Journal Pre-proof

Machine learning and COSMO-RS integration for predicting anthocyanin extraction from berries using eutectic solvents

Leonardo M.de Souza Mesquita, João A.P.Coutinho, Filipe H.B.Sosa

PII: S0308-8146(26)02191-6

DOI: https://doi.org/10.1016/j.foodchem.2026.150033

Reference: FOCH 150033

To appear in: *Food Chemistry*

Received date: 11 January 2026

Revised date: 15 April 2026

Accepted date: 9 June 2026

Please cite this article as: L.M.d.S. Mesquita, J. A.P.Coutinho and F. H.B.Sosa, Machine learning and COSMO-RS integration for predicting anthocyanin extraction from berries using eutectic solvents, *Food Chemistry* (2024), https://doi.org/10.1016/j.foodchem.2026.150033

Machine Learning and COSMO-RS Integration for Predicting Anthocyanin Extraction from Berries Using Eutectic Solvents

Leonardo M. de Souza Mesquita[1], João A. P. Coutinho[2], Filipe H. B. Sosa[2]*

[1]Department of Botany, Institute of Bioscience, University of São Paulo, Rua do Matão 277, São Paulo, SP, 05508-090, Brazil

[2]CICECO, Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal

Abstract

The development of efficient, green extraction processes for bioactive compounds from fruits requires strategies that account for both solvent–solute interactions and the inherent chemical heterogeneity of natural biomass. We present a hybrid computational framework integrating the COSMO-RS with machine learning (ML) to predict anthocyanin yields from diverse berry matrices using deep eutectic solvents (DES). A dataset of 299 experimental points spanning 15 biomass types was used to train and validate seven ML algorithms; Gradient Boosting emerged as the most accurate (R^2 = 0.92). The model combines COSMO-RS–derived solvent descriptors, key process variables, and a single experimental biomass descriptor: the anthocyanin content obtained with ethanol. The model was validated with independent literature data and new extraction experiments using different DES and berry matrices, showing strong

agreement between predicted and experimental yields. This approach minimizes experimental screening and enables rapid optimization of sustainable extraction processes for complex plant materials.

Keywords: berries, deep eutectic solvents, molecular descriptor; COSMO-RS; machine learning modelling; deep learning.

1. Introduction

The design of sustainable extraction processes has long centered on the rational selection of solvents that align with green extraction principles (de Souza Mesquita, Contieri, e Silva, et al., 2024). Historically, this selection relied heavily on empirical testing, often involving hazardous or non-renewable solvents. However, recent advances in computational modeling now offer a more efficient and environmentally conscious alternative. Among these, the Conductor-like Screening Model for Real Solvents (COSMO-RS) has emerged as a particularly promising tool for *in silico* solvent screening (Anantharaj & Banerjee, 2010; Lorenzo-Llanes et al., 2025; Pontes et al., 2025). By leveraging quantum chemical calculations, COSMO-RS predicts thermodynamic properties, such as activity coefficients, solubility parameters, and partition coefficients, for both pure solvents and complex mixtures. This enables researchers to assess, before experimentation, the capacity of a given solvent system to effectively solubilize (and thus extract) a target compound of interest. In the context of natural product recovery, such as anthocyanins from berries, which hold significant nutritional and economic value, this predictive capability reduces the need for trial-and-

error experimentation. It accelerates the development of truly green, scalable, and economically viable extraction processes. The design of sustainable extraction processes hinges on the rational selection of solvents that fulfill green chemistry criteria, like low toxicity, biodegradability, renewability, and high efficiency (Chemat et al., 2012). In this context, deep eutectic solvents (DESs) have gained significant attention as designer green solvents. DESs are typically formed by mixing a hydrogen bond acceptor (HBA) with at least one hydrogen bond donor (HBD), leading to strong specific interactions, primarily hydrogen bonding, that induce significant negative deviations from ideality (Abbott et al., 2003). These interactions result in the formation of a eutectic mixture whose melting point is substantially lower than that of the individual components. Their tunable physicochemical properties, low cost, and biocompatibility make them ideal candidates for natural product extraction (when properly designed). Our research group has leveraged COSMO-RS as a predictive framework to guide the design and selection of DESs for high-performance anthocyanin recovery. In one study, a nicotinamide-based DES was identified as optimal for extracting anthocyanins from grape pomace (de Souza Mesquita, Viganó, Veggi, et al., 2024). COSMO-RS screening revealed a highly negative natural logarithm of the activity coefficient at infinite dilution (ln $\gamma\infty$), indicating strong solute–solvent affinity among thousands of candidate systems. This prediction was experimentally validated: the DES enabled the recovery of 21 $mg_{anthocyanins}.g_{biomass}^{-1}$, with malvidin and peonidin aglycones as the predominant compounds, outperforming conventional ethanol–water mixtures. A second example focused on *jabuticaba* (*Plinia cauliflora*), a Brazilian berry rich in cyanidin-based anthocyanins (de Souza Mesquita et al., 2023). COSMO-RS was used to screen a large library of potential DESs, highlighting choline chloride (as HBA) paired with acidic HBDs, particularly lactic acid, as promising

for anthocyanin solubilization. The selected choline chloride–lactic acid DES yielded a threefold increase in anthocyanin extraction compared to ethanol–water mixtures, confirming the predictive power of COSMO-RS in tailoring eutectic solvents for specific phytochemical targets. Together, these cases demonstrate how integrating COSMO-RS with the modular design of DESs accelerates the development of efficient, green, and economically viable extraction processes, matching molecular-level insights with practical biorefinery goals.

Although COSMO-RS is an excellent tool for screening vast libraries of solvent candidates and identifying optimal DES systems based on molecular interactions, it has a key limitation: it does not account for process-level variables that critically influence extraction efficiency. COSMO-RS's thermodynamic framework is agnostic to process variables, such as solid-liquid ratio, extraction time, temperature, biomass type, or particle size, which can dramatically affect yields. To bridge this gap, machine learning (ML) offers a powerful complementary approach. When combined with COSMO-RS, ML models can exploit thermodynamically grounded descriptors to generate a solvent "fingerprint" that captures its intrinsic molecular interactions. These descriptors, such as predicted activity coefficients, polarity-related parameters, or σ-profile-derived features, enable the correlation of solvent characteristics with experimental extraction performance. As a result, this hybrid framework validates whether a solvent is thermodynamically suitable and could predict the expected extraction yield across a range of process parameters and a set of matrices for that specific solvent system. Such data-driven strategies have already shown promise beyond solvent selection, for instance, in predicting enzyme activity (Sosa et al., 2023), DES physicochemical

properties (Omar & Sadeghi, 2022), stability, or reaction kinetics in non-conventional media (Vittor et al., 2024).

Thus, in the context of anthocyanin recovery, the integration of COSMO-RS–guided DES design with ML-based process modelling represents a promising strategy to achieve green, high-efficiency extraction that accounts for both molecular affinity and operational variables. To this end, we compiled a database of published studies on DES-based anthocyanin extraction, encompassing 299 experimental data points across 15 distinct biomass matrices (Table S1). Seven different ML algorithms, namely, Gradient Boosting (GB), Logistic Regression optimized with Stochastic Gradient Descent (SGD), Support Vector Machines (SVM), Artificial Neural Networks (ANN), AdaBoost (AdB), k-Nearest Neighbors (kNN), and Random Forest (RF), were systematically evaluated. The best-performing model successfully captured the complex relationship between extraction conditions (e.g., time, temperature, and solid-liquid ratio) and solvent composition, achieving a coefficient of determination exceeding 0.9 ($R^2 > 0.9$). These seven algorithms were selected to represent a diverse spectrum of learning paradigms, ranging from linear (SGD-optimized Logistic Regression) and instance-based (kNN) to ensemble tree-based (RF, GB, AdB) and kernel- or network-driven (SVM, ANN), thereby enabling a robust comparative assessment of predictive accuracy, generalization capability, and sensitivity to nonlinear interactions among extraction variables. Given the complex, non-additive effects of solvent composition, hydration level, and process parameters on anthocyanin yield, this ensemble approach ensured comprehensive evaluation of both interpretable and high-capacity models within a sustainable extraction framework.

To validate the predictive capacity of this hybrid framework, we selected four globally available and commercially relevant berry matrices, namely, strawberry, blackberry, raspberry, and blueberry, chosen for their well-documented anthocyanin profiles, economic significance, and accessibility as representative feedstocks for anthocyanin extraction. The objective of this task is therefore to demonstrate an *in silico* guided workflow that couples COSMO-RS solvent screening with ML-driven process optimization to achieve high-yield anthocyanin extraction from real biomass using tailored deep eutectic solvents.

