPy Python library to retrieve solubility data, which was then systematically cleaned, standardized, and evaluated against literature datasets. The final dataset consisted of 32,874 entries, categorized into salts (811 compounds) and non-salts (32,063 compounds), with 368 data points matching literature sources for validation. To utilize the extracted data and other known literature datasets, an improved model for predicting aqueous and non-aqueous solubility was developed.

For predicting aqueous solubility, I used a curated dataset compiled from four distinct sources. Our approach integrates chemical descriptors, fingerprints, and functional groups, leading to high predictive accuracy. The model was tested on the Huuskonen dataset (1,282 unique organic compounds), achieving an R^2 value of 0.92 and an MAE of 0.40, outperforming existing solubility prediction methods.

For multi-solvent solubility prediction, the model was designed to incorporate nine different solvents while treating temperature as a key variable. By accounting for solvent-specific interactions and temperature-dependent solubility variations, this approach provides a more comprehensive and adaptable framework for predicting solubility across diverse chemical environments.

Kurzfassung

Die Wasserlöslichkeit ist eine Schlüsseleigenschaft chemischer Verbindungen, die ihre potenziellen Anwendungen bestimmt, von der Arzneimittelentwicklung bis zur Materialwissenschaft. Um die aktuellen Methoden zur Vorhersage der Löslichkeit weiter zu verbessern, verfolgte ich drei sich ergänzende Ansätze: (1) Hinzufügen weiterer Daten und Darstellungen, (2) Verfeinerung der Modellentwicklung mit zusätzlichen Funktionen (3) Erweiterung auf ein Mehr-Lösungsmittelsystem.

Im ersten Projekt wurde eine Hochdurchsatz-Pipeline entwickelt, um Löslichkeitsdaten aus der PubChem-Datenbank unter Verwendung von SMILES-Darstellungen (Simplified Molecular Input Line Entry System) zu extrahieren, zu bereinigen und zu validieren. Insgesamt wurden 50 Millionen SMILES mithilfe der Python-Bibliothek PubChem verarbeitet, um Löslichkeitsdaten abzurufen, die dann systematisch bereinigt, standardisiert und anhand von Literaturdatensätzen ausgewertet wurden. Der endgültige Datensatz bestand aus 32.874 Einträgen, kategorisiert in Salze (811 Verbindungen) und Nichtsalze (32.063 Verbindungen), wobei 368 Datenpunkte zur Validierung mit Literaturquellen übereinstimmen. Um die extrahierten Daten und andere aus der Literatur bekannte Datensätze nutzen zu können, wurde ein verbessertes Modell zur Vorhersage der Wasser- und Nicht-Wasserlöslichkeit entwickelt.

Zur Vorhersage der Wasserlöslichkeit habe ich einen kuratierten Datensatz verwendet, der aus vier verschiedenen Quellen zusammengeführt wurde. Unser Ansatz integriert chemische Deskriptoren, Fingerabdrücke und funktionelle Gruppen, was zu einer hohen Vorhersagegenauigkeit führt. Das Modell wurde am Huuskonen-Datensatz (1.282

einzigartige organische Verbindungen) getestet und erreichte einen R^2-Wert von 0,92 und einen MAE von 0,40, womit es bestehende Methoden zur Vorhersage der Löslichkeit übertraf.

Für die Vorhersage der Löslichkeit mehrerer Lösungsmittel wurde das Modell so konzipiert, dass es neun verschiedene Lösungsmittel einbezieht und dabei die Temperatur als Schlüsselvariable behandelt. Durch die Berücksichtigung solcher Lösungen, spezifischer Wechselwirkungen und temperaturabhängiger Löslichkeitsschwankungen bietet dieser Ansatz einen umfassenderen und anpassungsfähigeren Rahmen für die Vorhersage der Löslichkeit in unterschiedlichen chemischen Umgebungen.

1. Introduction

1.1 Importance of the property solubility

Aqueous solubility is a fundamental property essential for investigating and applying chemical compounds across various scientific disciplines, including chemistry, biology, and materials sciences. In materials science, the solubility of chemicals affects how materials are constructed and determines the effectiveness of active functional components. An example of the importance of solubility can be described for the development of organic semiconductors. Organic semiconductor processing is a challenging task due to the uniqueness of the compounds, particularly when they are insoluble in typical solvents. For these materials, traditional solution-based techniques are not practical, and it is necessary to use more complex and costly alternatives. One method is vapor deposition, where organic semiconductor materials are evaporated in a vacuum and deposited onto a substrate. This is typically done to produce the thin, uniform films necessary for high-performance devices, including organic solar cells, light-emitting diodes (OLEDs), and field-effect transistors. However, vapor deposition is not only expensive but also requires intricate equipment and stringent process control to ensure proper deposition rates, substrate temperatures, and film uniformity. Additionally, the high capital cost of deposition systems and the requirement for specialized facilities render this process prohibitively expensive for large-scale manufacturing. Consequently, even though vapor deposition can be pertinent for some high-performance organic semiconductors, it is, nonetheless, less feasible for general commercial use compared to the solution-processing method, which, in turn, fuels the quest for new, scalable fabrication techniques for such materials.

Being able to dissolve a chemical substance also plays a crucial role, for example, in designing novel drugs, as solubility has a significant impact on the bioavailability of drugs and their distribution[2]. Solubility is one of the most essential features that directly defines the extent and speed of the drug, depending on different doses and bioavailability. For instance, solubility evaluates the drug's potential to interact with other drugs, either for effective synergy or for interacting in a harmful way when administered together. When a drug is absorbed into the bloodstream, it can then be distributed to different body parts, delivering its active ingredients to cells to achieve the desired result. Usually, drugs, when in a free state, can act on one or more organs and are typically cleared from the systemic circulation through various mechanisms, such as renal, hepatic, or pulmonary expiration. A drug that exhibits poor solubility and tends to accumulate at a specific site in the body may lead to local toxicity or raise concerns about potential respiratory effects.[3] To begin with, the solubility of a drug, that is its ability to dissolve in a liquid, must be ensured so that the intestine can take it up. For this reason, hydrophobic drugs are delivered using particles or droplets to enhance their ability to disperse within the fluid and thus reach their targets more quickly. While the solubility of commercially available building blocks is often known, at least for a few solvents, the investigation of newly designed compounds needs the determination of the compounds' solubility if the solubility affects the desired application. [4]

Predicting solubility across diverse chemical systems remains a significant challenge in computational chemistry. Recent advancements in machine learning and deep learning have led to improved predictions of solubility. However, existing approaches primarily focus on single-solvent systems, limiting their applicability in multi-solvent environments where solubility behavior is more complex. Furthermore, while molecular fingerprint-based and graph-based models have demonstrated promising results, they often lack domain adaptation capabilities, experimental validation, and explainability**.** To address these gaps, this study aims to develop a robust multi-solvent solubility prediction framework that incorporates solute-solvent-specific features and leverages Message Passing Neural Networks (MPNN) and Hybrid MPNN techniques to enhance accuracy, transferability, and interpretability.

1.2 Availability of solubility databases

Accurate determination and prediction of solubility are essential in the fields of drug development, materials science, and chemical process design. Solubility is one of the critical physicochemical properties studied using machine learning models for accurate prediction, but its availability in reliable, comprehensive, and easily accessible databases remains a challenge. This section provides an overview of the current landscape of solubility databases, highlighting the available datasets in the literature and discussing the limitations and gaps that hinder their applicability in advanced research.

Several databases have been developed to compile solubility data for various classes of compounds in different solvent types—some of the well-known sources of solubility data are listed below.

1. AqSolDB[6] is a comprehensive, curated dataset comprising aqueous solubility data for 9,982 unique compounds. Developed by the Autonomous Energy Materials Discovery (AMD) research group, this dataset consolidates information from nine publicly available aqueous solubility datasets. The dataset reports aqueous solubility data under standard conditions, with a temperature of 25°C (298.15 K) and a pH that varies but is generally close to neutral (~7.0). It can be downloaded from the GitHub repository AqSolDB. [7]

2. Alex Avdeef [8] conducted intrinsic solubility measurements across multiple laboratories using both the CheqSol and saturation shake-flask (SSF) methods. The research analyzes a dataset comprising 233 intrinsic solubility values (log S_0) obtained via the CheqSol method for 145 drug-like molecules, as well as 838 log S_0 values determined primarily through the SSF method for 124 of these molecules. The solubility values in this study range from -1.0 to -10.6 log molar units, with an average of -3.8 log molar units. Intrinsic solubility measurements are typically performed at standard laboratory conditions, which are

approximately 25 °C (298.15 K). Intrinsic solubility refers to the solubility of the neutral form of a compound. The full text of the study is available for download in PDF format from the ADMET and DMPK journal's website.

3. The Aquasol Database of Aqueous Solubility, Version 5, compiled by Samuel H. Yalkowsky and Richard M. Dannenfelser, [9] is a comprehensive collection of solubility data for organic non-electrolytes. It includes over 10,000 solubility values for more than 4,000 different compounds, with recommended solubility values provided for the 20-40°C temperature range. Specific pH values are not explicitly mentioned in the provided information. The Aquasol Database is available through the University of Arizona's research portal. [10]

4. The Handbook of Aqueous Solubility Data, Second Edition (2010) by Samuel H. Yalkowsky, Yan He, and Parijat Jain, [11], is a comprehensive compilation of solubility data for organic compounds. It includes over 18,000 solubility values for more than 4,600 compounds, covering pharmaceuticals, pollutants, nutrients, herbicides, pesticides, and other industrial and energy-related substances. The solubility data are primarily reported at 25 °C (298.15 K), which is standard for aqueous solubility measurements. The pH values vary depending on the compound's characteristics. For weak acids and bases, solubility can be pH-dependent, and the handbook provides data across different pH ranges to account for these variations. The handbook is available for purchase through various retailers.

5. The PHYSPROP Database, developed by Peter Howard and William Meylan in September 1999, is a comprehensive collection of physical and chemical property data for over 41,000 chemicals. The database includes vapor pressure data for over 2,000 chemical compounds, measured over a temperature range of 20°C to 30°C. Specific pH values are not explicitly mentioned in the provided information. However, intrinsic solubility measurements are

typically performed at standard laboratory conditions, approximately 25°C (298.15 K). The PHYSPROP Database is included in the U.S. Environmental Protection Agency's (EPA) Estimation Programs Interface Suite™ (EPI Suite™). EPI Suite™ is a free software package that provides access to various environmental and chemical property estimation tools, including the PHYSPROP Database.

6. Huuskonen[12] developed predictive models for aqueous solubility using molecular connectivity, shape, and atom-type electrotopological state (E-state) indices. The dataset comprised 1,297 organic compounds. The aqueous solubility values (log S, where S is in mol/L) were measured at temperatures ranging from 20 to 25°C. The pH conditions for these measurements were not specified in the study. The AqSolDB GitHub repository includes curated aqueous solubility data, citing Huuskonen's work at this source, GitHub. [13]

7. Bergström [14] developed computational models to predict the aqueous solubility of drug-like molecules. The study utilized a dataset of 85 structurally diverse drug-like compounds with experimentally determined intrinsic solubility values. These solubility measurements were conducted at a controlled temperature of 25 °C (298.15 K). The intrinsic solubility, referring to the solubility of the neutral form of the compounds, thus, specific pH values were not a primary variable in this context. Data can be found at the reference website DLS-100 solubility data. [15]

8. Delaney [16] developed a computational model titled "ESOL: Estimating Aqueous Solubility Directly from Molecular Structure,". The study compiled a dataset of 1,128 organic compounds with experimentally determined aqueous solubility values. These values are expressed as log solubility in moles per liter (log S). The specific temperature at which the solubility measurements were conducted is not detailed in the abstract. Typically, such measurements are performed under standard laboratory conditions, often at a temperature of approximately 25°C (298.15 K). The study does not specify the pH conditions under which

the solubility data were obtained. Aqueous solubility measurements are typically conducted in pure water or buffered solutions at a neutral pH, although exact conditions may vary. The raw data file is available at the DeepChem GitHub repository. [17]

9. The FreeSolv database, developed by Mobley and Guthrie[18], provides a curated collection of experimental and calculated hydration free energies for small neutral molecules in water. This resource is invaluable for understanding solvation properties and interactions of molecules in aqueous environments. The hydration free energies in the FreeSolv database are determined under standard conditions, typically at 298.15 K (25°C) and pH 7. These conditions are standard for aqueous solubility measurements and are consistent across the dataset. The database is available for download from the Mobley GitHub repository [19]

10. BigSolDB [20] is an extensive dataset containing 54,273 experimental solubility values for 830 unique compounds across 138 solvents, including organic solvents and water. The data spans temperatures from 243.15 K to 403.15 K at atmospheric pressure. The dataset does not specify pH values, as it includes both organic solvents and water, where pH may not be a relevant factor. Download the BigSolDB dataset in CSV format from the given reference [20]

11. Cui et al., 4, titled "Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning," a dataset of 9,943 compounds encoded by molecular fingerprints. These models were further evaluated using 62 recently published novel compounds, demonstrating superior performance over existing tools and human experts. The solubility measurements were conducted at standard laboratory conditions, typically at 25°C (298.15 K). The study does not specify the pH conditions under which the solubility data were obtained. The authors have made the dataset and related resources available for public access at the given source.[21]

12. Boobier[22] developed computational models to predict the solubility of compounds in both organic solvents and water. The researchers curated five open-access solubility datasets for

their analysis, focusing on neutral solutes in single-component solvents. These datasets include solubility measurements in various organic solvents and water. The solubility data were collected from the Open Notebook Science Challenge aqueous solubility dataset and the Reaxys database. The study does not specify the exact temperature and pH conditions under which the solubility measurements were conducted. However, it is common practice to measure solubility at standard laboratory conditions, typically at 25°C (298.15 K) and at a neutral pH, unless otherwise specified. The datasets compiled for this study are publicly available and can be accessed through the supplementary information provided in the original publication.[23]

13. Panapitiya[24] developed a general model capable of predicting the solubility of a broad range of organic molecules. The researchers utilized the largest currently available solubility dataset, which encompasses a diverse range of organic molecules with experimentally determined solubility values. The study does not specify the exact temperature and pH conditions under which the solubility measurements were conducted. However, it is common practice to measure solubility at standard laboratory conditions, typically at 25 °C (298.15 K) and at a neutral pH, unless otherwise specified. The dataset used in this study is publicly available and can be accessed through the supplementary information provided in the original publication. [25,26,27,6]

14. Sumeen Lee[5] "Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks" evaluates solubility prediction using two approaches: molecular fingerprints with physicochemical descriptors and graph convolutional neural networks (GCNs). The dataset comprises solubility measurements of various solute-solvent pairs, conducted at standard laboratory conditions (~25 °C or 298.15 K), although the exact temperature is not explicitly specified. Intrinsic solubility is usually calculated at neutral pH. The dataset can be obtained from the GitHub repository. [28]

Apart from literature data sources for solubility, various online platforms such as OCHEM, PubChem, eChem, and similar databases provide extensive solubility data, enabling researchers to access, analyze, and integrate experimentally measured solubility values. Some of them are listed below.

1. The Online Chemical Modeling Environment (OCHEM) is a comprehensive web platform designed for data storage, model development, and the publication of chemical information. It offers access to a vast collection of chemical and biological data, including experimentally measured properties such as water solubility. Each record typically includes information about the measured property, the associated chemical compound, and experimental conditions such as temperature and pH. For water solubility data, the temperature is often specified, and pH conditions are noted when available. OCHEM contains over 4 million records for 685 properties, including water solubility, collected from more than 21,000 sources. For detailed information, please visit the OCHEM website. [29]

2. eChemPortal is an open-source chemical property database developed by the Organisation for Economic Co-operation and Development (OECD). Solubility data were extracted after applying the filters “experimental studies” and “water solubility”. Detailed information is available on the echem portal. [30]

3. ChemSpider is a free chemical structure database providing access to over 130 million structures from hundreds of data sources. It offers a variety of search options, including structure, identifier, and property searches. While ChemSpider provides a wealth of information on chemical compounds, including physical and chemical properties, it does not consistently include solubility data for all compounds. The availability of solubility information varies depending on the specific compound and the data sources integrated into ChemSpider. Detailed information is available on the ChemSpider portal. [31]

4. PubChem, another source of data maintained by the National Center for Biotechnology Information (NCBI), is a free and publicly accessible chemical database. It contains over 111

million chemical compounds, providing extensive data on chemical structures, properties, bioactivity, safety, and patents. The database supports API integration through RESTful services, allowing programmatic access for searching, retrieving, and analyzing chemical information. Researchers and developers can utilize PubChem's API for seamless data extraction and computational applications. All data is freely available via the PubChem website.

The four described open-access databases provide solubility data, as well as data on other chemical and physical properties. While they are very user-friendly, they are often compiled from many different sources using different methods, which makes the data inconsistent and less reliable compared to the literature data.

Although these databases have been developed, some challenges still exist, as highlighted in the following section: Data Fragmentation. Results are mostly fragmented in the literature, hence making it challenging to locate comprehensive datasets for specific compound-solvent systems. Inconsistencies and Reliability: Different experimental conditions (e.g., temperature, pH, purity) and variations in methodology lead to inconsistencies in reported solubility values. Limited Coverage: Most databases focus on specific types of compounds. For instance, organic molecules in water form complex mixtures, ionic liquids, and less-studied solvents that are untreated. Lack of Modern Experimental Data: Most datasets rely on older studies, leaving gaps for modern materials and emerging applications. Accessibility Issues: While some resources are freely accessible, such as PubChem and ChemSpider, others require subscription access, including IUPAC SDS.

The lack of reliable and sufficient solubility data impacts the performance of the quality prediction model. Advanced machine learning and generative AI have massive potential for solubility prediction, but require immense, high-quality datasets for training the models. The lack of these often leads to incorrect predictions or an inability to optimize processes such as drug formulation and material synthesis.

To overcome these challenges, efforts are needed to collaborate on the development of experimental protocols, standardize reporting, and establish open-access, centralized databases where experimental and computational solubility data can be integrated. This should be complemented by data curation algorithms and AI-driven predictions to fill the gaps in experimental data, with a strong emphasis on data availability and validation to ensure reproducibility in solubility studies.

1.3 Options to experimentally define solubility

Solubility measurements can be categorized into two main types: thermodynamic solubility and kinetic solubility. Each type represents a different aspect of a solute's interaction with a solvent, and this distinction determines the most suitable experimental methods for accurate solubility assessment.

Thermodynamic solubility, which represents the maximum solubility of a compound at equilibrium, can be measured using methods such as High-Performance Liquid Chromatography (HPLC), the Gravimetric method, and Nephelometry-based light-scattering analysis. Kinetic solubility, on the other hand, refers to the rate at which a compound dissolves and is typically assessed using techniques like UV-Vis spectroscopy, the shake flask method, and real-time turbidity measurements.

High-performance liquid chromatography (HPLC). It is a powerful analytical method used for determining the solubility of compounds, especially in the pharmaceutical and chemical industries. HPLC can be used to quantify the concentration of a dissolved substance in a solution, which is directly related to its solubility. In the development of the experiment to determine solubility, a saturated solution of the analyte is prepared by adding a small volume of the respective solvent to a sufficient amount of the compound, thereby yielding a suspension. The mixture is briefly swirled to achieve preliminary homogenization and then sonicated for at least 24 hours to facilitate complete solubilization. The solutions are allowed to equilibrate at room temperature for a specific period to ensure sufficient interaction between the compound and the solvent, with intermittent swirling to maintain equilibrium. Once equilibration is achieved, the mixture undergoes high-speed centrifugation to separate undissolved particles. The resulting supernatant, containing only the dissolved compound, is carefully collected and passed through a fine membrane filter to remove any residual particulates or interfering matter.

During the preparation of the saturated solution, a set of standard solutions is prepared by serial dilution of a stock solution, ensuring that the resulting concentrations fall within the detection range of the UV/Vis detector of the HPLC system. These standards are essential for constructing a calibration curve, which will later be used to quantify the solubility of the test compound. Once the saturated solution reaches equilibrium, the supernatant is carefully filtered to remove undissolved particles and injected into the HPLC system. Chromatograms are recorded, and the retention time and peak area of the compound of interest are noted. The same procedure is applied to the standard solutions to generate a calibration curve by plotting peak area versus concentration. Based on the data points of the calibration standards, a calibration line is obtained through linear regression, providing the solubility value, typically expressed in milligrams per milliliter (mg/mL). To ensure reliability, the experiment is repeated multiple times, and the results are averaged.

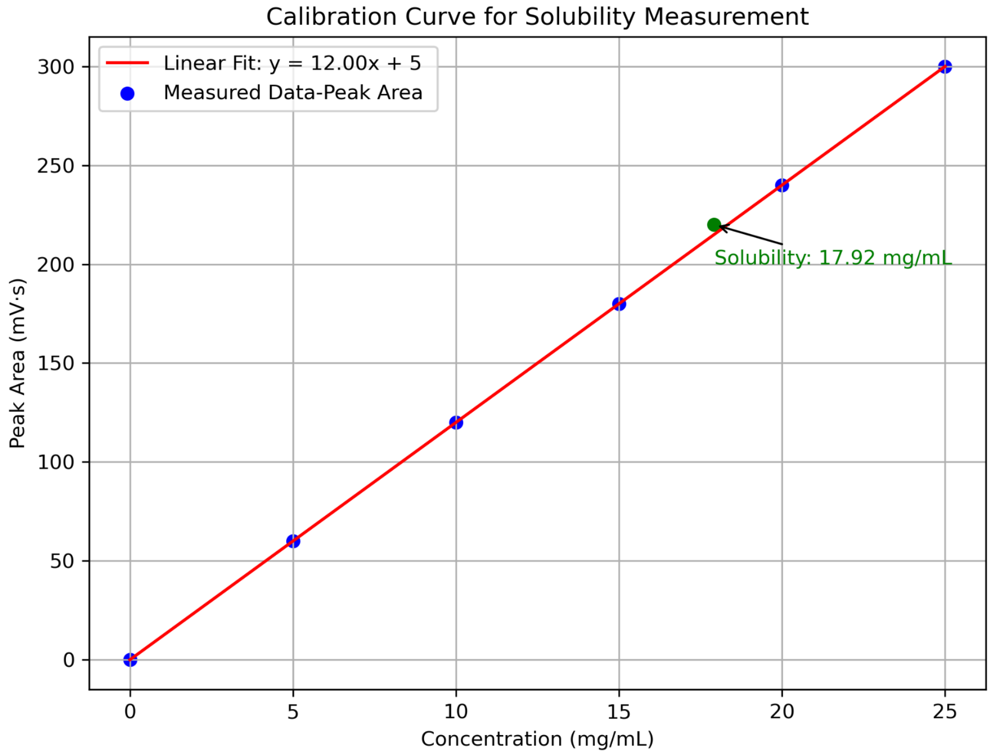


Figure 1. Calibration curve (red) obtained by linear regression of the data points from external standards. Using the fit equation, the concentration of an unknown sample (represented by the green dot) can be calculated. The y-axis shows the Peak Area measured by the HPLC.

Assume we measure the peak area of a saturated solution and find it to be 220 milli-absorption units per second.

From the regression equation

(1) $y=12x+5$ (y is the Peak area 220 m-AU)

x is the concentration of the compound in mg/mL (unknown compound which has a peak area of 220 a.u)

Rearrange the regression equation to calculate the x, which is the solubility

(2) $x = \frac{y-5}{12}$

Substitute the value of y = 220

(3) $\quad x = \frac{220-5}{12}$

$= 17.92$ mg/mL

The solubility of the unknown compound is approximately 17.92 mg/mL.

Gravimetric analysis: This method involves determining the amount of a compound dissolved by measuring the mass of the undissolved residue. A known amount of the compound is weighed on an analytical balance and introduced into a measured volume of the solvent in a clean and dry container. The mixture is continually stirred to optimize the compound's exposure to the solvent for efficient dissolution. Stirring is performed in a temperature-controlled environment to ensure reproducibility and account for the temperature-dependent solubility of the compound. It is then allowed to equilibrate for many hours to ensure complete dissolution to the maximum extent.

When this solution is at equilibrium, filtration of the undissolved residue and the saturated solution is carried out using a previously weighed filter paper or membrane filter. Filtration must be carried out without loss of any of the dissolved material. The filtering paper containing the undissolved residue is then dried to constant weight in a desiccator or drying oven to remove residual solvent. The filter paper is then weighed again on an analytical balance, with the mass of the undissolved residue being determined by subtracting the initial weight of the filter paper.

The concentration of the dissolved compound in the solvent is determined by subtracting the mass of the undissolved residue from the initial mass of the compound and then dividing this value by the volume of the solvent used. This indicates the solubility of the compound, typically expressed in units of milligrams per milliliter (mg/mL). The operation is repeated several times to establish the accuracy and reproducibility of the result.

The data obtained using this method must be repeated under various conditions to assess the effects of temperature, solvent type, and stirring duration on the compound's solubility. This gravimetric method is particularly suited for compounds that do not require complex analytical instrumentation, delivering reliable results under controlled conditions. The compound's properties influence the choice of analytical instruments, as different compounds may require specific methods based on their solubility, chemical structure, and sensitivity.

Nephelostar[32] is a specialized nephelometric plate reader designed for high-throughput solubility analysis by detecting light scattering from suspended particles in solution. It is widely used in pharmaceutical, chemical, and materials science research to assess compound solubility under various conditions efficiently. Unlike traditional solubility measurement techniques such as HPLC or UV-Vis spectroscopy, which rely on direct quantification of dissolved molecules, nephelometry measures the turbidity of a solution by analyzing the intensity of scattered light. The principle behind this technique is that higher turbidity indicates a greater presence of undissolved particles, while a clear solution suggests higher solubility.

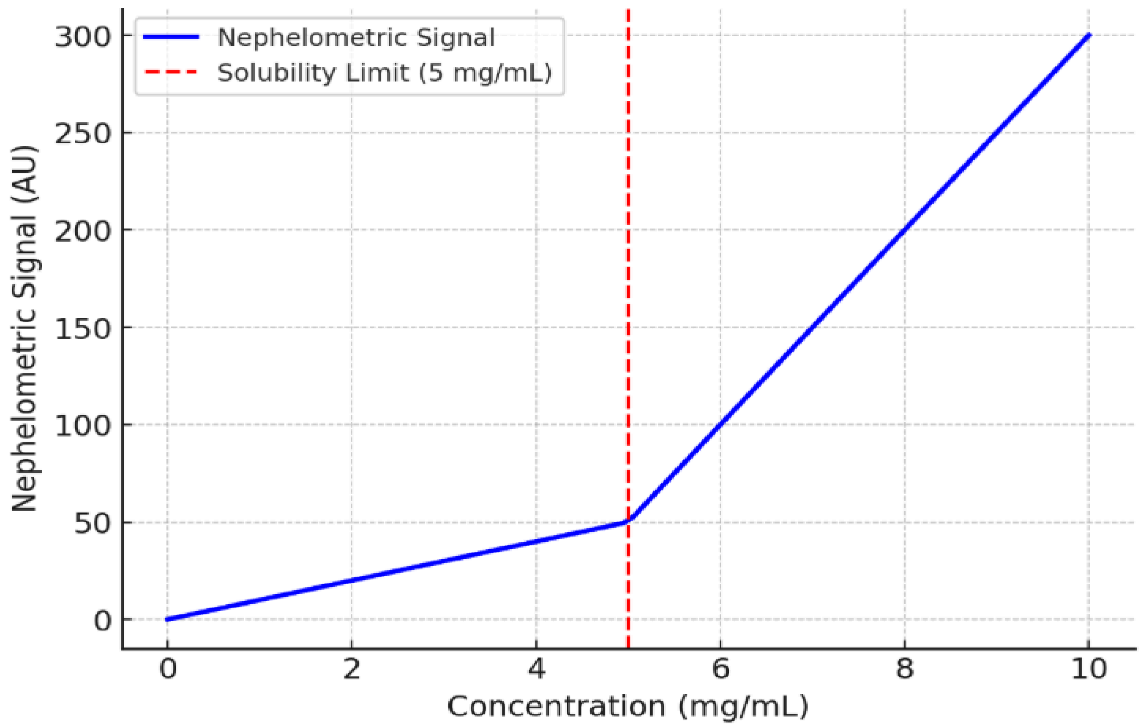


Figure 2. Graph determined using nephelometry (NepheloStar). The X-axis represents the compound concentration (mg/mL), and the Y-axis represents the Nephelometric signal (light scattering intensity)

At low concentrations, the solution remains clear, producing a minimal nephelometric signal due to the absence of undissolved particles. As the concentration increases, the solution remains transparent until it reaches the solubility limit, indicated by the red dashed line. At this critical point, excess solute begins to precipitate out, leading to the formation of undissolved particles that scatter light significantly. This results in a sharp increase in the nephelometric signal, marking the transition from a fully dissolved state to a suspension. The solubility value is determined as the concentration just before this rapid rise in light scattering, representing the maximum amount of solute that can dissolve under the given conditions.

Each solubility measurement technique has its advantages and disadvantages. For example, Gravimetry requires a large amount of substance. HPLC is a tedious process and also demands high amounts of substance. Nephelometry is a fast technique; however, it measures kinetic solubility rather than thermodynamic solubility.

The ability to predict solubility accurately could offer numerous benefits to the current process of manually determining the solubility of new compounds. Firstly, predictions can help design potentially interesting molecules, preventing the synthesis of compounds that do not exhibit the expected properties. Secondly, time- and resource-consuming, laborious measurements can be replaced by suitable predictions (at least in part).

