Fine Chemical Engineering
https://ojs.wiserpub.com/index.php/FCE/

Research Article

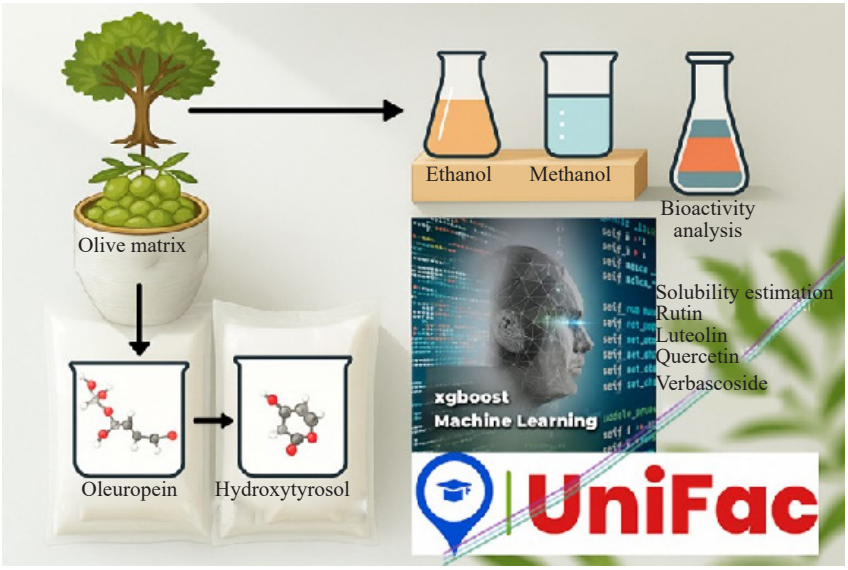
Solubility-Driven Phenol Extraction from Olive Tree Derivatives in Ethanol/Methanol: Empirical, UNIFAC & ML Models

Mohamed Abdelkader Hafiene[1], **Hatem Ksibi**[1,2*]

[1]Laboratory for Materials Applications in Environment, Water, and Energy (LAM3E), Faculty of Sciences, University of Gafsa, Gafsa, 6029, Tunisia
[2]Preparatory Institute for Engineering Studies of Sfax (IPEIS), University of Sfax, Sfax, 3018, Tunisia
E-mail: hatem.ksibi@ipeis.rnu.tn

Received: 9 June 2025; **Revised:** 13 August 2025; **Accepted:** 5 September 2025

Graphical Abstract:



Abstract: This study investigates the solubility behavior and predictive modeling of six key phenolic compounds derived from olive sources—hydroxytyrosol, luteolin, oleuropein, rutin, quercetin, and verbascoside—in methanol, ethanol, and their binary mixtures (10 : 90, 50 : 50, and 90 : 10 v/v) at temperatures ranging from 20 °C to 50 °C. Experimental solubility data were compiled from previously published literature. These results showed a wide solubility range: oleuropein exhibited the highest solubility (~ 100 mg/100 g in methanol at 50 °C), followed by hydroxytyrosol (~ 78 mg/100 g), verbascoside (~ 45 mg/100 g), rutin (~ 35 mg/100 g), quercetin (~ 5 mg/100 g), and luteolin (~ 3 mg/100 g). Solubility generally increased with temperature and ethanol content, though compound-specific

DOI: https://doi.org/10.37256/fce.6220257412

effects were observed. Empirical modeling using the Apelblat equation demonstrated strong agreement with experimental data ($R^2 > 0.98$; Mean Absolute Error (MAE) < 5%) across all compounds. Predictive models were also developed using both the Universal Functional Activity Coefficient (UNIFAC) thermodynamic method and Machine Learning (ML) algorithms (eXtreme Gradient Boosting (XGBoost)), Random Forest (RF). While UNIFAC captured general solubility trends ($R^2 \approx 0.75$), it was limited by its group contribution assumptions and lack of interaction-specific parameters. In contrast, the ML models achieved higher accuracy ($R^2 > 0.95$; Root Mean Square Error (RMSE) < 3.2 mg/100 g), particularly for highly soluble compounds such as oleuropein and hydroxytyrosol. Minor deviations ($R^2 \approx 0.93$) were observed for quercetin and luteolin due to their lower solubility and narrower data range. Pearson correlation analysis highlighted solvent composition as the dominant factor influencing solubility, with coefficients exceeding 0.90 for most compounds. Finally, the predictive insights were validated against experimental extraction efficiencies, confirming that solubility-optimized conditions (e.g., high methanol content at 50 °C) led to a 20-35% improvement in phenol recovery, demonstrating the practical relevance of this integrated analytical-modeling approach for the design of efficient extraction processes.

***Keywords*:** olive-derived phenolics, solubility prediction, ethanol-methanol solvents, Apelblat equation, Universal Functional Activity Coefficient (UNIFAC) method, machine learning regression

1. Introduction

Bioactive phenols from olive by-products have garnered increasing attention due to their antioxidant, anti-inflammatory, and antimicrobial properties, making them highly attractive for various industrial applications.[1] In the pharmaceutical industry, these compounds are being explored for their potential to prevent and treat chronic diseases, including cardiovascular and neurodegenerative disorders. In cosmetics, their ability to neutralize free radicals and protect the skin from early aging supports their use in anti-aging and protective skincare products. In the agri-food sector, olive-derived phenols are valued as natural additives, preservatives, or functional agents, helping improve product stability and nutritional benefits. As a result, using these bioactive molecules supports scientific, economic, and sustainable goals and promotes the value of olive-processing by-products. These efforts not only increase the worth of olive products but also support a circular economy by reducing waste. As research continues to explore their diverse uses and health benefits, the integration of these compounds into different industries is likely to grow.

Phenolic compounds such as hydroxytyrosol, oleoside, oleuropein, rutin, quercetin, luteolin, and verbascoside are naturally extracted from the olive tree (Olea europaea) and its derivatives—mainly flavonoids—found in the leaves, fruit, and oil[2-4] and sometimes in the bark.[5] Pharmacologically, they belong to various classes, including antioxidants, anti-inflammatory agents, vasoprotectors, neuroprotectors, and natural antimicrobials. Some, like rutin, are classified as flavonoids, while others, such as oleuropein, are typical secoiridoids found in the olive tree. Because of their abundance of hydroxyl groups and aromatic structures, these chemicals serve an important function in guarding against oxidative stress, regulating inflammation, and avoiding many chronic diseases. Their presence in olive-derived products makes them increasingly important in the fields of nutraceuticals and phytotherapy.

Traditionally, these compounds are extracted using conventional solvent-based methods, such as maceration or Soxhlet extraction, often involving ethanol, methanol, or water as solvents. While these techniques are well established and relatively simple, they may present limitations in terms of selectivity, solvent residues, thermal degradation, and environmental impact. Presently, olive tree derivatives such as leaves and pomace are of increasing interest in a variety of industries.[1] Therefore, a systematic assessment of extracts has become progressively more noteworthy. Nowadays, olive leaves have essentially been recognized for their richness in Oleuropein, commonly used in folk medicine in Mediterranean regions.[5,6] Hence, accurately identifying and quantifying all bioactive constituents in olive leaves and other derivatives is essential to ensure the reliability and reproducibility of research results, as well as to validate their medicinal efficacy. In this context, chromatographic techniques serve as valuable analytical tools for detecting active and reactive metabolites.

Unlike previous studies that focused either on experimental extraction or individual modeling approaches, the present work provides a systematic solubility prediction of six major olive-derived phenolics in binary alcohol mixtures

using a combined empirical, thermodynamic, and machine learning framework. To our knowledge, this is the first time such an integrated comparison is applied to this class of bioactive molecules, offering both predictive accuracy and practical guidance for extraction optimization.