2. Materials and Methods

2.1. Chemicals and DES preparation

To experimentally validate the hybrid ML framework, the trained model was used to predict total anthocyanin content for a randomly selected set of DESs and operating conditions. Choline chloride (ChCl, ≥98%, Sigma-Aldrich) was selected as the fixed HBA. The hydrogen HBDs evaluated included: acetic acid (P.A., Merck), propionic acid (>99.5%), glycolic acid (>99.5%), anhydrous glycerol, levulinic acid (98%), *p*-toluene sulfonic acid (98%), sorbitol (99%), and anhydrous glucose, all sourced from Sigma-Aldrich unless otherwise specified. For comparative purposes, anhydrous ethanol (Synth) and ultrapure water (resistivity ≥18.2 MΩ·cm, obtained from a Merck Milli-Q® purification system) were also used as reference solvents in extraction trials. The DES was prepared by mixing the HBA and the HBD in the predefined molar ratio (1_{HBA}:2_{HBD}). The mixture was transferred into sealed glass vials containing a magnetic stirring bar and then heated in a paraffin bath at 333.15 ± 0.01 K under continuous stirring for approximately 2 hours, or until a homogeneous, transparent, and colorless liquid phase was obtained. After preparation, the water content of the mixtures was determined

using an 870 KF Titrino Metrohm volumetric titrator. The measured values were subsequently taken into account in the preparation of the DES aqueous solutions.

2.2. Anthocyanin samples, extraction and quantification

Berries (strawberry, blackberry, raspberry, and blueberry) were sourced from a Brazilian supermarket and certified organic producers. Samples were lyophilized for 96 h until constant weight was achieved, then ground to a fine powder using a domestic blender and stored at -40 °C until further processing.

All extractions were performed in a controlled ultrasonic bath operating at 37 kHz (135 W), serving as a reference system. Extraction parameters were fixed based on predictions from a hybrid machine learning (ML) framework: solid-liquid ratio of 0.1 $g_{biomass} \cdot mL_{solvent}^{-1}$, extraction time of 120 min, and a DES containing 35% (v/v) water. A total of 159 extraction experiments were conducted to validate the model across selected berry matrices and DES formulations. After extraction, the mixtures (DES+biomass after extraction) were immediately centrifuged at 14,000 rpm for 15 min at 4 °C. Supernatants were carefully collected—avoiding residual solids—and stored at –20 °C until analysis. Total anthocyanin content (TAC) was determined using the differential pH method, which relies on the absorbance shift of anthocyanins in two buffer systems (pH 1.0 and pH 4.5), following the protocol of Giusti and Wrolstad (Giusti & Wrolstad, 2001).

2.3. COSMO-RS

The COSMO-RS model was employed as a tool to generate the input parameters for solvent description. COSMO-RS was used to calculate the chemical potential, misfit

interaction energy, hydrogen-bonding interaction energy (H-bond), van der Waals interaction energy (VdW), and activity coefficients of each component that composes the solvent mixture (HBA, HBD, and water). These descriptors were selected due to their direct physical meaning and their ability to capture the key intermolecular interactions governing DES behavior. While chemical potentials and activity coefficients describe thermodynamic non-ideality and phase affinity, the decomposition into misfit, hydrogen-bonding, and van der Waals contributions enables a mechanistic interpretation of electrostatic, specific, and dispersion interactions, respectively.

These parameters, obtained for all components that make up the solvent system, were then used as input features for the machine learning models. Initially, the structures of the HBAs and HBDs were compared with the COSMObase2025. Compounds not found in the database were optimized for their lowest-energy conformers using the TmoleX 2025 (Graphical User Interface to the TURBOMOLE Quantum Chemistry Program Package) at the BP86/TZVP level of theory. Vibrational frequency analyses were performed for the Turbomole-generated files to confirm the absence of imaginary frequencies (Santiago et al., 2023).

Once the *.cosmo* files of all compounds were obtained, they were used as input in the COSMOtherm V 25.0.0 package (BP_TZVP_25 parametrization) for mixture calculations using the known solvent compositions at 25 °C (Paduszyński, 2017)**.**

2.4. Machine Learning (ML) and Data Visualization by Orange software

In this study, the Orange software package (version 3.39), a visual programming environment, was used to assess the performance of seven machine learning (ML) algorithms: Gradient Boosting (GB), Logistic Regression optimized with Stochastic

Gradient Descent (hereafter referred to as SGD), Support Vector Machines (SVM), Artificial Neural Networks (ANN), AdaBoost (AdB), k-Nearest Neighbors (kNN), and Random Forest (RF) (Demšar et al., 2013; Rodríguez-Pérez & Bajorath, 2020). A train–test split was applied to ensure proper model development and evaluation, with 75% of the data used for training and the remaining reserved for validation. This separation is essential to prevent overfitting and to obtain an unbiased estimate of predictive performance. Each experimental entry consisted of 19 input features and 1 output variable. The inputs included: (i) process variables (extraction time, temperature, and solid–liquid ratio), (ii) one biomass descriptor (total anthocyanin content obtained using ethanol), and (iii) COSMO-RS-derived descriptors for each component in the mixture (HBA, HBD, and water), namely chemical potential, hydrogen-bonding interaction energy (H-bond), van der Waals interaction energy (VdW), and activity coefficient. The output variable was the total anthocyanin content. The ethanol-extraction yield was included as an additional descriptor to allow the models to qualitatively capture differences in anthocyanin content among various biomasses. This choice was motivated by the fact that anthocyanin levels differ not only between biomass types but also within the same biomass due to environmental and agronomic factors such as climate, rainfall variation, cultivation practices, and cultivar-specific characteristics. To obtain an initial overview of model behavior, all algorithms were first run using the default Orange parameters (Table S2). For a more rigorous assessment of generalization performance and to make the most of the available dataset, a 5-fold stratified cross-validation procedure was employed.

Orange was also used to examine the influence of different input variables on model predictions through SHAP (Shapley additive explanations) values. In addition, the FreeViz visualization method was applied to explore multidimensional relationships and improve the interpretability of the dataset.

2.5. Statistical performance

In this study, five performance evaluation metrics were employed: coefficient of determination (R^2), Spearman correlation (r_s) mean square error (MSE), mean absolute error (MAE) and root mean square error (RMSE) in order to assess the predictive accuracy of the different models. These metrics provide complementary perspectives on model behavior, capturing aspects such as average deviation, sensitivity to large errors, overall predictive agreement, and the proportion of variance explained by the model (Ritter & Muñoz-Carpena, 2013). Their definitions are presented below:

$$\boldsymbol{MAE} = \frac{\sum_{i=1}^{N} |\boldsymbol{O_i} - \boldsymbol{P_i}|}{\boldsymbol{N}} \quad (1)$$

$$\boldsymbol{MSE} = \frac{\sum_{i=1}^{N} (\boldsymbol{O_i} - \boldsymbol{P_i})^2}{\boldsymbol{N}} \quad (2)$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\boldsymbol{O_i} - \boldsymbol{P_i})^2}{\boldsymbol{N}}} \quad (3)$$

$$\boldsymbol{R^2} = \frac{\sum_{i=1}^{N} (\boldsymbol{P_i} - \underline{\boldsymbol{O}})^2}{\sum_{i=1}^{N} (\boldsymbol{O_i} - \underline{\boldsymbol{O}})^2} \quad (4)$$

$$\boldsymbol{r_s} = \boldsymbol{1} - \frac{\boldsymbol{6} \sum_{i=1}^{N} \boldsymbol{D^2}}{\boldsymbol{n}(\boldsymbol{n^2} - \boldsymbol{1})} \quad (5)$$

where $\boldsymbol{O_i}$ and $\boldsymbol{P_i}$ represent the observed values and model estimates for a sample of size *N*, while $\underline{\boldsymbol{O}}$ denotes the mean of the observed values, and *D* is the difference between the two ranks of each observation.

3. Results

3.1. Experimental dataset

The experimental database (Table S1) was assembled based on total anthocyanin content data from different types of biomass, predominantly berries (grape, cherry, strawberry, cranberry, blackberry, blueberry, bilberry, and others), using deep eutectic solvents assisted by an ultrasonic bath. A total of 299 experimental data points on total anthocyanin content (TAC) were collected and organized to train machine learning models. TAC was selected because the types of anthocyanins present vary between biomass sources; therefore, total quantification was chosen to provide a unified measure and to evaluate the effect of DES and the process conditions across different matrices. The experimental data, collected from the literature and obtained in this study, included 15 types of biomass, with grape representing the largest group (~18.8% of the data), followed by raspberry (13.4%), jabuticaba (11%), and haskap berry (10%). Regarding the DES, among the HBA components, choline chloride (ChCl) was predominant (~71% of the data), while citric acid accounted for ~10% of the HBD components (Figure 1). The temperature range of the dataset points was 20–80 °C, with a solid-liquid ratio of 0.01–0.1 $g_{biomass}.mL_{solvent}^{-1}$ and extraction time between 10 and 180 min (Figure S1-S3). The complete list of collected data, including the biomass, the DES, and the aqueous solutions considered in this study for the ML-based model, along with the process conditions, is provided in the Electronic Supplementary Material (Table S1). The dataset covers a wide range of TAC values, varying from 0.01 to 19.80 mg total anthocyanins per g of dry biomass (Figure S4).