1.4 Theoretical calculations

1.4.1 Traditional methods

A general understanding of solubility prediction was gained through the general solubility equation,[33], which incorporates solute-solvent interactions via parameters such as activity coefficients, and enables predictions under various conditions. Hildebrand and Hansen solubility parameters[34,35] describe the properties of solvents and solutes. Hansen developed this idea by introducing three solubility parameters that take into account hydrogen bonding, polarity, and dispersion forces, which depend on the different properties of the compound. In effect, this set of parameters provides a more detailed illustration of the various reactions between the solute and solvent that contribute to solubility. For example, δh captures the contribution of hydrogen bonds, which is a key for predicting the solubility of polar compounds, while δp accounts for dipole-dipole interactions. The resulting Hansen solubility spheres will enable a mapping of solubility in a three-dimensional space, providing a valuable tool for both predicting and visualizing compatibility between solvents and solutes.

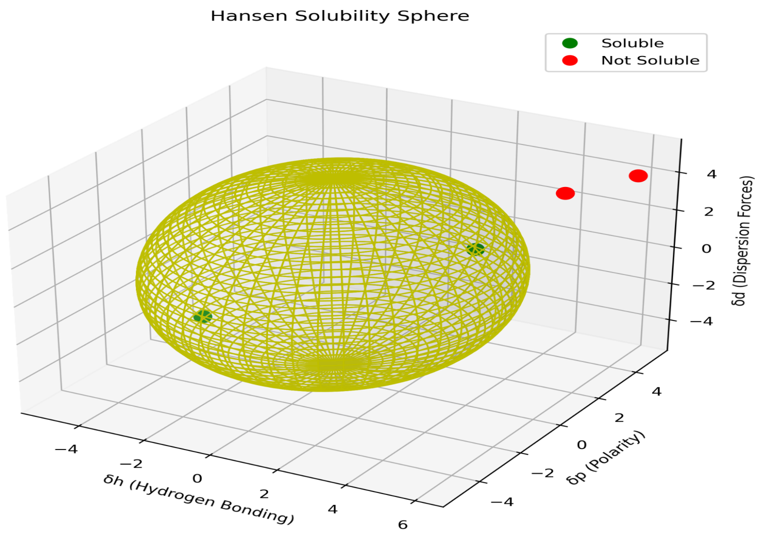


Figure 3. Plot representing solubility prediction using Hansen solubility parameters. It includes a 3D sphere model with labeled axes for hydrogen bonding (δh), polarity (δp), and dispersion forces (δd), illustrating solubility mapping and solvent-solute interactions. The yellow transparent sphere represents the solubility space. Green points (inside sphere) indicate soluble compounds. Red points (outside the sphere) indicate insoluble compounds.

In (Conductor-like Screening model for real solvents) COSMOS-RS[45,46] quantum chemistry, tools are used in the form of a conductor-like screening model for real solvents to forecast solubilities in complex systems. It considers the environment, electrostatic interactions, and molecular structure. These are subsequently extended to real solvent environments, including very critical factors such as electrostatic interactions, hydrogen bonding, and van der Waals forces. Unlike all previous methods, COSMO-RS also encompasses information on the molecular structure of the solute and the environmental conditions surrounding it, such as temperature and pressure. It does so by providing an all-encompassing method for predicting solubility in multicomponent systems, yielding information on how variations in the molecular environment influence solubility.

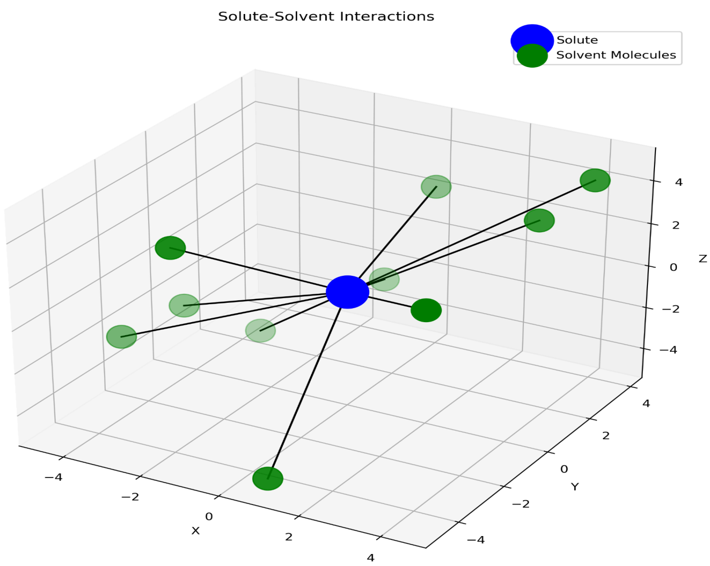


Figure 4. 3D Plot representing solute-solvent interactions: A solute molecule (blue) is surrounded by solvent molecules (green), with black lines representing potential interactions between the solute and solvent molecules. This model illustrates the spatial arrangement of molecules and highlights the importance of molecular interactions in determining solubility. X-axis: Represents the horizontal position of the molecules along the first dimension. Y-axis: Represents the horizontal position of the molecules along the second dimension. Z-axis: Represents the vertical position of the molecules along the third dimension.

1.4.2 ML methods

Recently, significant progress has been made in the field of solubility prediction, with a focus on utilizing statistical methods supported by rigorous data analysis to uncover hidden patterns and correlations within solubility data. Statistical models use advanced algorithms to identify key factors influencing solubility, thereby enhancing predictive accuracy and machine learning methods[36]. These data-driven approaches leverage machine learning techniques to process vast datasets, recognize complex patterns, and generate accurate predictions. The majority of solubility predictions in the past have been made using overly simplistic models that only considered a small number of molecular characteristics, such as molecular weight, log P (partition coefficient), and others, including atom counts and ring counts. 37 These early models were insightful, but they frequently failed to account for the complex interactions that determine solubility. Traditional experimental methods for determining solubility, while accurate, are resource-intensive and time-consuming, with throughput limitations. Computational approaches, such as machine learning, emerge as strong alternatives that provide rapid and efficient predictions based on the chemical and physical properties of compounds. This section describes the application of machine learning in the prediction of solubility, describing methodologies, strengths, limitations, and future directions of this technique.

Machine learning is a part of artificial intelligence that enables computers, with the help of statistical methods, to learn patterns from data and make predictions or decisions without prior programming. Regarding solubility, ML utilizes large chemical datasets to build predictive models, which can estimate the solubility of compounds under various conditions across different solvents.

It typically follows the supervised learning approach for developing an ML-based solubility prediction model, where a dataset of compounds with experimentally determined solubility is used to train the model. Once trained, the model predicts the solubility of new compounds by analyzing their molecular descriptors, structure, and fingerprint.

1.4.3 ML techniques

Linear regression: It was among the first ML models applied to predict solubility due to its simplicity and interpretability. It assumes a linear relationship between molecular descriptors (such as molecular weight, logP, and hydrogen bond donors/acceptors) and solubility. However, solubility is often influenced by complex, nonlinear interactions, which limit the accuracy of linear models.[16]

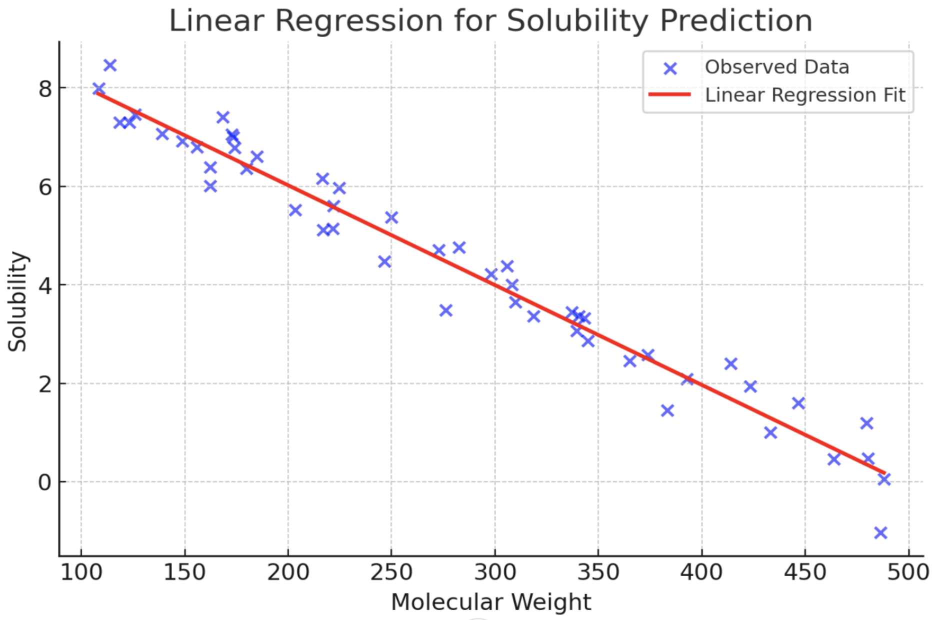


Figure 5. In this scatter plot, the x-axis represents molecular weight, and the y-axis indicates solubility. It demonstrates a linear relationship between solubility and molecular weight.

Logistic regression: It is typically used for binary classification, such as predicting whether a compound is soluble (1) or insoluble (0) based on molecular descriptors like molecular weight (MW) and LogP.

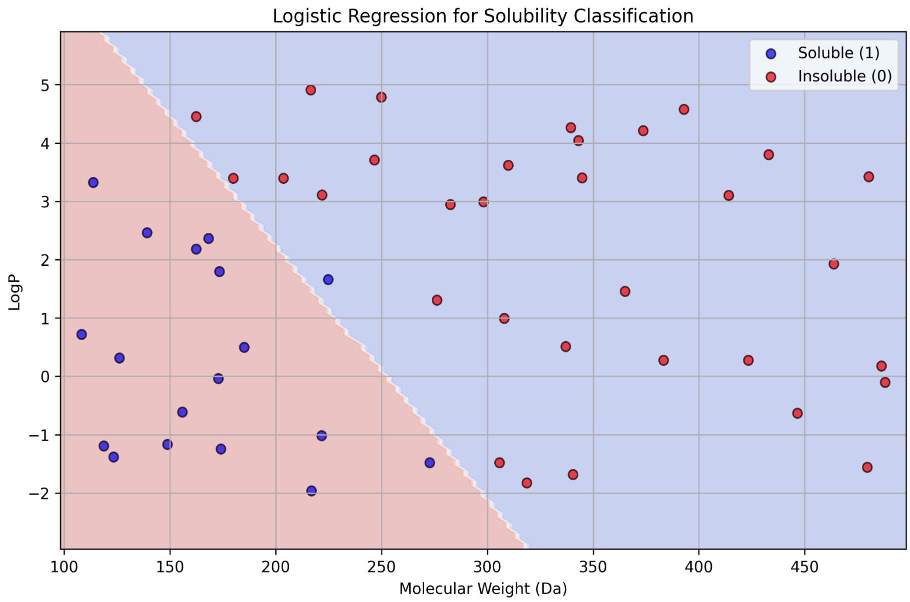


Figure 6. In this scatter plot, the x-axis represents molecular weight, and the y-axis shows LogP. It displays a logistic regression model that classifies compounds as soluble (blue) or insoluble (red) based on molecular weight and LogP. The decision boundary (shaded region) separates the two classes, illustrating how solubility prediction can be achieved using molecular descriptors.

Decision trees and random forests: Decision trees partition the dataset into subsets based on feature thresholds, creating a tree-like structure for predictions. Random forests improve upon decision trees by aggregating the results of multiple trees, reducing overfitting, and enhancing accuracy. These methods are particularly effective for solubility prediction since solubility often exhibits nonlinear relationships with molecular and physical properties. Tree-based methods help capture complex interactions between features.

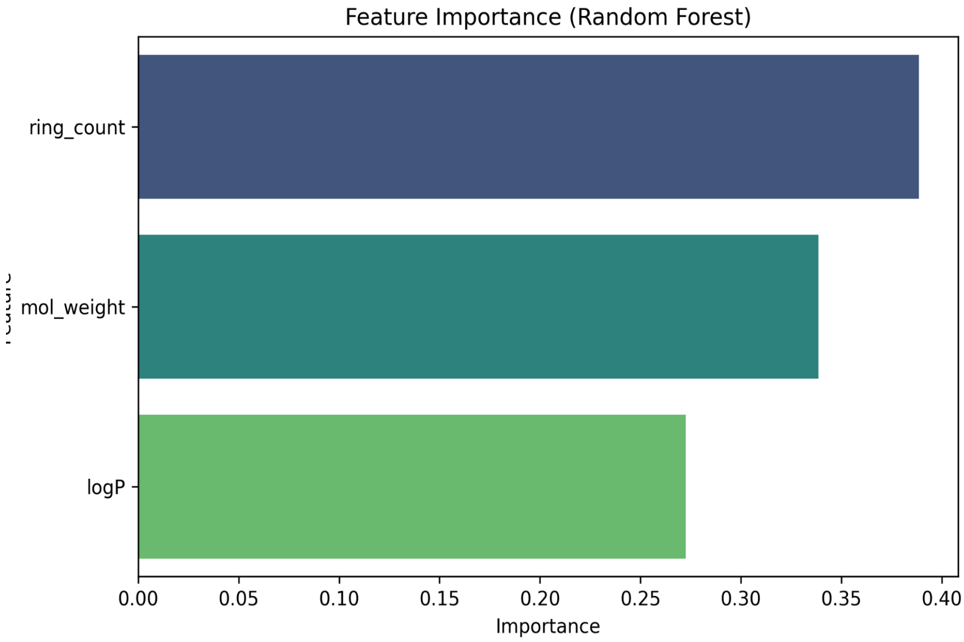


Figure 7. In this bar chart, the x-axis represents the importance, and the y-axis shows the descriptors random forest support to identify which features are more important in predicting the model.

Support vector machines: SVM is a supervised learning algorithm used to analyze data for classification and regression. The primary goal of SVM is to find the **optimal hyperplane** that best separates different classes (in classification) or makes the most accurate prediction (in regression).

Classification with SVM: SVM is often used for classification tasks, where the goal is to separate data points into different categories. For solubility prediction, the task could be classifying compounds as **soluble** or **insoluble** based on various features (e.g., chemical structure, molecular weight, temperature, etc.).

Regression with SVM: In solubility prediction, SVM can also be applied for regression tasks, where the algorithm predicts a continuous value (e.g., the exact solubility of a compound in a given solvent). [22]

Neural networks: Their capability to capture complicated nonlinear relationships makes neural networks, particularly deep learning architectures, popular in solubility predictions. MLP and CNN have been used in solubility predictions by extracting features from molecular graphs and other structural representations.[22]

Graph neural networks (GNNs) are a class of deep learning models specifically designed to operate on graph-structured data, unlike traditional neural networks, which operate on Euclidean data (e.g., text, images). GNNs operate on non-Euclidean data, such as molecules, social networks, and recommendation systems. They leverage graph connectivity to learn representations by recursively aggregating and transforming node features from neighbors.

Type of GNN architecture:

1) Basic graph neural network: The fundamental GNN model follows a message-passing framework where node features are updated by aggregating information from their neighbors.

2) Message passing neural networks (MPNNs): MPNNs are a generalization of GNNs, explicitly designed for graph-based learning. It consists of a message function that computes messages from neighboring nodes. Aggregation function: Aggregates messages from neighbors. Update function: Updates the node representation.[38]

3 Hybrid GNNs: Hybrid GNNs integrate multiple types of GNN architectures or combine GNNs with other machine learning models. Some hybrid approaches include:

Graph convolutional networks (GCN) + MPNN: Using GCN layers for global feature extraction and MPNN for local message passing.[39]

Attention-based GNNs: Integrating transformer-like self-attention mechanisms to assign importance to different neighbors dynamically.

GNN + Recurrent neural networks (RNNs): Used in time-dependent applications such as traffic prediction or molecular dynamics simulations.

Ensemble method or combined methods: It works based on gradient boosting, including XGBoost and LightGBM, and combines the predictions of multiple weak learners to produce a robust model. These methods are highly effective for tabular datasets containing molecular descriptors.[40]

In the case of aqueous solubility, the use of molecular descriptors, such as diverse molecular descriptors characterizing the chemical structure or its properties, has been shown to give the most accurate prediction so far. Sorkun et al.[36] achieved notable predictive accuracy using a dataset with 4,399 unique data points, a feature set comprising 123 descriptors, and a consensus model that included Random Forest, XGBoost, and an artificial neural network. Despite the development of improved models over the past few years, accurate solubility prediction remains a significant challenge, and further improvements to the current models are urgently needed.

Experimental solubility data often contain inconsistencies due to variations in measurement conditions, techniques, and lab errors. Noisy data can mislead machine learning models, thereby reducing prediction accuracy and making model generalization more challenging. The vast number of possible molecular structures creates challenges in ensuring the training dataset is representative of real-world compounds. Limited or biased training data can lead to poor model performance when predicting solubility for novel compounds that fall outside the training distribution.

Solubility spans several orders of magnitude, making it difficult for models to learn a single predictive pattern. Highly soluble and poorly soluble compounds may require different molecular features to be accurately predicted, complicating model development. Addressing these challenges requires high-quality datasets, robust preprocessing techniques, and advanced model architectures that can generalize well across diverse molecular structures.

1.5 Cheminformatics as a basis for solubility calculation

The choice of input data format has a significant impact on the performance of ML models. Common representations include:

1.5.1 Molecular descriptors:

The molecular descriptor is typically a numerical value that describes the extent of a specific structural or physicochemical property of molecules. These are essential inputs to machine learning models for solubility prediction, as the algorithms understand and make predictions of molecular behavior based on such features. The classification of such descriptors can broadly be categorized into three major classes: physicochemical, topological, and quantum chemical descriptors.

The physicochemical descriptors include basic molecular properties, such as molecular weight, polar surface area, logP (hydrophobicity), and hydrogen bond donor/acceptor counts. These descriptors capture the fundamental chemical characteristics driving solubility, such as polarity and intermolecular interactions.

Topological descriptors describe the connectivity and arrangement of atoms in a molecule. Examples include, among others, molecular graph indices such as the Wiener and Zagreb indices, as well as adjacency matrices. The kind of descriptor supplies information regarding the dimensional structure and complexity of a molecule that directly impacts its solubility.

The quantum chemical descriptors are calculated using quantum mechanics and include dipole moments, the energy gap between the HOMO and LUMO, and electronic distribution. These are computationally expensive descriptors, but they provide very detailed information concerning the reactivity of molecules and their interactions.

Selecting the proper molecular descriptors is a crucial step in developing accurate predictive models. The modern approach utilizes automated feature selection techniques or employs deep learning methodologies that directly learn the representation of molecules; hence, molecular descriptors are foundational in cheminformatics and predictive modeling.

1.5.2 SMILES

SMILES stands for Simplified Molecular Input Line Entry System and is a notation mainly used to represent chemical structures in a compact, human-readable form. It is a linear string format that describes the structure of a molecule by encoding its connectivity, arrangement of atoms, bonds, and stereochemistry. For example, ethane is represented as "CC," and water is defined as "O." SMILES strings are instrumental in cheminformatics and machine learning applications, as they are much easier to parse and process computationally.

One of the significant advantages of SMILES is its simplicity and efficiency in encoding even complex molecules. It offers a standard format for the storage and exchange of molecular information, therefore enabling the creation of big chemical databases. Furthermore, it includes notations that can represent stereochemistry, thereby distinguishing between isomers. Extended versions, like canonical SMILES and isomeric SMILES, ensure consistency and uniqueness of representation.

SMILES have played a significant role in the development of cheminformatics tools, mainly in machine learning-based predictions of solubility. Representation of molecules as SMILES enables researchers to employ text-based methods, such as NLP and graph-based representation methods, for meaningful pattern extraction. The strings derived from the use of SMILES are then used to create molecular descriptors and embeddings, which enable the construction of predictive models not only in solubility but also in drug discovery and material science

1.5.3 Canonical SMILES

One molecule can have multiple valid SMILES strings, depending on the order in which atoms and bonds are described. To ensure consistency, a Canonical SMILES is generated — this is a unique and standardized representation for each molecule. Canonical SMILES are produced by applying a specific algorithm that orders atoms and bonds in a deterministic manner, ensuring that the same molecule always results in the exact canonical string, regardless of how it was initially drawn or processed. Importantly, canonical SMILES ensure a consistent representation of chemical structures across databases and tools (PubChem, RDKit, Open Babel), and are essential for data deduplication. Different tools or sources might provide the same molecule in different orders, but canonicalization collapses these into a single, unique entry. Furthermore, they enhance reproducibility in cheminformatics workflows, ensuring that model inputs are consistently maintained and reliable.

Example of Ethanol:

valid SMILES: CCO

Canonical smiles: CCO

Example of Dimethyl Ether:

valid SMILES: COC

Canonical smiles: COC

The canonical form ensures that the same molecule is represented consistently across different systems, thereby eliminating ambiguity in how molecules are described.

1.5.4. Graph representation

One of the most common ways to model a molecule as a graph is through a graph representation, in which atoms are represented as nodes and chemical bonds are represented as edges. This form captures topological and structural relationships within a molecule, making it useful for several applications, such as predicting the solubility of substances.

Components of molecular graphs:

Nodes (Vertices): Each node of the graph is related to one atom in the molecule. The node features typically include atomic properties such as atomic number, valence electrons, hybridization state, formal charge, and aromaticity.

Edges represent the bonds between atoms. The bond-specific feature includes the following: Bond type: single, double, triple, aromatic, bond length, bond polarity

An adjacency matrix is a **square matrix** used to represent a graph. If a molecule has n atoms, the matrix is of size n×n, where each element Aij represents the bond between atoms i and j.

Aij=1 if there is a bond between atoms i and j

Aij=0 if there is no bond

For weighted graphs, the matrix elements can represent **bond types** (such as single, double, and triple bonds) instead of just 1s and 0s.

1.5.5 Fingerprint description/descriptors

Molecular fingerprints are the compact, numerical representations of chemical structures widely used in cheminformatics. They are an essential tool for similarity searching, clustering, and property prediction tasks, including solubility prediction. Fingerprints encode specific structural features, functional groups, or patterns within a molecule into machine-readable formats, making them highly effective for computational analysis.

There are several types of molecular fingerprints. Some of the well-known structural key-based fingerprints include MACCS keys and PubChem fingerprints. These are binary vectors where each bit represents either the presence or absence of some predefined substructure. Path-based fingerprints, such as Daylight and RDKit fingerprints, capture molecular features based on linear paths of atoms. On the other hand, circular fingerprints, including Morgan fingerprints (ECFP), consider the local environment of each atom through iterative extension to atomic neighborhoods and have proven highly suitable for machine learning tasks. Pharmacophore fingerprints encode three-dimensional chemical features such as hydrogen bond donors and hydrophobic regions, while 3D fingerprints represent the spatial arrangement of atoms, often used in receptor-ligand interaction studies.

The other benefits of fingerprints are that they are compact, computationally efficient, and versatile for both 2D and 3D structures. Fingerprints in the solubility prediction constitute a vital input feature of machine learning models that capture structural information relevant to solubility. They enable similarity-based predictions by comparing molecular structures and identifying key substructures that influence solubility, thereby enhancing the accuracy of the predictive model.

2. Aim of the work

The goal of this research is to develop a cutting-edge, end-to-end solubility prediction platform by compiling a high-quality and diverse dataset from various sources, ensuring the development of a robust and reliable predictive model. A significant component of this work involves building an automated data pipeline that systematically collects, processes, and cleans solubility data from the largest open chemical database, PubChem. This pipeline ensures the integration of big data, resolving inconsistencies, problems with missing data, and data redundancy issues, thereby increasing the reliability of solubility predictions.

Additionally, this work utilizes state-of-the-art machine learning techniques to enhance the predictive power of solubility models, enabling them to generalize effectively across the richness of chemical spaces. Unlike traditional solubility prediction models that focus primarily on aqueous solubility, this study extends its scope to predict solubility in other solvents by incorporating solute-solvent interactions. Temperature is also introduced as an essential variable, allowing for the prediction of solubility under various thermodynamic conditions.

By combining data-driven approaches with fundamental chemical principles, this work aims to bridge the gap between theoretical predictions and experimental observations through a more general and practical approach to solubility estimation. The developed framework has broad implications in drug discovery, materials science, and chemical engineering, where accurate prediction of solubility is crucial for formulation design, reaction optimization, and process development. Ultimately, this work contributes to the further development of computational solubility modeling, enabling more effective applications in both academic and industrial contexts.

3. Main part

In this section, an advanced solubility prediction framework is presented, leveraging curated datasets and state-of-the-art machine learning techniques. Our approach emphasizes careful data preprocessing, ensuring high-quality molecular representations for improved model accuracy. Additionally, we conduct a thorough applicability domain analysis to assess the generalization capability of our models and provide error handling for incorrect molecular inputs. To validate the performance, we compare various machine learning models on benchmark datasets, ensuring robust and interpretable predictions of solubility. Through this systematic approach, we aim to enhance the reliability and applicability of solubility prediction in real-world scenarios.

3.1 Data curation and preprocessing

To ensure high-quality and unbiased solubility predictions, we first merge data from multiple curated sources, creating a diverse and comprehensive dataset. Next, we eliminate duplicate entries within the training set to prevent redundancy and bias. Finally, we remove any molecules in the test set that also appear in the training set, ensuring a fair evaluation and preventing data leakage. This preprocessing pipeline enhances data integrity and model reliability.

Table 1: Details of the five different datasets used to generate a unique dataset to train and test the model

Dataset	Authors	Dataset Size	Use	Duplicates	Unique	Reference
A	BNN Lab	900	Training	4	898	[22]
B	Gihan	11862	Training	261	11724	[24]
C	Xian Zeng	9942	Training	347	9750	[4]
D	Sorkun	6154	Training	471	5907	[36]
E	Huuskonen	1291	Testing	18	1282	[12]

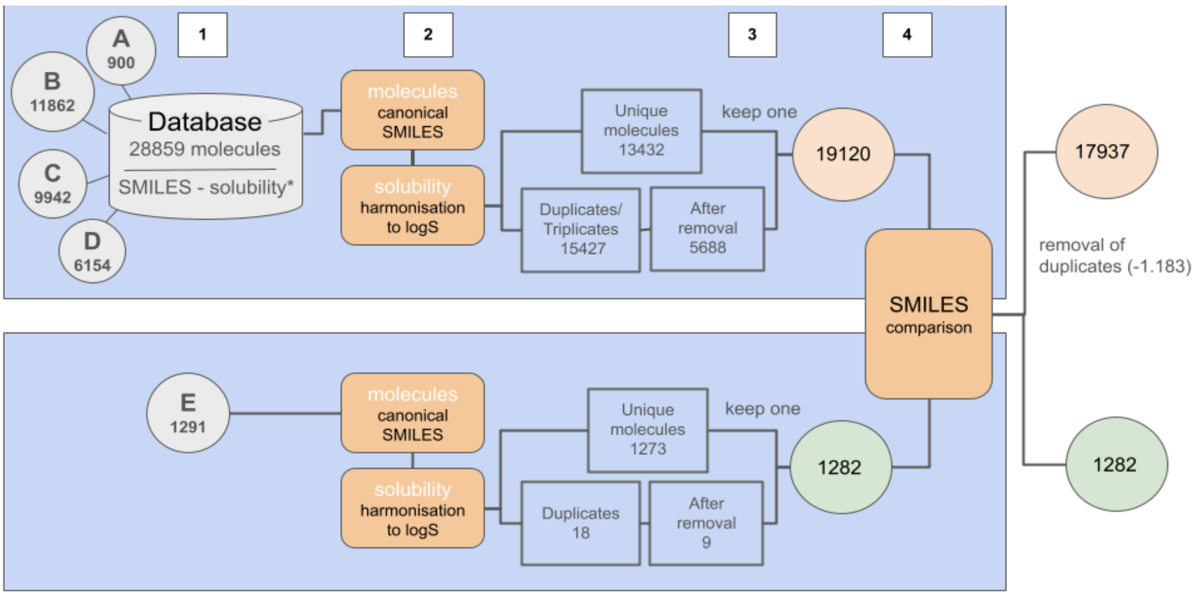


Figure 8. Schematic presentation of the preprocessing pipeline for training and test data consisting of (1) data collection, (2) generation of canonical SMILES and solubility harmonisation, (3) removal of duplicates and data cleaning in the training and test data sets, (4) comparison of the test to the training dataset and removal of duplicates from the latter training set.

3.1.1 Merging data from four sources

We started with data from four different sources: **A**, **B**, **C**, and **D**.

Total number of samples (molecules + solubility value) after merging datasets A-D	
Training data: 28,859 samples	Test data: 1,291 samples

3.1.2 Removing duplicates from training and test data

To remove duplicates, we first divided the data into **unique data** (data with no duplicates) and **duplicate data (**data that has duplicate entries). From this step, we identified:

Unique samples and duplicates in training and test data	
Training data: 13,432 (unique), 15,427 (duplicates)	Test data:1,273 (unique), 18 (duplicates)

3.1.3 Handling duplicate data based on solubility differences

To handle duplicates, we first considered solubility differences between duplicates. If the difference between the highest and lowest solubility values for a set of duplicates exceeded 0.50, we removed the duplicates, as they could lead to high average values. After applying this threshold rule:

Remaining duplicate data in training: 14,044 samples. From the remaining duplicates in the training data, we calculated the average solubility value and identified 5,688 samples that were not duplicates. We then added the unique 13432 + 5688 samples and their corresponding values, resulting in 19,120 unique samples in the training data and 1,282 unique samples in the test data.

Remaining duplicate data in testing: No significant duplicates were left to remove in the test set.