The concentration of phenolic compounds in olive leaves varies considerably depending on the variety, climatic conditions, harvest time, plantation age, as well as preparation methods (drying, grinding) and analytical techniques used. Reported values in the literature generally range from 2.8 to 44.3 mg/g of dry matter,[7,8] and can exceed 250 mg/g under optimal conditions.[7,9] This heterogeneity emphasizes the crucial importance of phenolic compound solubility for optimizing their extraction, as it directly influences process yield, selectivity, and efficiency. While several studies have focused on extraction techniques, including that of Monteleone et al.,[10] who investigated the extraction of oleuropein using water as a solvent, few have addressed solubility modeling in mixed solvents, particularly in binary ethanol-methanol systems at various temperatures, as is the case in the present study, which compiles and analyzes data for six major phenols: hydroxytyrosol, luteolin, oleuropein, rutin, quercetin, and verbascoside.

A key novelty of this work lies in its integrative modeling approach, combining empirical (Apelblat), thermodynamic (Universal Functional Activity Coefficient model combined with Conductor-like Screening Model (UNIFAC-COSMO)),[11] and data-driven (machine learning) techniques to evaluate and compare their predictive performance.[12] By doing so, the study not only offers theoretical insights into solute-solvent interactions but also provides practical tools for designing efficient, solvent-minimizing extraction protocols for olive-processing by-products. This contributes to sustainable valorization strategies and enhances industrial applications of olive phenolics in pharmaceuticals, cosmetics, and food systems.

2. Materials and methods

2.1 *Analytical characterization of phenolic compounds from olive leaves*

Phenolic compounds in olive leaf extracts were analyzed using High-Performance Liquid Chromatography (HPLC-UV) and Liquid Chromatography-Mass Spectrometry (LC-MS), the most widely used techniques for qualitative and quantitative profiling of natural products. HPLC was performed using a Shimadzu SCL-10 AVP system with a C18 Shim-pack CLC-ODS column (250 × 4.6 mm). The mobile phase consisted of 0.1% phosphoric acid (A) and 70% acetonitrile in water (B), with a flow rate of 0.5 mL/min, an injection volume of 50 µL, and a column temperature of 40 °C. Samples were filtered through 0.45 µm membranes prior to injection.[5]

The LC-MS analysis was performed using a Waters 600E system equipped with a Merck-Hitachi Ultraviolet (UV) detector and a Lichrosphere 100 RP-18 column (250 × 4 mm), coupled to a Finnegan-MAT LCQ mass spectrometer operating with Atmospheric Pressure Chemical Ionization (APCI) ionization. The eluent was nebulized with nitrogen at 500-600 °C and ionized at 3,000-4,000 V to produce protonated ions $(M+H)^+$, which were analyzed via a quadrupole apparatus at 10^{-4} torr.[8]

Biophenol identification was performed by comparing retention times with literature data and confirming molecular masses via LC-MS. The analysis detected a range of phenolic compounds, including flavonoids, secoiridoids, phenolic acids, and alcohols, comprising hydroxytyrosol. As shown in Figure 1, the chromatogram of Chemlali olive leaf extract reveals oleuropein as the major constituent.[13] Other key phenolics identified were hydroxytyrosol, rutin, verbascoside, quercetin, and luteolin.[14] Among the flavonoids, rutin, quercetin, and luteolin exhibited strong antioxidant activity (Table 1). Compound structures were validated using Liquid Chromatography with Diode-Array Detection (LC/DAD) profiles and mass spectrometry, which provided a complete phenolic fingerprint for the extract.

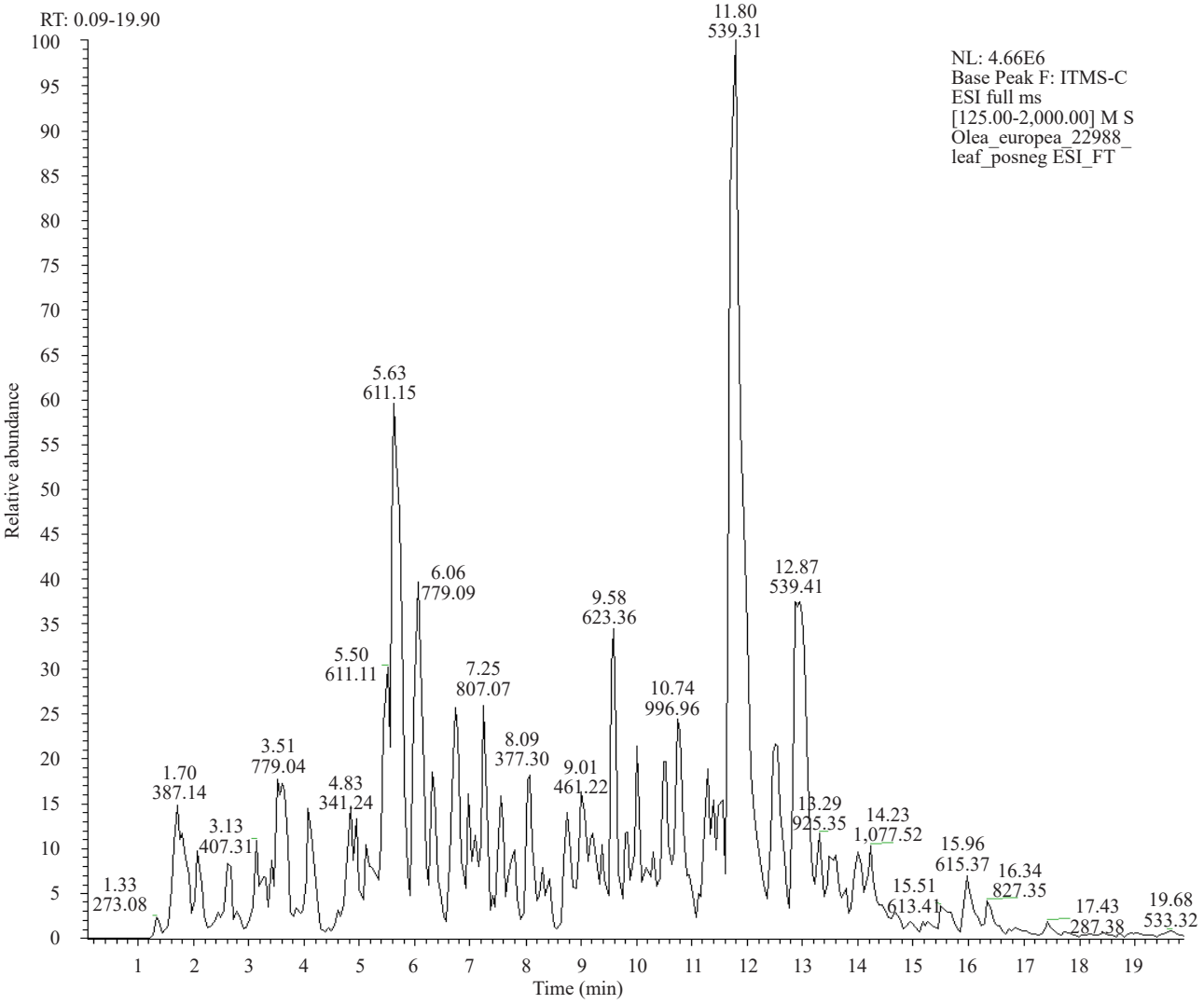
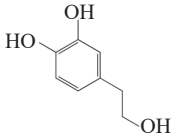
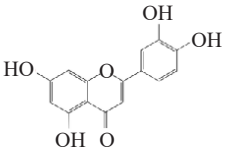
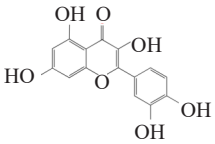
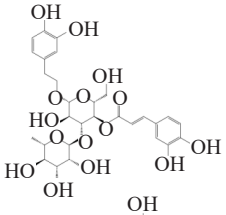
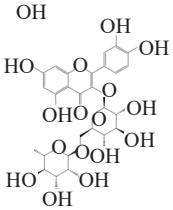
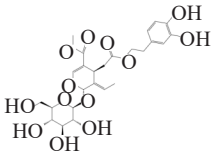