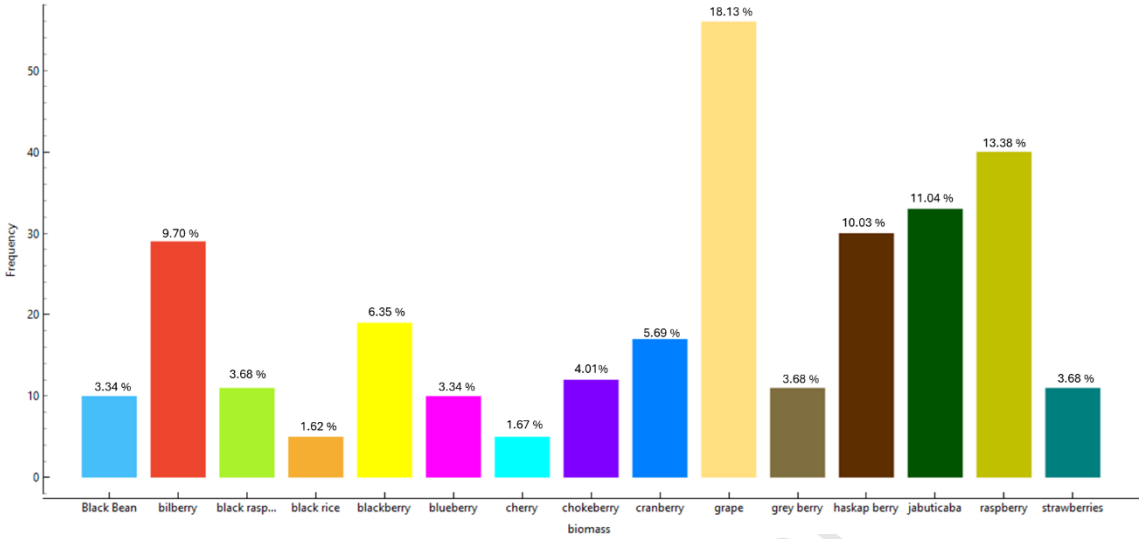


Figure 1. Distribution of berry biomass sources across the training dataset for the machine learning model. Colors represent the following biomass types: Black Bean (Light Blue), bilberry (Red), black raspberry (Lime Green), black rice (Orange), blackberry (Yellow), blueberry (Magenta), cherry (Cyan), chokeberry (Purple), cranberry (Blue), grape (Pale Yellow), grey berry (Light Brown), haskap berry (Dark Brown), jabuticaba (Dark Green), raspberry (Olive Green) and strawberry (Teal).

3.2. Total anthocyanin content correlation with descriptors

It is well established that both the solvent and the operational conditions exert a strong influence on the extraction yield of total anthocyanins. In recent years, COSMO-RS has been increasingly employed to support solvent selection for biomass and non-biomass processing (Eckert & Klamt, 2002; Hizaddin et al., 2022; Mohan et al., 2024; Sicaire et al., 2018). While this approach provides relevant molecular-level insights, its predictive capacity is intrinsically qualitative. COSMO-RS evaluates the solubility of representative solutes in different solvents but does not account for the additional mechanisms and process parameters that govern real extraction systems (Oliveira et al.,

2021). Factors such as extraction mode, temperature, solid-liquid ratio, and the structural complexity of biomass, simplified in COSMO-RS as isolated solutes, play decisive roles in practice. As a result, extraction behavior emerges from a combination of phenomena that extend beyond the model's descriptive scope.

This challenge becomes even more pronounced when dealing with heterogeneous biomass materials. Even within the same botanical family, despite apparent morphological and taxonomic similarities, considerable compositional variability is common, particularly in phytochemical profiles, cell wall architecture, and matrix composition. This variability is further amplified by differences in edaphic conditions (e.g., soil type, nutrient availability, pH) and geographic origin (e.g., climate, altitude, sunlight exposure), which significantly influence the biochemical composition of berry crops (Omwango et al., 2024). Although all materials classified as berries share certain general characteristics, such as high-water content, acidic pH, and the presence of polyphenol-rich vacuoles, their structural composition in terms of proteins, carbohydrates (e.g., pectins, cellulose, hemicellulose), and elemental makeup differs markedly (B. Yang et al., 2011). These differences directly affect solvent penetration, mass transfer kinetics, and anthocyanin solubilization in DES systems. Predicting anthocyanin yields from berry matrices in aqueous DES therefore, constitutes a nontrivial task, requiring an integrated approach that combines molecular descriptors of both biomass and solvent with key experimental extraction parameters. To address this complexity, the present study incorporated a broad set of descriptors. These included three process variables (time, temperature, and solid/liquid ratio), one biomass descriptor (anthocyanin content extracted by ethanol), and five COSMO-RS molecular descriptors (misfit energy – MF, van der Waals energy – vdW, hydrogen-bonding energy

– Hb, chemical potential – μ) together with the activity coefficient (γ) for each component (water, HBA, and HBD), yielding a total of 19 descriptors per mixture.

Initially, the dataset was subsequently projected into a two-dimensional space using the FreeViz algorithm (Figure 2). This visualization enabled exploration of the multivariate relationships among descriptors. The results reveal no direct or intuitive alignment between descriptor vectors and extraction performance, indicating that simple linear relationships do not characterize the system.

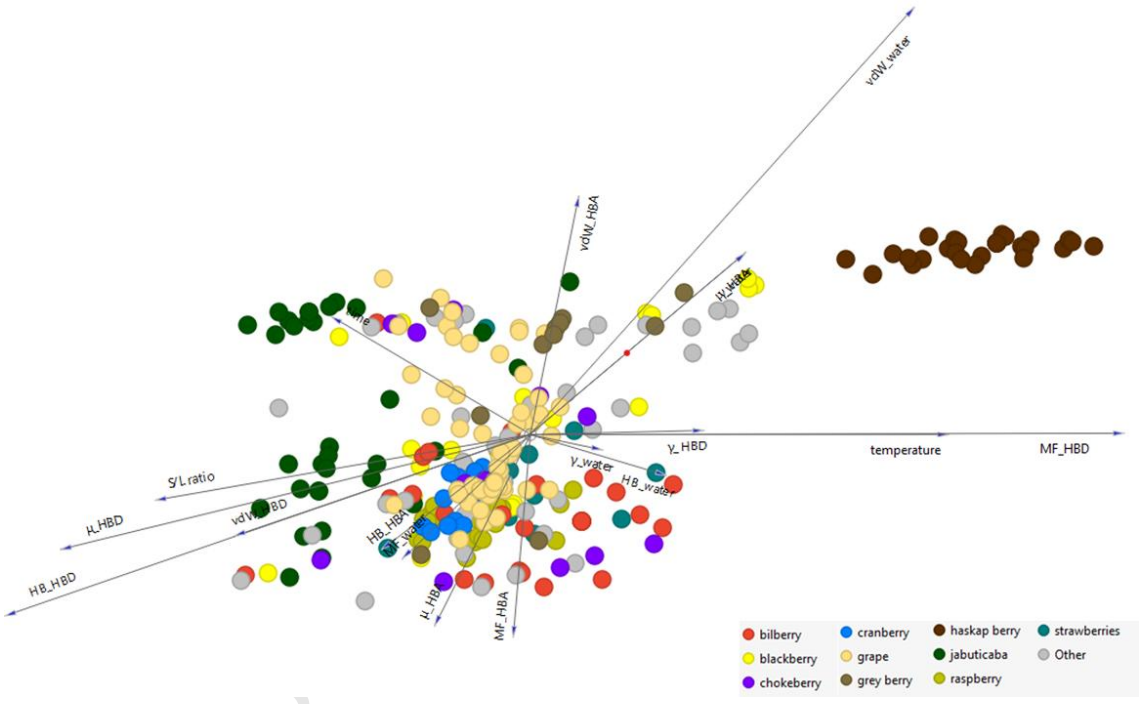


Figure 2. FreeViz diagram illustrating the similarity of records in the multidimensional space of solvent parameters and process conditions.

Although a cluster of points corresponding to haskap berry extractions (brown markers) is noticeable, this grouping does not reflect biomass-dependent behavior. Instead, it corresponds to extractions performed exclusively with a single DES composed of citric acid and maltose under varying conditions (MacLean et al., 2021). The cluster,

therefore, represents the chemical uniformity of the solvent rather than any intrinsic trait of the haskap berry matrix.

Overall, the absence of clear trends linking descriptors and anthocyanin yield highlights the nonlinear and multifactorial nature of the underlying phenomena. This scenario strongly supports the application of ML models to capture these complex interactions and improve predictive capability in DES-based extraction systems.

3.3. Data mining

To cover different machine learning paradigms and assess their suitability for predicting anthocyanin extraction yields, seven machine learning algorithms were selected: Support Vector Machines (SVM), AdaBoost (AdB), Random Forest (RF), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), k-Nearest Neighbors (kNN), and Artificial Neural Networks (ANN). These models can be broadly grouped according to their methodological similarities. Tree-based ensemble models (RF, GB, and AdB) share the ability to capture nonlinear relationships and complex feature interactions, making them particularly suitable for heterogeneous datasets (Hastie, Tibshirani, Friedman, et al., 2009). Among them, RF is known for its robustness against overfitting and reduced sensitivity to noise, whereas GB, by sequentially constructing decision trees aimed at minimizing residual errors, typically achieves higher predictive accuracy at the expense of increased sensitivity to hyperparameter selection (Jun, 2021). AdB, while effective at reducing bias in relatively simple problems, may exhibit inferior performance in the presence of strong nonlinearity and noisy data. In contrast, models such as SVM and SGD rely on more linear or quasi-linear principles, limiting their flexibility in representing complex interactions, whereas kNN is highly sensitive to local

data density and therefore becomes sensitive to high dimensionality (Hastie, Tibshirani, & Friedman, 2009). ANNs, in turn, offer high nonlinear modelling capacity but require larger datasets and careful tuning to avoid instability and overfitting (Roadknight et al., 1997).