Total number of samples after removal of duplicates	
Training data: 19,120 samples	Test data: 1,282 samples

3.1.4 Comparison of training and test data, removal of duplicates

To ensure that the training data does not contain duplicate SMILES (which could overlap with the test data), we converted both the training and test data SMILES to canonical SMILES. After comparing the canonical SMILES between the training and test data, we found 1,183 matching SMILES. These matching SMILES were removed from the training data.

Final unique samples (molecule + solubility value)	
Training data: 17,937 samples	Test data: 1,282 records

3.2 Generation of descriptors

Following the idea of using molecules and their descriptors for training neural networks, we created an initial set of four fundamental descriptors representing essential molecular characteristics using RDKit (exact molecular weight, water-octanol partition coefficient logP, aromatic proportion, and rotatable bonds), which form the foundation of our feature representation. To systematically investigate the impact of including additional descriptors on the model's performance, we incrementally expanded the descriptor set. As we progressed, we continually introduced new descriptors into our feature space, each chosen to capture specific chemical attributes and properties. The step-by-step expansion of our descriptor set provided insights into the optimal feature space for our predictive modeling task, allowing us to refine and enhance our solubility predictions. Within the extended descriptor set, four different descriptor types used at different stages of our study can be described (Table 2): (1) 125 classical descriptors, including topological, physicochemical, and electronic properties. These descriptors provide a comprehensive characterization of molecular structures and properties (Appendix, Section 5.3.4). (2) Different molecular fingerprints with varying bit lengths, ranging from 128 to 1024 bits (radius of 2), were considered in our model. Molecular fingerprints enable the capture of fine-grained structural information at various levels of granularity (Appendices, section 5.1.4). (3) We included binary representations of the presence or absence of specific functional groups as descriptors, as we expected them to be a potentially critical property of the molecular structures, referring to their solubility. An overview of the prevalence and diversity of functional groups present in the training dataset, along with a summary of their influence on the solubility of molecules, is included in Section.

5 of the Appendices. Along with these descriptors, we added 38 molecular descriptors, including charge, double bonds, atoms' connectivity, valence electrons, hybridization types, bond types, and chirality features (Appendices, section 5.3.5). These descriptors aim to capture the detailed structural and electronic characteristics of the molecules, thereby improving predictive accuracy. A complete descriptors list added in Appendices (5.1 to 5.5)

3.3 Model building

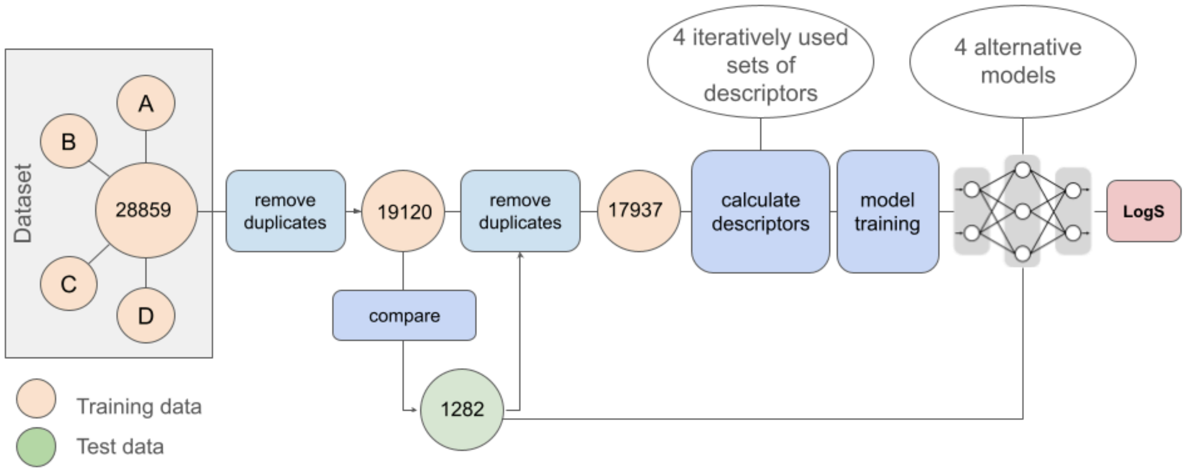


Figure 9. Schematic description of the workflow followed in this work, consisting of dataset preparation for training and testing, calculation of descriptors, training of the model (according to four different basic models), and the prediction of the solubility in logS.

We employed four different machine learning models, each designed to capture specific nuances in our solubility prediction task (Table 2) (details are provided in Appendices 6.2). (1) The ensemble learning technique, **Random Forest,** was chosen because of its ability to handle a variety of data types and identify non-linear relationships. Hyperparameters, such as tree depth and forest size, were given special consideration when building our random forest model. (2) **XGBoost** was used because of its excellent prediction abilities. We employed hyperparameter tuning to optimize crucial parameters, including the learning rate, tree depth, and number of estimators (i.e., trees in the ensemble) (3). **Artificial neural networks** (ANNs) are renowned for their ability to recognize patterns. Our ANN architecture consists of multiple hidden layers, each composed of various neurons with distinct activation mechanisms. To ensure the network's optimal learning and generalization, hyperparameter tuning included the learning rate and dropout. (4) **Message passing neural networks** employed both a standard Message-Passing Neural Network (MPNN) and a hybrid MPNN architecture for solubility prediction. The hybrid MPNN comprises seven layers, including a combination of message-passing layers, global pooling, integration mechanisms for additional physical property features, and fully connected layers. This architecture was designed to enhance the predictive accuracy by capturing both molecular graph-level information and physical property data. Comparative evaluations were conducted to assess the performance improvements achieved with the hybrid MPNN compared to the standard MPNN approach. The selection of message-passing functions, layer configurations, and other crucial parameters was addressed during the hyperparameter tuning phase. The optimal hyperparameters for all models are presented in the Appendices (Section 4).

To improve the predictive performance and interpretability of our XGBoost model, we employed the Least Absolute Shrinkage and Selection Operator (LASSO) regularization technique, which applies L1 regularization to reduce the coefficients of less essential features towards zero, allowing for feature selection. To evaluate the model, we perform five-fold cross-validation, which ensures that our model generalizes well and does not overfit the training data. One fold was reserved for the validation set for each iteration, and the remaining folds served as the training set[45]. For each fold, the model's performance was evaluated using three specific metrics: mean absolute error (MAE), root mean squared error (RMSE), and R^2 coefficient of determination. The model's performance was then evaluated across the entire dataset by aggregating the performance metrics across all five cross-validation splits.

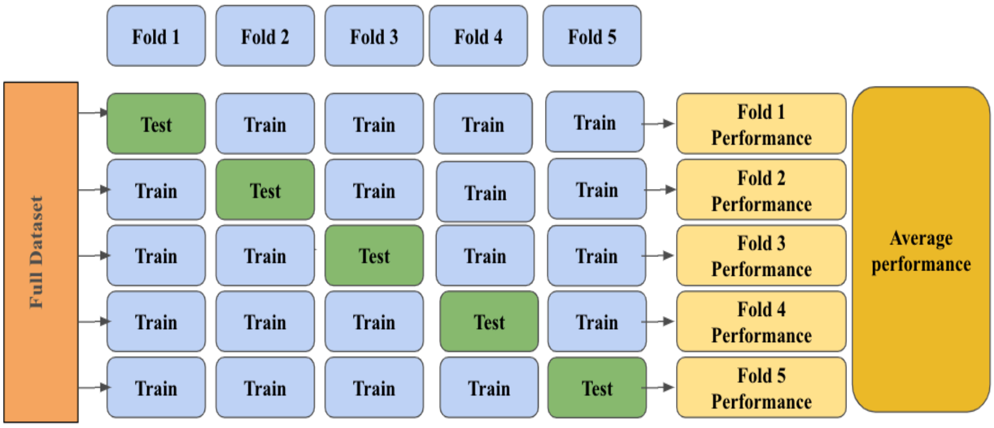


Figure 10 Illustration of 5-Fold Cross-Validation: The dataset is divided into five equal subsets (folds). In each iteration, one fold is used as the test set while the remaining four serve as the training set. The process repeats five times, ensuring each fold is used for testing once. The final model's performance is averaged across all iterations (source image from reference[45]).

In our work, we adopt the concept of using molecular descriptors as described previously by others. [36, 37, 41] To improve the currently available models in terms of accuracy and suitability for a wide range of chemical substance classes, we collected additional key aspects that we expected to enhance solubility prediction. The main changes included (1) the application of traditional machine learning algorithms like XGBoost and Random Forest, along with neural networks and (hybrid) message-passing neural networks from graph neural networks. (2) We further extended the molecular descriptors used in a stepwise approach; (3) we included four datasets from different established citations as a training set and compared them to the previously used test dataset published by Huuskonen[12].

3.4. Evaluation matrix

Our analysis focused on MAE, RMSE, and R^2 as essential metrics for comparing the performance of the models.

3.4.1 Mean absolute errors (MAE)

MAE measures the average error between the predicted values (Y_pred) and the actual values (Y_true).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |Ytrue, i - Ypred, i|$$

Interpretation: MAE reflects the average error in the same units as the target variable. Lower values indicate better predictive performance

3.4.2 Root mean square error (RMSE)

RMSE measures the square root of the average squared differences between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Ytrue, i - Ypred, i)2}$$

Interpretation: RMSE emphasizes larger errors more than MAE due to the squaring operation, making it sensitive to outliers. Lower values represent better performance.

3.4.3 Coefficient of determination (R^2)

The R^2 metric, also known as the coefficient of determination, is a statistical measure commonly used to evaluate the goodness of fit of a regression model. It indicates how well the model's predictions align with the actual data.

$$R2 = 1- \frac{\Sigma\left(yi-\hat{yi}\right)2}{\Sigma(yi-\underline{y})2}$$

It indicates how well the model captures the variability of the target variable, with higher values reflecting better performance.

3.4.3 Comparative analysis with different sets of features

Using XGBoost, Random Forest regression, ANNs, and MPNNs, four models using a particular set of descriptors were trained (Figure 2, Table 2). Our analysis focused on MAE and RMSE as key metrics for comparing the performance of the models and descriptors (Appendix, Section 6.2). We evaluated the models' cross-validation results to determine how well they generalize to unseen data, and we further investigated the subtleties of each model’s configuration, considering the impact of the quantity and variety of descriptors on the model's complexity. We started our investigation with XGBoost, due to its ability to capture complex, non-linear relationships and mitigate overfitting through built-in regularization. Its robustness, efficiency, and adaptability, combined with extensive hyperparameter tuning, make it an ideal choice for our dataset's characteristics. Using XGBoost with four descriptors (molecular weight, logP, aromatic proportion, and rotatable bonds, XGB-4) gave an R^2 score of 0.87, in combination with an RMSE of 0.72 and MAE of 0.56. The extension of the descriptor set to 17 descriptors (XGB-17) improved the model's performance to a level comparable to the best models currently available, as referenced in the same test data from Huuskonen (see Tables 2 and 3). Further increases in the number of descriptors yielded a systematic improvement in the performance of the XGBoost model; however, the improvements from 17 to 125 descriptors and the subsequent ones were lower than the improvement made in the first step (an increase from 4 to 17). The last improvements, introduced by the 128-bit fingerprint descriptor and the addition of 7 functional groups, are minimal (not represented in the number of digits). The increase in fingerprint size to 512 and 1024 bits (Table 2, entries 7 and 8) led to a decrease in performance on the test dataset, likely due to the introduction of redundant or irrelevant features. This causes overfitting. The best results were finally achieved with XGB in combination with 298 descriptors (Table 2). This model (further named XGB-298), yielding an MAE value of 0.40, an RMSE value of 0.55, and an R^2 score of 0.92, was used for further investigation and comparison with three other known models in the literature. Neither Random Forest (MAE = 0.49, RMSE = 0.64, R^2 = 0.90) nor ANN (MAE = 0.52, RMSE = 0.70, R^2 = 0.88) could compete with the model XGBoost-298, even when the same set of descriptors was used. Initial attempts with a standard MPNN (six-layer architecture) didn’t perform well, achieving an R^2 of 0.82, a MAE of 0.68, and an RMSE of 0.76. However, the inclusion of a hybrid MPNN significantly improved performance and surpassed that of the standard MPNN. With an MAE of 0.46, an RMSE of 0.63, and an R^2 of 0.90, the hybrid MPNN model yielded nearly comparable results to the XGBoost model.

Table 2: Comparison of test set performance on different models and combinations of descriptors given by the metrics MAE, RMSE, and R^2.

Entry	**Model Name**	**MAE**	**RMSE**	**R^2**	**Number of descriptors/fingerprint version**				
					Descr[a]	**FP[b]**	**FGs[c]**	**Feat.[d]**	**Layers**
1	XGB-4	0.56	0.72	0.87	4	-	-	-	-
2	XGB-17	0.44	0.58	0.91	17	-	-	-	-
3	XGB-125	**0.40**	**0.55**	**0.92**	125	-	-	-	-
4	XGB-253	**0.40**	**0.55**	**0.92**	125	128	-	-	-
5	XGB-260	**0.40**	**0.55**	**0.92**	125	128	7	-	-
6	XGB-298	**0.40**	**0.55**	**0.92**	125	128	7	38	-
7	XGB-682	0.41	**0.55**	**0.92**	125	512	7	38	-
8	XGB-1194	0.41	**0.55**	**0.92**	125	1024	7	38	-
9	RANDOM FOREST	0.49	0.64	0.90	125	128	7	38	-
10	MPNN	0.68	0.76	0.82	-	-	-	-	6
11	Hybrid MPNN	0.46	0.63	0.90	125	-	7	38	7
12	ANN	0.52	0.70	0.88	125	128	7	38	6

[a]Model includes the given number of descriptors; [b]Bits of Fingerprint: [c]FGs = number of functional groups included; [d]Additional selected descriptors included.

3.4.3 Standard Deviations recorded for different metrics

Table 3: Comparison of test set performance on different models and combinations of descriptors given by the metrics MAE, RMSE, and R^2, including the Standard Deviations (SD).

Entry	Model Name	MAE	SD	RMSE	SD	R^2	SD
1	XGB-4	0.5750	0.0068	0.7407	0.0089	0.8684	0.0031
2	XGB-17	0.4576	0.0040	0.5999	0.0055	0.9137	0.0016
3	XGB-125	0.4165	0.0030	0.5656	0.0046	0.9233	0.0013
4	XGB-253	0.4164	0.0013	0.5642	0.0035	0.9237	0.0009
5	XGB-260	0.4135	0.0030	0.5615	0.0039	0.9244	0.0010
6	XGB-298	0.4122	0.0027	0.5632	0.0043	0.9239	0.0009
7	XGB-682	0.4164	0.0028	0.5629	0.0033	0.9246	0.0009
8	XGB-1194	0.4227	0.0034	0.5671	0.0053	0.9229	0.0014
9	RANDOM FOREST	0.5010	0.0029	0.6447	0.0030	0.9003	0.0009
10	MPNN	0.6645	0.0060	0.7511	0.0021	0.8122	0.0035
11	Hybrid MPNN	0.4521	0.0034	0.61878	0.0028	0.8924	0.0018
12	ANN	0.5316	0.0062	0.7115	0.0060	0.8786	0.0021

Performance metrics (Mean Absolute Error [MAE], Root Mean Square Error [RMSE], and Coefficient of Determination [R^2]) along with their respective standard deviations (SD) for the evaluated model on 5-fold validation on the models (Random Forest, XG Boosting (XGB), Artificial Neural Network (ANN), Messages passing neural network (MPNN) and Hybrid Messages passing neural network (H-MPNN).

3.4.4 Comparative analysis with literature results on test data.

To the best of our knowledge, there are currently 11 studies that predict aqueous solubility and utilize the Huuskonen dataset as a reference test dataset (Table 3). We compared the results of our model with results from the literature using the same test dataset.[12] In our first comparison, we found our model to be comparatively powerful as the best model from the literature, which was published in 2020 by Sorkun *et al.* (Table 3). We then compared the models and workflows in more detail and found one main difference in the preparation of the training and test datasets. While we harmonized the data sets used for training and testing by transforming all molecules into canonical SMILES without stereochemical information, Sorkun *et al.* used the InChIKey from stereochemical SMILES in the training set and SMILES without stereochemical information in the test set. This difference in data preparation has a significant impact on identifying potential overlaps between training and test data. In the canonical SMILES approach without stereochemical information, molecules with defined stereochemistry in either of the training datasets yield the same canonical SMILES code as the same molecule in the test set without specific stereochemical annotation, and duplicates are consequently removed. In this way, e.g., a double bond that is given in either *Z* or *E* annotation (or a mixture of both) in one dataset is to be considered as the same molecule in another dataset if the double bond is not explicitly annotated and just given as any double bond. In contrast, the InChIKey approach assigns different InChIKey codes to molecules with and without assigned isomer details. Consequently, the approach of Sorkun *et al.* includes molecules that might be the same but are described with fewer isomer details in the test data than in the training data (and vice versa). Our review of the dataset used by Sorkun *et al.* identified 133 duplicate samples where such a correlation might be an issue. Therefore, we reproduced the model once with and without these 133 data points in the training set. With the data included, we were able to produce the published results and obtain the values presented in Table 3 from the literature. After removing the 133 data points from the training dataset, we were unable to reproduce the original results and obtained an R^2 value of 0.87, along with MAE and RMSE values of 0.54 and 0.73, respectively. We can think of two possible explanations for the results of the reproduction of the literature-known work:

Either the overlap in test and training sets caused an improvement of the performance of the model and their removal gives a more reliable and unbiased estimation of the generalization performance of the model, or the decrease in training set size due to the removal of the 133 samples caused the model's performance drop. The first conclusion is more likely, considering that the 133 data points represent only a small portion of the overall training dataset. Consequently, a comparison of the models, as shown in Table 3, should be based on the data obtained from the reproduced results of the model by Sorkun *et al.,* excluding the 133 data points. Taking this into account, we were able to improve upon the currently best results achieved in previous work (by Yan and Gasteiger, as well as Lusci *et al.*) in terms of the MAE and RMSE values. We can now compete with the currently highest R^2 value of 0.92.

Table 4: Comparison of our model XGB-298 with previous work in the literature

Entry	Method	Test dataset size	MAE	RMSE	R^2	Year	Reference+
1	ANN	1294	-	0.71	0.88	2000	Huuskonen[46]
2	ANN	1291	-	0.62	0.91	2001	Tetko et al[44]
3	ANN	1294	0.68	0.59	0.92	2003	Yan and Gasteiger[47]
4	MLR	1290	0.68	0.87	0.71	2004	Delaney[16]
5	MLR	1294	0.52	0.63	0.90	2004	Hou et al.[48]
6	SVM	1290	0.43	0.60	-	2007	Schroeter et al. [49]
7	MLR	1290	0.72	0.94	0.73	2012	Ali et al.[50]
8	UG-RNN	1026	0.46	0.60	0.91	2013	Lusci et al.[37]
9	MLR	1290	0.93	1.15	0.68	2017	Daina et al. [51]
10	ANN	1297	-	0.65	0.90	2018	Bjerrum and Sattarov[52]
11a	Consensus	1290	(0.35)	(0.53)	(0.93)	2020	Sorkun et al.[36]
11b*			0.54*	0.73*	0.87*	2024*	rework from Sorkun et al*

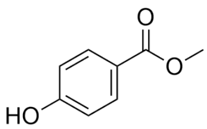
Entry	Method	Test dataset size	MAE	RMSE	R^2	Year	Reference+
12	**XGB-298**	**1282**	**0.40**	**0.55**	**0.92**	**2024**	**this work**

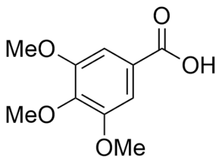
*Results obtained with the model and test data of Sorkun *et al.* after the exclusion of data referring to the same SMILES code in training and test datasets. **ANN**, artificial neural network, **MLR,** multiple linear regression, SVM, **support** vector machine, **UG RNN,** Undirected graph recursive neural networks, **RF,** random forest; **XGB,** Extreme Gradient Boosting; **Consensus,** an ensemble of ANN, RF, and XGB. Best results are given in bold; best results in past developments are underlined.

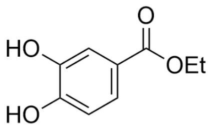
3.4.5 Comparative analysis with lab experimental values, other online predictors, and our prediction

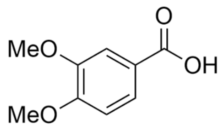
In addition to the standard analysis of predictive models and comparison with models that utilize the same test dataset as Huuskonen, described above, we conducted a comparative evaluation of our model against well-established solubility prediction tools. To this aim, we selected five standard compounds that are readily available, for which literature solubility values are available, and for which solubility data can be determined in our labs. The combination of solubility values from the literature and our labs (detail given in experimental part) as reference values allowed us to gain additional confidence in the correctness of the literature values and allowed us to identify potentially problematic compounds, for that the measurement details such as the pH might be crucial for the results but perhaps not included to the literature data. We selected the models of VCC labs,[53,54] Sorkun,[53,54], and Chembcpp[55] for comparative analysis, as they are well-known (e.g., the VCC labs model is included in DrugBank) and available in the form of a web service. The comparison was achieved through statistical analysis and by determining the models that are closest to either the literature value or the experimentally determined values (Table 4, bold). Across the five compounds, our model achieved an average mean absolute error (MAE) of 0.88, compared to 2.04 for VCCLAB, 1.56 for Sorkun, and 1.30 for Chembcpp. Although this statistical analysis only considers a small number of compounds, the values confirm the better performance of XGBoost-298 compared to the other models, at least within the given compound scope. In addition, in identifying the models that best fit the literature or experimental results, XGBoost-298 yielded better results (4 values fit best) than Chembcpp (2 values fit best), VCC (1 value fits best), and the Sorkun model (0 values fit best).

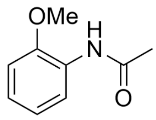
Table 5: Comparison of solubility values with the literature, experiments, and predictive models


1


2


3


4


5

Structure	VCC[56] [g/L][a]	Sorkun[36] [g/L][a]	Chembcpp [g/L][a]	XGB-298 [g/L][b]	Exp. Lit [g/L][c]	Exp. lab [g/L][d]
1	3.64	5.39	3.18	**2.26**	2.50[57]	2.42
2	2.07	1.68	2.12	**2.97**	2.48[58]	3.10
3	3.55	2.88	1.63	**2.32**	insoluble[59]	2.12
4	1.66	2.40	3.89	**0.79**	0.50[60]	7.20
5	3.78	4.24	1.85	**5.37**	17.40[61]	12.0

a predicted solubility in g/L from a reference model (value calculated from the given information in logS); b. results gained from our model; c. experimental values extracted from different literature sources; d. experimental values determined in our labs (see the Experimental part for details on the method used). Highlighted in bold: Values for which the given model performs best in comparison to the other predictive models, either referring to the experimental data from the literature or the experimental data determined in our labs.

3.4.6 Comparative analysis with train and test data

Our model, XGB-298, exhibits a high level of predictive accuracy, resulting in minimal variation between actual and predicted values (Figure 11a). The model's performance on the testing dataset, serving as an indicator of its robustness, supports this finding. The model's predictions (blue) and actual solubility values (orange) closely match, demonstrating the model's ability to generalize beyond the training data (Figure 11b). The minor variations in the testing dataset are an indication of how well the model predicts outcomes in real-world situations. The consistent correlation between predicted and actual values across both datasets highlights the model's suitability for applications where precise solubility estimations are essential.

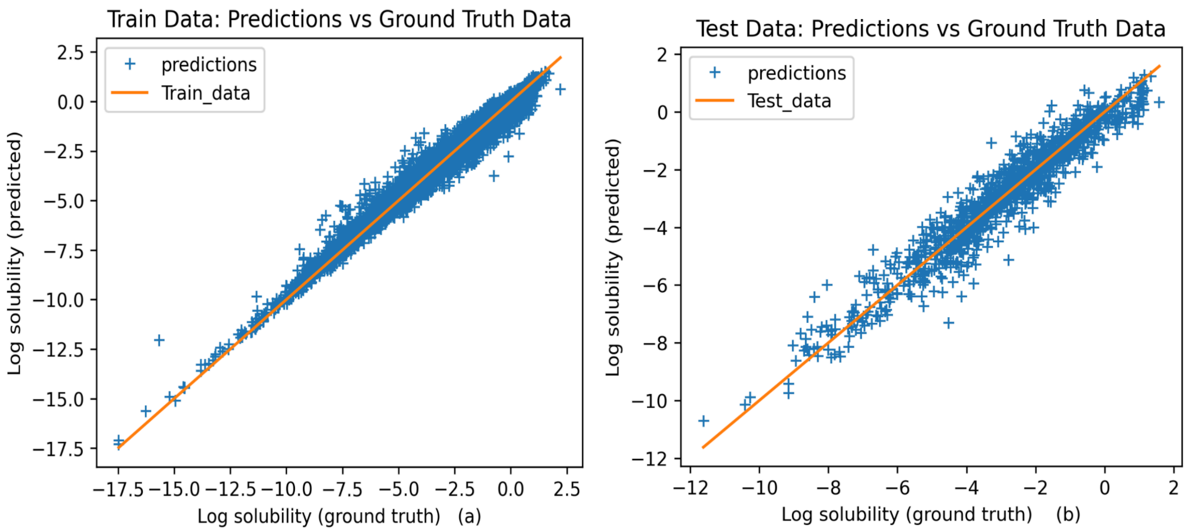


Figure 11. The model's predictive power and ability to generalize to new data are demonstrated through a plot comparing its performance. **a.** evaluation of training data; **b.** evaluation of and testing of data. Blue = actual solubility values; orange = predicted solubility values by our model.

3.4.7 Duplicates removed from training data for reproduction of the Sorkun results

Sorkun et al. provided a dataset comprising 6,154 training data points and 1,291 test data points. They used InChI keys to preprocess and identify unique compounds in their dataset. After preprocessing, they obtained 4,399 unique training data points and 1,290 unique test data points. Their model was trained on this preprocessed data, achieving a Mean Absolute Error (MAE) of 0.35. We aimed to reproduce Sorkun's work while extending it by preprocessing the dataset using canonical SMILES instead of InChI keys. This preprocessing step enabled us to identify 133 compounds that were common to both the test and training datasets.

Adjustments to the Training Data: To ensure the test data remains independent and untainted by information leakage, we excluded these 133 overlapping compounds from the training

data. After removal, the training dataset was reduced to 4,275 unique data points. The test dataset remained unchanged, consisting of 1,290 unique data points.
Model Training and Evaluation: We trained the model on the revised training dataset (4,275 points) using the same methodology as Sorkun. The resulting performance metrics showed an MAE of 0.54, which is significantly higher than Sorkun's reported MAE of 0.35. This result highlights the potential impact of the overlap between the training and test datasets. The inclusion of overlapping compounds in Sorkun's training data likely contributed to their lower MAE, as the model could effectively "memorize" part of the test set during training.

Key observations

1. Data leakage and its impact: The presence of overlapping compounds (133 matches) between the training and test datasets likely introduced a form of data leakage in Sorkun's approach. By removing these compounds, our results demonstrate the model's true generalization capability.
2. Significant performance difference: After ensuring the independence of the test data, the MAE increased to 0.54. This considerable difference underscores the importance of careful dataset curation to avoid overestimating model performance.
3. Reproducibility and transparency: To ensure transparency and reproducibility, the list of the 133 overlapping compounds and the preprocessed datasets will be made available in a GitHub repository for reference.

Conclusion and significance: Our findings highlight the importance of rigorous preprocessing to prevent data leakage and ensure the robustness of model evaluation. While Sorkun's reported MAE (0.35) is competitive, our revised methodology suggests that this result may be partially due to data overlap. The adjusted MAE of 0.54 provides a more accurate benchmark for comparing against other literature.

3.4.8 Comparison analysis with recent challenge for JCIM paper

For the comparison analysis with recent challenges in solubility prediction, we evaluate our model's performance against existing state-of-the-art approaches. In our study, we aim to benchmark our solubility prediction model against recent solubility prediction challenges, such as those hosted in the Journal of Chemical Information and Modeling (JCIM) [43]. Two datasets are provided in the challenge: set1, comprising 100 compounds, and set2, comprising 32 compounds, both of which have intrinsic solubility in logS. We have tested our model's predictions on this dataset, removing overlapping data points from the training set to ensure fairness, and ensuring that no samples from Set 1 or Set 2 remain. As a result, the training dataset was reduced from 17,937 to 17,884 after eliminating 37 matching data points from Set 1 and 16 from Set 2.

Our model predicted RMSE values of 1.03 and 1.05, respectively, which are lower than the average RMSE reported in the published studies (1.14 and 1.62, as cited in the referenced publication). Our model correctly predicted solubility within ±0.5 logS units for 38% of compounds in Set 1 and 31% in Set 2 (published average is 40% vs 30% within the same 0.5-unit tolerance).

We compared our results with those of other participants in the competition and visualized the outcomes alongside the results from 37 different participants for the set 2 compounds. We achieved the lowest error, demonstrating the best performance among all participants.