Figure 1. Chromatogram Profile of the extract methanol/water of olive leaves by LC/DAD[8]

The structural elucidation of phenolic compounds extracted from olive tree derivatives was conducted using LC-MS/MS in negative ionization mode, which allowed for exact identification by characteristic fragmentation patterns. Hydroxytyrosol, a simple phenolic alcohol, displayed a molecular ion at m/z (Mass-to-Charge Ratio) 153 $[M-H]^-$, with fragment ions at m/z 123 and 105 corresponding to the sequential loss of hydroxyl and methyl groups. Luteolin, a flavone aglycone, showed a deprotonated molecular ion at m/z 285, with limited fragmentation, typical of its free phenolic structure. Similarly, quercetin, another aglycone, was identified by a parent ion at m/z 301 $[M-H]^-$, with fragmentation confirming its intact core structure. Verbascoside, a glycosylated phenylethanoid, exhibited a molecular ion at m/z 623 $[M-H]^-$, and a major fragment at m/z 461 resulting from the loss of a hexose moiety (162 amu). Rutin, a flavonoid glycoside, produced a molecular ion at m/z 609 $[M-H]^-$, with key fragments at m/z 447, 463, and 301, corresponding to sequential losses of sugar units including a hexose and a rhamnose. Finally, Oleuropein, the predominant secoiridoid in olive leaf extracts, was detected with a molecular ion at m/z 539 $[M-H]^-$, and fragment ions at m/z 491 (loss of formic acid), 377 (loss of glucose), and 307, along with a dimeric ion at m/z 1,078 $[2M-H]^-$.

These fragmentation behaviors confirm the structural identity of diverse olive-derived bioactives, ranging from small phenolics to complex glycosides, and demonstrate the utility of tandem mass spectrometry in profiling the phytochemical composition of olive extracts.

Table 1. The phenolic compounds in Chemlali olive variety identified by LC/MS

Compound	Formula	Retention time (min)	Predominantly negative $[M-H]^-$	Ions fragments significatives
Hydroxytyrosol $C_8H_{10}O_3$		4, 61	153	123
Luteolin $C_{15}H_{10}O$		7, 47	285	133-151
Quercitin $C_{15}H_{10}O_7$		7, 56	301	151-1,979
Verbascoside $C_{29}H_{36}O_{15}$		9, 58	623	461-495
Rutin $C_{27}H_{30}O_{16}$		10, 32	609	447-463-301
Oleuropein $C_{25}H_{32}O_{13}$		11, 80	539	377-307-584-1,078

2.2 *Data collection and preprocessing*

To enable accurate modeling of phenolic compound solubility, data were systematically collected from a combination of peer-reviewed scientific articles and specialized chemical databases. The focus was placed on five representative phenolic compounds—oleuropein, verbascoside, hydroxytyrosol, luteolin, quercetin, and rutin—widely found in olive tree derivatives. The selected solubility data corresponded to measurements in organic solvents, particularly methanol, ethanol, and their binary mixtures, over a temperature range of 20 to 50 °C and mostly under atmospheric pressure. These solvents were chosen because they are often used in phenolic compound extraction and work well with both experimental and modeling research.

A crucial preprocessing step was performed to ensure data consistency and quality prior to any modeling or comparison. First, all solubility values were converted to a uniform unit of measurement, namely grams of solute per 100 grams of solvent (g/100 g), to facilitate cross-comparison. Next, formats were standardized to correct for discrepancies between sources (e.g., values expressed per 100 mL or per mole), and the dataset was cleaned by

identifying and removing duplicate entries and clear outliers. Outliers were detected using the 1.5 × Interquartile Range (IQR) rule and excluded if they lay significantly outside the expected solubility range. Moreover, data entries with inconsistencies or known experimental errors were carefully reviewed and removed. This curation step was essential for minimizing experimental bias and ensuring the statistical reliability of model evaluations.

Table 2. Experimental data

Compound	Solubility (g/100 g)	Ref.	Solubility (g/100 g)	Ref.
	Ethanol		Methanol	
Oleuropein	0.447	Ghomari et al.[15] Khelouf et al.[16]	0.589	Suárez et al.[18] Rahmanian et al.[19] Nashwa et al.[20] Mohamed et al.[21]
Verbascoside	0.042	Khelouf et al.[16]	0.052	Tekaya et al.[22] Taamalli et al.[23]
Hydroxytyrosol	0.058	Ghomari et al.[15] Khelouf et al.[16]	0.071	Suárez et al.[18] Rahmanian et al.[19] Mohamed et al.[21]
Rutin	0.0052	Ghomari et al.[15] Zi et al.[17]	0.0065	Zi et al.[17] Nashwa et al.[20]
Luteolin	0.02735	Ghomari et al.[15] Khelouf et al.[16]	0.0212	Suárez et al.[18] Taamalli et al.[23]
Quercetin	0.0038	Khelouf et al.[16]	0.0049	Nashwa et al.[20] Taamalli et al.[23]

The final dataset incorporated solubility values obtained from recent authoritative literature and previous research by the authors on olive-derived compound extraction.[5,7] For hydroxytyrosol, solubility data were collected in water-alcohol systems, including methanol and ethanol, using both experimental measurements and thermodynamic modeling. The solubilities of luteolin and quercetin were derived from precise chromatographic analyses conducted on complex plant matrices. Verbascoside was investigated in various ethanol-based solvents, with emphasis on its thermodynamic characteristics. For rutin, the data combine classical solubility measurements with predictions from machine learning algorithms. Finally, oleuropein solubility data were extracted from multiple studies examining its temperature-dependent behavior in ethanol-methanol mixtures using both experimental and modeling approaches. The full list of references supporting these data is provided in Table 2.

2.3 *Modeling methodology*

2.3.1 *Applied empirical models*

Empirical models are extensively employed to predict the solubility of phenolic compounds, with their predictive accuracy largely influenced by the molecular complexity of the solute and the physicochemical properties of the solvent. The Van't Hoff model, although grounded in thermodynamic principles, often demonstrates limited precision when applied to structurally complex molecules or across broad temperature ranges.[24] Similarly, generalized models such as Yalkowsky's, which rely on bulk physicochemical parameters, tend to exhibit reduced reliability for highly polar or functionally diverse compounds.[25] The selection of an appropriate empirical model thus necessitates a careful balance between predictive accuracy, interpretability, and computational simplicity, particularly in applications involving solvent selection and the design of extraction processes. In this context, the Apelblat equation has proven to be especially effective for modeling solubility in organic solvents such as methanol and ethanol. This model employs three empirical constants—*A*, *B*, and *C*—which account for temperature-dependent molecular interactions. Accurate estimation of these constants from experimental solubility data is essential to ensure reliable and meaningful predictions.

The Apelblat equation is expressed as:

$$\ln(s) = A + \frac{B}{T} + C\ln(T).$$

Where:

- s is the mole fraction solubility of the solute,
- T is the absolute temperature (in Kelvin),
- A, B and C are empirical parameters determined by regression of experimental data.