To ensure statistical robustness and to evaluate the generalization capability of the models, a structured data-splitting strategy was adopted, dividing the dataset into training, validation, and test subsets. The training set was used to fit the internal model parameters, allowing the algorithms to learn the underlying patterns in the experimental data. An independent test set, corresponding to 15% of the total dataset, was reserved exclusively for the final performance assessment, ensuring an unbiased evaluation on unseen data. During the intermediate stages of hyperparameter optimization and model selection, the coefficient of determination (R^2) was used as the primary metric. For the final evaluation, additional metrics, including MSE, RMSE, MAE, and MAPE, were also reported (Table 1), enabling a comprehensive assessment of different error characteristics associated with the predictions.

Table 1. Mean square error (MSE), Root mean square error (RMSE), Mean absolute error (MAE), Mean absolute percentage error (MAPE) and Coefficient of determination (R^2) for different ML algorithms used in this study.

Model	MSE	RMSE	MAE	MAPE (%)	R^2
GB	1.50	1.22	0.78	49.41	0.92
RF	2.30	1.52	0.89	74.77	0.88
ANN	2.32	1.52	0.95	133.43	0.88
SGD	2.41	1.55	1.04	75.06	0.87
AdB	3.17	1.78	1.04	39.92	0.83
SVM	4.67	2.16	1.28	145.83	0.76

kNN	5.45	2.33	1.27	86.02	0.72

The comparative analysis of performance metrics (Table 1 and Figure 3) clearly indicates that ensemble-based models, particularly GB and RF, achieved the best overall results. The GB model emerged as the most accurate, exhibiting the lowest MSE (1.50), RMSE (1.22), and MAE (0.78), together with the highest R^2 value (0.92), demonstrating a strong ability to explain the variability of the experimental data. This superior performance can be attributed to the nature of the input data, which comprises 19 heterogeneous features, including process operating conditions, COSMO-RS-derived physicochemical descriptors, and the total anthocyanin content extracted using ethanol. The combination of these variables results in a highly nonlinear input space with complex cross-interactions, a scenario in which GB is particularly effective due to its sequential correction of residual errors and its ability to capture subtle contributions from less dominant variables.

A more detailed inspection of the results in Table 1 and Figure 3 further supports these observations. Although RF also achieved high performance (R^2 = 0.88), its error metrics were consistently higher than those of GB, suggesting that the bagging strategy of RF, while robust, is less effective at capturing fine-grained patterns present in the data. ANN and SGD exhibited intermediate performance, with R^2 values comparable to RF but accompanied by higher MAE and MAPE, indicating greater error dispersion and lower predictive stability. AdB showed a relatively low MAPE but higher MSE and RMSE, suggesting sensitivity to isolated significant prediction errors. Finally, SVM and kNN presented the poorest overall performance, with low R^2 values and high error metrics, highlighting their limitations in modelling complex, high-dimensional multivariate systems.

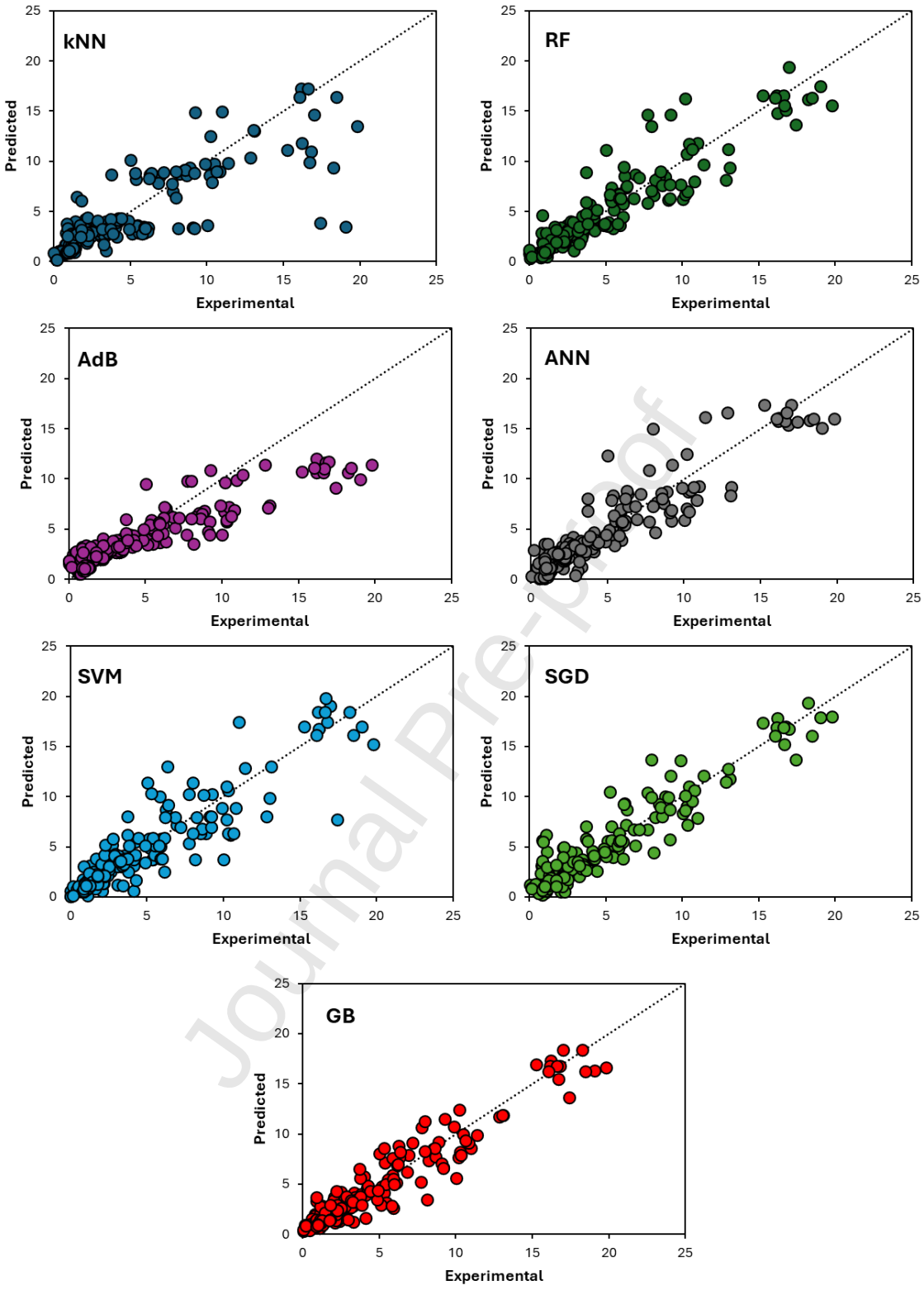


Figure 3. Experimental *versus* predicted total anthocyanin content (mg/g) using k-Nearest Neighbors (kNN), Random Forest (RF), AdaBoost (AdB), Artificial Neural

Networks (ANN), Support Vector Machines (SVM), with Stochastic Gradient Descent (SGD) and Gradient Boosting (GB).

Given that Gradient Boosting (GB) was identified as the most promising model for correlating the input features with the extraction yield, further model interpretation was conducted using SHAP analysis and permutation-based feature importance (Figure 4).