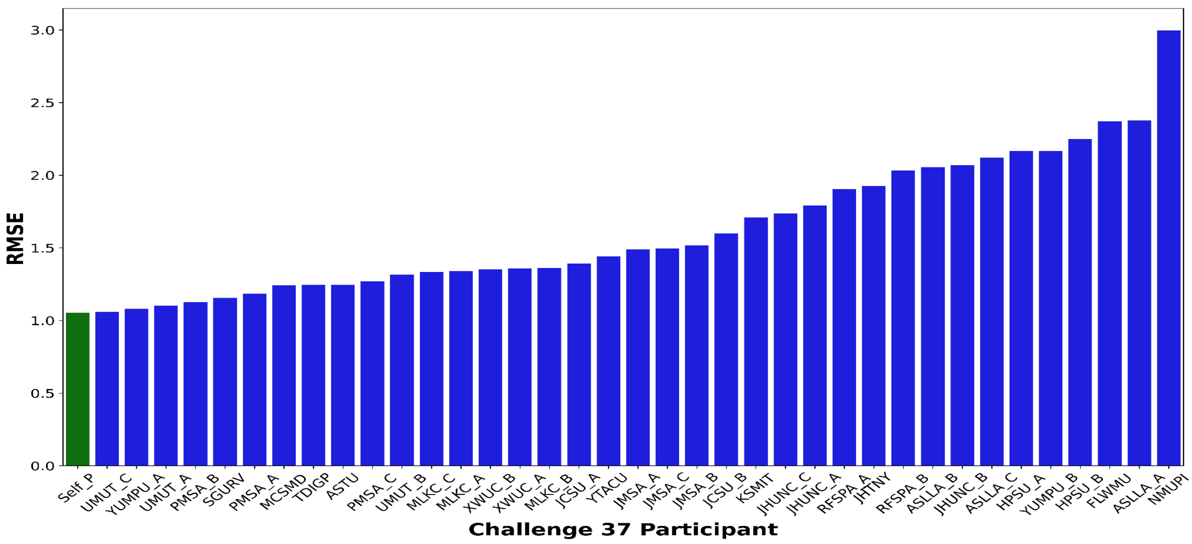


Figure 12. RMSE with 37 participant prediction results: The bar chart visualizes the RMSE values of different participants,[43] with our prediction ('Self_P') highlighted in green as it achieved the lowest RMSE (1.05).

3.5 Applicability domain analysis

To ensure the reliability of predictions, we implemented an Applicability Domain (AD) analysis using a combination of visualization and quantitative techniques. The AD defines the chemical space in which the predictive model operates with confidence, helping to identify compounds for which predictions may be unreliable.

3.5.1 t-SNE-based visualization

We employed t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of molecular feature space and visualize the distribution of compounds. This allowed us to visually assess whether a new molecule falls within the dense region of training data or lies in an extrapolated space, highlighting potential outliers.

3.5.2 Threshold optimization for applicability domain

To establish a robust threshold for distinguishing in-domain and out-of-domain compounds, we used the 95th percentile of the Mahalanobis distance in the training set. The Mahalanobis distance is a multi-feature metric that accounts for correlations between features and provides a scale-invariant technique for estimating how far away a point is from the distribution mean. Compared to Euclidean distance, it considers the inherent covariance structure, making it particularly effective in outlier detection. By setting the threshold at the 90th percentile, we are only marking molecules that are significantly different from the training distribution as out-of-domain, which improves the validity of our determination of applicability domain

3.5.3 Determination of applicability fomain

To define the applicability domain (AD) of our model, we utilized Principal Component Analysis (PCA). This dimensionality reduction technique preserves the most significant variance in the data and the overall structure of the dataset. By projecting the high-dimensional molecular feature space onto a lower-dimensional representation, PCA enables a more interpretable assessment of whether a new compound falls within the distribution of the training set. This transformation is particularly useful in calculating Mahalanobis distance, a measure of how far a given molecule is away from the center of the training set. With PCA, we have a robust and efficient method of establishing the model's applicability domain while reducing noise and redundancy in the feature space.

3.5.4 Validation with different molecular weights

To validate the effectiveness of our AD methods, we tested three different SMILES representing molecules with varying molecular weights (180 and 1202). The results demonstrated distinct classification outcomes, aligning with expectations based on molecular size and chemical diversity.

This applicability domain analysis is integrated into our web framework, allowing users to determine whether a given molecule falls within the model’s chemical space before making predictions. The detailed implementation can be found in our repository under

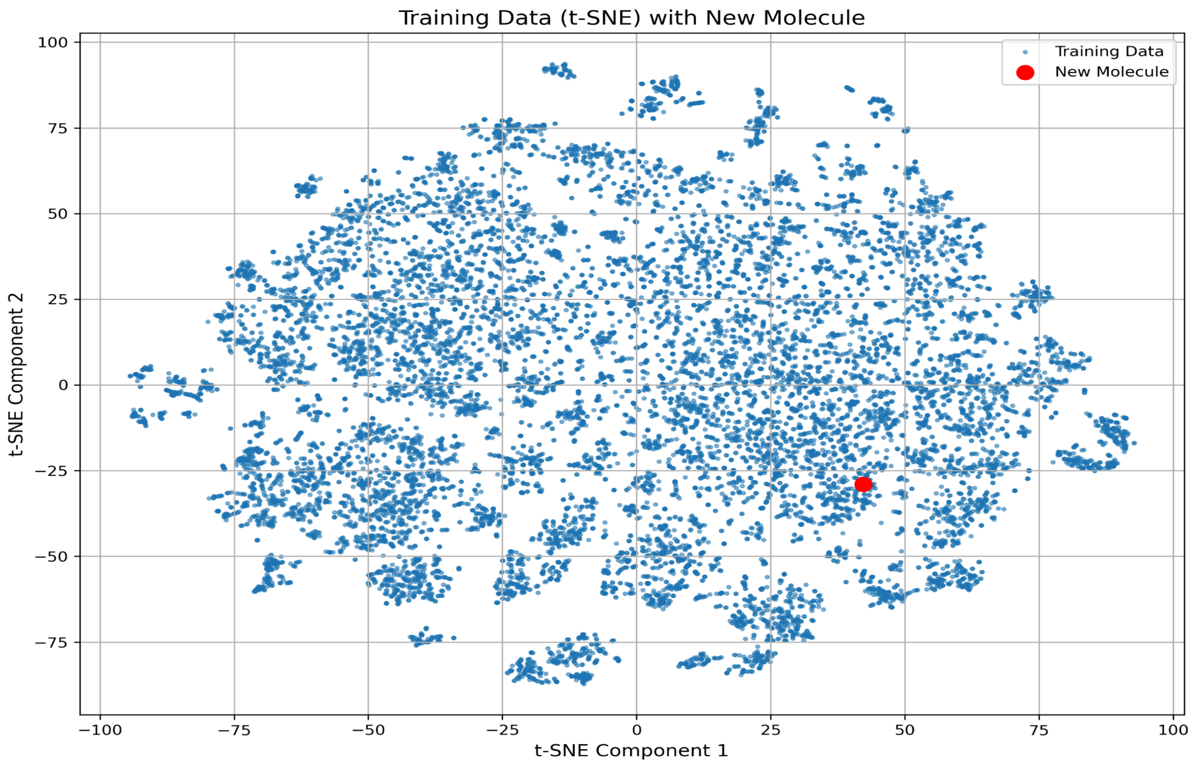


Figure 13. The scattered plot shows how the new molecule falls within the training space, confirming its inclusion in the applicability domain. Clustered points represent the training space, and the new molecule, positioned within this region, indicates that the model can reliably predict its properties. (SMILES: CC(=O)OC1=CC=CC=C1C(=O)O, molecular weight: 180 g/mol).

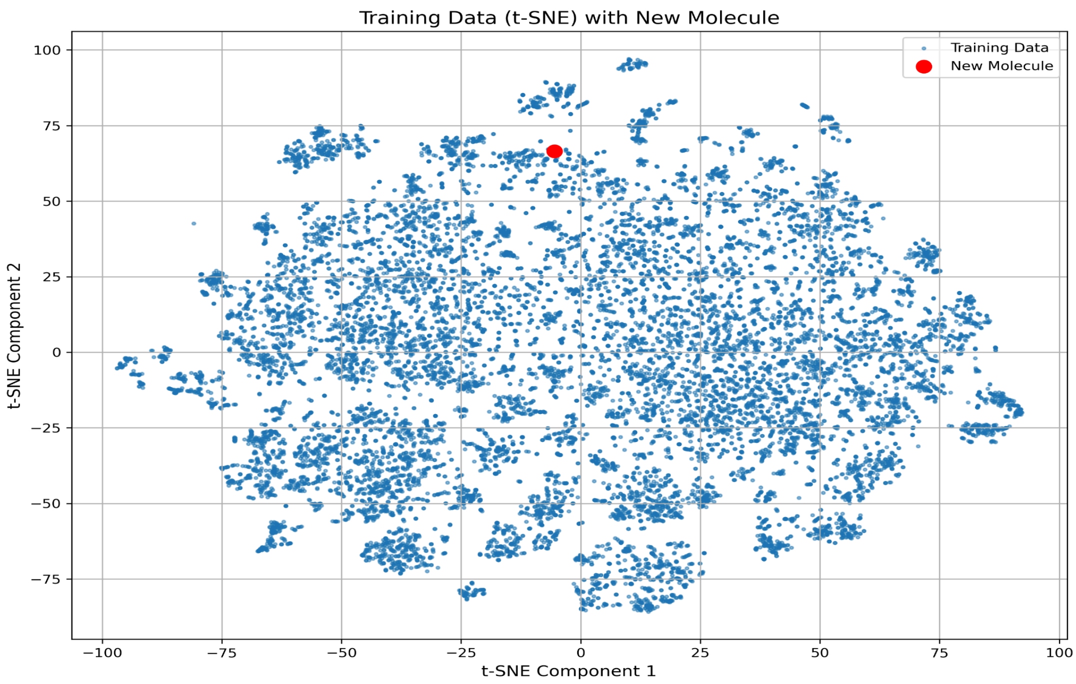


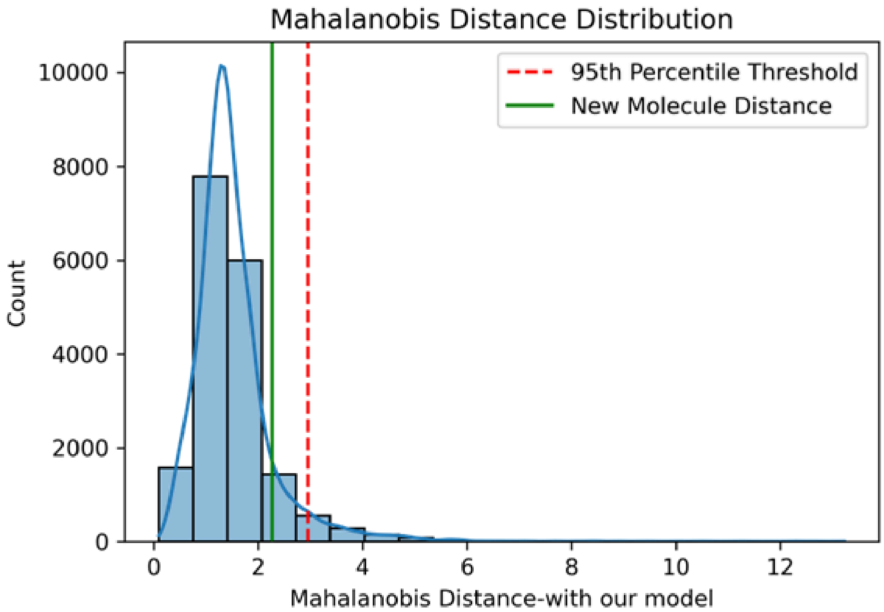
Figure 14. The scatter plot illustrates that the new molecule is positioned at the boundary of the training space, confirming that it lies outside the model's applicability domain. Clustered points represent the training space, while the new molecule's placement suggests that the model cannot reliably predict its properties (SMILES: CCC1C(=O)N(CC(=O)N(C(C(=O)NC(C(=O)N(C(C(=O)NC(C(=O)NC(C(=O)N(C(C(=O)N(C(C(=O)N(C(C(=O)N(C(C(=O)N1)C(C(C)CC=CC)O)C)C(C)C)C)CC(C)C)C)CC(C)C)C)C)C)CC(C)C)C)C(C)C)CC(C)C)C)C, molecular weight 1202 g/mol).

3.5.5 Comparison: our model vs. reference model

To evaluate the generalizability of our model, we compared its applicability domain with the reference model (Table 3, entry 11) using the Mahalanobis distance framework. The applicability domain was determined based on the distribution of Mahalanobis distances within the t-SNE space, with the threshold set at the 95th percentile of the training set distances. The figure below presents a comparative visualization, where the Figure 15 corresponds to our model, and the Figure 16 represents the reference model. The red dashed line in each plot indicates the applicability threshold, while the solid line represents the Mahalanobis distance of a new molecule. Notably, in our model, the new molecule falls inside the applicability domain, whereas in the reference model, the same molecule is outside the domain. This indicates that our model has a broader applicability domain than the reference model, potentially making it more robust for predicting diverse chemical structures. The results suggest that our model generalizes better than the reference model, thereby enhancing its reliability for real-world applications.

Figure 15. The plot shows the Mahalanobis distance distribution for a set of molecules, highlighting the applicability domain for our model. The red dashed line represents the 95th percentile threshold, above which molecules are considered outside the applicability domain. The green solid line indicates the distance for a new molecule predicted by our model, which lies within the applicability domain (distance = 2.27, threshold = 2.96)

.



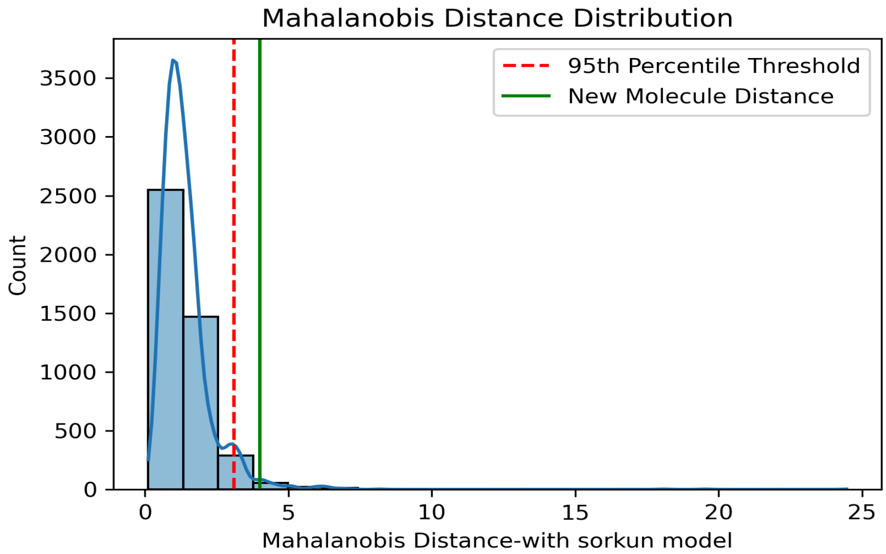


Figure 16. The plot shows the Mahalanobis distance distribution for a set of molecules, highlighting the applicability domain for our model. The red dashed line represents the 95th percentile threshold, above which molecules are considered outside the applicability domain. The green solid line indicates the distance for a new molecule predicted by our model, which is outside the applicability domain (distance = 3.103, threshold = 2.96).

3.5.6 Comparison of convex hull area with reference data:

To compare our data with reference data, we utilized the convex hull area, a useful geometric tool for understanding the spatial distribution of data points. It represents the smallest convex shape that encloses all the given data points in a particular dataset. The convex hull area for our dataset is 27,411, whereas the reference training set has a convex hull area of 16,485. This indicates that our dataset has a broader spread in the reduced-dimensional space, suggesting higher variance or diversity in the data distribution. A larger convex hull area implies that the dataset covers a broader range of feature variations compared to the reference dataset.

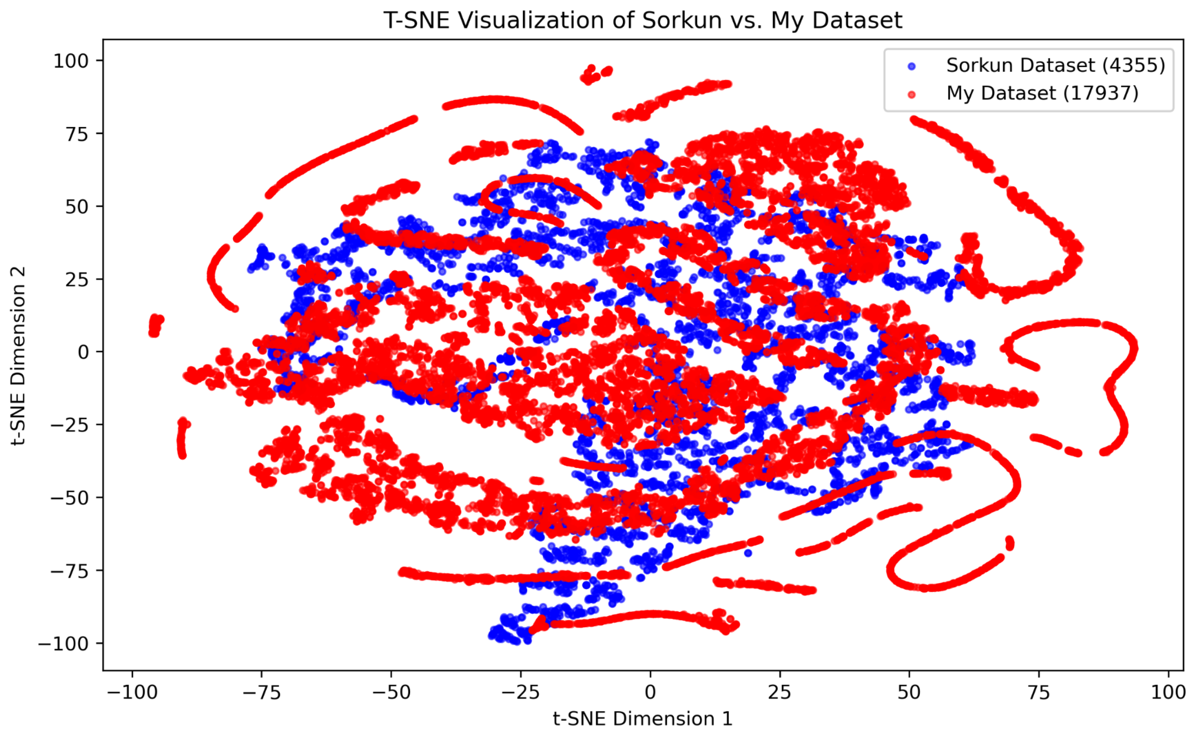


Figure 17: The scatter plot t-SNE (t-distributed Stochastic Neighbor Embedding) visualization showing the reference train set (blue) and our dataset (red), with a comparison based on convex hull area.

3.5.7 Expanding the dataset and impact on applicability

Applicability domain determines the generalization capability of a predictive model. One way to assess this expansion is through t-SNE visualization and convex hull analysis. We compare three training datasets—AB (8,147 samples), ABC (15,395 samples), and ABCD (17,937 samples) to examine how adding more data increases coverage in the chemical space. We use t-SNE (t-distributed Stochastic Neighbor Embedding) to project high-dimensional molecular features into a 2D space, providing a visual representation of chemical diversity.

To quantify dataset expansion, we calculate the convex hull area, which represents the minimum enclosing boundary of data points in the t-SNE space. The convex hull areas for our datasets are as follows: AB Dataset: 22,107; ABC Dataset: 27,034; ABCD Dataset:

27,512. These results highlight that increasing dataset size enhances the model's applicability and robustness, making it more reliable for broader chemical predictions.

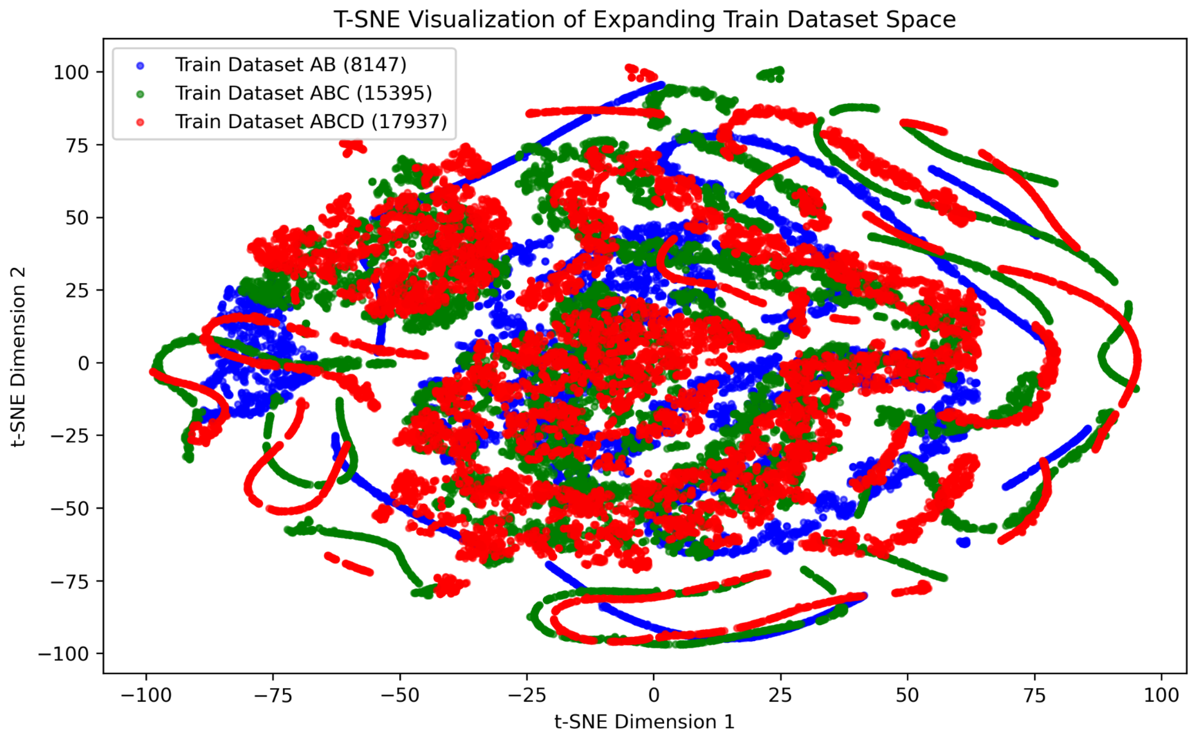


Figure 18. This scattered t-SNE visualization shows the expanding chemical space across three datasets: AB (blue), ABC (green), and ABCD (red). The increasing diversity in the chemical space, as indicated by the larger convex hull areas, demonstrates the model's expanding applicability as more data is included.

3.5.8 Model prediction on JCIM challenge data with applicability domain

To evaluate the applicability domain of the solubility prediction model on real-world data, we utilized the JCIM Journal of Chemical Information and Modeling [43] competition dataset, which comprises diverse and challenging compounds. The second test set in the competition consisted of 32 "difficult" drugs known for poor interlaboratory reproducibility (SD ~0.62 log unit), primarily from the SSF solubility determination method. Nearly a third of these compounds exhibited intrinsic solubility below one µM, contributing to their high variability. Additionally, several molecules, such as amiodarone, clofazimine, and itraconazole, occupy sparsely populated regions of chemical space, with few structurally similar compounds known. The complexity of this dataset makes it an ideal benchmark for testing the robustness and generalizability of our predictive model.

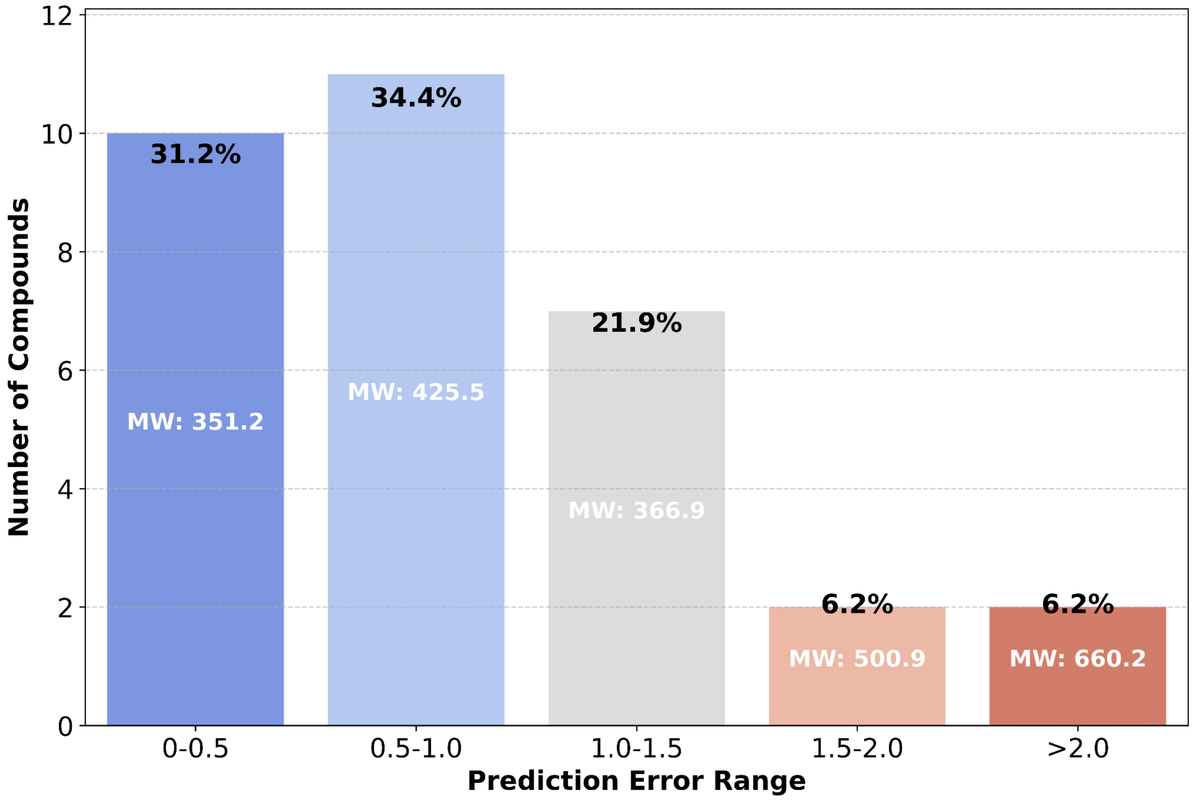


Figure 19. The bar chart illustrates the distribution of absolute prediction errors for solubility estimation, categorized into predefined error ranges. The height of each bar represents the number of compounds within each error bin, while the percentage of compounds in each category is displayed near the top of the bars. Additionally, the average molecular weight (MW) of compounds in each bin is shown at the center of the bars, providing insights into the correlation between molecular weight and prediction accuracy. This visualization helps to assess the model's performance across different molecular weight ranges, highlighting areas where predictions are more accurate and where improvements are needed.

3.6 Web framework

We developed a web framework that allows the prediction of solubility based on one or multiple SMILES entered by the user. The service is available at the given reference solubility prediction.[62]

SMILES validation: The framework automatically checks the validity of the entered SMILES. If an invalid SMILES is provided, an error message is displayed, prompting the user to correct it before proceeding.

Applicability domain check: Users have the option to evaluate the applicability domain of the given SMILES, allowing them to assess whether the compound falls within the trained model's chemical space.

To enhance the understanding of molecular structures on the Streamlit website, **2D and 3D visualization** of compounds was integrated. This allows users to analyze the molecular geometry that influences solubility.

Single or multiple molecule input: Users can input a single molecule as a SMILES string or upload a *.csv file containing multiple SMILES strings for batch processing.

logS Calculation: The framework calculates the **logS** (logarithm of solubility) values based on the provided SMILES strings.

Conversion to mol/L and g/L: The calculated logS values are then converted into solubility values in **mol/L** and **g/L** for more straightforward interpretation.

Comparison with PubChem data: The framework retrieves solubility values from **PubChem** and allows users to compare the predicted solubility values with existing solubility data for the molecules.

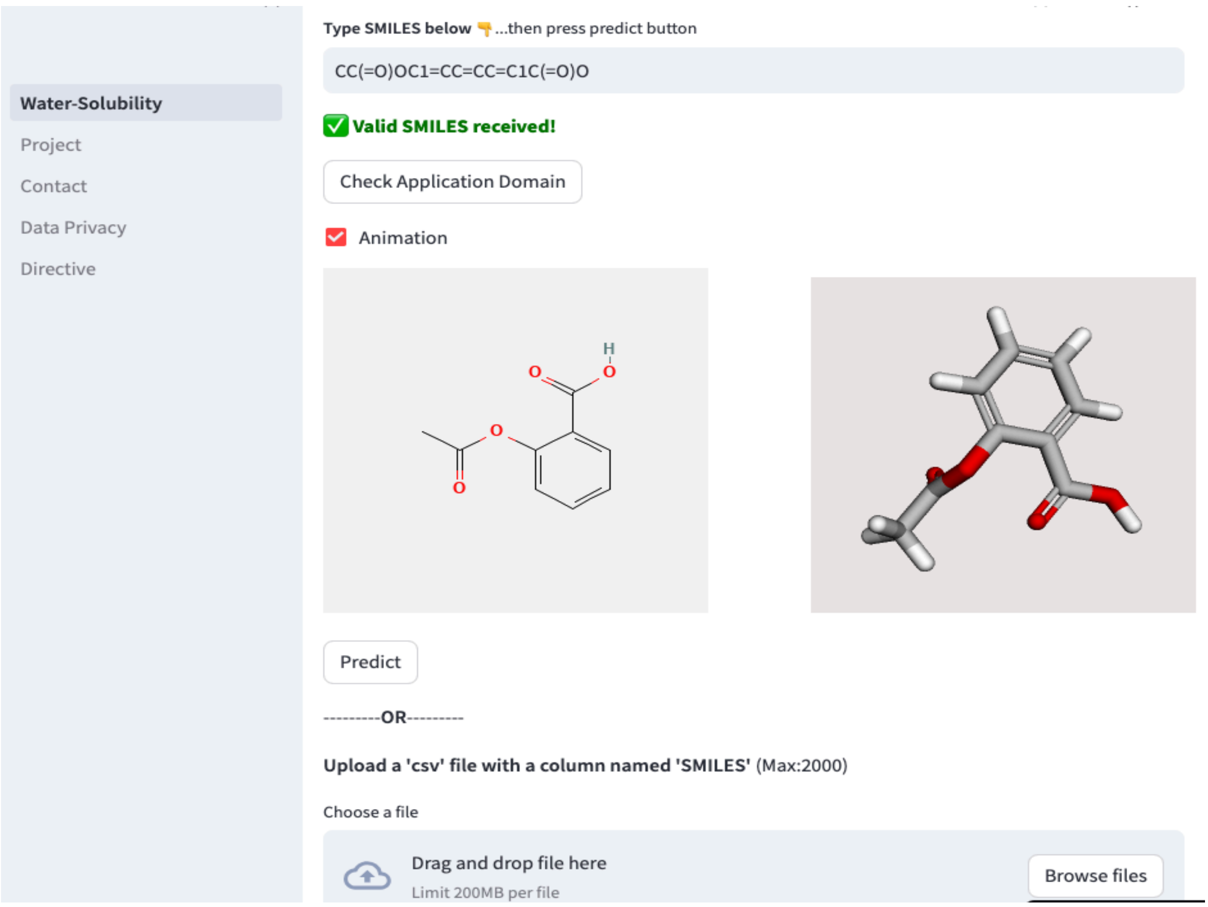


Figure 20. The web service enables the prediction of the solubility of unknown compounds using our model. It supports the prediction of solubility values for both single molecules and mixtures of molecules.