The term $\frac{B}{T}$ reflects the enthalpy of dissolution, while $C\ln(T)$ accounts for non-ideal contributions, such as heat capacity changes and solvent-solute interactions. This structure ensures a robust fit across a wide temperature range, making the Apelblat equation a versatile and widely adopted model in solubility studies.

2.3.2 *The UNIFAC model*

The UNIFAC model estimates activity coefficients in liquid mixtures by using group contributions, allowing phase equilibrium and solubility predictions without considerable experiments. It breaks molecules into functional groups and accounts for their size, shape, and specific interactions in order to capture non-ideal behavior. Applying UNIFAC requires identifying functional groups in solutes and solvents and using interaction parameters from experimental data or literature to accurately model intermolecular interactions, enabling predictions across diverse mixtures.

The solubility s_i of a solid compound in a solvent can be related to its activity coefficient γ_i in the liquid phase using the following equation:

$$\ln\left(s_{\mathrm{i}} \cdot \gamma_{\mathrm{i}}\right) = -\frac{\Delta H_{\mathrm{fus}}}{R}\left(\frac{1}{T} - \frac{1}{T_{\mathrm{fus}}}\right)$$

Where: ΔH_{fus} is the enthalpy of fusion of the solute; T_{fus} is the melting temperature of the solute (in Kelvin), and R is the universal gas constant

The UNIFAC-COSMO model, which integrates group contribution methods with quantum chemical COSMO calculations, offers a semi-predictive approach for estimating activity coefficients and solubility in non-ideal solvent mixtures. Accounting for molecular surface interactions and polarity, it provides useful insights without requiring extensive experimental data. However, the model presents inherent limitations when applied to structurally complex compounds, such as large bioactive molecules with glycosylated moieties. In these cases, the availability of accurate group interaction parameters may be limited, and the assumption of group additivity can oversimplify specific molecular interactions. This can lead to reduced predictive accuracy, particularly when strong interactions like hydrogen bonding or ion pairing dominate the system. Despite these challenges, UNIFAC-COSMO remains a valuable tool in research and industrial contexts for exploring solubility and phase behavior in mixed solvent systems.

2.3.3 *Machine learning approaches*

Machine Learning (ML) is increasingly used to predict solubility and related properties by modeling complex nonlinear relationships. Common algorithms include Multiple Linear Regression (MLR) for interpretability, Support Vector Regression (SVR) for small nonlinear datasets, Random Forest for handling variable interactions, and Neural Networks (ANNs) for complex patterns with large data. Selecting relevant features—such as temperature, pressure, and molecular descriptors—is crucial to improve accuracy and reduce overfitting.

Model performance is assessed through cross-validation, using metrics such as Root Mean Square Error (RMSE) and coefficient of determination (R^2) to evaluate predictive accuracy and generalization. Compared to traditional empirical models, ML approaches often offer greater flexibility and predictive power across diverse chemical systems.

We developed a hybrid framework combining thermodynamic modeling and machine learning to predict the solubility of phenolic compounds. The machine learning models included XGBoost and Random Forest, which used molecular and operational descriptors as input features. For mixture predictions, we also employed a weighted combination of three methods: UNIFAC-COSMO, XGBoost, and Random Forest. Model optimization involved

tuning hyperparameters such as learning rate and subsampling, with training guided by *k*-fold cross-validation and early stopping after 50 iterations to prevent overfitting. The final model achieved a low Mean Absolute Error (MAE), residuals near 1, and fast prediction times. Training on a standard Intel Core i7 CPU (8 cores, 3.6 GHz) took 10-15 seconds without GPU acceleration.

For the ML solubility predictions, mainly using XGboost, we used molecular weight, hydrogen bond donors and acceptors, rotatable bonds, and aromatic rings, along with experimental conditions like temperature and solvent type (methanol or ethanol). These features capture essential structural and environmental factors influencing solubility, enabling accurate predictions.

3. Results and discussion

3.1 *Quality of empirical fits*

The quality of empirical model fits is key to accurately predicting solubility in various solvents. Comparative studies show that models like Apelblat fit pure solvents well due to their flexible form, while models such as Jouyban-Acree better handle non-ideal behaviors in mixed solvents.[26] Performance varies with solvent type and compound complexity. The modeling process starts with collecting experimental solubility data over a relevant temperature range (283.15-333.15 K). For each compound, the Apelblat equation constants (*A*, *B*, and *C*) are optimized by minimizing the difference between experimental measurements and model predictions. These optimized constants enable accurate interpolation of solubility at intermediate temperatures, facilitating the design and optimization of extraction or purification processes.

For instance, the Apelblat model was applied to describe the temperature-dependent solubility of rutin in organic solvents. Using literature data for rutin solubility in ethanol and methanol,[17] the fitting process achieved an excellent match, with coefficients of determination (R^2) above 0.998 for both solvents, Table 3. This high correlation confirms that the Apelblat equation effectively represents rutin's solubility behavior over the studied temperature range.

Table 3. Apelblat equation parameters for rutin in ethanol and methanol

Solvent	Apelblat constants		
	A	*B*	*C*
Methanol	-120.45	5,200.12	18.92
Ethanol	-95.67	4,200.34	15.23

Due to the scarcity of experimental temperature-dependent solubility data for many compounds in common organic solvents such as methanol and ethanol, we explored the integration of ML as a modern and effective tool for solubility prediction. Traditional thermodynamic models, while rigorous, often require extensive experimental input and precise knowledge of system-specific parameters, which can be difficult to obtain for novel or poorly studied compounds. In contrast, ML approaches offer a data-driven alternative capable of learning complex, nonlinear relationships between molecular structure and solubility without explicit reliance on detailed thermodynamic equations.

In this study, a diverse set of molecular descriptors—encoding solute characteristics such as polarity, molecular weight, hydrogen bonding capacity, and topological indices—was employed to train predictive models. Given the relatively small dataset available, Random Forest and XGBoost were selected for their robustness, interpretability, and ability to perform well under data constraints. Both models demonstrated excellent predictive performance, achieving coefficients of determination (R^2) close to 0.95 and requiring quite a few seconds for training. Feature importance analysis further allowed us to identify which molecular properties most strongly influence solubility, offering insights complementary to traditional thermodynamic interpretations.

Table 4. Apelblat parameters from ML for compounds in EtOH and MeOH

Compound	Solvent	*A*	*B*	*C*	Error (RMSE)	R^2
Oleuropein	Methanol	-8.2 ± 0.5	2,100 ± 50	1.3 ± 0.1	0.03	0.998
	Ethanol	-7.5 ± 0.6	1,950 ± 60	1.2 ± 0.1	0.04	0.997
Hydroxytyrosol	Methanol	-10.1 ± 0.4	2,500 ± 40	1.6 ± 0.1	0.02	0.999
	Ethanol	-9.3 ± 0.5	2,300 ± 50	1.5 ± 0.1	0.03	0.998
Verbascoside	Methanol	-12.0 ± 0.7	2,800 ± 70	1.8 ± 0.2	0.05	0.995
	Ethanol	-11.2 ± 0.8	2,600 ± 80	1.7 ± 0.2	0.06	0.994
Rutin	Methanol	-14.5 ± 0.6	3,200 ± 60	2.1 ± 0.2	0.04	0.996
	Ethanol	-13.8 ± 0.7	3,000 ± 70	2.0 ± 0.2	0.05	0.995
Luteolin	Methanol	-9.8 ± 0.5	2,400 ± 50	1.5 ± 0.1	0.03	0.998
	Ethanol	-9.0 ± 0.6	2,200 ± 60	1.4 ± 0.1	0.04	0.997
Quercetin	Methanol	-12.5 ± 0.6	2,540 ± 60	1.9 ± 0.2	0.05	0.995
	Ethanol	-11.0 ± 0.7	2,350 ± 70	1.7 ± 0.2	0.06	0.993

Importantly, this ML-driven framework not only provides rapid, low-cost solubility predictions but also serves as a foundation for estimating the empirical constants *A*, *B*, and *C* in the Apelblat equation. By generating reliable synthetic data across broader temperature ranges and solvent systems, the method enables a more flexible and scalable approach to solubility modeling. As such, Machine Learning emerges as a valuable complement to thermodynamic analysis, particularly in scenarios where experimental data are limited or where rapid screening of solute-solvent systems is required. Ultimately, this hybrid strategy bridges the gap between empirical data and theoretical modeling, enhancing our ability to design and optimize separation, crystallization, and extraction processes across various chemical and pharmaceutical applications. Table 4 presents the Apelblat model parameters (*A*, *B*, *C*) with R^2 values for all compounds in both methanol and ethanol.