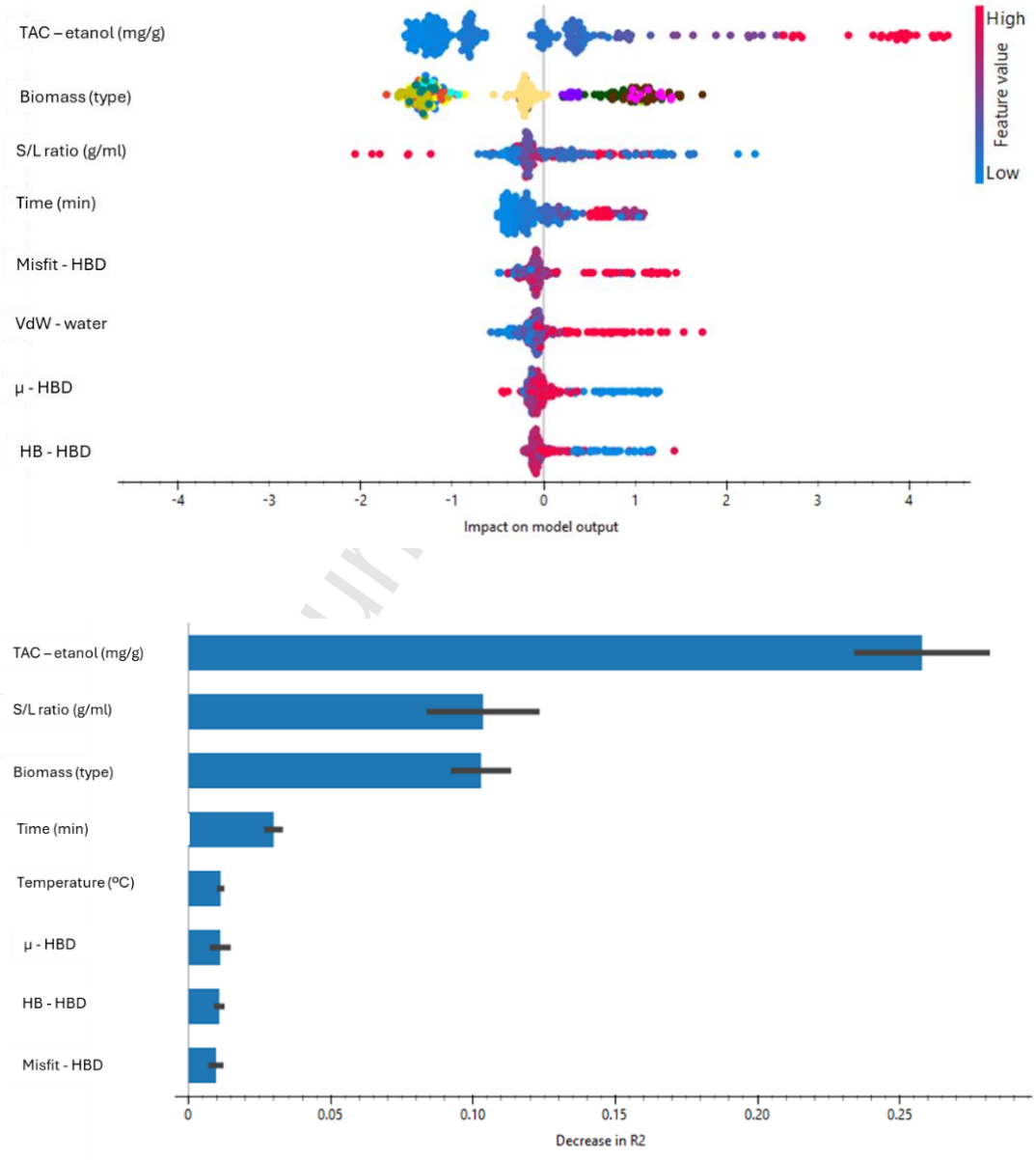


Figure 4. SHAP summary plot (top) and feature importance based on decrease in R^2 (bottom) for the eight most relevant features in the Gradient Boosting (GB) model. In the SHAP plot, each point represents an individual prediction, where color indicates low (blue) to high feature (pink) values. Biomass types are represented by distinct colors: Black Bean (Light Blue), bilberry (Red), black raspberry (Lime Green), black rice (Orange), blackberry (Yellow), blueberry (Magenta), cherry (Cyan), chokeberry (Purple), cranberry (Blue), grape (Pale Yellow), grey berry (Light Brown), haskap berry (Dark Brown), jabuticaba (Dark Green), raspberry (Olive Green) and strawberry (Teal).

The results indicate that anthocyanin extraction yield is predominantly governed by biomass-related and process-related variables, as well as by properties associated with the hydrogen bond donor (HBD) component of the solvent. Among all input features, the total anthocyanin content obtained with ethanol (TAC-ethanol) showed the highest feature importance. This reflects its role as a proxy for the maximum extractable anthocyanin in each biomass sample. Ethanol was used as a reference solvent due to its well-established efficiency in solubilizing anthocyanins (Nour et al., 2013; Y. Yang & Kilmartin, 2025). However, the inclusion of TAC–ethanol as a predictor should not be regarded as a limitation to a priori screening. On the contrary, ethanol extraction represents one of the simplest, fastest, and most accessible experimental assays to assess biomass extractability, especially when compared to more complex and labor-intensive descriptors such as cell wall composition. In this context, TAC–ethanol provides a practical and reliable first-level screening metric. Consequently, TAC–ethanol serves as an effective proxy for the intrinsic anthocyanin content of the biomass, indicating whether a given sample is inherently rich or poor in extractable anthocyanins,

while enabling subsequent optimization of solvents and process variables in a rational manner.

Consistently, the solid-liquid ratio and the biomass type also emerge as dominant variables, reinforcing the role of mass transfer limitations, matrix structure, and biomass-specific interactions in controlling the extraction process. Process conditions such as extraction time and temperature exhibit a secondary, yet non-negligible, influence on the predicted yield, suggesting that their role is primarily associated with fine-tuning the extraction efficiency rather than defining its upper limit.

Regarding solvent molecular descriptors, a clear asymmetry is observed between the roles of the hydrogen bond acceptor (HBA) and hydrogen bond donor (HBD) components. Notably, none of the COSMO-RS descriptors associated with the HBA appear among the eight most relevant features, indicating that variations in HBA properties do not significantly affect extraction yield within the present dataset. In contrast, several HBD-related descriptors, including chemical potential, hydrogen-bond interaction energy, and mismatch parameters, display a measurable impact on model predictions. This suggests that the solvent effect is primarily governed by its hydrogen-donating ability, which can be rationalized by the stabilization and solvation of anthocyanins once mass transfer constraints are overcome. To assess the potential bias from the predominance of choline chloride (ChCl) as HBA, an additional analysis excluding ChCl-based systems (87 data points) was performed. The results (Figure S6) showed that HBA-related descriptors remain of low importance, indicating that their limited influence is not solely due to reduced variability.

To further investigate and exemplify the relative influence of key variables, the effect of TAC-ethanol was compared with that of temperature as a representative secondary factor. The analysis of Individual Conditional Expectation (ICE) plots (Figure 5) further corroborates the feature attribution results obtained from SHAP and permutation importance analyses.

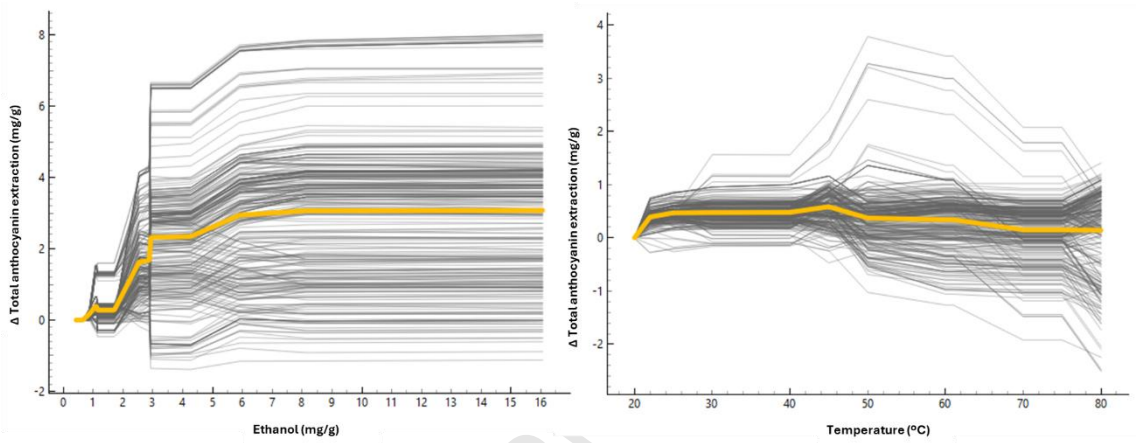


Figure 5. Individual Conditional Expectation (ICE) plots illustrating the effect of total anthocyanin content with ethanol - TAC_ethanol (left) and extraction temperature (right) on the predicted anthocyanin extraction yield obtained from the Gradient Boosting model. Grey lines represent individual conditional responses for each observation, while the yellow line denotes the average effect across all samples.

The ICE profiles for TAC-ethanol display a pronounced and systematic positive effect on the predicted extraction yield, characterized by a steep increase followed by a clear plateau, indicating that this variable strongly defines the maximum extraction yield. In contrast, the ICE curves associated with temperature exhibit a markedly flatter and more heterogeneous behavior, with only modest changes in model response across the explored range and no consistent trend among individual samples. This comparatively weak and dispersed effect explains why temperature does not appear

among the most influential features in the SHAP analysis. Overall, the ICE results confirm that TAC-ethanol exerts a more substantial influence on model output than temperature, reinforcing the interpretation that the intrinsic extractability of the biomass dominates the extraction yield, while temperature plays a secondary, fine-tuning role.

3.4. Evaluation of model predictive performance based on experimental literature data

Although it is common to evaluate model performance using metrics such as R^2, MSE, MAE, MAPE, and RMSE, these indicators do not capture a model's extrapolative ability and remain limited to the dataset under study. Even when reserving part of the data for post-training validation, the evaluation remains constrained to the original dataset and may not fully reflect the model's performance on entirely new data.

To assess the generalization capability of the seven models in predicting anthocyanin extraction yields from different berries under varied processing conditions and solvents, eight optimized anthocyanin extraction datapoints, previously reported in the literature and excluded from the initial dataset, were selected (Pires et al., 2024). The systems included grape, raspberry, cranberry, blueberry, strawberry, chokeberry, and black rice. These systems involved DES composed of choline chloride combined with carboxylic acids, alcohols, or sugars (Figure 6). Experimental conditions varied widely, with temperatures ranging from 30 to 80 °C, solid/liquid ratios of 0.02–0.1 $g_{biomass}.mL_{solvent}^{-1}$, and extraction time between 12 and 65 min, representing a diverse set of scenarios.