3.7 PubChem data sourcing

In this section, we present a PubChem Data Sourcing pipeline to extract, curate, and preprocess molecular datasets for solubility prediction and cheminformatics applications. Our approach is to leverage PubChem's vast chemical database, employing aggressive cleaning, standardization, and filtering techniques to deliver high-quality molecular properties. For enhanced dataset reliability, we conduct structural validation, duplicate removal, and multi-source consistency checks.

Furthermore, we systematically compare the PubChem-derived data with literature-reported solubility data, exploring potential discrepancies and differences. The comparison offers insight into the differences in experimental conditions, measurement techniques, and data representation across various sources. By clarifying these differences, we aim to refine our dataset selection criteria. Through the systematic procedure, we ensure that the curated dataset is a reliable and reproducible foundation for solubility prediction and cheminformatics studies.

From industrial processes to everyday household items, performance, safety, and sustainability all depend on water solubility. It is quite apparent that water solubility data is important, yet it is often fragmented across several records, this renders reliable and comprehensive information elusive and cumbersome to retrieve.

PubChem is an open resource that is among the largest repositories of chemical data. PubChem compiles information about 111 million chemical substances, including their structures, physical properties, and biological properties. Nevertheless, the process of curating, cleaning, and standardizing water solubility data from such a vast database requires meticulous attention to detail to ensure the data is accurate and usable.

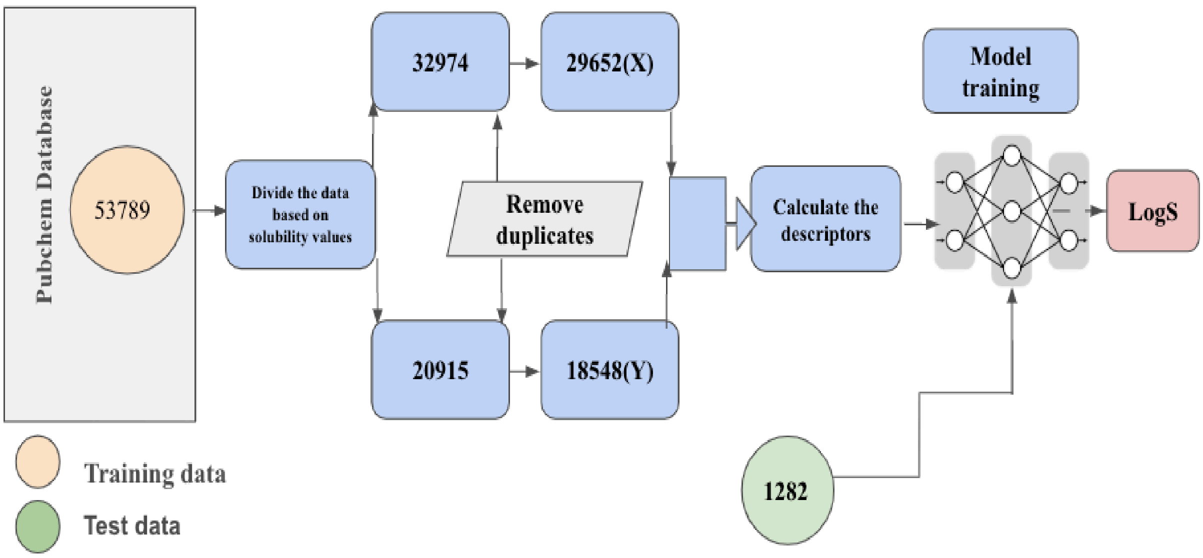


Figure 21. From PubChem's vast database of 50 million molecules to actionable insights: Curated 53,789 solubility records, split into datasets X (29,652) and Y (18,548), validated against literature data (X: 357 matches, Y: 498 matches). With X data demonstrating lower error, predictive capacity was tested using the Husskonnen dataset, achieving an MAE of 0.91 and R^2 of 0.63. A robust workflow integrating data curation, validation, and prediction for reliable solubility analysis.

3.7.1 Data retrieval process

We developed a systematic approach to extract and analyze solubility data from the PubChem database, beginning with a comprehensive dataset containing over 111 million chemical compounds. The primary data source was the PubChem FTP database, specifically the CID-SMILES mapping file, which provides compound identifiers (CIDs) paired with their corresponding SMILES (Simplified Molecular Input Line Entry System) representations.

Given the substantial size of the dataset, we implemented a carefully designed three-phase processing pipeline:

Phase 1: Initial data preparation

The raw data underwent initial preprocessing to ensure data quality and usability. We extracted CID-SMILES pairs from the source file and implemented rigorous cleaning procedures to eliminate duplicate entries, malformed SMILES strings, and incomplete records. This refined dataset was preserved in a structured CSV format, establishing a reliable foundation for subsequent analysis.

Phase 2: Scalable processing implementation

To manage the computational demands of processing 50 million compounds, we developed an efficient framework for chunked processing. The dataset was systematically divided into manageable segments of 10,000 compounds each. This segmentation strategy proved optimal for maintaining reasonable memory utilization, enabling efficient error recovery, facilitating progress monitoring,, and ensuring system stability during extended processing runs.

Phase 3: Automated data collection

We leveraged the PubChemPy library to retrieve solubility data for each compound systematically. Our automated collection system processed the segmented data sequentially, incorporating several key features: intelligent query management for efficient interaction with the PubChem database, systematic data validation at each processing stage, progressive data storage with clearly defined naming conventions, and automated error handling and recovery mechanisms

3.7.2 API integration and management

Interacting with PubChem's API (PUG View) required careful attention to various technical constraints. To control request rate management, we implemented controlled delays between successive queries to comply with API rate limitations, which allowed no more than five records per second. Additionally, we adjusted query intervals to maintain optimal throughput while preventing server rejection. We monitored request patterns to ensure consistent access within allowed limits (Details in Appendix 6.4).

The retrieved data for each segment was stored in intermediate CSV files, creating a robust audit trail and enabling easy verification of the collection process. This methodical approach ensured data integrity while maintaining efficient processing speeds for the large-scale dataset.

3.7.3 Data preprocessing

A comprehensive data cleaning protocol was implemented to ensure the integrity and consistency of the solubility data. Initial quality control measures focused on identifying and removing incomplete or ambiguous data entries. First step: systematic identification and removal of entries lacking solubility values, elimination of records with missing critical fields, and verification of data completeness across all essential parameters. We excluded qualitative solubility descriptions (e.g., "poorly soluble", "highly soluble") and retained exclusively quantitative measurements. A rigorous unit standardization process was implemented to ensure consistency across all measurements,, including the removal of records with ambiguous or non-standard unit annotations.

Along with unit verification, we standardized the units and converted them to g/L to make them uniform and standard. We converted mg/mL to g/L (1:1 conversion), µg/mL to g/L (1:1000 reduction), and mg/L to g/L (1:1000 reduction). The initial process began by hitting the PubChem API for 50 million SMILES (Simplified Molecular Input Line Entry System) strings sourced from the PubChem database.

3.7.4 Data consolidation and segmentation

From the whole PubChem dataset, we obtained a dataset of 66,639 data points containing solubility information.

Total number of samples (molecules + solubility value) after curating 50 million smiles
Total data points 66,639

3.7.4.1 Removing records which does not have quantitative solubility

To refine the dataset, data points with non-quantitative solubility descriptions such as "poorly soluble" or "very soluble" were excluded, leaving 53,789 records for further analysis.

More clean data with solubility
Total data points 53,789

3.7.4.2 Divided the data based on the solubility values

Once the data was consolidated, the 53,789 solubility data points were categorized into two segments based on the solubility values given in the dataset.

The total number of samples divided	
Data points 32,974 samples contained precise solubility values (Data X)	Data points 20,915 samples contained solubility in qualitative terms (e.g., "greater than" or "less than"). (Data Y)

3.7.4.3 Handling duplicate data

Once the data is divided, check for duplicates and remove them.

Total number of samples after removal of duplicates	
Unique data points: 29,652, sample data X	Unique data points: 18,652, sample data Y

The final phase of data processing involved integrating and organizing the cleaned data. Compilation of validated measurements from all processed segments, verification of data consistency across combined datasets, and implementation of final quality control checks. Once we consolidated the data, the remaining 53,789 solubility data points were divided into two segments based on the nature of the solubility information: quantitative solubility values. A total of 32,974 data points provided precise solubility values and removed duplicate entries. This segment was reduced to 29,652 unique records. Other segments of the data,, which have relative solubility values,, contain a total of 20,915 data points, that describe solubility in relative terms, such as "greater than" or "less than." Following the removal of duplicate entries, this segment contained 18,548 unique records. This segmentation formed the foundation for subsequent analyses, ensuring a clear distinction between specific and relative solubility data for model evaluation and validation.

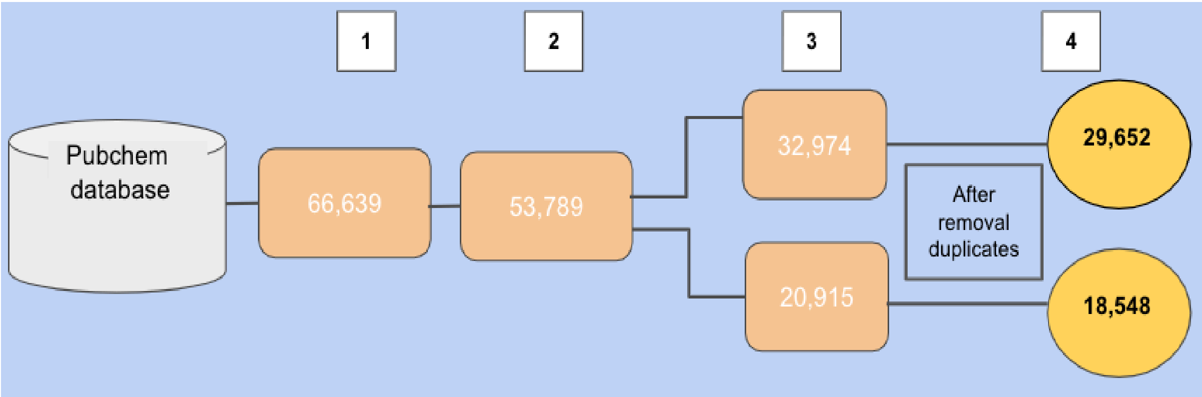


Figure 22. Schematic summary of the preprocessing pipeline for source data from Pubchem, consisting of (1) data collection, (2) cleaning based on the quantitative or qualitative value of solubility, (3) dividing the data based on the value of solubility (precise value and in the range greater than or lower), and (4) removal of duplicates

3.7.5 Data verification and validation

To ensure data reliability, we implemented several validation steps, including the assessment of solubility values against established physical limits and the identification and review of statistical outliers, as well as the literature-based validation of extreme values,, and the removal of physically implausible measurements.

3.7.6 Data evaluation

Table 6: Details of 5 different datasets used to compare the solubility values with the PubChem database

Dataset	Authors	Dataset Size	Use	Duplicate	Unique	Reference
A	BNN Lab	900	Training	4	898	[22]
B	Gihan	11862	Training	261	11724	[24]
C	Xian Zeng	9942	Training	347	9750	[4]
D	Sorkun	6154	Training	471	5907	[36]
E	Huuskonen	1291	Testing	18	1282	[12]

This evaluation aimed to determine the overlap and consistency of the processed data with established solubility values. Among the 29,652 unique records, 357 data points matched values from the literature reference set in the quantitative solubility values.

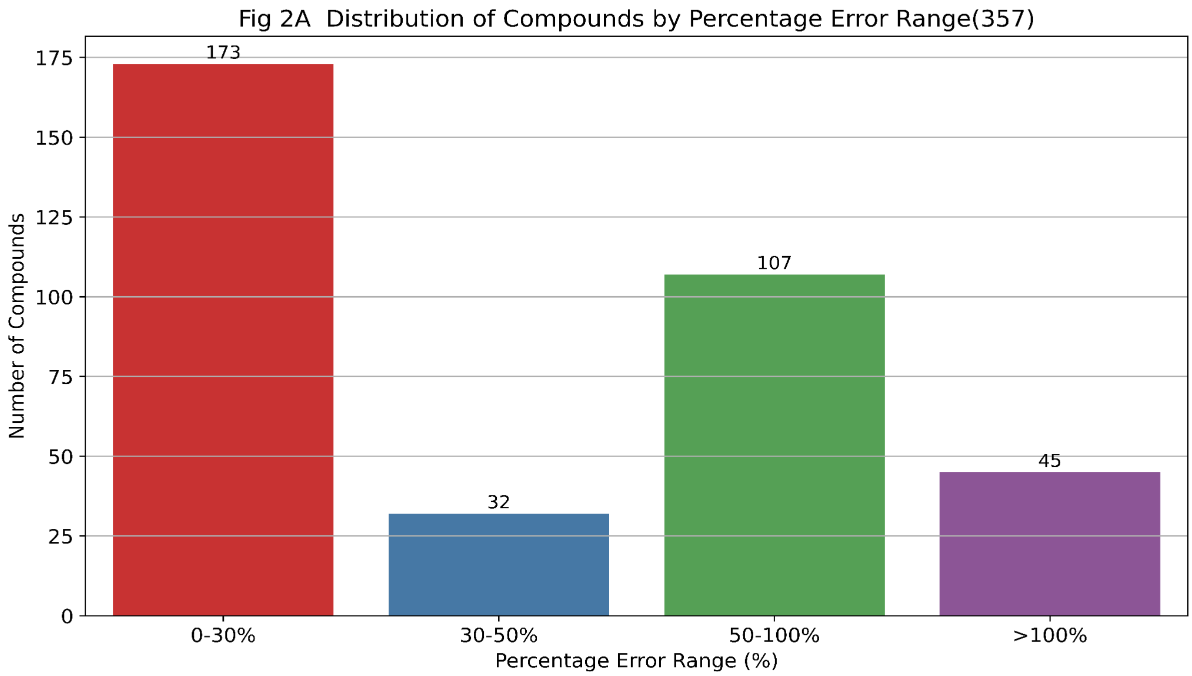


Figure 23. Distribution error between the match of X data solubility records with literature values. The y-axis represents the number of records, while the x-axis categorizes the solubility error into bins: 0–10%, 10–25%, 25–50%, 50–100%, and more than 100%.

For relative solubility values out of 18,548 unique records, 498 data points showed consistency with the literature reference set.

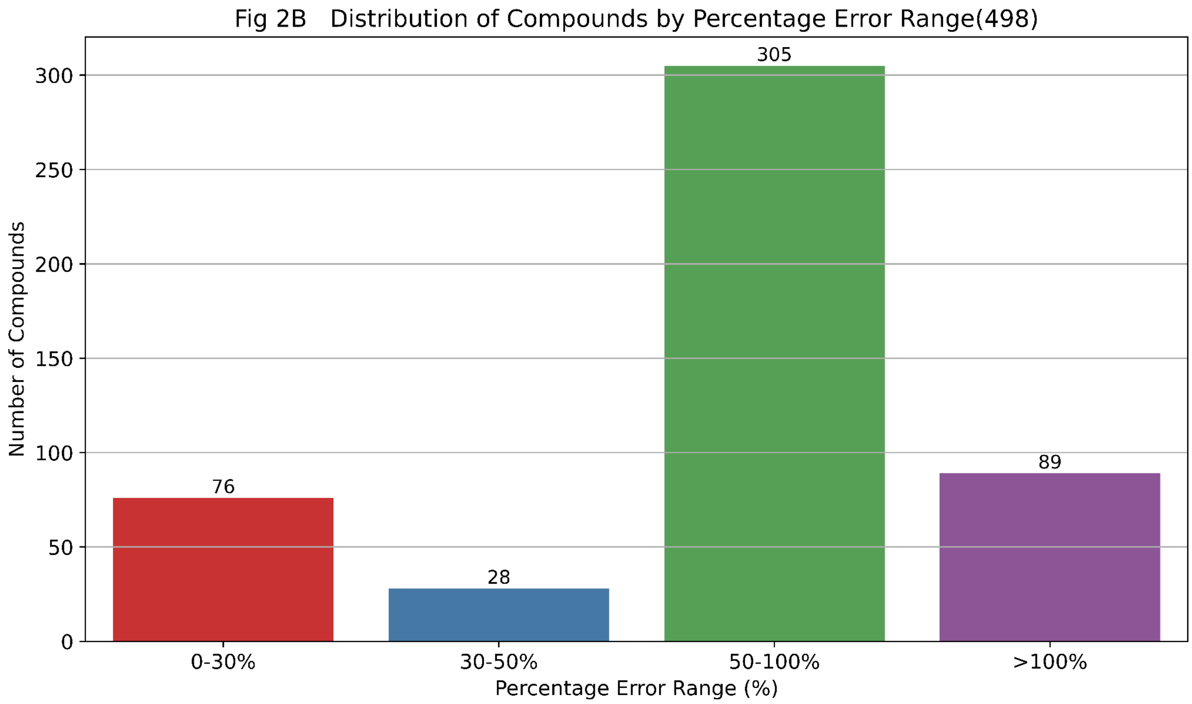


Figure 24. Distribution error between the match of Y data solubility records with literature values. The y-axis represents the number of records, while the x-axis categorizes the solubility error into bins: 0–10%, 10–25%, 25–50%, 50–100%, and more than 100%.

This comparative analysis highlights the alignment of the processed solubility dataset with existing literature. Although the number of exact matches was limited, these findings validate the dataset's potential utility for machine learning applications in solubility prediction. The curated dataset provides a robust foundation for predictive modeling, particularly in addressing challenges associated with solubility estimation across diverse chemical compounds.

3.7.7 Comparison with variation of data

The finalized dataset, ,consisting of 29,652 entries,, was divided into two categories: with salt (801) and without salt (28,851). This division was crucial for evaluating the predictive capabilities of the dataset under various conditions. Predictions were generated for both subsets (with salt and without salt) using the Huuskonen benchmark test dataset as a reference. Results were evaluated to understand the prediction accuracy, precision, and overall reliability. The predictive performance of the finalized dataset was compared against results derived from Reaxys data. A similar methodology was applied to ensure consistency in comparisons, focusing on statistical measures such as R^2, RMS,E and MAE; data from relevant literature was used as an additional benchmark for comparison. Historical prediction models and data trends were analyzed to gauge the relative performance of the finalized dataset. Details are given in Table 2. The potential predictive capability of data sourced from PubChem is critically assessed. Statistical and qualitative analyses were conducted to assess the alignment between PubChem-based predictions and benchmark datasets and literature values.

Table 7: Comparison of test set performance with different data set combinations using the metrics MAE (Mean absolute error), RMSE (Root mean square error), and R^2 (Coefficient of determination) applied to the model XGB. We also divided the data into I (Ionized form) and NI (non-ionized form) to analyze the variation in the result.

Entry	Included Databases						No FP[c]	MAE	RMSE	R^2
	PubChem[a]		Reaxys[b]		Literature[c]					
	I	NI	I	NI	I	NI				
1			6293					1.77	2.22	0.18
2				5953				0.71	0.92	0.79
3			5200	5953				1.27	2.30	0.27
4	811		6293		3022			1.32	1.62	0.36
5		28851(X)		5953		14915		0.40	0.56	0.92
6		28851(X)						0.92	1.24	0.62
7	811	28851(X)						0.91	1.24	0.63
8	1004	17544(Y)						1.12	2.24	0.48
9	1004							1.74	4.71	0.13
10		17544(Y)						1.11	2.09	0.49
11	28851	17544(Y)						1.78	2.23	0.19

3.7.8 Comparison with literature data

Data from relevant literature served as an additional benchmark for comparison. Historical prediction models and data trends were analyzed to evaluate the relative performance of the finalized dataset. A detailed comparison of the Huuskonen test dataset with other datasets from the literature was conducted. Table 8 provides a comprehensive breakdown of this comparison, including the origins of the datasets compared, which highlight datasets from experimental studies, computational predictions, and hybrid approaches. Trends in performance variation were noted, particularly how the Huusskonen test dataset performed under varying conditions and its alignment with other datasets in terms of prediction quality. The comparison revealed that the Häusskonen dataset consistently demonstrated competitive or superior prediction accuracy when benchmarked against similar datasets in the literature, validating its reliability and applicability. This comparison offered more profound insight into the predictive capability of PubChem data. While PubChem-derived predictions did not perform exceptionally well, they were evaluated against previous predictions to assess their consistency and potential as a data source for predictive tasks.

Table 8: Comparison of curated Pubchem data on the test dataset, which was used in previous work in the literature.

Entry	**Method**	**Test dataset size**	**MAE**	**RMSE**	**R^2**	**Year**	**Reference**
1	ANN	1294	-	0.71	0.88	2000	Huuskonen[46]
2	ANN	1291	-	0.62	0.91	2001	Tetko et al[63]
3	ANN	1294	0.68	0.59	0.92	2003	Yan and Gasteiger[47]
4	MLR	1290	0.68	0.87	0.71	2004	Delaney[16]

Entry	**Method**	**Test dataset size**	**MAE**	**RMSE**	**R^2**	**Year**	**Reference**
5	MLR	1294	0.52	0.63	0.90	2004	Hou et al.[48]
6	SVM	1290	0.43	0.60	-	2007	Schroeter et al. [49]
7	MLR	1290	0.72	0.94	0.73	2012	Ali et al.[50]
8	UG-RNN	1026	0.46	0.60	0.91	2013	Lusci et al.[37]
9	MLR	1290	0.93	1.15	0.68	2017	Daina et al. [51]
10	ANN	1297	-	0.65	0.90	2018	Bjerrum and Sattarov[51,52]
11a	Consensus	1290	(0.39)	(0.53)	(0.93)	2020	Sorkun et al.[36]
11b*			0.54*	0.73*	0.87*	2024*	rework from Sorkun et al*
12	**XGB-298**	**1282**	**0.91**	**1.24**	**0.63**	**2025**	**With Pubchem data(X)**

*Results obtained with the model and test data of Sorkun *et al.* after the exclusion of data referring to the same SMILES code in training and test datasets.**ANN**, artificial neural network, **MLR,** multiple linear regression, SVM, **support** vector machine **UG RNN,** Undirected graph recursive neural networks, **RF,** random forest; **XGB** Extreme Gradient Boosting; **Consensus,** an ensemble of ANN, RF, and XGB. Best results are given in bold; best results in past developments are underlined.

3.8 Multi-solvent prediction

In this work, we present a multi-solvent solubility prediction platform to enable accurate solubility prediction in nine diverse solvents and the effect of temperature variation. We utilize curation and preprocessing of solvent-specific solubility data in a uniform and trustworthy format to enable predictive modeling. We include solute and solvent molecular descriptors, as well as temperature, as primary features, so that our models can capture the subtle relationships between solubility, solvent, and temperature properties.

Additionally, we perform comparative solubility variation analysis across solvents and temperatures, discerning the function of polarity, hydrogen bonding, solvation effects, and thermal fluctuations. Through stepwise examination of solvent- and temperature-dependent trends, we refine our prediction techniques to optimize model precision and generalizability.

Through implementing this framework, we aim to provide a systematic, data-driven approach to solubility prediction, offering insight into solute-solvent interactions across varying conditions.

3.8.1 Data curation and preprocessing

To ensure high-quality and unbiased multi-solvent solubility predictions, we carefully curated the dataset, eliminating duplicate entries to prevent redundancy and bias. This preprocessing pipeline enhances data integrity and model reliability.

We developed a multi-solvent solubility prediction model using a dataset comprising 54,272 data points, each representing a combination of solute, solvent, and temperature. The dataset was processed to focus on nine solvents, each having at least 200 unique solutes, resulting in a working dataset of approximately 32,707 data points. Molecular descriptors were calculated separately for both solutes and solvents, capturing their physicochemical properties, while temperature was incorporated as a critical environmental variable.

Table 9: Details of the datasets used to generate a unique dataset to train and test the model

Authors	Dataset Size	Use	Duplicate	Unique	Reference
Krasnov	54273	Train and Test	1369	52904	Krasnov[20]

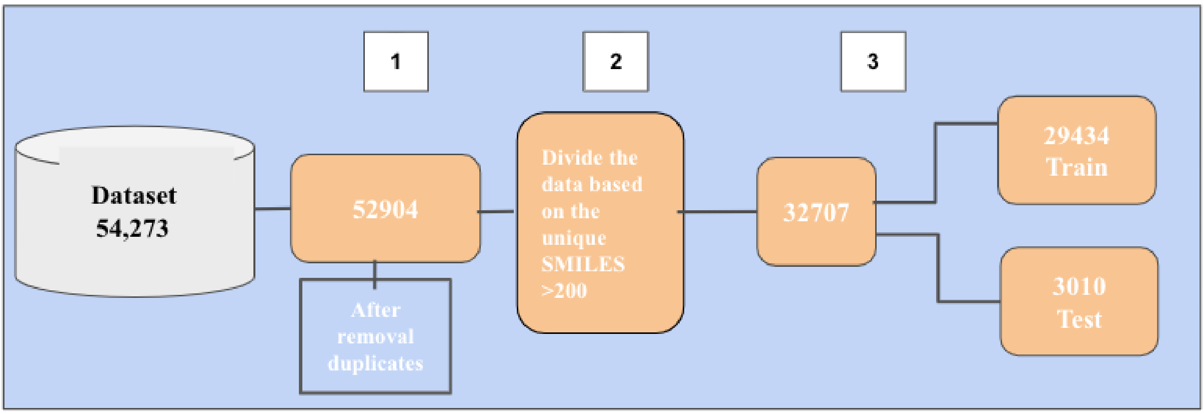


Figure 25. Schematic presentation of the preprocessing pipeline for training and test data consisting of (1) removal of duplicates, (2) selection of the dataset based on unique SMILES present across solvents, and (3) dividing the data into train and test sets to evaluate the performance of the model.

3.8.1.1 Removing duplicates from the dataset

Duplicates in machine learning datasets can significantly impact model performance, generalizability, and interpretability. During training, the presence of duplicate samples can introduce bias, causing the model to memorize specific patterns rather than learning to generalize to unseen data. While this may lead to artificially high accuracy on the training set, it often results in poor real-world performance. To address this, we first remove the duplicates. This initial step allowed us to systematically identify and manage redundant records, ensuring a more balanced and representative dataset for model training.

Unique samples in the dataset after removing the duplicates
Total unique data 52,904

Unique samples across the solvents
Total data points 32,707

3.8.1.2 Remove outliers from the unique dataset

For better representation of the data, we remove outliers

Total number of samples after removal of outliers
Total data points 32,444

3.8.1.3 Dividing the data into a train and test set

To evaluate the model,, we have separated the data into a train-test set with a a 90:10 ratio.