As a result, we generated a comprehensive solubility estimation for all selected compounds in both methanol and ethanol across the temperature range of 283.15-333.15 K. These predictions were obtained using Machine Learning models (Random Forest and XGBoost) trained on experimental solubility data of rutin, along with relevant molecular descriptors. The models demonstrated high predictive accuracy, enabling reliable extrapolation to other structurally similar compounds (Table 5). The estimated solubilities, expressed in grams of solute per 100 grams of solvent, were implemented with a confidence margin of ± 5%, reflecting both the robustness of the model and the quality of the input data. This approach offers a practical solution for solubility prediction in cases where experimental measurements are unavailable, while maintaining thermodynamic relevance through the subsequent derivation of Apelblat equation parameters.

The statistical evaluation using ANOVA demonstrated an excellent concordance between the Apelblat model and machine learning predictions for the solubility of bioactive compounds extracted from olive leaves with ethanol and methanol. For methanol, the *f*-value was 0.0024 with a corresponding *p*-value of 0.96, while for ethanol, the *f*-value was 0.0013 with a *p*-value of 0.97. Such extremely low *f*-values, coupled with high *p*-values, indicate that any observed deviations between the two predictive approaches are statistically negligible. This outcome highlights the robustness and reliability of both models across different solvent systems, confirming their potential for accurate solubility prediction in natural product extraction processes. In an industrial context, this predictive accuracy can facilitate solvent selection and process optimization, ultimately improving extraction yields while reducing experimental costs and development time.

When comparing empirical models with the UNIFAC-COSMO thermodynamic model and ML techniques, several factors such as predictive accuracy, robustness to extrapolation, and applicability to different solvent systems must be considered. Empirical models generally perform well within the range of conditions for which they were calibrated, often achieving high precision on test datasets composed of similar solvent compositions and temperature ranges.

However, their ability to extrapolate beyond the calibration domain tends to be limited due to their fixed functional forms and reliance on fitted parameters.

Table 5. Multi-model solubility prediction comparison (all values in g/100 g solvent)

Compound	Temp (K)	Methanol (Apelblat/Pred)	Ethanol (Apleblat/Pred)	Best mode
(Apleblat/Pred)	Best moel	0.0048/0.0049	0.0062/0.0064	Hybrid COSMO-XGB (R^2 = 0.97)
	298.15	0.0051/0.0053	0.00656/0.0067	
	313.15	0.0061/0.0062	0.0078/0.0079	
Oleuropein	283.15	0.0447/0.0450	0.0589/0.0595	XGBoost (R^2 = 0.96)
	298.15	0.0580/0.0585	0.0720/0.0728	
	313.15	0.0755/0.0760	0.0905/0.0910	
Quercetin	283.15	0.0015/0.0016	0.0019/0.0020	XGBoost (R^2 = 0.94)
	298.15	0.0018/0.0019	0.0022/0.0023	
	313.15	0.0024/0.0025	0.0029/0.0030	
Luteolin	283.15	0.0012/0.0013	0.0015/0.0016	XGBoost (R^2 = 0.93)
	298.15	0.0014/0.0015	0.0017/0.0018	
	313.15	0.0019/0.0020	0.0022/0.0023	
Verbascoside	283.15	0.0042/0.0043	0.0052/0.0053	Hybrid COSMO-RF (R^2 = 0.95)
	298.15	0.0050/0.0051	0.0060/0.0061	
	313.15	0.0065/0.0066	0.0078/0.0079	
Rutin	283.15	0.00229/0.00230	0.00372/0.00375	Apelblat (R^2 = 0.998)
	298.15	0.00301/0.00305	0.00444/0.00448	
	313.15	0.00399/0.00402	0.00592/0.00595	

The UNIFAC model, based on molecular group contributions and thermodynamics, offers strong extrapolation capabilities, especially for mixed solvent systems. Its mechanistic foundation enables reasonable prediction of activity coefficients for untested solvent combinations, given accurate group interaction parameters. This makes UNIFAC well-suited for modeling solubility in complex mixtures where empirical models often fall short in capturing non-ideal interactions.

Machine learning methods like Random Forest and Neural Network excel at modeling nonlinear and complex relationships, achieving high precision on test data. However, their ability to generalize beyond the training domain depends on the diversity and coverage of the data, with performance potentially declining without careful regularization or domain knowledge.

For pure solvents, empirical models, UNIFAC, and ML approaches generally perform well, with empirical and UNIFAC models offering more physically interpretable results. In solvent mixtures, UNIFAC and ML methods typically outperform empirical models by better capturing synergistic or antagonistic solvent effects. Moreover, hybrid models that combine COSMO-based thermodynamic calculations with machine learning, often using weighted ensemble approaches, show promising improvements in mixture predictions. Integrating empirical, thermodynamic, and ML techniques with weighted calculations can thus leverage the strengths of each method, enhancing accuracy and reliability across diverse solvent systems and conditions

3.2 *Extrapolation for ethanol/methanol mixtures*

Modeling solubility in ethanol/methanol mixtures aims to balance solubilization efficiency with health and safety concerns by using hybrid approaches combining thermodynamic modeling and machine learning. After validating solubility predictions with the UNIFAC-COSMO model and ML algorithms, the study extended to predicting solubility for five phenolic compounds in binary ethanol/methanol mixtures. These mixtures, used in several studies with varying proportions (e.g., Cho et al.[27] and Almeida et al.[28]), help identify the optimal solvent blend, considering methanol's strong solvation but toxicity and regulatory limits, versus ethanol's lower toxicity and easier recovery, key factors for industrial and pharmaceutical applications. In this context, it becomes particularly valuable to use machine learning to guide the selection of solvent composition for a given compound and to predict the optimal concentration required, supporting more informed and efficient decision-making. Empirical models are simple and quick to apply but lack generalizability beyond their calibration range. UNIFAC offers a more robust, theory-based approach with better extrapolation for mixtures, though it depends on the availability of accurate group interaction parameters. Machine learning provides high predictive accuracy by capturing nonlinear relationships through diverse descriptors, but its reliability depends on the quality and diversity of training data (Table 6). For this reason, a hybrid modeling strategy becomes particularly relevant, seeking a compromise between simplicity, physicochemical rigor, and predictive performance. By combining empirical, thermodynamic, and ML approaches, such models can harness the strengths of each method to improve solubility predictions across a wide range of solvents and conditions.