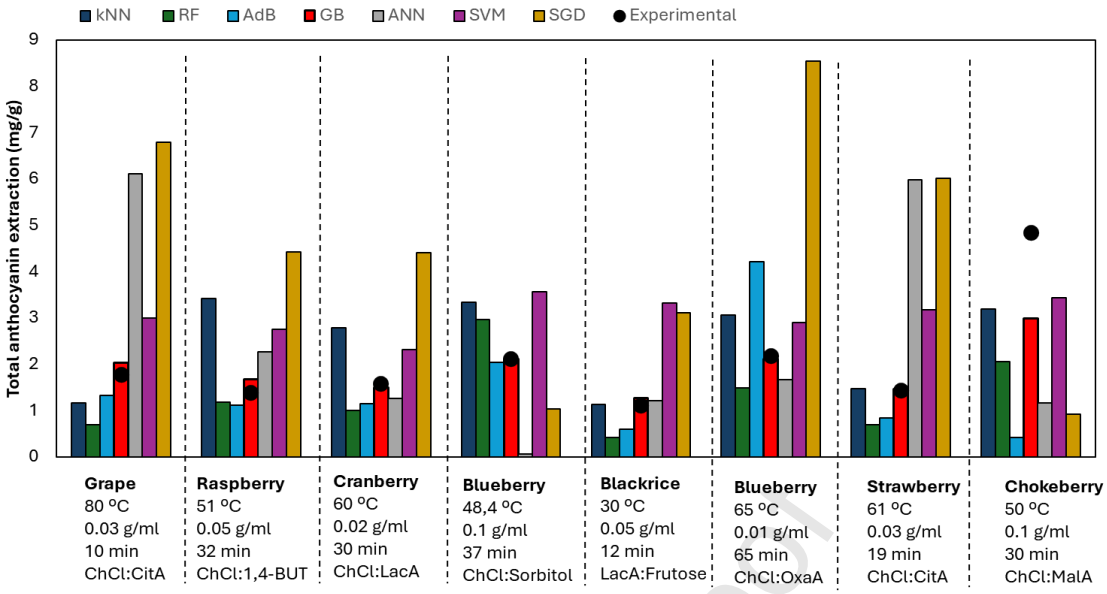


Figure 6. Experimental versus predicted total anthocyanin content (mg g^{-1}) using different machine learning models (kNN, RF, AdaBoost, GB, ANN, SVM, and SGD) for extractions with DESs under various conditions reported in the literature and not included in the training dataset. ChCl denotes choline chloride; 1,4-BUT, 1,4-butanediol; LacA, lactic acid; OxaA, oxalic acid; and MalA, malic acid.

As previously observed, analysis of model performance in predicting anthocyanin extraction revealed that GB not only achieved the highest R^2, MSE, and MAE values but also exhibited a very strong Spearman correlation (r_s = 0.96), indicating that its predictions accurately preserve the rank order of experimental responses (Table 2). Models such as RF (r_s = 0.86) and kNN (r_s = 0.87), while showing relatively lower R^2, maintain good monotonic correlation with observed data, suggesting that they can still identify conditions yielding higher or lower extraction, even if the magnitude of predictions is imprecise. In contrast, linear or low-complexity algorithms such as SGD (r_s = −0.48), SVM (r_s = 0.80), and AdB (r_s = 0.71) demonstrate limited ability to capture

overall trends, whereas the ANN (r_s = 0.30) fails to even learn the relative ordering of responses.

Table 2. Coefficient of determination (R^2), Spearman correlation (r_s) Mean square error (MSE), Mean Absolute Error (MAE) and Root mean square error (RMSE) for different ML algorithms used to predict TAC.

Model	R^2	r_s	MSE	MAE	RMSE
GB	0.88	0.96	0.08	0.06	0.27
RF	0.30	0.86	0.24	0.16	0.49
ANN	0.08	0.30	1.22	0.34	1.10
SGD	0.20	-0.48	2.60	0.60	1.61
AdB	0.01	0.71	0.52	0.18	0.72
SVM	0.15	0.80	0.34	0.23	0.59
kNN	0.19	0.87	0.23	0.16	0.48

These results confirm that GB is the most suitable model, providing predictions that are both quantitatively accurate and consistent in ranking. Notably, GB accurately predicted anthocyanin extraction from black rice (Figure 6), indicating that the model learned generalizable patterns of the extraction process beyond berries. This outcome highlights the robustness and extrapolative capacity of GB, suggesting that it can be reliably applied to other anthocyanin sources, even when the chemical matrix differs substantially from the training set.

3.5. *GB screening prediction capability*

Once trained and validated to predict TAC yields from DES-based extractions across varying operating conditions and berry types, the GB model was applied to a real-world case study to evaluate its practical robustness and generalization capacity. Four berry species (strawberry, blueberry, raspberry, and blackberry) were selected for this

external validation. Critically, these samples originated from geographical regions and harvest batches entirely distinct from those in the training dataset (Table S1), ensuring no overlap in source material. This deliberate design introduces natural variability in key intrinsic properties, including anthocyanin profiles, concentrations, cell wall architecture, and co-extracted matrix constituents, all of which are known to influence extraction efficiency.

Despite these differences and without any model retraining or recalibration, the GB framework delivered highly consistent predictions, with minimal deviation between observed and estimated TAC values (Table 3, Figure 7). The model's strong performance under such partially extrapolative conditions confirms that it captures the underlying mechanisms governing anthocyanin solubilization and mass transfer in aqueous DES systems, rather than overfitting to idiosyncrasies of the original data. This outcome underscores the model's generalizability and practical relevance for predicting extraction yields from novel or uncharacterized berry biomass, supporting its integration into sustainable process design and solvent screening workflows. Initially, ethanol extractions were performed on all selected berries, as the anthocyanin yield obtained with ethanol constitutes one of the required model inputs. To isolate the effect of the solvent, extraction conditions were fixed (Table S3), enabling a direct comparison of solvent performance. Subsequently, a screening of new DES was conducted by combining 220 different HBDs with choline chloride as HBA at a 1:2 molar ratio. This strategy was motivated by the strong influence of the HBD nature on anthocyanin extraction yields, as previously identified through SHAP analysis and feature importance classification.

The selection of HBDs was based on those reported in the DES database compiled by (Odegova et al., 2024). Choline chloride was chosen as the HBA due to its well-established suitability for food-related applications, supported by its low toxicity, biodegradability, strong hydrogen-bonding capacity, and favorable interactions with polar phenolic compounds such as anthocyanins (European Commission, 2026). From the predicted dataset, eight choline chloride–based DES, encompassing carboxylic acids, alcohols, and sugars, were randomly selected, together with pure water as a reference solvent. Importantly, this selection intentionally included both DES predicted to exhibit high extraction yields and those associated with poor performance, allowing for a more comprehensive validation of the model beyond favorable scenarios.

The predictive performance of the GB model for each berry type is summarized in Table 3 and Figure 7. Across all tested berries, the model achieved high Spearman rank-order correlation coefficients (r_s = 0.92–0.99), indicating a consistent ability to capture the relative influence of solvent composition on anthocyanin extraction performance, meaning it correctly ranks solvents from low to high efficiency. Notably, blueberry exhibited a marked discrepancy between metrics: while the coefficient of determination (R^2) was relatively low (0.31), the rank correlation remained very high (r_s = 0.95). This pattern suggests that the model reliably predicts the ordering of solvent effectiveness for blueberry but does not accurately estimate the absolute anthocyanin yields, as observed for the other samples. Thus, we hypothesize that this behavior stems from fundamental differences in anthocyanin chemistry. Blueberry anthocyanins are predominantly derived from delphinidin, malvidin, and peonidin aglycones with greater polarity compared to the pelargonidin- and cyanidin-based anthocyanins dominant in strawberry, raspberry, and blackberry (Mesquita et al., 2023). These structural

differences influence solubility, stability, and interactions with DES, particularly in aqueous systems, introducing nonlinear and matrix-specific effects that are more challenging to model in absolute terms. However, because the *relative* solvation trends across solvents remain consistent, the model retains strong ranking capability even when absolute yield prediction is less precise. This observation underscores an important and essential distinction in model utility: for applications prioritizing solvent screening or comparative performance (e.g., green solvent selection), rank-based accuracy may be more relevant than absolute yield prediction—especially when dealing with chemically complex or structurally distinct biomass.

Table 3. Coefficient of determination **(**R^2), Spearman correlation (r_s) Mean square error (MSE), Mean Absolute Error (MAE) and Root mean square error (RMSE) for different berries using the GB model.

Berry	R^2	r_s	MSE	MAE	RMSE
Strawberry	0.95	0.99	0.01	0.02	0.09
Blueberry	0.31	0.95	0.09	0.07	0.30
Raspberry	0.64	0.99	0.02	0.03	0.13
Blackberry	0.62	0.92	0.14	0.10	0.37

In contrast, strawberry shows both high R^2 (0.95) and r_s (0.99), indicating strong agreement between predicted and experimental values in both magnitude and ordering. Raspberry and blackberry exhibit intermediate behavior, with moderate R^2 values (0.64 and 0.62, respectively) but still very high r_s values (≥0.92). This trend reinforces the interpretation that the GB model is particularly robust in capturing relative solvent performance across different berry matrices, even when absolute prediction accuracy varies.