Total number of samples after removal of duplicates	
Training data: 29,434 samples	Test data: 3,010 samples

3.8.1.4 Finding the unique across solvents and selection of the data

We selected unique data points while ensuring the maximum possible coverage. Specifically, we selected 200 unique data points for each solvent, resulting in a total of nine solvents

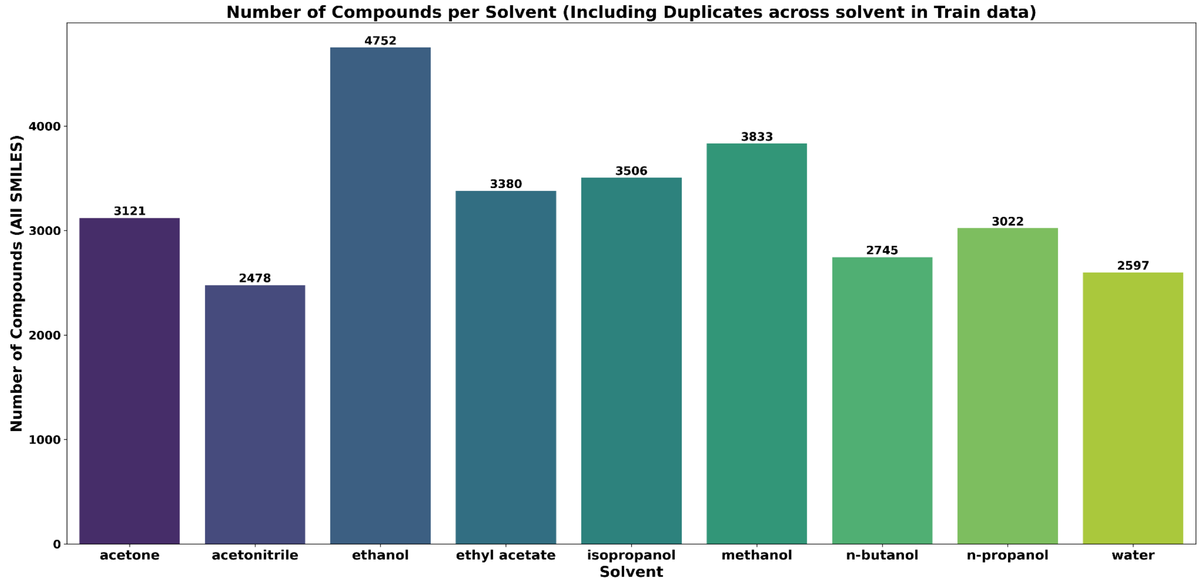


Figure 26. Distribution of compounds across solvents in the training dataset. Each bar represents a specific solvent, showing the total number of compounds associated with it (including duplicates across different temperatures). This visualization highlights the solvent-specific compound diversity present in the training data, offering insight into how well each solvent is represented.

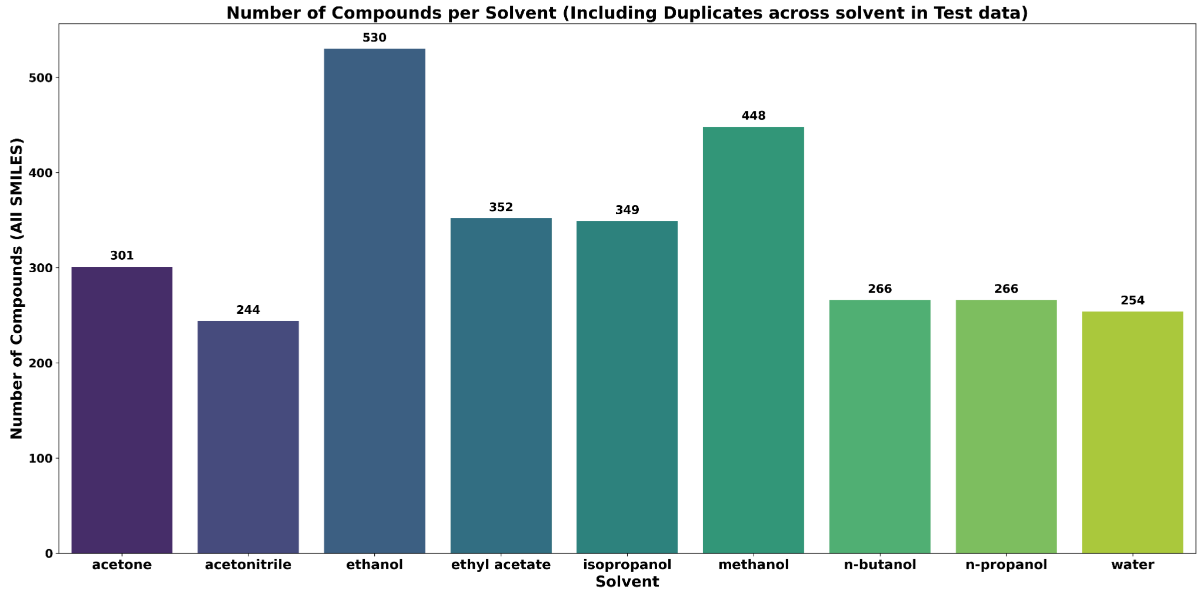


Figure 27. Distribution of compounds across solvents in the test dataset. Similar to the training data, each bar shows the total number of compounds for each solvent. This visualization helps evaluate the solvent and compound coverage in the test set, ensuring it adequately represents chemical and solvent diversity.

3.8.2 Feature Generation

For each solute and solvent**,** molecular descriptors **were** computed using the cheminformatics tools **from the** RDKit. These descriptors capture a wide range of physicochemical properties, including molecular weight, LogP (Octanol-Water Partition Coefficient), topological polar surface area (TPSA), number of hydrogen bond donors, number of hydrogen bond acceptors, rotatable bonds, aromatic ring count, fraction of sp3 carbon atoms (Fsp3), and electrotopological state indices.

3.8.3 Expansion of the feature space

To systematically investigate the impact of including additional descriptors on the model's performance, we incrementally expanded the descriptor set. This stepwise augmentation allowed us to assess how the inclusion of a higher number of descriptors influences the accuracy and predictive power of our model. As we progressed, we continually introduced new descriptors into our feature space, each chosen to capture specific chemical attributes and properties.

In addition to the primary molecular descriptors, we further enhanced feature richness by exploring solvent-specific statistical properties, such as the mismatchsolvent polarity index, dielectric constant, and combined interaction features, such as the difference in LogP between the solute and solvent, or polarity mismatch[64]. Hybrid features, representing interaction terms between solute and solvent descriptors, were also incorporated to capture synergistic effects (e.g., solute TPSA × solvent LogP)[65]. Temperature-derived interaction features, such as temperature-normalized descriptors or temperature × polarity terms, were introduced to account for temperature-dependent solvation dynamics [66]Both solute descriptors and solvent descriptors were included as separate feature blocks in the model, ensuring the model learns the contributions of both chemical entities to the overall solubility behavior. Temperature was directly included as an additional numeric feature, allowing the model to capture thermal effects on solubility (Miller & Zhang, 2022).

3.9 Model building

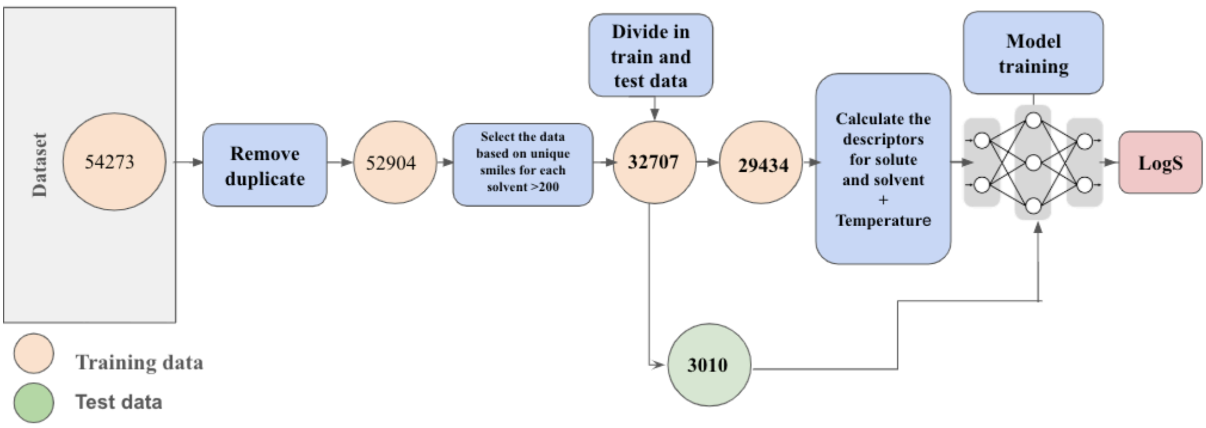


Figure 28: Workflow of multi-solvent solubility prediction. Training a predictive model for solubility involves merging the molecular descriptors of both the solute and the solvent, alongside temperature data. These features are used to train a machine learning model, which then predicts the solubility of the solute in the given solvent under specific temperature conditions.

To develop a robust and generalizable predictive model for solubility across multiple solvents and temperatures, we followed a structured pipeline, described in Figure 28. This figure provides a step-by-step overview of the modeling workflow, encompassing all essential stages from data preprocessing and descriptor generation to model training and solubility prediction.

The initial stage involves data preparation, as detailed in Section 3.8.1. Data curation and preprocessing, comprising unique combinations of solute, solvent, and temperature, form the foundation of the modeling process. Each entry in the dataset is encoded using molecular descriptors extracted separately for solutes and solvents via cheminformatics libraries such as RDKit. These descriptors encapsulate key physicochemical properties like molecular weight, topological polar surface area (TPSA), number of rotatable bonds, hydrogen bonding capacity, LogP, aromaticity, and fraction of sp^3-hybridized carbon atoms (Fsp^3). These features provide an intrinsic chemical representation necessary for learning structure–property relationships.

In addition to these primary descriptors, the workflow incorporates an expansion of the feature space to account for more nuanced interactions. The feature blocks for solutes, solvents, and temperature are then concatenated into a unified input vector, as shown in the figure, creating a rich feature space that encodes both individual molecular properties and environmental interactions.

In the modeling stage, machine learning and deep learning algorithms are trained on this comprehensive feature set to learn the quantitative relationship between these features and the experimentally measured solubility. The model learns to recognize how certain descriptor patterns correlate with solubility under specific solvent and temperature conditions.

The final trained model, once validated, can predict the solubility of a new compound in a target solvent at a specific temperature.

This prediction framework is not only accurate but also scalable to a wide range of chemical spaces, offering insight into solute–solvent–temperature interactions.

3.9.1 Comparison with variations of features on test data

Table 10: Comparison of test set performance on different models and combinations of descriptors given by the metrics MAE, RMSE, and R^2, across the solvent

Entry	Model Name	MAE	RMSE	R^2	Number of descriptors/fingerprint version				
					Descr. [a]	FP[b]	FGs[c]	Feat.[d]	Layers
1	XGB-5	0.67	0.89	0.39	5	-	-	-	-
2	XGB-20	0.64	0.85	0.44	20	-	-	-	-
3	XGB-125	0.62	0.83	0.59	125	-	-	-	-
6	XGB-298	0.60	0.82	0.51	125	128	7	38	
12	ANN	0.71	0.84	0.46	-	-	-	-	6
9	RANDOM FOREST	0.68	0.88	0.43	125	128	7	38	
10	MPNN	0.74	0.87	0.44	125	128	7	38	6
11	Hybrid MPNN	0.66	0.83	0.48	125	128	7	38	7

[a]Model includes the given number of descriptors; [b]Type of Fingerprint: [c]FGs = number of functional groups included; [d]Additional selected descriptors included.

3.9.2 Comparative results across the solvents

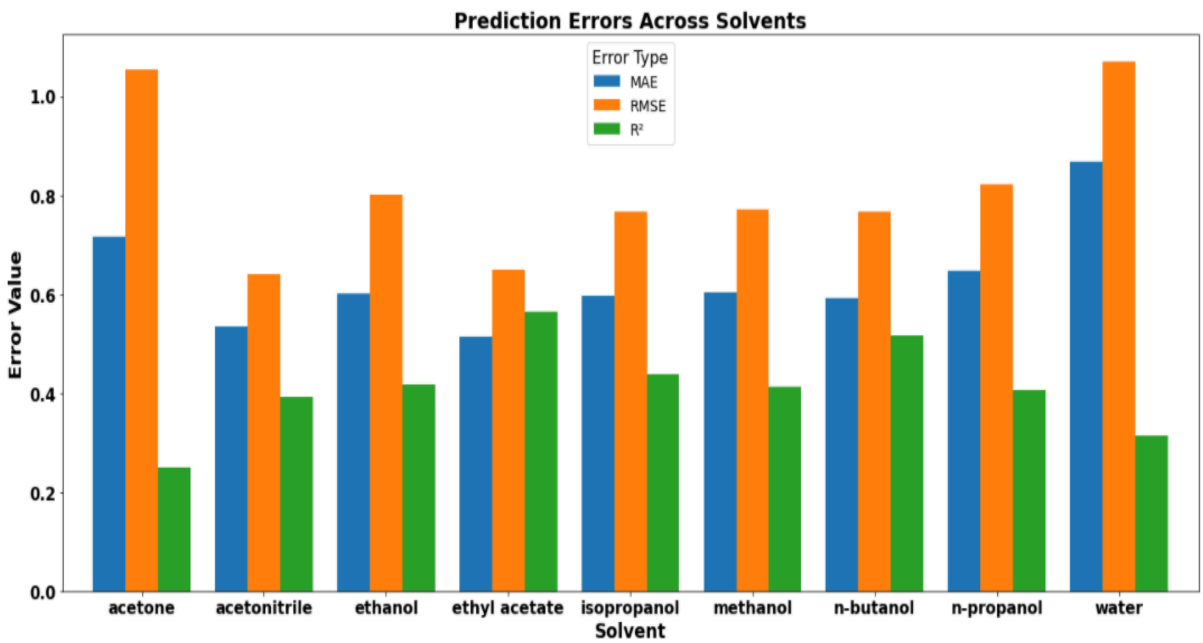


Figure 29 Prediction Errors Across Solvents: The bar chart illustrates the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 values for solubility predictions in different solvents. Bold solvent names indicate distinct categories, while the y-axis represents error values. The figure highlights variations in prediction accuracy across solvents.

3.9.3 Prediction error with temperature

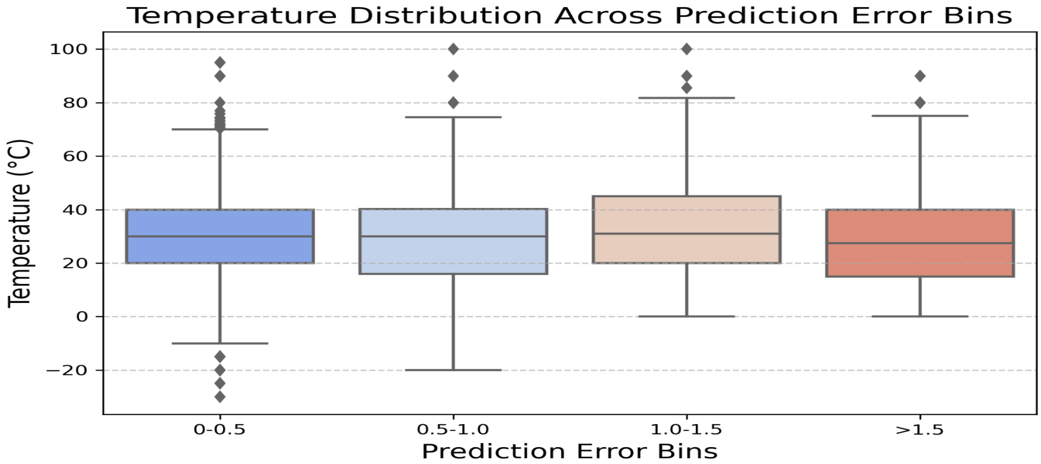


Figure 30. Prediction Errors Across Temperature. Distribution of temperature across different prediction error bins. The plot highlights variations in temperature within each bin, showing the spread and potential outliers.

3.9.4 Prediction error compared with molecular weight

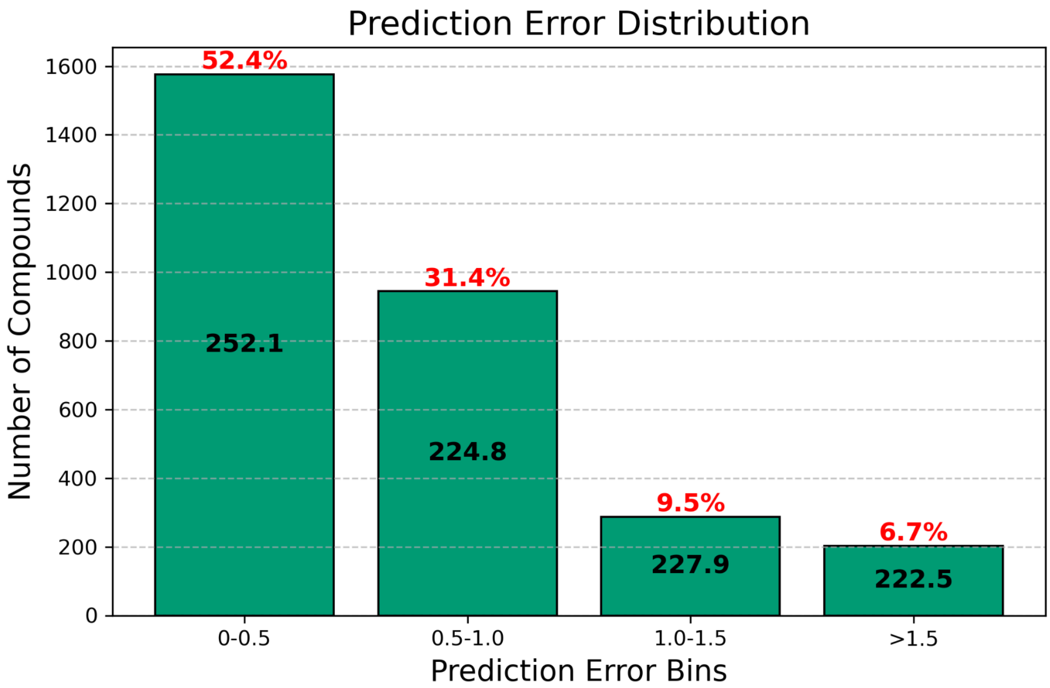


Figure 31. Prediction Errors Across the Average Molecular Weight. The bar chart illustrates the distribution of absolute prediction errors for solubility estimation, categorized into predefined error ranges. The height of each bar represents the number of compounds within each error bin, while the percentage of compounds in each category is displayed near the top of the bars. Additionally, the average molecular weight (MW) of compounds in each bin is shown at the center of the bars, providing insights into the correlation between molecular weight and prediction accuracy. This visualization helps to assess the model's performance across different molecular weight ranges.

4. Conclusion

This work explored novel computational approaches to solubility prediction, addressing key challenges in both single-solvent and multi-solvent scenarios while leveraging large-scale data acquisition from PubChem to enhance predictive modeling. By integrating advanced machine learning algorithms, feature engineering techniques, and multi-modal data fusion, the research significantly improved solubility predictions across diverse chemical environments.

For aqueous solubility prediction, the developed models achieved a Mean Absolute Error (MAE) of 0.40 and an R^2 of 0.92, demonstrating a substantial improvement over traditional methods and existing literature benchmarks. These results highlight the model's ability to accurately capture complex molecular interactions and enhance the predictive reliability of solubility estimation, particularly for drug-like compounds.

Beyond water, this work successfully extended solubility prediction to nine different solvents, incorporating solvent-specific interactions and thermodynamic effects by explicitly including temperature as a variable in the model. The multi-solvent models achieved an MAE of 0.60, demonstrating strong predictive performance across different solvent systems and temperature conditions. These results confirm the feasibility of scalable, data-driven solubility modeling for broader applications in the chemical and pharmaceutical industries. However, while the models effectively integrate temperature variations, further validation across a broader range of temperature conditions and less common solvent systems could enhance their generalizability.

A crucial aspect of this research was the curation and preprocessing of large-scale PubChem data, ensuring high-quality and diverse training material for solubility prediction models. The study demonstrated that incorporating extensive molecular descriptors and physicochemical properties into machine learning models significantly boosts predictive performance. However, data inconsistencies and missing values within public datasets posed challenges, emphasizing the need for rigorous data validation and experimental verification.

Limitations and future research directions

While this research achieved notable advancements in solubility prediction, several limitations remain:

1. **Generalization to additional solvents:** The models performed well for the nine studied solvents, but their performance across a broader solvent space remains an open challenge. Expanding the dataset and applying transfer learning techniques could improve generalization.
2. **Experimental validation:** The study relied on existing datasets, and although rigorous model evaluation was performed, experimental validation of predictions remains a crucial next step. Collaborating with experimental chemists to generate high-quality benchmark datasets will enhance the reliability of the model.
3. **Interpretability and explainability:** Despite achieving high predictive accuracy, the black-box nature of machine learning models limits interpretability. Future research should integrate explainable AI (XAI) techniques to provide deeper insights into the key molecular features that drive solubility.
4. **Data quality and standardization:** Although PubChem provides extensive data, inconsistencies and measurement errors pose challenges. Future efforts should focus on automated data curation, outlier detection, and experimental standardization to further improve data reliability.

Final thought

This thesis presents a scalable and data-driven framework for solubility prediction, demonstrating the power of machine learning in advancing molecular property predictions.

By addressing the identified limitations and incorporating emerging computational techniques, future research can further refine predictive accuracy and expand solubility modeling applications in drug discovery, material science, and chemical engineering.

5. Experimental part

5.1 General

All transfers of liquids were performed using VWR® HIGH PERFORMANCE (20 µL - 200 µL and 100 µL - 1000 µL) microliter pipettes. An AGILENT 1100 HPLC system equipped with a diode array detector and a VDSphere 100 C18-E (5 µm, 250 x 4.0 mm) column from VDS OPTILAB was used for HPLC analytics. The following program was applied:

HPLC (VDSpher 100, C18-E, 5 µm, 250 x 4.0 mm; 20 °C, injector volume 5.0 µl; ACN/H2O 10:90 (0-1 min), 10:90 to 99:1 (1-6 min), 99:1 (6-9 min), 99:1 to 10:90 (9-10 min), 10:90 (10-13 min), 1 mL/min; λ = 254 nm). Each sample was injected twice to avoid injection errors, and the mean was used for the calculations. For each measurement, the area under the curve (AUC) was determined and correlated to the concentration.

5.2 Calibration curve

For each of the compounds, an 8-point calibration (9-point calibration for compound **1**) was performed before the determination of their maximum solubility. Therefore, a solution with a concentration of 1.0 mg/mL in water was prepared as a standard solution: About 5 mg of substance (for compound-specific descriptions, see further sections) was weighed into a 10 mL crimp vial. Bi-distilled water was added using microliter pipettes (amount according to the exact weight of substance), the vial was closed, and the solution was shaken for 24 h at 25°C. This standard solution was diluted with acetonitrile to obtain concentrations between 0.02 (additionally 0.01 mg/mL for **1**) and 1 mg/mL:

a. 1.0 mg/mL: pure standard solution
b. 0.8 mg/mL: 800 µL standard solution + 200 µL of MeCN
c. 0.6 mg/mL: 600 µL standard solution + 400 µL of MeCN
d. 0.4 mg/mL: 400 µL standard solution + 600 µL of MeCN
e. 0.2 mg/mL: 200 µL standard solution + 800 µL of MeCN
f. 0.1 mg/mL: 100 µL standard solution + 900 µL of MeCN
g. 0.05 mg/mL: 50 µL standard solution + 950 µL of MeCN
h. 0.02 mg/mL: 20 µL standard solution + 980 µL of MeCN
i. 0.01 mg/mL: 10 µL standard solution + 990 µL of MeCN

All samples were measured using the above-mentioned conditions in HPLC. For each compound, this whole procedure was repeated 2-4 times to gain sufficient replicates. All values were taken into account for the calculation of the calibration curves.

5.3 Experimental determination of solubility

Each substance was weighed into a 1.5 mL Eppendorf vial, equipped with a magnetic stirring bar (10 x 3mm), and stirred at 25°C for 24 h. The vial was centrifuged with a VWR® MicroStar 12 (10.000 rpm, 10 min), and 100 µL of the supernatant was diluted with 900 µL of MeCN and measured on HPLC according to the standard method. In case the values (area under the curve) were close to the maximum values of the calibration curve, additional dilutions were prepared. Therefore, 100 µL of the first dilution was again diluted with 900 µL of MeCN. For each compound, the procedure to determine the maximum solubility was performed 2-4 times. The area under the curve was determined for 9 different concentrations. Subsequently, a calibration curve was created assuming a linear relationship (Eq 1) or a saturation curve (Eq 2).

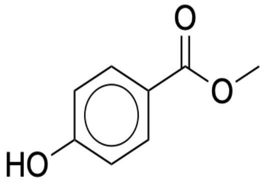
(1) $y = a + b * conc;$
(2) $y = a * conc / (b + conc);$

Furthermore, the R^2 was calculated as an indication of the goodness of the fit. In addition, the root mean squared error (RSS) was calculated, which directly quantifies the differences between the fit and the observed data.

Table 11: SMILES representation of the compounds along with experimental, literature, and predicted solubility values

		Solubility [g/L]		
Smiles	**Cpd Label**	**Experimental**	**Literature**	**Predicted**
COC(=O)c1ccc(cc1)O	1	2.59 ± 0.14	2.50[57]	2.27
COc1cc(C(=O)O)cc(OC)c1OC	2	0.95 ± 0.033	2.48[58]	2.97
CCOC(=O)c1ccc(O)c(O)c1	3	4.20 ± 0.66	insoluble[59]	2.32
OC(c1cc(OC)c(OC)cc1)=O	4	1.085 ± 0.078	0.50[60]	0.79
CC(Nc1ccccc1OC)=O	5	11.09 ± 0.18	17.40[61]	5.37

Methyl 4-hydroxybenzoate (1)

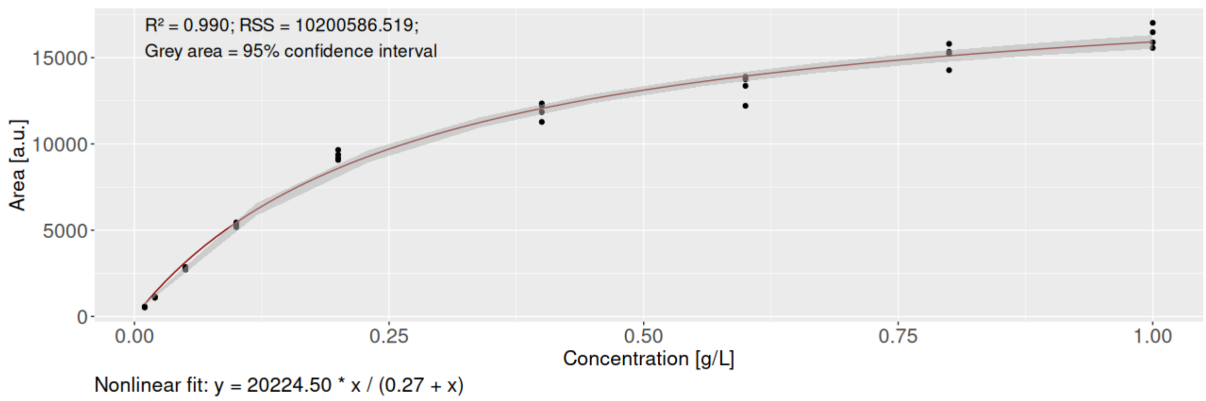


As a standard solution, methyl 4-hydroxybenzoate was dissolved in bi-distilled water according to the general procedure for the calibration. Here, 6.88 mg were dissolved in 6.88 mL of bi-distilled water (replicates: 5.11 mg in 5.11 mL, 5.86 mg in 5.86 mL, 7.79 mg in 7.79 mL). To determine the solubility, 9.19 mg (replicates: 12.57 mg, 24.1 mg, 18.89 mg) of the compound were dissolved in 1.0 mL of bi-distilled water according to the procedure for the determination of the solubility and measured on HPLC.

Table 12: HPLC calibration data and curve for compound 1, including the concentration (in g/L) and area under the curve (AUC). The graph shows the data points and the resulting linear fit.

Concentration [g/L]	Standard [1mg/mL]	Standard [uL]	MeCN [µL]	AUC
1	1	500	0	15565.0, 16473.5, 17016.0, 15890.5
0.8	1	800	200	14276.0, 15337.0, 15802.5, 15226.5
0.6	1	600	400	12210.5, 13366.5, 13748.0, 13892.5
0.4	1	400	600	11274.5, 11851.0, 12346.0, 12213.5
0.2	1	200	800	9070.0, 9377.0, 9653.0, 9194.5
0.1	1	100	900	5173.5, 5298.5, 5450
0.05	1	50	950	2724.5, 2719, 2866
0.02	1	20	980	1103.5, 1085.5, 1162
0.01	1	10	990	546, 524.5, 567
Back calculation				
Concentration [g/L]			**AUC [a]**	
2.59 ± 0.14			101810, 102630, 102070, 102880, 98630, 99590, 99060, 99560, 95340, 95190, 98240, 97890, 95400, 95560, 97630, 97460	

[a] For the determination of the maximum solubility, all values of the double injection were taken into account.



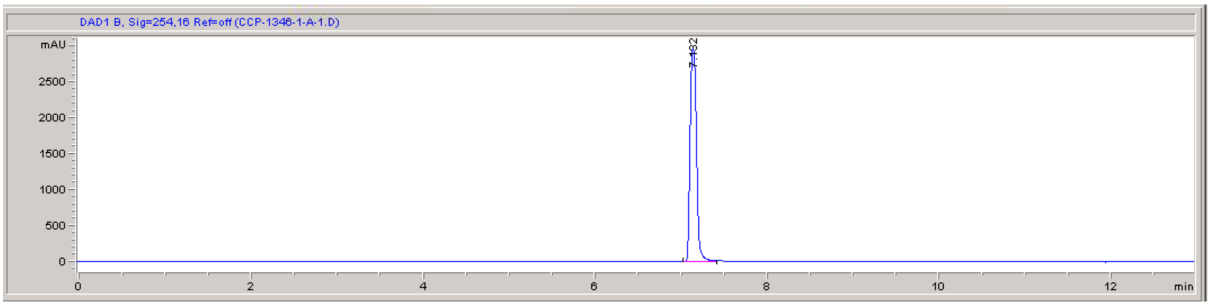
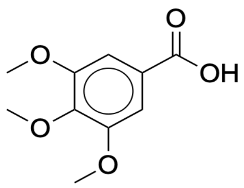


Figure 32: Exemplary chromatogram obtained for the measurement of compound **1** at a concentration of 1.0 mg/mL at λ = 254 nm. The area under the curve was found to be 15597.