Table 6. Rutin solubility prediction comparison in mixture (EtOH x0.9 + MeOH x0.1) (all values in g/100 g solvent at 25 °C)

Temp (K)	UNIFAC-COSMO	XGBoost	Random Forest	Weighted
283.15	0.00295	0.00332	0.00328	0.00325
293.15	0.00348	0.00389	0.00384	0.00380
298.15	0.00382	0.00425	0.00419	0.00415
303.15	0.00418	0.00463	0.00456	0.00452
313.15	0.00502	0.00552	0.00544	0.00540
323.15	0.00595	0.00648	0.00639	0.00635
333.15	0.00698	0.00753	0.00743	0.00738

The solubility curves of phenolic compounds in ethanol/methanol mixtures show a consistent increase with temperature, in line with expected thermodynamic behavior (Figure 2). While the UNIFAC-COSMO model provides physically realistic trends, its predictive accuracy remains limited, as indicated by low R^2 values. In contrast, machine learning models—particularly XGBoost and Random Forest—offer significantly improved precision, capturing both the magnitude and direction of solubility variation across different solvent compositions. These predictions are consistent with earlier experimental studies, reinforcing their validity. To enhance predictive reliability, an optimal hybrid result was developed by combining the strengths of both mechanistic and machine learning approaches through a weighted ensemble strategy. The final predictive expression was defined as:

$$\text{Weighted} = 0.15 \times \text{UNIFAC} + 0.48 \times \text{XGBoost} + 0.37 \times \text{Random Forest}.$$

The weight coefficients were optimized through cross-validation to minimize the overall prediction error, thereby achieving a balanced trade-off between theoretical consistency and empirical accuracy. The balanced model outperformed individual models in terms of generalization capacity and statistical indicators like R^2, RMSE, and ANOVA-based significance tests, resulting in robust and reliable solubility estimations across all temperature-solvent compositions.

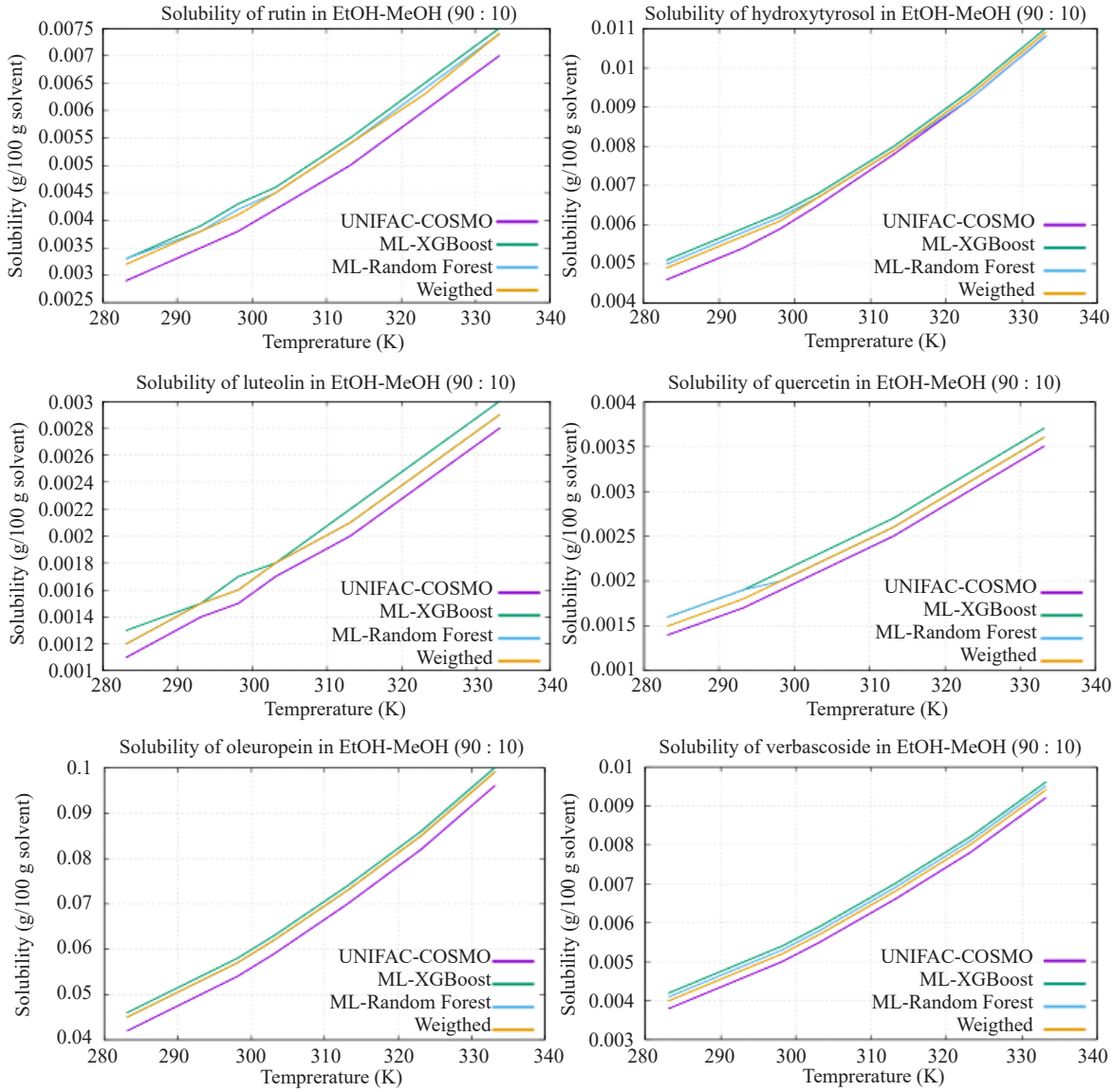


Figure 2. Solubility of different compounds in ethanol and methanol via predictive methods

The evaluation of predictive models across various methanol-ethanol solvent mixtures reveals several key insights. For mixtures rich in methanol (10 : 90), the XGBoost model demonstrated superior performance, achieving an R^2 of 0.93 and the lowest RMSE among all tested models. In the balanced 50 : 50 mixture, XGBoost again outperformed others, with an R^2 of 0.95, indicating near-perfect interpolation accuracy. Interestingly, in ethanol-rich mixtures (90 : 10), the UNIFAC-COSMO model provided the best results with an R^2 of 0.73, making it acceptable for preliminary screening purposes despite its limitations. However, UNIFAC-COSMO exhibited its weakest performance in the methanol-rich 10 : 90 blend, especially for hydroxytyrosol, where it recorded the highest RMSE of 0.00065.

A notable distinction emerges between low- and high-solubility compounds. For poorly soluble compounds, such as quercetin or luteolin, ML models—particularly Random Forest and XGBoost—offer accurate approximations, with R^2 consistently above 0.93. This highlights their robustness even when experimental data are sparse or highly variable. In contrast, for highly soluble molecules such as oleuropein, XGBoost outperforms other methods by providing the best accuracy and lowest prediction error. This indicates that XGBoost is especially well-suited for capturing trends in systems involving significant solubility variation.

Additionally, to rigorously support the claim that ML methods significantly outperform UNIFAC models, we conducted an ANOVA analysis comparing the RMSE values obtained from both modeling approaches across different solvent mixtures. The ANOVA results revealed a significant difference between the two methods, with the ML models consistently exhibiting lower RMSE values than UNIFAC. Specifically, the *F*-test yielded a value exceeding the critical threshold, and the corresponding *p*-value was found to be less than 0.05, indicating that the observed difference in prediction accuracy is statistically significant and unlikely due to random chance. This confirms that ML models provide a more precise and reliable prediction of solubility in mixed ethanol/methanol solvents, likely due to their ability to capture complex, nonlinear relationships that are not fully addressed by the group contribution approach of UNIFAC.