Consistent with these observations, despite a reduction in predictive accuracy for blueberry, the model remains robust, particularly in capturing relative solvent performance. This limitation is likely associated with the higher chemical complexity and distinct anthocyanin profile of this matrix, which is not explicitly described in the model inputs, as the model was intentionally designed to rely on simplified descriptors to enhance general applicability and reduce data requirements. More broadly, the predictive reliability of the model is stronger for comparative and ranking purposes than for absolute yield estimation in chemically distinct systems. Nevertheless, the consistently high rank correlation across all cases confirms that the model provides reliable guidance for solvent screening, even when applied to complex or extrapolative scenarios.

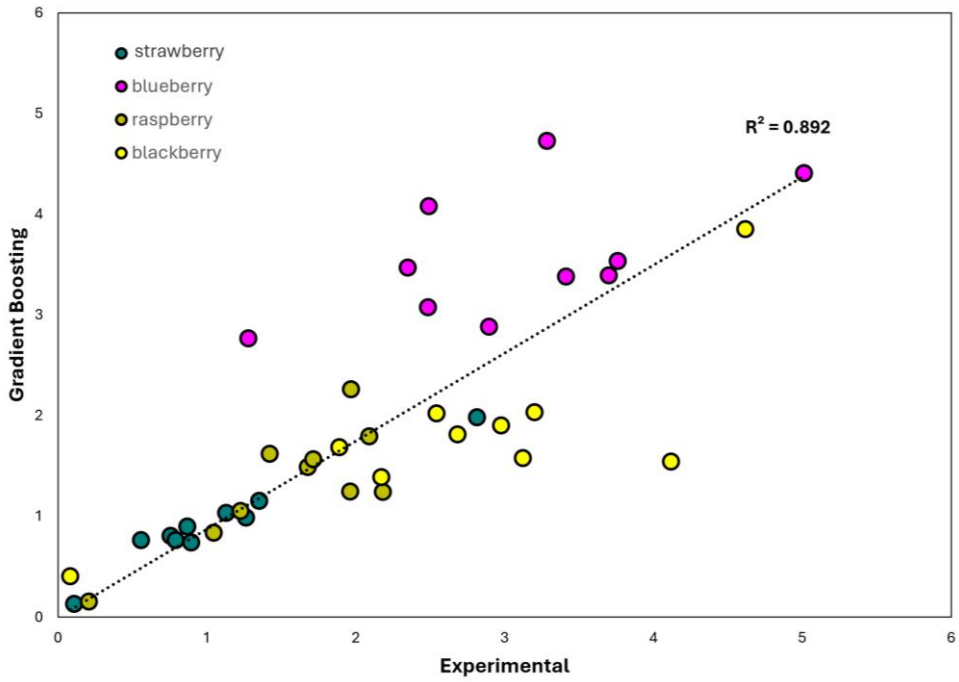


Figure 7. Comparison between experimental and GB-predicted total anthocyanin content yields (mg g^{-1}) for strawberry (Teal), blueberry (Magenta), raspberry (Olive Green), blackberry (Yellow), using selected deep eutectic solvents (DES), see individual dataset at Figure S7.

4. Conclusions

This work demonstrates that a hybrid framework combining COSMO-RS–derived thermodynamic descriptors, key process variables, and a minimal experimental biomass descriptor enables robust, predictive modelling of anthocyanin extraction across chemically diverse berry matrices. Critically, the model was validated on biomass samples that differ not only in species but also in intrinsic chemical composition, reflecting variations in anthocyanin profile, cell wall architecture, and co-matrix constituents driven by botanical origin, cultivar, and edaphoclimatic conditions. Despite this heterogeneity, the approach achieves high predictive accuracy, underscoring its robustness in real-world scenarios where biomass is inherently variable. The total anthocyanin content extracted with ethanol (TAC-ethanol) served as a highly effective, single experimental descriptor of biomass matrix nature. This allows the model to generalize across distinct berry types without requiring detailed biochemical characterization of each sample.

Among the machine learning algorithms tested, Gradient Boosting delivered the best performance, with consistently high Spearman rank correlations (r_s = 0.92–0.99) and strong coefficients of determination. Feature importance analysis confirmed that biomass-related factors dominate model predictions, while solvent behavior is primarily modulated by hydrogen bond donor properties, as captured by COSMO-RS. Together, these results establish COSMO-RS + ML as a transferable, physics-informed platform for rapid solvent screening and process optimization in sustainable biorefining, capable of handling the inherent chemical diversity of natural biomass while minimizing experimental burden.

CRediT authorship contribution statement

Leonardo M. de Souza Mesquita: Conceptualization, Methodology, Investigation, Data curation, Writing–original draft. **João A. P. Coutinho**: Writing–review & editing, Supervision, Funding acquisition. **Filipe H. B. Sosa:** Conceptualization, Methodology, Investigation, Data curation, Writing–original draft, Visualization, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was developed within the scope of the project CICECO Aveiro Institute of Materials, UID/50011/2025 (DOI 10.54499/UID/50011/2025) & LA/P/0006/2020 (DOI 10.54499/LA/P/0006/2020), financed by national funds through the FCT/MCTES (PIDDAC)a and within the scope of the Young Researcher Project supported by the São Paulo Research Foundation (FAPESP) under grant number 2023/16744-0 and fellowship 2025/01561-3. Filipe H. B. Sosa acknowledges FCT – Fundação para a Ciência e a Tecnologia, I.P., for the researcher contracts CEECIND/07209/2022 under the Scientific Employment Stimulus - Individual Call 2022.

Appendix A. Supplementary data

Supplementary data to this article can be found online at XXX. **Table S1.** Literature experimental data - Literature experimental data used to develop the dataset for anthocyanin extraction, **Table S2. List of parameters used in each machine learning algorithm** - List of parameters and hyperparameters used in each machine learning algorithm, **Table S3. Extraction Conditions** - Experimental extraction conditions, including operational variables and system composition, **Figure S1. Workflow of machine learning framework for anthocyanin extraction prediction** - Workflow of the machine learning framework developed for anthocyanin extraction prediction, **Figure S2 -** Frequency distribution of anthocyanin yield values, **Frequency distribution of anthocyanin yield values**, **Figure S3. Frequency distribution of temperature values** - Frequency distribution of temperature values across different biomass types**, Figure S4. Frequency distribution of solid liquid ratios values** - Frequency distribution of solid-liquid ratio values across different biomass types, **Figure S5. Frequency distribution of time values** - Frequency distribution of extraction time values across different biomass types, **Figure S6. Feature importance based on decrease in R^2 for the Gradient Boosting (GB) model using a reduced dataset excluding systems where choline chloride (ChCl) is used as the HBA (87 data points)** - Feature importance based on the decrease in R^2 for the Gradient Boosting (GB) model using a reduced dataset excluding systems with choline chloride as HBA, **Figure S7. Experimental versus predicted Total anthocyanin extraction (mg/g) using GB model in DES non included in the training dataset** - Comparison between experimental and predicted total anthocyanin extraction (mg/g) using the GB model for DES not included in the training dataset.

Data availability

Data will be made available on request.

References

Abbott, A. P., Capper, G., Davies, D. L., Rasheed, R. K., & Tambyrajah, V. (2003). Novel solvent properties of choline chloride urea mixtures. *Chemical Communications*, (1), 70–71.

Anantharaj, R., & Banerjee, T. (2010). COSMO-RS-based screening of ionic liquids as green solvents in denitrification studies. *Industrial & Engineering Chemistry Research*, *49*(18), 8705–8725.

Chemat, F., Vian, M. A., & Cravotto, G. (2012). Green extraction of natural products: concept and principles. *International Journal of Molecular Sciences*, *13*(7), 8615–8627.

de Souza Mesquita, L. M., Contieri, L. S., e Silva, F. A., Bagini, R. H., Bragagnolo, F. S., Strieder, M. M., Sosa, F. H. B., Schaeffer, N., Freire, M. G., Ventura, S. P. M., Coutinho, J. A. P., Rostagno, M. A. (2024). Path2Green: introducing 12 green extraction principles and a novel metric for assessing sustainability in biomass valorization. *Green Chemistry*, *26*(19), 10087-10106.

de Souza Mesquita, L. M., Contieri, L. S., Sosa, F. H. B., Pizani, R. S., Chaves, J., Viganó, J., Ventura, S. P. M., & Rostagno, M. A. (2023). Combining eutectic solvents and pressurized liquid extraction coupled in-line with solid-phase extraction to recover, purify and stabilize anthocyanins from Brazilian berry waste. *Green Chemistry*, *25*(5), 1884–1897.

de Souza Mesquita, L. M., Viganó, J., Veggi, P., Contieri, L. S., Sosa, F. H. B., Vera de Rosso, V., Ventura, S. P. M., & Rostagno, M. A. (2024). Techno-Economic analysis of an

efficient anthocyanin extraction process from grape pomace using eutectic solvents – A critical panorama regarding drying techniques and reusability of solvents. *Separation and Purification Technology*, *347, 127647*.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., et al. (2013). Orange: data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

Eckert, F., & Klamt, A. (2002). Fast solvent screening via quantum chemistry: COSMO-RS Approach. *AIChE Journal*, *48*(2), 369–385.