3,4,5-Trimethoxybenzoic acid (2)

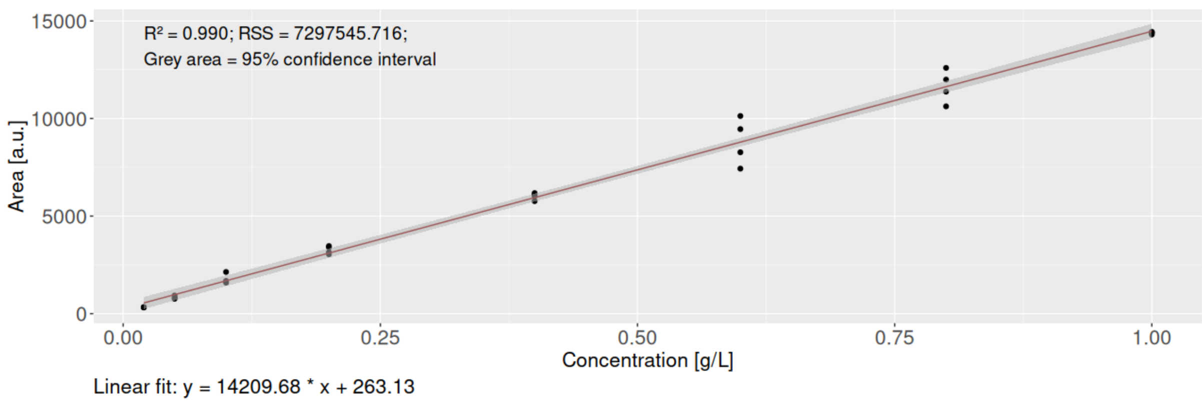


As a standard solution, 3,4,5-trimethoxybenzoic acid was dissolved in bi-distilled water according to the general procedure for the calibration. Here, 4.56 mg were dissolved in 4.56 mL of bi-distilled water (replicates: 6.24 mg in 6.24 mL, 5.72 mg in 5.72 mL, 7.97 mg in 7.97 mL). To determine the solubility, 5.56 mg (replicates: 6.14 mg, 10.17 mg) of the compound was dissolved in 1.0 mL of bidistilled water according to the procedure for the determination of the solubility and measured on HPLC.

Table 13: HPLC calibration data for compound two, including the concentration (in g/L) and the area under the curve (AUC). The graph shows the data points and the resulting linear fit.

Concentration [mg/mL]	Standard [1mg/mL]	Standard [uL]	MeCN [µL]	AUC
1	1	500	0	14303.5, 14379.5, 14444.5, 14403.5
0.8	1	800	200	12594.0, 11377.5, 10623.0, 11993.0
0.6	1	600	400	10128.5, 8269.5, 7431.5, 9457.5
0.4	1	400	600	6177.0, 5765.5, 6087.0, 6015.0
0.2	1	200	800	3398.5, 3463.5, 3150.0, 3040.5
0.1	1	100	900	1677.5, 2136.0, 1588.5
0.05	1	50	950	823.0, 929.0, 762.0
0.02	1	20	980	337.0, 320.0, 318.5
Back calculation				
Concentration [g/L]			**AUC [a]**	
0.95 ± 0.033			16450, 16420, 16260, 16270, 16210, 16210, 16150, 16150, 16550, 16630, 16270, 16260, 15730, 15740, 15070, 15020	

[a] For the determination of the maximum solubility, all values of the double injection were taken into account.



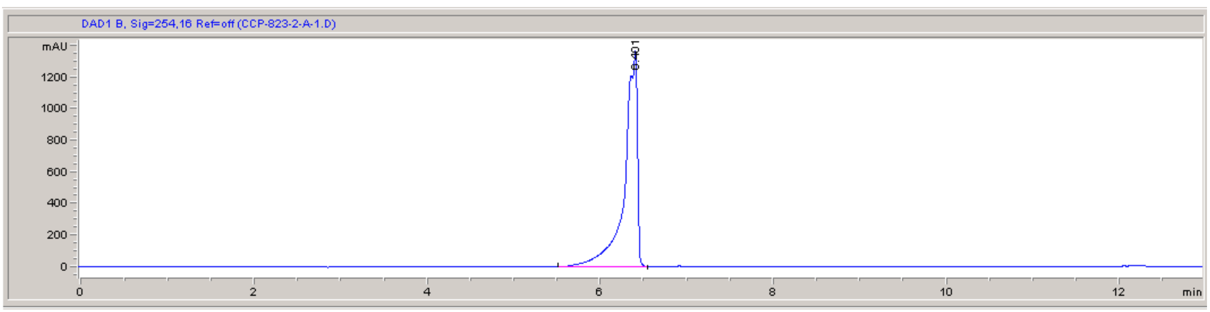
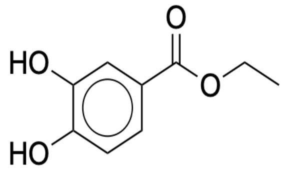


Figure 33: Exemplary chromatogram obtained for the measurement of compound **2** at a concentration of 1.0 mg/mL at λ = 254 nm. The area under the curve was found to be 14367.

Ethyl 3,4-Dihydroxybenzoate (3)

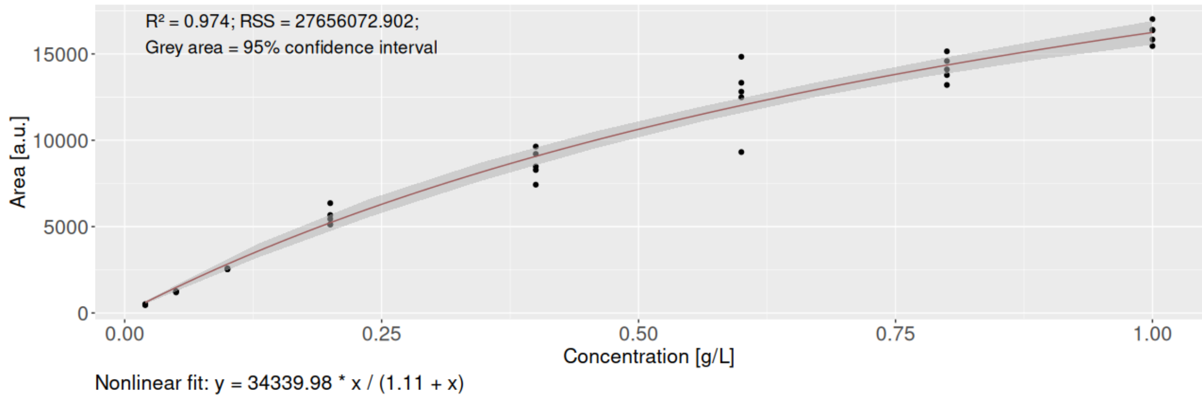


As a standard solution, ethyl 3,4-dihydroxybenzoate was dissolved in bi-distilled water according to the general procedure for the calibration. Here, 6.30 mg were dissolved in 6.30 mL of bi-distilled water (replicates: 5.87 mg in 5.87 mL, 6.46 mg in 6.46 mL, 5.70 mg in 5.70 mL, 5.64 mg in 5.64 mL). To determine the solubility, 7.64 mg (replicates: 4.72 mg, 13.99 mg, 12.61 mg) of the compound were dissolved in 1.0 mL of bi-distilled water according to the procedure for the determination of the solubility and measured on HPLC.

Table 14: HPLC calibration data for compound 3, including the concentration (in g/L) and the area under the curve (AUC). The graph shows the data points and the resulting linear fit.

Concentration [mg/mL]	Standard [1mg/mL]	Standard [uL]	MeCN [µL]	AUC
1	1	500	0	15835.0, 16397.0, 17011.5, 16364.0, 15455.5
0.8	1	800	200	13201.5, 13776.5, 14586.0, 15149.0, 14099.0
0.6	1	600	400	9316.0, 13328.0, 14836.0, 12810.0, 12488.0
0.4	1	400	600	8481.0, 8279.5, 7427.0, 9207.5, 9634.0
0.2	1	200	800	5122.5, 5670.5, 6359.0, 5441.0, 5132.5
0.1	1	100	900	2530.0, 2529.0, 2584.5
0.05	1	50	950	1228.5, 1192.0, 1261.0
0.02	1	20	980	459.5, 470.5, 510.5
Back calculation				
Concentration [g/L]			**AUC [a]**	
4.20 ± 0.66			103520, 103410, 104260, 103460, 99220, 99310, 99000, 99020, 98460, 98540, 102290, 102660, 97150, 96390, 98990, 98580, 100400, 98800, 107500, 103800, 102200, 102500, 98600, 98700	

a For the determination of the maximum solubility, all values of the double injection were taken into account.



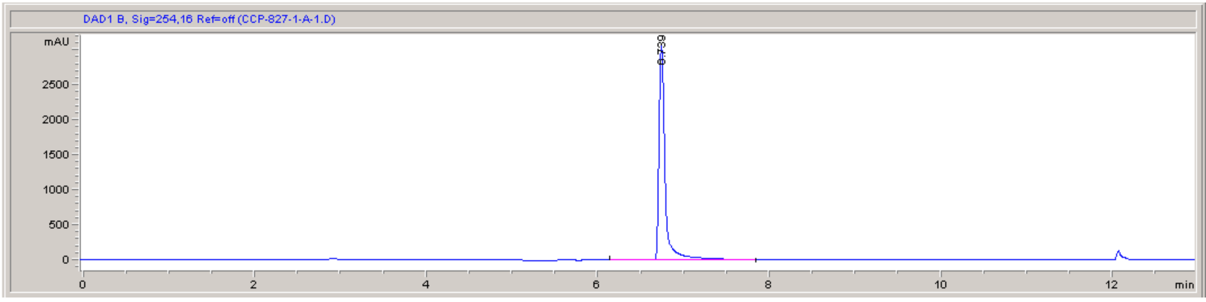
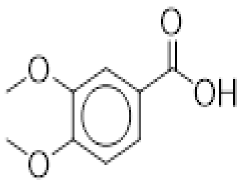


Figure 34: Exemplary chromatogram obtained for the measurement of compound **3** at a concentration of 1.0 mg/mL at λ = 254 nm. The area under the curve was found to be 15760.

3,4-Dimethoxybenzoic acid (4)



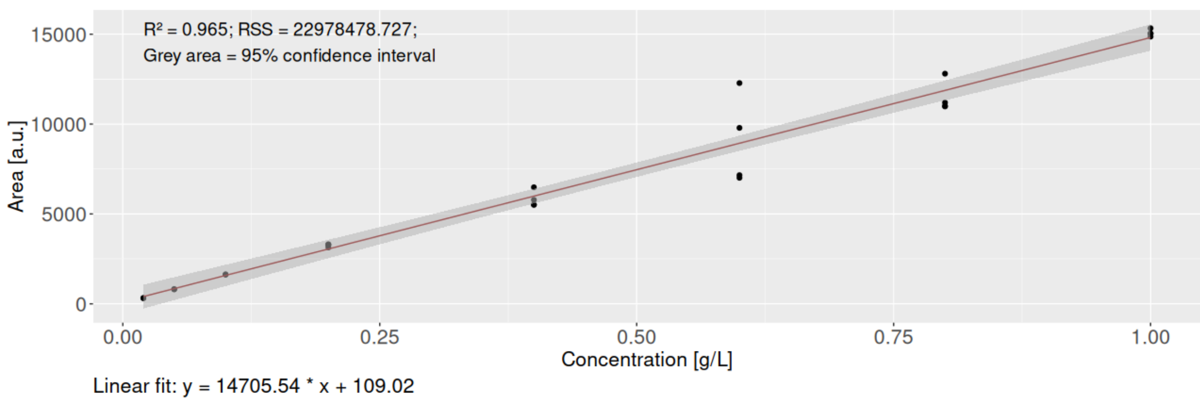
As a standard solution, 3,4-dimethoxybenzoic acid was dissolved in bi-distilled water according to the general procedure for the calibration. Here, 4.98 mg were dissolved in 4.98 mL of bi-distilled water (replicates: 5.22 mg in 5.22 mL, 6.79 mg in 6.79 mL, 6.66 mg in 6.66 mL). To determine the solubility, 6.34 mg (replicates: 7.18 mg, 26.90 mg, 29.36 mg) of the compound were dissolved in 1.0 mL of bi-distilled water according to the procedure for the determination of the solubility and measured on HPLC.

Concentration [mg/mL]	Standard [1mg/mL]	Standard [uL]	MeCN [µL]	AUC
1	1	500	0	15044.0, 15030.0, 14864.0, 15331.0
0.8	1	800	200	11001.0, 11194.0, 10983.0, 12803.5
0.6	1	600	400	12281.0, 7157.0, 7011.0, 9788.5
0.4	1	400	600	5515.0, 5782.0, 5513.5, 6483.0

0.2	1	200	800	3308.0, 3296.5, 3134.5, 3258.0
0.1	1	100	900	1635.5, 1625.5
0.05	1	50	950	810.0, 813.0
0.02	1	20	980	318.5, 320.0
Back calculation				
Concentration [g/L]			**AUC [a]**	
1.085 ± 0.078			17080, 17080, 17450, 17440, 18670, 18510, 16750, 16700, 15750, 15810, 18330, 18330, 15730, 15750, 16910, 16280	

[a] For the determination of the maximum solubility, all values of the double injection were taken into account.

Table 15: HPLC calibration data for compound 4, including the concentration (in g/L) and the area under the curve (AUC). The graph shows the data points and the resulting linear fit.



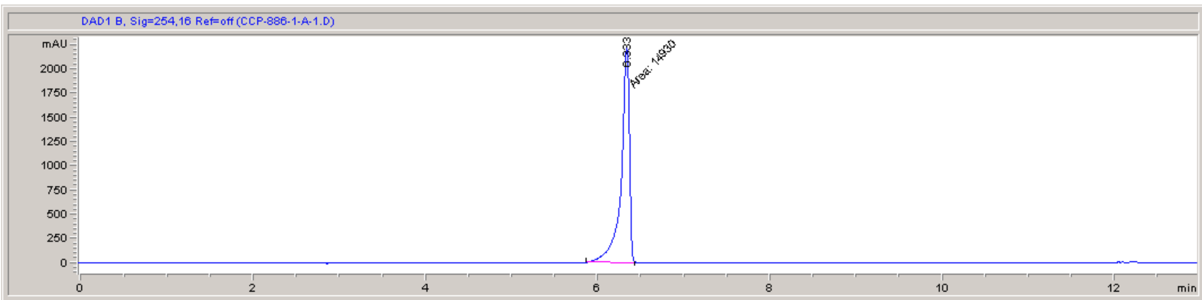
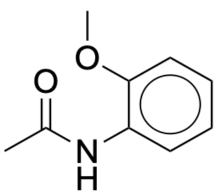


Figure 35: Exemplary chromatogram obtained for the measurement of compound **4** at a concentration of 1.0 mg/mL at λ = 254 nm. The area under the curve was found to be 14930.

N-(2-methoxyphenyl)acetamide (5)



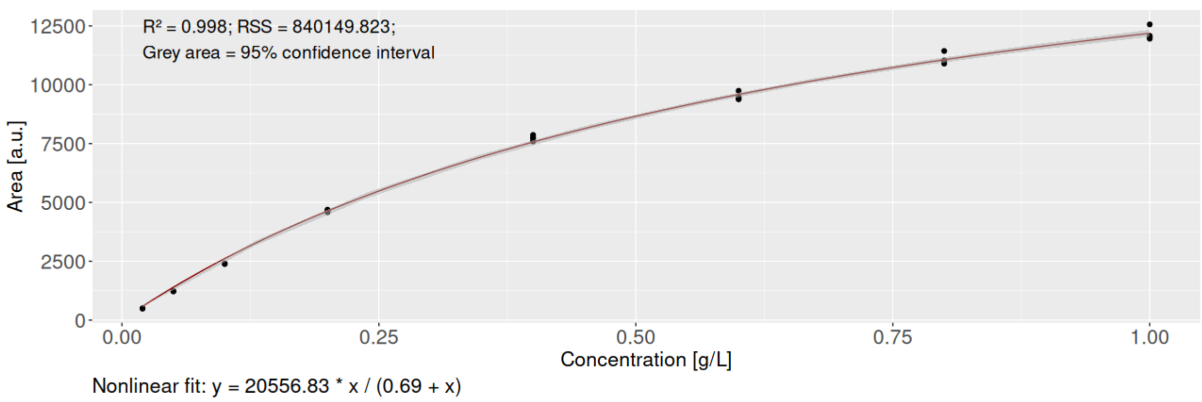
As a standard solution, N-(2-methoxyphenyl)acetamide was dissolved in bi-distilled water according to the general procedure for the calibration. Here, 5.19 mg were dissolved in 5.19 mL of bi-distilled water (replicates:

6.80 mg in 6.80 mL, 6.18 mg in 6.18 mL, 7.28 mg in 7.28 mL). To determine the solubility, 34.77 mg (replicate: 48.83 mg) of the compound was dissolved in 1.0 mL of bi-distilled water according to the procedure for the determination of the solubility and measured on HPLC. For this compound only the dilution with a factor of 100 could be used for determination due to the high solubility

Table 16: HPLC calibration data for compound nine, including the concentration (in g/L) and the area under the curve (AUC). The graph shows the data points and the resulting linear fit.

Concentration [mg/mL]	Standard [1mg/mL]	Standard [uL]	MeCN [µL]	AUC
1	1	500	0	12562.0, 11954.0, 12077.5, 12013.0
0.8	1	800	200	11433.5, 10897.0, 11017.5, 11029.0
0.6	1	600	400	9479.5, 9384.5, 9392.5, 9741.0
0.4	1	400	600	7590.5, 7767.5, 7704.0, 7865.5
0.2	1	200	800	4651.0, 4585.0, 4691.0, 4628.0
0.1	1	100	900	2380.5, 2412.5
0.05	1	50	950	1234.0, 1219.5
0.02	1	20	980	490.5, 499.5
Back calculation				
Concentration [g/L]			**AUC [a]**	
11.09 ± 0.18			291800, 290900, 283600, 284000, 287900, 284200, 283200, 280200	

[a] For determination of the maximum solubility, all values of the double injection were taken into account.



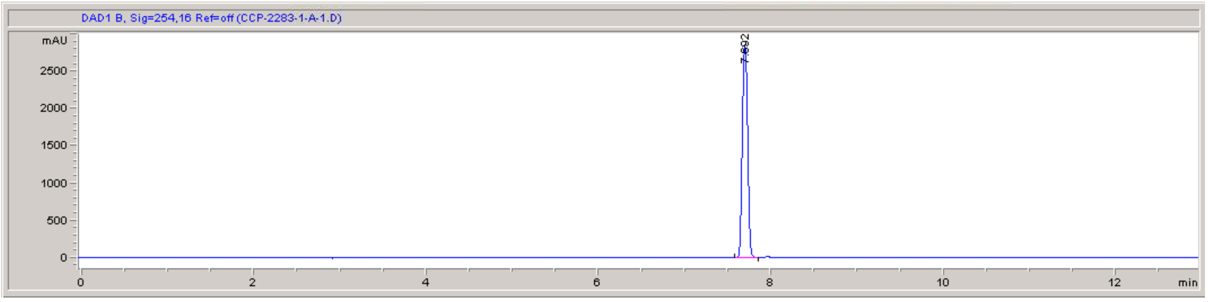


Figure 36: Exemplary chromatogram obtained for the measurement of compound **5** at a concentration of 1.0 mg/mL at λ = 254 nm. The area under the curve was found to be 12572; the corresponding concentration was calculated to be 5.4 g/L (retention time: 7.22)

6. Appendix

6.1 Documentation for the set of descriptors

6.1.1 List of 4 descriptors

1. MolLogP: Logarithm of the octanol-water partition coefficient, indicating molecule lipophilicity.
2. MolWt: Molecular weight of the compound, representing molecular size.
3. NumRotatableBonds: Number of rotatable bonds, indicating molecule flexibility.
4. Aromatic Proportion: The proportion of aromatic atoms in the molecule, reflecting stability.

6.1.2 List of 17 descriptors

1. MolLogP: Logarithm of the octanol-water partition coefficient, indicating molecule lipophilicity.
2. MolWt: Molecular weight of the compound, representing molecular size.
3. NumRotatableBonds: Number of rotatable bonds, indicating molecule flexibility.
4. Aromatic Proportion: The proportion of aromatic atoms in the molecule, reflecting stability.
5. RingCount: Total number of rings, characterizing structural complexity.
6. TPSA: topological polar surface area, indicating absorption and permeability.
7. NumHDonors: Number of hydrogen bond donors, affecting solubility and interactions.
8. NumSaturatedRings: Number of saturated rings, indicating molecule rigidity.
9. NumAliphaticRings: Number of aliphatic rings, relating to hydrophobic properties.
10. NumHAcceptors: Number of hydrogen bond acceptors, influencing solubility.
11. NumHeteroatoms: Number of heteroatoms (non-carbon atoms), impacting polarity and reactivity.
12. MaxPartialCharge: Maximum partial charge on an atom, reflecting charge distribution.
13. FpDensityMorgan1: Density of Morgan fingerprints (radius 1), representing molecular complexity.
14. NumValenceElectrons: Total valence electrons, indicating electronic properties.
15. NHOHCount: Number of -OH and -NH groups, important for hydrogen bonding.
16. FractionCSP3: Fraction of sp3-hybridized carbons, reflecting 3D molecular character.
17. SP_Bonds: Count of sp-hybridized bonds, indicating linear or triple-bonded structures.

6.1.3 List of 125 descriptors

1. MaxAbsEStateIndex: The maximum absolute value of the E-state index, a measure of electron distribution.
2. MaxEStateIndex: The maximum E-state index in the molecule.
3. MinAbsEStateIndex: The minimum absolute value of the E-state index.
4. MinEStateIndex: The minimum E-state index in the molecule.
5. QED: Quantitative estimate of drug-likeness, a metric used to assess the drug-likeness of a molecule.
6. SPS: Synthetic Accessibility Score, predicting the ease of synthesis for a compound.
7. MolWt: Molecular Weight, the total mass of the molecule.
8. HeavyAtomMolWt: The molecular weight of heavy atoms (non-hydrogen atoms).
9. ExactMolWt: The exact molecular weight considering isotopic distribution.
10. NumValenceElectrons: Total number of valence electrons in the molecule.
11. NumRadicalElectrons: Number of radical electrons in the molecule.
12. MaxPartialCharge: Maximum partial charge on any atom in the molecule.
13. MinPartialCharge: Minimum partial charge on any atom in the molecule.
14. MaxAbsPartialCharge: Maximum absolute value of the partial charge on any atom.
15. MinAbsPartialCharge: Minimum absolute value of the partial charge on any atom.
16. FpDensityMorgan1: Morgan fingerprint density with radius 1, a circular fingerprint used to encode molecular structure.
17. FpDensityMorgan2: Morgan fingerprint density with radius 2.
18. FpDensityMorgan3: Morgan fingerprint density with radius 3.
19. BCUT2D_MWHI: BCUT metric using molecular weight as the property.
20. BCUT2D_MWLOW: BCUT metric for low molecular weight property.
21. BCUT2D_CHGHI: BCUT metric using high charge as the property.
22. BCUT2D_CHGLO: BCUT metric for low charge property.
23. BCUT2D_LOGPHI: BCUT metric using high LogP as the property.
24. BCUT2D_LOGPLOW: BCUT metric for low LogP property.
25. BCUT2D_MRHI: BCUT metric using high molar refractivity as the property.
26. BCUT2D_MRLOW: BCUT metric for low molar refractivity property.
27. AvgIpc: Average Information Content (Ipc) of the molecule.
28. BalabanJ: Balaban’s J index, a topological descriptor.
29. BertzCT: Bertz complexity index, a measure of molecular complexity.
30. Chi0: The first-order connectivity index.
31. Chi0n: The first-order connectivity index with nitrogen.
32. Chi0v: The first-order valence connectivity index.
33. Chi1: The second-order connectivity index.
34. Chi1n: The second-order connectivity index with nitrogen.
35. Chi1v: The second-order valence connectivity index.
36. Chi2n: The third-order connectivity index with nitrogen.
37. Chi2v: The third-order valence connectivity index.
38. Chi3n: The fourth-order connectivity index with nitrogen.

39. Chi3v: The fourth-order valence connectivity index.
40. Chi4n: The fifth-order connectivity index with nitrogen.
41. Chi4v: The fifth-order valence connectivity index.
42. Hall-Kier Alpha: Hall-Kier alpha modification of the shape index.
43. Ipc: Information content index, a molecular descriptor.
44. Kappa1: First Kappa shape index.
45. Kappa2: Second Kappa shape index.
46. Kappa3: Third Kappa shape index.
47. LabuteASA: Labute's approximation to molecular surface area.
48. PEOE_VSA1: PEOE charge on molecular surface area, bin 1.
49. PEOE_VSA10: PEOE charge on molecular surface area, bin 10.
50. PEOE_VSA11: PEOE charge on molecular surface area, bin 11.
51. PEOE_VSA12: PEOE charge on molecular surface area, bin 12.
52. PEOE_VSA13: PEOE charge on molecular surface area, bin 13.
53. PEOE_VSA14: PEOE charge on molecular surface area, bin 14.
54. PEOE_VSA2: PEOE charge on molecular surface area, bin 2.
55. PEOE_VSA3: PEOE charge on molecular surface area, bin 3.
56. PEOE_VSA4: PEOE charge on molecular surface area, bin 4.
57. PEOE_VSA5: PEOE charge on molecular surface area, bin 5.
58. PEOE_VSA6: PEOE charge on molecular surface area, bin 6.
59. PEOE_VSA7: PEOE charge on molecular surface area, bin 7.
60. PEOE_VSA8: PEOE charge on molecular surface area, bin 8.
61. PEOE_VSA9: PEOE charge on molecular surface area, bin 9.
62. SMR_VSA1: SMR surface area, bin 1.
63. SMR_VSA10: SMR surface area, bin 10.
64. SMR_VSA2: SMR surface area, bin 2.
65. SMR_VSA3: SMR surface area, bin 3.
66. SMR_VSA4: SMR surface area, bin 4.
67. SMR_VSA5: SMR surface area, bin 5.
68. SMR_VSA6: SMR surface area, bin 6.
69. SMR_VSA7: SMR surface area, bin 7.
70. SMR_VSA8: SMR surface area, bin 8.
71. SMR_VSA9: SMR surface area, bin 9.
72. SlogP_VSA1: SlogP surface area, bin 1.
73. SlogP_VSA10: SlogP surface area, bin 10.
74. SlogP_VSA11: SlogP surface area, bin 11.
75. SlogP_VSA12: SlogP surface area, bin 12.
76. SlogP_VSA2: SlogP surface area, bin 2.
77. SlogP_VSA3: SlogP surface area, bin 3.
78. SlogP_VSA4: SlogP surface area, bin 4.
79. SlogP_VSA5: SlogP surface area, bin 5.
80. SlogP_VSA6: SlogP surface area, bin 6.
81. SlogP_VSA7: SlogP surface area, bin 7.

82. SlogP_VSA8: SlogP surface area, bin 8.
83. SlogP_VSA9: SlogP surface area, bin 9.
84. TPSA: Topological polar surface area, a predictor of drug absorption.
85. EState_VSA1: E-state surface area, bin 1.
86. EState_VSA10: E-state surface area, bin 10.
87. EState_VSA11: E-state surface area, bin 11.
88. EState_VSA2: E-state surface area, bin 2.
89. EState_VSA3: E-state surface area, bin 3.
90. EState_VSA4: E-state surface area, bin 4.
91. EState_VSA5: E-state surface area, bin 5.
92. EState_VSA6: E-state surface area, bin 6.
93. EState_VSA7: E-state surface area, bin 7.
94. EState_VSA8: E-state surface area, bin 8.
95. EState_VSA9: E-state surface area, bin 9.
96. VSA_EState1: VSA_E-state, bin 1.
97. VSA_EState10: VSA_E-state, bin 10.
98. VSA_EState2: VSA_E-state, bin 2.
99. VSA_EState3: VSA_E-state, bin 3.
100. VSA_EState4: VSA_E-state, bin 4.
101. VSA_EState5: VSA_E-state, bin 5.
102. VSA_EState6: VSA_E-state, bin 6.
103. VSA_EState7: VSA_E-state, bin 7.
104. VSA_EState8: VSA_E-state, bin 8.
105. VSA_EState9: VSA_E-state, bin 9.
106. FractionCSP3: Fraction of carbon atoms that are sp3 hybridized.
107. HeavyAtomCount: The number of heavy atoms (non-hydrogen atoms).
108. NHOH Count: The number of -OH and -NH groups.
109. NOCount: The number of nitrogen and oxygen atoms.
110. NumAliphaticCarbocycles: The number of aliphatic carbocycles.
111. NumAliphaticHeterocycles: The number of aliphatic heterocycles.
112. NumAliphaticRings: The number of aliphatic rings.
113. NumAromaticCarbocycles: The number of aromatic carbocycles.
114. NumAromaticHeterocycles: The number of aromatic heterocycles.
115. NumAromaticRings: The number of aromatic rings.
116. NumHAcceptors: The number of hydrogen bond acceptors.
117. NumHDonors: The number of hydrogen bond donors.
118. NumHeteroatoms: The number of heteroatoms (non-carbon and non-hydrogen atoms).
119. NumRotatableBonds: The number of rotatable bonds.
120. NumSaturatedCarbocycles: The number of saturated carbocycles.
121. NumSaturatedHeterocycles: The number of saturated heterocycles.
122. NumSaturatedRings: The number of saturated rings.
123. RingCount: The total number of rings in the molecule.