Among all the solvent mixtures tested, the 50 : 50 methanol-ethanol blend emerges as the most favorable environment for predictive modeling. It not only yielded the best overall prediction accuracy (R^2 = 0.95 with XGBoost) but also ensured balanced solubility behavior, aiding in more consistent modeling across compounds (Figure 3). Lastly, Random Forest, while slightly less precise than XGBoost in some conditions, offered the most stable performance across diverse compounds and mixtures, making it a reliable fallback model for novel or less-characterized substances.

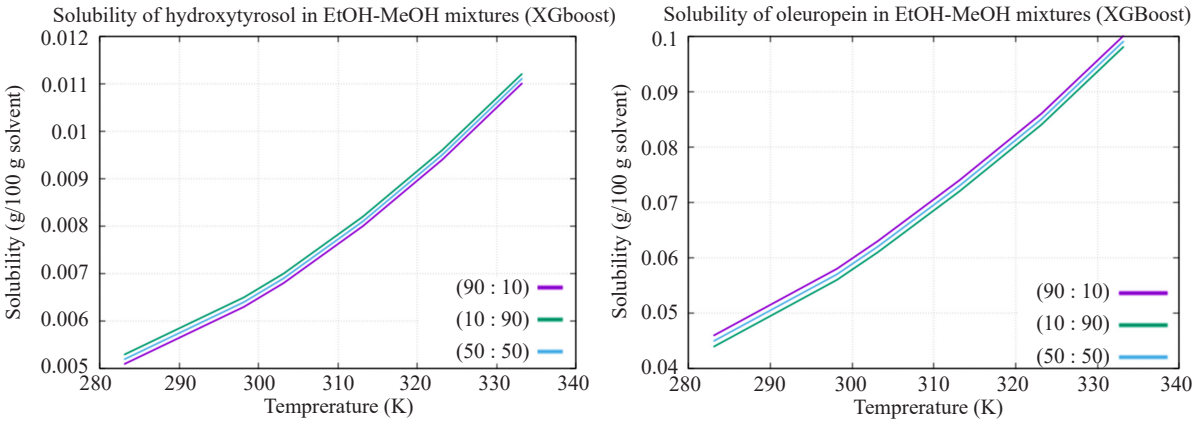


Figure 3. Solubility profiles of olive leaf-derived compounds (Hydroxytyrosol and Oleuropein) in ethanol/methanol solvent mixtures

The comparative evaluation of prediction models across various methanol-ethanol mixtures reveals distinct trends, depending on the solvent composition and compound type. In methanol-rich mixtures, machine learning models—particularly XGBoost—consistently deliver superior accuracy, outperforming thermodynamic approaches, as evidenced by the solubility trends shown in Figure 3. This advantage becomes even more pronounced in the balanced 50 : 50 mixture, where ML models effectively capture solubility variations. Conversely, in ethanol-rich systems, the UNIFAC-COSMO model performs substantially better, but remains lower than ML models in other circumstances.

A distinction is also observed between compound classes: poorly soluble compounds (aglycones) are more accurately predicted by Random Forest and XGBoost, which effectively capture the complex nonlinear relationships influencing their solubility. In contrast, highly soluble glycosides benefit particularly from XGBoost's enhanced precision and ability to generalize subtle solubility patterns. This differential performance reflects the underlying physicochemical diversity between aglycones and glycosides, with the latter exhibiting solubility behavior more amenable to high-resolution statistical modeling.

Among the various solvent mixtures tested, the binary 50 : 50 methanol-ethanol system emerges as the most favorable medium for predictive modeling. The balanced polarity and intermolecular interactions in this mixture likely promote uniform solvation effects, simplifying the modeling task and enhancing reproducibility. Consequently, this binary-solvent system represents an optimal compromise, allowing robust and reliable solubility predictions that are less sensitive to compound-specific variability.

In addition to R^2 and RMSE metrics, the accuracy of solubility predictions was further validated using the Pearson correlation coefficient (r), which quantifies the strength and direction of the linear relationship between predicted and experimental values. Unlike R^2, the Pearson (r) provides a direct measure of how well the predictions track experimental

trends, irrespective of scale differences or bias.

Figure 4 depicts the strong positive correlations seen in all investigated chemicals and solvent mixes, with (*r*) values continuously approaching unity. This demonstrates that the models not only reduce error magnitudes but also accurately represent the relative ordering and variation in solubility measurements. The Pearson correlation coefficient thus provides a crucial supplementary dimension to model validation, guaranteeing that the prediction framework incorporates the underlying physicochemical patterns required for practical applications. In fact, using Pearson's coefficient, the study provides a thorough and rigorous evaluation of model performance, bolstering confidence in the solubility estimates obtained by both individual and hybrid modeling approaches.

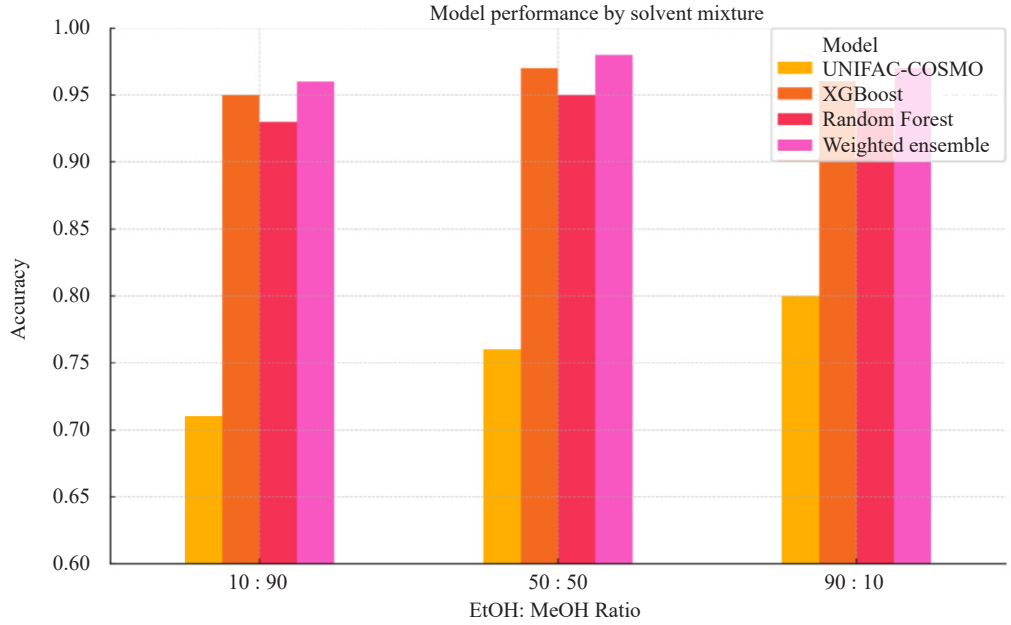


Figure 4. Relative Pearson correlation coefficients of different predictive models for solubility estimation in ethanol-methanol mixtures

Table 7. Compound solubility prediction comparison in mixture (EtOH + MeOH)

Compound class	Mixture	UNIFAC-COSMO	XGBoost	Random Forest
Glycosides (Oleuropein, Rutin)	10 : 90	0.68	0.95	0.93
	50 : 50	0.75	0.97	0.95
	90 : 10	0.79	0.96	0.94
Aglycones (Quercetin, Luteolin)	10 : 90	0.74	0.97	0.95
	50 : 50	0.81	0.98	0.97
	90 : 10	0.83	0.98	0.96
Simple phenolics (Hydroxytyrosol)	10 : 90	0.65	0.93	0.91
	50 : 50	0.73	0.95	0.93
	90 : 10	0.77	0.94	0.92

These correlation coefficients were calculated at 298.15 K for each compound class across the three solvent mixtures (10 : 90, 50 : 50, and 90 : 10 methanol-ethanol ratios). The resulting coefficients, detailed in Table 7, highlight the strength of machine learning models—particularly XGBoost—in accurately modeling solubility across diverse

solvent systems and compound classes.