European Commission. (2026). *Food and Feed Information Portal Database – Feed Additives Details (POL-FEED-IMPORT-1245)*. Available at: https://ec.europa.eu/food/food-feed-portal/screen/feed-additives/search/details/POL-FEED-IMPORT-1245

Giusti, M. M., & Wrolstad, R. E. (2001). Characterization and measurement of anthocyanins by uv-visible spectroscopy. *Current Protocols in Food Analytical Chemistry*, 1, F1.2.1-F1.2.13.

Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). *The elements of statistical learning*. Springer, New-York.

Hizaddin, H. F., Wazeer, I., Huzaimi, N. A. M., El Blidi, L., Hashim, M. A., Lévêque, J.-M., & Hadj-Kali, M. K. (2022). Extraction of phenolic compound from model pyrolysis oil using deep eutectic solvents: computational screening and experimental validation. *Separations*, *9*(11), 336.

Jun, M.-J. (2021). A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area. *International Journal of Geographical Information*

Science, *35*(11), 2149–2167.

Lorenzo-Llanes, J., Palomar, J., Escalona, N., & Canales, R. I. (2025). COSMO-RS-based solvent screening and experimental analysis for recovering added-value chemicals from the bio-oil aqueous phase. *Separation and Purification Technology*, *369*, 133104.

MacLean, A. M. G., Silva, Y. P. A., Jiao, G., & Brooks, M. S. (2021). Ultrasound-assisted extraction of anthocyanins from Haskap (Lonicera caerulea L.) berries using a deep eutectic solvent (des) des extraction of anthocyanins from Haskap berries. *Food Technology and Biotechnology*, *59*(1), 56–62.

Mesquita, L. M. D. S., Contieri, L. S., Sanches, V. L., Kamikawachi, R., Sosa, F. H. B., Vilegas, W., Rostagno, M. A. (2023). Fast and green universal method to analyze and quantify anthocyanins in natural products by UPLC-PDA. *Food Chemistry*, *428*(July), 136814.

Mohan, M., Demerdash, O. N., Simmons, B. A., Singh, S., Kidder, M. K., & Smith, J. C. (2024). Physics-Based Machine Learning Models Predict Carbon Dioxide Solubility in Chemically Reactive Deep Eutectic Solvents. *ACS Omega*, *9*(17), 19548–19559.

Nour, V., Stampar, F., Veberic, R., & Jakopic, J. (2013). Anthocyanins profile, total phenolics and antioxidant activity of black currant ethanolic extracts as influenced by genotype and ethanol concentration. *Food Chemistry*, *141*(2), 961–966.

Odegova, V., Lavrinenko, A., Rakhmanov, T., Sysuev, G., Dmitrenko, A., & Vinogradov, V. (2024). DESignSolvents: an open platform for the search and prediction of the physicochemical properties of deep eutectic solvents. *Green Chemistry*, *26*(7), 3958–3967.

Oliveira, G., Farias, F. O., Sosa, F. H. B., Mafra, M. R. (2021). Green solvents to tune the

biomolecules' solubilization in aqueous media: an experimental and in silico approach by COSMO-RS. *Journal of Molecular Liquids*, 117314.

Omar, K. A., & Sadeghi, R. (2022). Physicochemical properties of deep eutectic solvents: A review. *Journal of Molecular Liquids*, *360*, 119524.

Omwango, E., Onguso, J., Ochora, J., Kirira, P., Kinyua, Z., & Mandela, E. (2024). Phytochemical variability of selected medicinal plants from different agro-climatic zones in Kenya. *Biochemical Systematics and Ecology*, *117*, 104915.

Paduszyński, K. (2017). An overview of the performance of the COSMO-RS approach in predicting the activity coefficients of molecular solutes in ionic liquids and derived properties at infinite dilution. *Phys. Chem. Chem. Phys.*, *19*(19), 11835–11850.

Pires, I. V., da Silva, L. H. M., Rodrigues, A. M. da C., & Saldaña, M. D. A. (2024). Natural deep eutectic solvents for anthocyanin extraction from agricultural sources: Process parameters, economic and environmental analysis, and industrial challenges. *Comprehensive Reviews in Food Science and Food Safety*, *23*(6), e70057.

Pontes, P. V. A., Ferreira, A. M., Coutinho, J. A. P., Andreaus, J., Costa Lopes, A. M., & Sosa, F. H. B. (2025). COSMO-RS assisted selection of eutectic solvents for lignin dissolution and enhanced laccase activity. *International Journal of Biological Macromolecules*, *320*, 146048.

Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*, 33–45.

Roadknight, C. M., Balls, G. R., Mills, G. E., & Palmer-Brown, D. (1997). Modeling complex environmental data. *IEEE Transactions on Neural Networks*, *8*(4), 852–862.

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models

using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*(10), 1013–1026.

Santiago, R., Bordón Sosa, F. H., Díaz, I., González-Miquel, M., & Pereira Coutinho, J. A. (2023). Predicting Partition Coefficients in Organic Biphasic Systems Using COSMO-RS. *Industrial & Engineering Chemistry Research*, *62*(43), 17905–17913.

Sicaire, A.-G., Filly, A., Vian, M., Fabiano-Tixier, A.-S., & Chemat, F. (2018). Cosmo-RS-Assisted Solvent Screening for Green Extraction of Natural Products. In *Handbook of Green Chemistry* (pp. 117–138). John Wiley & Sons, Ltd.

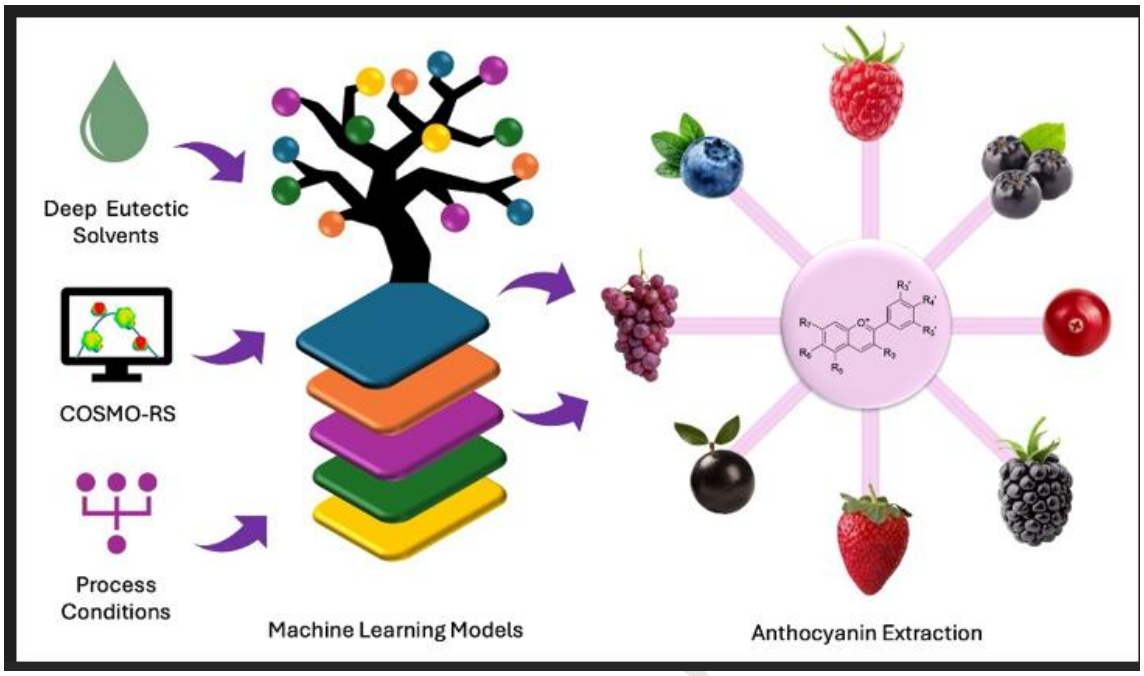
Sosa, F. H. B., Kilpeläinen, I., Rocha, J., & Coutinho, J. A. P. (2023). Recovery of superbase ionic liquid using aqueous two-phase systems. *Fluid Phase Equilibria*, *573*(June), 113857.

Vittor, L., Duarte, T., Bel, S., & Tavares, F. W. (2024). *Assessing Viscosity in Sustainable Deep Eutectic Solvents and Cosolvent Mixtures : An Artificial Neural Network-Based Molecular Approach*. *ACS Sustainable Chemistry & Engineering, 12*(21), 7987–8000.

Yang, B., Zheng, J., & Kallio, H. (2011). Influence of origin, harvesting time and weather conditions on content of inositols and methylinositols in sea buckthorn (Hippophaë rhamnoides) berries. *Food Chemistry*, *125*(2), 388–396.

Yang, Y., & Kilmartin, P. A. (2025). Advancing anthocyanin extraction: Optimising solvent, preservation, and microwave techniques for enhanced recovery from merlot grape marc. *Food Chemistry*, *472*, 142648.

Graphical abstract



Highlights

- Hybrid COSMO-RS and machine learning framework predicts anthocyanin yields from chemically diverse berry matrices.
- Gradient Boosting algorithm achieved highest predictive accuracy ($R^2 = 0.92$) across 15 berry types.
- Solvent–solute interactions captured via COSMO-RS hydrogen-bonding descriptors drive extraction efficiency.
- Approach allows rapid, data-driven optimization of sustainable anthocyanin extraction protocols