124. MolLogP: The logarithm of the partition coefficient (LogP) of the molecule.
125. MolMR: The molar refractivity of the molecule.

6.1.4 Fingerprint descriptors

Morgan fingerprints, also known as Extended Connectivity Fingerprints (ECFP), define molecular substructures based on a circular neighborhood around each atom. The radius determines how many bond hops are considered from each atom.

Molecular fingerprints are binary or vector representations of the molecular structure and are used in cheminformatics for similarity searching.

Types of fingerprints used in this study for the analysis of solubility prediction

64-bit Fingerprints: Optimized for lightweight applications with limited memory.

Use Case: Suitable for small datasets or quick similarity searches.

Bit Density: lower due to the smaller size, which may lead to more hash collisions.

128-bit fingerprints

Purpose: Standard size for moderate computational tasks.

Use case: Commonly used in substructure searching and diversity analysis.

Advantage: Balances memory usage and representational power.

256-bit fingerprints

Purpose: Enhanced molecular diversity representation.

Use case: Preferred for larger datasets or higher-resolution similarity searches.

Advantage: More bits reduce the chance of collisions, improving performance.

512-bit fingerprints

Purpose: High-resolution molecular encoding.

Use case: Best suited for deep learning or applications requiring high granularity.

Radius = 0 → Only the atom itself is considered.

Radius = 1 → Atom and its immediate neighbors.

Radius = 2 → Atom, its neighbors, and their neighbors (up to 2 bonds away)

Advantage: Superior representation of molecular features, ideal for large and complex datasets.

6.1.5 List of 38 feature-engineered descriptors

1. charge: The total charge of the molecule.
2. many_double_bonds: The count of double bonds in the molecule.
3. atoms_degree_0: Number of atoms with zero degree (not connected to other atoms).
4. atoms_degree_1: Number of atoms with one connection.
5. atoms_degree_2: Number of atoms with two connections.
6. atoms_degree_3: Number of atoms with three connections.
7. atoms_degree_4: Number of atoms with four connections.
8. atoms_degree_5: Number of atoms with five connections.
9. atoms_degree_6: Number of atoms with six connections.
10. atoms_valence_0: Number of atoms with zero valence electrons.
11. atoms_valence_1: Number of atoms with one valence electron.
12. atoms_valence_2: Number of atoms with two valence electrons.
13. atoms_valence_3: Number of atoms with three valence electrons.
14. atoms_valence_4: Number of atoms with four valence electrons.
15. atoms_valence_5: Number of atoms with five valence electrons.
16. atoms_valence_6: Number of atoms with six valence electrons.
17. atom_hybridization_S: Number of atoms with S orbital hybridization.
18. atom_hybridization_SP: Number of atoms with SP orbital hybridization.
19. atom_hybridization_SP2: Number of atoms with SP2 orbital hybridization.
20. atom_hybridization_SP3: Number of atoms with SP3 orbital hybridization.
21. atom_hybridization_SP3D: Number of atoms with SP3D orbital hybridization.
22. atom_hybridization_SP3D2: Number of atoms with SP3D2 orbital hybridization.
23. atom_hybridization_UNSPECIFIED: Number of atoms with unspecified hybridization.
24. aromatic_atoms: Number of aromatic atoms in the molecule.
25. single_bonds: Number of single bonds in the molecule.
26. double_bonds: Number of double bonds in the molecule.
27. triple_bonds: Number of triple bonds in the molecule.
28. aromatic_bonds: Number of aromatic bonds in the molecule.
29. zero_bonds: Number of atoms with zero bonds.
30. conjugated_bonds: Number of conjugated bonds in the molecule.
31. bonds_in_ring: Number of bonds present in a ring structure.
32. chirality_none: Number of atoms without chirality.
33. chirality_any: Number of atoms with any chirality.
34. chirality_z: Number of atoms with Z chirality.
35. chirality_e: Number of atoms with E chirality.
36. n_atoms: Total number of atoms in the molecule.
37. n_bonds: Total number of bonds in the molecule.
38. n_rings: Total number of rings in the molecule

6.2. Model description and parameters

To find the best model for solubility prediction, the following structured steps were followed. These steps encompassed exploring multiple models, varying parameters, and testing the effectiveness of feature sets and fingerprint sizes. Functional groups, along with tuning the hyperparameters, achieve optimal results. To select the best model, we have started with

1. **Baseline models**:
   - Linear regression
   - Random forest
   - Support Vector Machines (SVMs)
2. **Advanced models**:
   - Neural networks: Feedforward and graph neural networks (GNNs)
   - Hybrid message passing neural networks (MPNNs)
   - Ensemble methods: Gradient boosting (e.g., XGBoost, LightGBM)

Once we finalized the best model, we tried to find the best features that could give a better prediction on the test data

- Started with **four features**, such as molecular weight and polarity
- Expanded to **17 features**, including hydrogen bond donors/acceptors, topological polar surface area, and logP
- Extended to **125 features**, leveraging detailed molecular descriptors from cheminformatics tools like RDKit or PubChem
- Additionally, 38 descriptors like "charge," "long chain," "double bonds," etc., were added.
- Added seven functional groups, whose coefficient correlation with solubility was high
- Generated fingerprints of varying lengths: **64, 128, 256, and 512 bits**, using Morgan or extended-connectivity fingerprints (ECFPs). Used RDKit libraries for implementation.

6.2.1 Description of the parameters provided for a machine learning model for XGBoost.

learning_rate (default: 0.3)

a. Also known as the *eta* parameter, it controls the step size at each iteration as the algorithm moves toward the minimum of the loss function.
b. Lower values make the learning process slower but potentially more accurate as the model learns more gradually.

n_estimators (default: 100)

c. The number of boosting rounds or trees to be built in the model.
d. Higher values increase model complexity and training time but can improve performance, provided overfitting is controlled.

max_depth (default: 6)

e. The maximum depth of each tree.
f. Controls model complexity by limiting the depth to which the trees can grow. A higher value can capture more relationships but risks overfitting.

min_child_weight (default: 1)

g. Minimum sum of instance weights (or the number of data points) needed in a child node.
h. Higher values prevent the algorithm from creating overly complex trees with nodes containing very few samples, thereby reducing the risk of overfitting.

gamma (default: 0)

i. Also called *min_split_loss*, it specifies the minimum loss reduction required for a split to occur.
j. A higher value makes the algorithm more conservative, requiring a higher gain to justify a split.

subsample (default: 1)

k. The fraction of training data randomly sampled for each boosting round.
l. Reducing this value can prevent overfitting, but it may increase bias. For example, 0.7 means 70% of the data is used in each round.

colsample_bytree (default: 1)

m. The fraction of features (columns) randomly sampled for each tree.
n. Reducing this value can improve generalization and prevent overfitting. For instance, 0.8 means 80% of features are used per tree.

scale_pos_weight (default: 1)

o. Balances the weights of positive and negative classes, making it useful for handling imbalanced datasets.
p. A value greater than 1 can assign more importance to the minority class, improving the model's focus on it.

List of the finalized hyperparameters for XGBoost

- learning_rate=0.01
- n_estimators=2000
- max_depth=8
- min_child_weight=6
- gamma=0
- subsample=0.7
- colsample_bytree=0.8
- scale_pos_weight=1

List of the finalized hyperparameters for Hybrid MPNN

- batch_size': 64
- learning_rate': 0.001
- mpnn_hidden_dim': 128
- optimizer: 'Adam'

6.3 Selection of functional group

To identify the most influential functional groups, we calculated the correlation coefficients between functional groups and solubility using a literature dataset. Out of 38 functional groups, the top 7 functional groups with the highest correlation coefficients were selected for further analysis. List of 38 functional groups along with their chemical representations in SMARTS format:

Polar functional groups

1. **Hydroxyl Group**: [OH]
2. **Amino Group**: [NH_2]
3. **Carboxyl Group**: C(=O)[OH]
4. **Carbonyl Group**: C=O
5. **Methoxy Group**: [CH_3]O

6. **Aldehyde**: C=O
7. **Ketone**: C(=O)[C]
8. **Ester**: C(=O)OC
9. **Amide Group**: C(=O)N
10. **Nitrile**: C#N
11. **Nitro Group**: [N+](=O)[O-]
12. **Thiol Group**: [SH]
13. **Sulfoxide**: SC
14. **Sulfone**: S(=O)C
15. **Phosphine**: P
16. **Isocyanate**: N=C=O
17. **Isothiocyanate**: N=C=S
18. **Azide**: N=[N+]=[N-]
19. **Imine**: C=NC
20. **Acyl Halide**: C(=O)[Cl,Br,I,F]
21. **Epoxide**: C1CO1
22. **Urea**: N(C=O)N
23. **Guanidine**: N=C(N)N
24. **Anhydride**: C(=O)OC(=O)
25. **Hydrazone**: C=NNC
26. **Carbamate**: OC(=O)N
27. **Peroxide**: COO

Non-polar functional groups

28. **Methyl Group**: $[CH_3]$
29. **Ethyl Group**: CC
30. **Chloro Group**: [Cl]
31. **Bromo Group**: [Br]
32. **Iodo Group:**[I]
33. **Fluoro Group**: [F]
34. **Aromatic Ring**: c
35. **Alkene**: C=C
36. **Alkyne**: C#C
37. **Alkyl**: [R]
38. **Thioether**: CSC

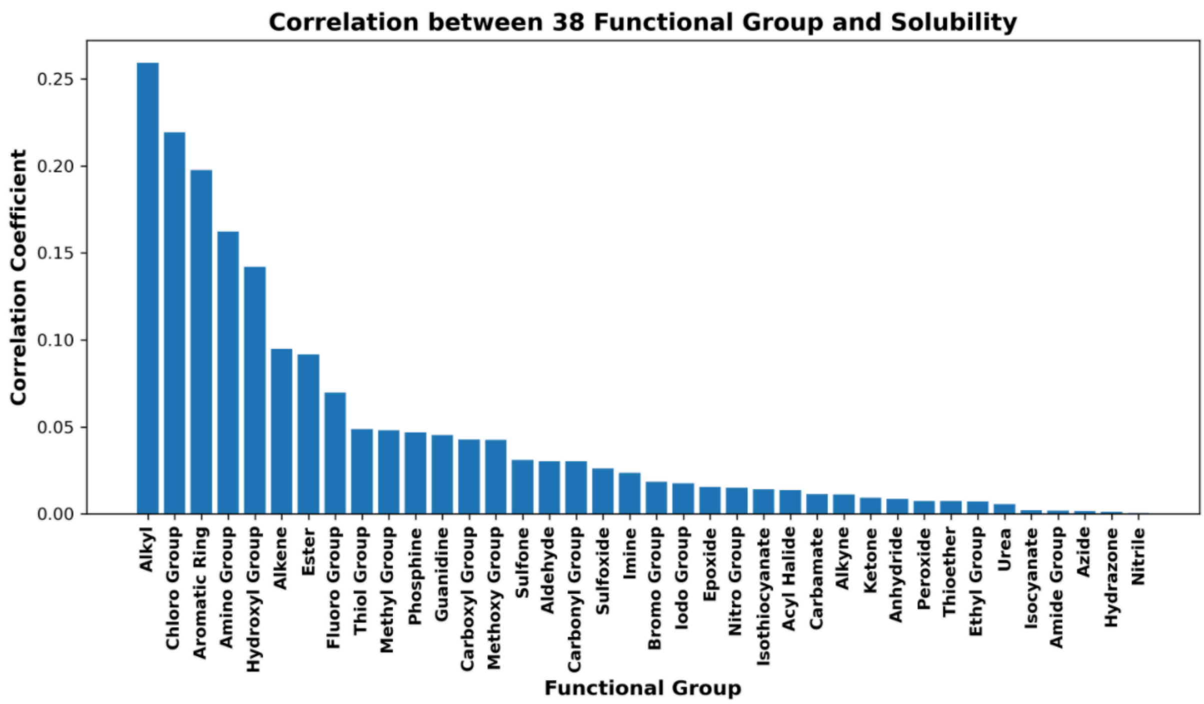


Figure A1: Correlation coefficients for 38 functional groups with solubility values. The x-axis represents the functional groups, and the y-axis shows the magnitude of the correlation coefficient. Higher values indicate a more substantial influence of the corresponding functional groups on solubility.

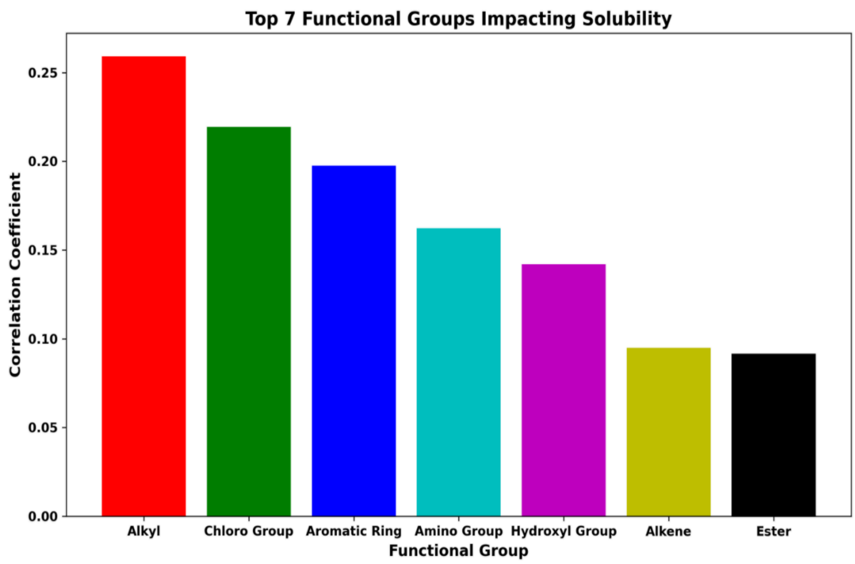


Figure A2: Correlation coefficients of the top 7 functional groups with the strongest influence on solubility. The x-axis represents the functional groups, and the y-axis indicates their corresponding correlation coefficients, highlighting the significant impact of these groups on solubility.

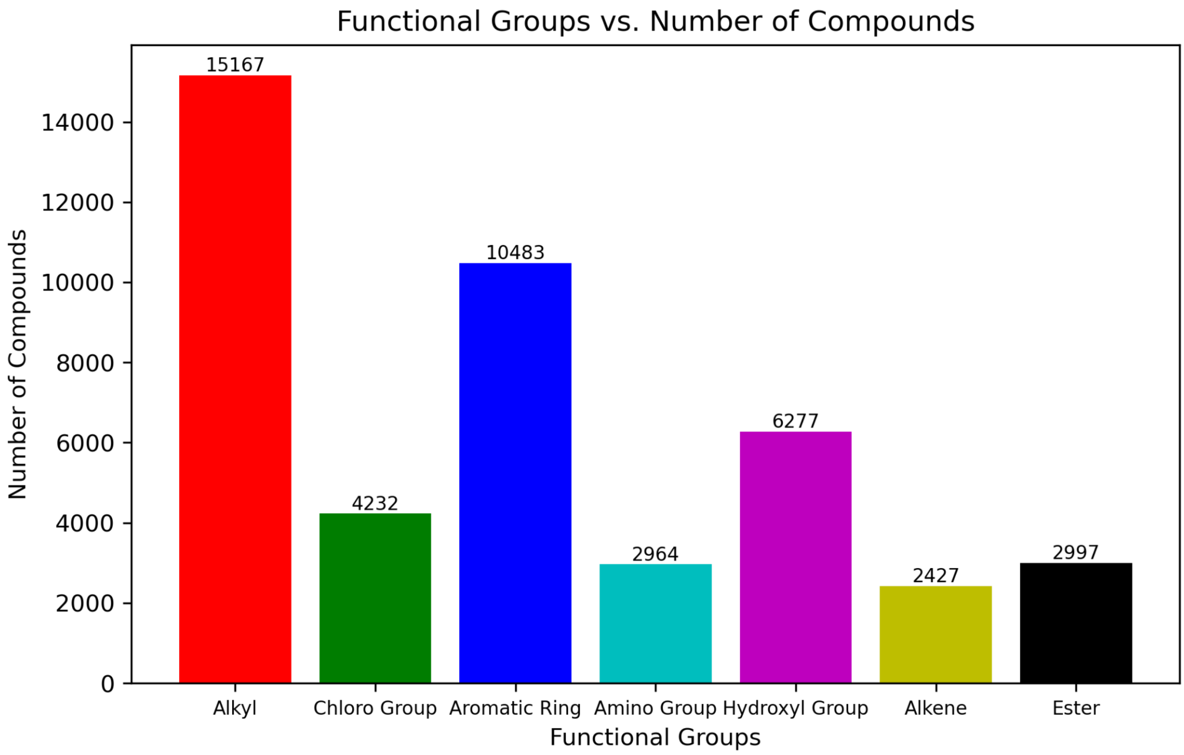


Figure A3. Distribution of functional groups within the training dataset used for our studies, including the summary of the types of functional groups used to extend the descriptors of our models and their counts (= number of compounds present in each group of the training dataset) in the collected training dataset of 17937 data.

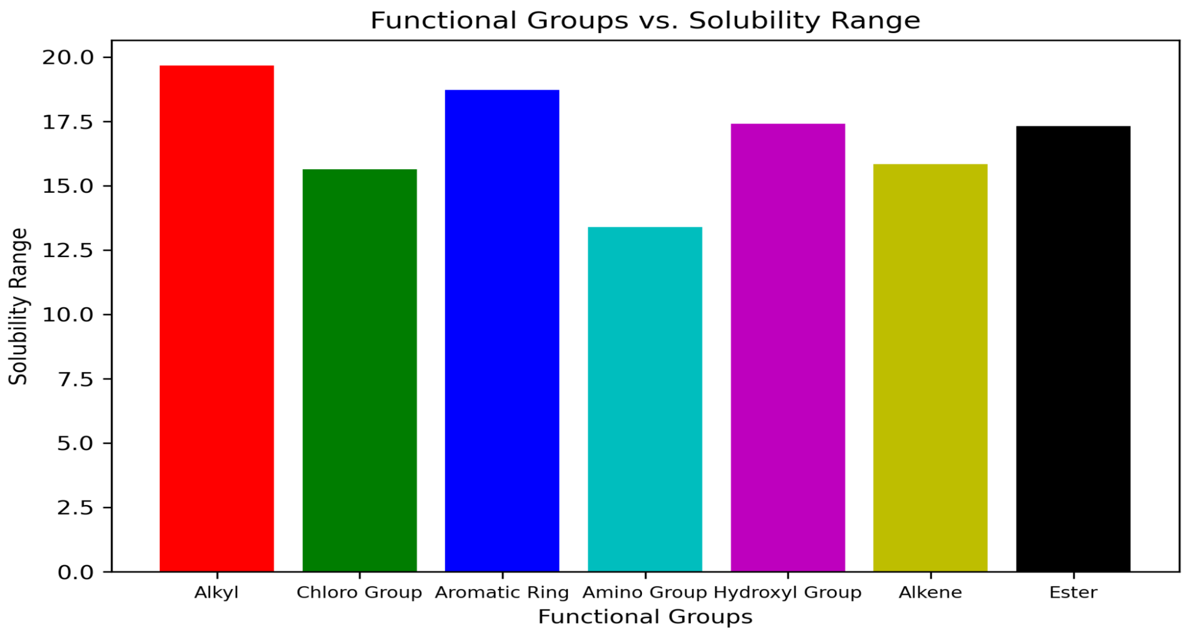


Figure A4. Relationship of solubility and functional groups within the training dataset. X-axis = functional groups found in these molecules; y-axis = solubility of various molecules. Each functional group is mapped against a range of solubility values, indicating how compounds containing a specific functional group tend to distribute in terms of solubility.

Table A1. Selected seven functional groups with an influence on solubility

Sr. No.	Group	Polar/Non-Polar	Example
1	Alky (-R)	Nonpolar	Ethane ($-CH_3CH_3$)
2	Chloro (-Cl)	Polar	Chloroform ($CHCl_3$)
3	Aromatic (C_6H_6)	Nonpolar	Benzene (C_6H_6)
4	Hydroxyl (-OH)	Polar	Ethanol (C_2H_5OH)
5	Amino ($-NH_2$)	Polar	Glycine ($H_2NCHCOOH$)
6	Alkene(C=C)	Nonpolar	Ethane (C_2H_4)
7	Ester (C(=O)OC)	Nonpolar	Ethyl acetate ($CH_3COOCH_2CH_3$)

6.4 API integration and management

The interaction with PubChem's API (Application Programming Interface) (PUG View) required careful consideration of several technical constraints, for detailed information on the types of data provided by this API is included in the supporting information.

Request Rate Management

- o Implementation of controlled delays between successive queries to comply with API rate limitations, not more than five records per second
- o Dynamic adjustment of query intervals to maintain optimal throughput while preventing server rejection
- o Monitoring of request patterns to ensure consistent access within allowed limits

Batch Processing Optimization

- o Careful calibration of batch sizes to maximize efficiency while adhering to API constraints
- o Implementation of batch splitting logic for oversized requests
- o Optimization of batch composition to minimize total query requirements

Resilience and Error Management

- o Development of robust error-handling mechanisms for various failure scenarios
- o Implementation of intelligent retry logic with exponential backoff for failed requests
- o Systematic logging of failed queries for manual review and reprocessing
- o Maintenance of transaction integrity during temporary service disruptions

7. References

1. Lee, J., *et al.* Green-solvent-processable organic semiconductors and future directions for advanced organic electronics. *J. Mater. Chem. A Mater. Energy Sustain.* **8**, 21455–21473 (2020).
2. Llompart, P., *et al.* Will we ever be able to predict solubility accurately? *Sci Data* **11**, 303 (2024).
3. Schwaiblmair, M. *et al.* Drug-induced interstitial lung disease. *Open Respir. Med. J.* **6**, 63–74 (2012).
4. Cui, Q., *et al.* Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Front. Oncol.* **10**, 121 (2020).
5. Lee, S., *et al.* Novel solubility prediction models: Molecular fingerprints and physicochemical features vs graph convolutional neural networks. *ACS Omega* **7**, 12268–12277 (2022).
6. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* **6**, 143 (2019).
7. Sorkun, M. C. *AqSolDB: A Curated Aqueous Solubility Dataset Contains 9,982 Unique Compounds*. (Github).
8. Avdeef, A. Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. *ADMET DMPK* **7**, 210–219 (2019).
9. Dannenfelser, R. & Yalkowsky, S. H. Database for aqueous solubility of nonelectrolytes. *Comput Appl Biosci* **5**, 235–236 (1989).
10. Website. https://experts.arizona.edu/en/publications/data-base-of-aqueous-solubility-for-organic-n.
11. Yalkowsky, S. H., He, Y. & Jain, P. *Handbook of Aqueous Solubility Data*. (CRC Press,

2016). doi:10.1201/EBK1439802458.

12. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777 (2000).
13. AqSolDB/data/dataset-E.csv at master · mcsorkun/AqSolDB. *GitHub* https://github.com/mcsorkun/AqSolDB/blob/master/data/dataset-E.csv.
14. Bergström, C. A. S. & Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int J Pharm* **540**, 185–193 (2018).
15. Website. https://research-portal.st-andrews.ac.uk/en/datasets/dls-100-solubility-dataset.
16. Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).
17. deepchem/datasets/delaney-processed · deepchem/deepchem. *GitHub* https://github.com/deepchem/deepchem/blob/master/datasets/delaney-processed.csv.
18. Mobley, D. L. & Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **28**, 711–720 (2014).
19. FreeSolv/database.txt at master · MobleyLab/FreeSolv. *GitHub* https://github.com/MobleyLab/FreeSolv/blob/master/database.txt.
20. Krasnov, L., Mikhaylov, S., Fedorov, M. & Sosnin, S. BigSolDB: Solubility dataset of compounds in organic solvents and water in a wide range of temperatures. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-qqslt.
21. https://frontiersin.figshare.com/articles/dataset/Data_Sheet_1_Improved_Prediction_of_Aqueous_Solubility_of_Novel_Compounds_by_Going_Deeper_With_Deep_Learning_ZIP/11833914/1?file=21663660.
22. Boobier, S., Hose, D. R. J., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat.*

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

23. BNNLab. BNNLab/Solubility_data: Leeds Solubility Data. doi:10.5281/zenodo.3686213.
24. Panapitiya, G., *et al.* Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **7**, 15695–15710 (2022).
25. Cui, Q. *et al.* Data_Sheet_1_Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. ZIP. Frontiers https://doi.org/10.3389/fonc.2020.00121.s001 (2020).
26. Sushko, I., *et al.* Online chemical modeling environment (OCHEM): a web platform for data storage, model development, and publishing of chemical information. *J. Comput. Aided Mol. Des.* **25**, 533–554 (2011).
27. Website. https://www.reaxys.com/#/search/quick.
28. Solubility_Prediction_GCN/data at main · mhlee216/Solubility_Prediction_GCN. *GitHub*. https://github.com/mhlee216/Solubility_Prediction_GCN/tree/main/data.
29. OCHEM home page. https://ochem.eu/home/show.do?render-mode=popup&restart=1&server=blue.L00.
30. eChemPortal - Home. https://www.echemportal.org/echemportal/content/participants.
31. Website. https://www.chemspider.com/.
32. NEPHELOstar Plus. *IMGEN Technologies* (2012).
33. Jain, N. & Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **90**, 234–252 (2001).
34. Hückel, W. Solubility of non-electrolytes. Von Prof. Joel H. Hildebrand. 203 Seiten. Reinhold Publishing Corporation, New York, 1936. Preisge b. $4,50. *Angew. Chem. Weinheim Bergstr. Ger.* **49**, 703–704 (1936).
35. Hansen, C. M. *Hansen Solubility Parameters: A User's Handbook, Second Edition*.

(Taylor & Francis, 2007).

36. Sorkun, M. C., Koelman, J. M. V. A. & Er, S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **24**, 101961 (2021).

37. Lusci, A., Pollastri, G. & Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53**, 1563–1575 (2013).

38. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Message Passing Neural Networks. In *Machine Learning Meets Quantum Physics,* 199–214 (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-030-40245-7_10.

39. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv [cs.LG]* (2016) doi:10.48550/ARXIV.1609.02907.

40. Jadama, A. F. & Toray, M. K. Ensemble Learning: Methods, Techniques, Application. (2024) doi:10.13140/RG.2.2.28017.08802.

41. Avdeef, A. Prediction of aqueous intrinsic solubility of drug-like molecules using Random Forest regression trained with the Wiki-pS0 database. *ADMET DMPK* **8**, 29–77 (2020).

42. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev. 54,* 355–366 (2002).

43. Llinas, A., Oprisiu, I. & Avdeef, A. Findings of the Second Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **60**, 4791–4803 (2020).

44. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).

45. Ultralytics. K-Fold Cross Validation. https://docs.ultralytics.com/guides/kfold-cross-validation/ (2023).

46. Huuskonen, J., Rantanen, J. & Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **35**, 1081–1088 (2000).

47. Yan, A. & Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **43**, 429–434 (2003).

48. Hou, T. J., Xia, K., Zhang, W. & Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on the atom contribution approach. *J. Chem. Inf. Comput. Sci.* **44**, 266–275 (2004).

49. Schroeter, T. S. *et al.* Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.* **21**, 485–498 (2007).

50. Ali, J., Camilleri, P., Brown, M. B., Hutt, A. J. & Kirton, S. B. In silico prediction of aqueous solubility using simple QSPR models: the importance of phenol and phenol-like moieties. *J. Chem. Inf. Model.* **52**, 2950–2957 (2012).

51. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness, and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717 (2017).

52. Bjerrum, E. J. & Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **8**, 131 (2018).

53. Website. https://aqsolpred.streamlit.app/.

54. Interactive ALOGPS calculations at the VCCLAB site. https://vcclab.org/web/alogps/.

55. Index-Home-ChemBCPP. http://chembcpp.scbdd.com/home/index/.

56. Tetko, I. V. *et al.* Virtual computational chemistry laboratory—design and description. *J. Comput. Aided Mol. Des.* **19**, 453–463 (2005).

57. Human Metabolome Database: Showing metabocard for Methylparaben (HMDB0032572). https://hmdb.ca/metabolites/HMDB0032572.
58. Human Metabolome Database: Showing metabocard for Eudesmic acid (HMDB0033839). https://hmdb.ca/metabolites/HMDB0033839.
59. Ethyl 3,4-dihydroxybenzoate (ethyl protocatechuate). *MedchemExpress.com* https://www.medchemexpress.com/ethyl-3-4-dihydroxybenzoate.html.
60. Navux Commerce Solutions. 3,4-Dimethoxybenzoic Acid. https://www.parchem.com/chemical-supplier-distributor/3-4-dimethoxybenzoic-acid-085313.
61. Navux Commerce Solutions. Acetyl-o-Anisidine. https://www.parchem.com/chemical-supplier-distributor/acetyl-o-anisidine-086188.
62. [No title]. https://aqua-solubility-prediction.streamlit.app/.
63. Balakin, K. V., Savchuk, N. P. & Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems, and solutions. *Curr. Med. Chem.* **13**, 223–241 (2006).
64. Klamt, A. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).
65. Reichardt, C. & Welton, T. *Solvents and Solvent Effects in Organic Chemistry*. (Wiley-VCH Verlag, Weinheim, Germany, 2010). doi:10.1002/9783527632220.
66. Marcus, Y. *The Properties of Solvents*. (John Wiley & Sons, Chichester, England, 1998).
67. PubChem. 3-Oxo-3-phenylpropionanilide. https://pubchem.ncbi.nlm.nih.gov/compound/70398#section=Computed-Properties.