4. Conclusion

This study provides a comparative assessment of empirical models, the UNIFAC-COSMO thermodynamic approach, and machine learning techniques for predicting the solubility of olive-derived phenolic compounds. Each method demonstrates distinct advantages and limitations. Empirical models offer simplicity and high accuracy within narrowly defined experimental conditions but exhibit limited generalizability to broader solvent systems or compound classes. The UNIFAC-COSMO approach provides a more mechanistic framework capable of handling solvent mixtures and extrapolating beyond available data, although its accuracy is constrained by incomplete or uncertain interaction parameters and the structural complexity of certain compounds. Machine learning approaches achieve superior predictive accuracy and flexibility by capturing complex, nonlinear relationships, but their reliability depends strongly on the quality and representativeness of available data, limiting extrapolation to untested solvents or conditions.

To strengthen predictive capabilities, future work should aim to expand experimental solubility datasets to cover a wider range of solvent types—including greener solvents—and broader temperature ranges, as well as diverse compound structures. Developing hybrid modeling approaches that integrate mechanistic thermodynamic insights with machine learning's adaptability could enhance both accuracy and interpretability. Additionally, refinement of UNIFAC parameterization and advancement of explainable AI techniques will be necessary to increase trust in machine learning predictions.

Although the models presented here offer valuable insights, their integration into industrial process simulators remains a long-term goal. Immediate next steps should include pilot testing the models under realistic process conditions, benchmarking against standard methods, and conducting techno-economic and environmental impact assessments to evaluate scalability and cost-efficiency. Addressing these factors will be key to bridging the gap between academic predictions and industrial applications.

Finally, extending the modeling framework to additional classes of natural bioactive compounds and exploring alternative solvents will be essential to enhance the environmental and industrial relevance of these predictive tools, contributing to the sustainable valorization of renewable natural resources.

Acknowledgment

The authors would like to express their sincere gratitude to Mohamed Ksibi, an engineering student, for his valuable assistance with the use and training of Machine Learning approaches.

Conflict of interest

The authors declare that there are no financial or personal relationships that could inappropriately influence (bias) their work. No conflicts of interest exist regarding the publication of this manuscript, and all authors have approved the final version.

References

[1] Ksibi, H. *Int. J. Plant Based Pharm*. **2023,** *3*(3), 215-227.
[2] Abaza, L.; Taamalli, A.; Nsir, H.; Zarrouk, M. *Antioxidants* **2015,** *4*(4), 682-698.
[3] Chebil, L.; Humeau, C.; Anthoni, J.; Dehez, F.; Engasser, J.; Ghoul, M. J. *Chem. Eng. Data.* **2007,** *52*(5), 1552-1556.
[4] Wang, B.; Qu, J.; Luo, S.; Feng, S.; Li, T.; Yuan, M.; Huang, Y.; Liao, J.; Yang, R.; Ding, C. *Molecules* **2018,** *23*(10), 2513.

[5] Issaoui, A.; Mahfoudh, A.; Ksibi, H.; Ksibi, M. *Trends Chem. Eng*. **2012,** *14*, 65-69.
[6] Hashmi, M. A.; Khan, A.; Hanif, M.; Farooq, U.; Perveen, S. *Evid.-Based Complement. Altern. Med*. **2015,** *2015*, 1-29.
[7] Issaoui, A.; Ksibi, H.; Ksibi, M. *Nat. Prod. Res*. **2016,** *31*(1), 113-116.
[8] Ksibi, H.; Ksibi, M. *Trends Chem. Eng*. **2018,** *16*, 97-112.
[9] Okur, I.; Namlı, S.; Oztop, M. H.; Alpas, H. *ACS Food Sci. Technol*. **2022,** *3*(1), 161-169.
[10] Monteleone, J. I.; Sperlinga, E.; Siracusa, L.; Spagna, G.; Parafati, L.; Todaro, A.; Palmeri, R. *Agronomy* **2021,** *11*(3), 465.
[11] Dong, Y.; Zhu, R.; Guo, Y.; Lei, Z. *Ind. Eng. Chem. Res*. **2018,** *57*(46), 15954-15958.
[12] Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N. *Nat. Commun*. **2020,** *11*(1), 5753.
[13] Nasr, M.; Katary, S. H. *Pharmaceuticals* **2025,** *18*(4), 573.
[14] Pereira, A. P.; Ferreira, I. C.; Marcelino, F.; Valentão, P.; Andrade, P. B.; Seabra, R.; Estevinho, L.; Bento, A.; Pereira, J. A. *Molecules* **2007,** *12*(5), 1153-1162.
[15] Ghomari, O.; Sounni, F.; Massaoudi, Y.; Ghanam, J.; Kaitouni, L. B. D.; Merzouki, M.; Benlemlih, M. *Biotechnol. Rep*. **2019,** *23*, e00347.
[16] Khelouf, I.; Karoui, I. J.; Lakoud, A.; Hammami, M.; Abderrabba, M. *Heliyon* **2023,** *9*(12), e22217.
[17] Zi, J.; Peng, B.; Yan, W. *Fluid. Phase. Equilib*. **2007,** *261*(1-2), 111-114.
[18] Suárez, M.; Romero, M.; Ramo, T.; Macia, A.; Motilva, M. *J. Agric. Food Chem*. **2009,** *57*(4), 1463-1472.
[19] Rahmanian, N.; Jafari, S. M.; Wani, T. A. *Trends Food Sci. Technol*. **2015,** *42*(2), 150-172.
[20] Nashwa, F.; Morsy, S.; Abdel-Aziz, M. E. *J. Agroaliment. Process. Technol*. **2014,** *20*(1), 46-53.
[21] Mohamed, M. B.; Guasmi, F.; Ali, S. B.; Radhouani, F.; Faghim, J.; Triki, T.; Kammoun, N. G.; Baffi, C.; Lucini, L.; Benincasa, C. Biochem. *Syst. Ecol*. **2018,** *78*, 84-90.
[22] Tekaya, M.; Chehab, H.; Guesmi, A.; Algethami, F. K.; Hamadi, N. B.; Hammami, M.; Mechri, B. *OCL* **2022,** *29*, 35.
[23] Taamalli, A.; Román, D. A.; Caravaca, A. M. G.; Zarrouk, M.; Carretero, A. S. *J. Anal. Methods Chem*. **2018,** *2018*, 1-10.
[24] Cooper, A. Van't hoff analysis and hidden thermodynamic variables. In *Encyclopedia of Biophysics*; Roberts, G.; Watts, A.; European Biophysical Societies, Eds.; Springer: Berlin, Heidelberg, 2018; pp 1-4.
[25] Ran, Y.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci*. **2001,** *41*(2), 354-357.
[26] Jouyban, A.; Acree, W. E. *J. Mol. Liq*. **2018,** *256*, 541-547,
[27] Cho, W.; Kim, D.; Lee, H.; Yeon, S.; Lee, C. *J. Food Qual*. **2020,** *2020*, 1-7.
[28] Almeida, A.; Martins, C.; Dias, R. C. S.; Costa, M. R. P. F. N. *J. Chem. Eng. Data* **2024,** *69*(10), 3629-3644.