

Lawrence Berkeley National Laboratory

LBL Publications

Title

Machine Learning-Driven Solvent Screening for Biobased 2,3-Butanediol Extraction

Permalink

<https://escholarship.org/uc/item/7n7505wm>

Journal

Industrial & Engineering Chemistry Research, 64(32)

ISSN

0888-5885

Authors

Edaugal, Justin P

Zhang, Difan

Lewis, Tyrell SA

et al.

Publication Date

2025-08-13

DOI

10.1021/acs.iecr.5c01569

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Machine Learning-Driven Solvent Screening for Bio-based 2,3-Butanediol Extraction

Justin P. Edaugal^{1, 2}, Difan Zhang^{*3, 4}, Tyrell Lewis^{1, 2}, Dupeng Liu^{*1, 2}, Vassiliki-Alexandra Glezakou⁴, Ning Sun^{*1, 2},

¹Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory, Emeryville, CA, United States.

²Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States.

³Physical Sciences Division, Pacific Northwest National Laboratory, Richland, WA, United States.

⁴Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States.

**Corresponding author:*

zhangd2@ornl.gov

pengduli@lbl.gov

nsun@lbl.gov

ABSTRACT:

Bio-based 2,3-butanediol (2,3-BDO) is a valuable biomass-derived chemical due to its versatility in being transformed into a wide variety of products. However, the separation and purification of 2,3-BDO from fermentation broth remains a significant challenge, owing to its high boiling point and hydrophilic nature. Herein, we developed a machine learning (ML)-based screening workflow that uses molecular calculations as training data and requires only a small number of experimental measurements for validation, to identify alternative solvent candidates for the liquid-liquid extraction (LLE) of 2,3-BDO from aqueous solution. In particular, 130 density functional theory (DFT) calculations with implicit solvation method not only built a correlation between computational partition coefficient and experimental distribution coefficient of 2,3-BDO but also parameterized an Extra-Trees ML model to screen the distribution coefficient for a wider range of 6717 organic solvents. The experimental measurements of merely 24 solvents were needed to validate the computational results. A list of 50 prioritized solvents were proposed for 2,3-BDO LLE and 7 additional experimental measurements were conducted to further verify our selected solvents. The impact of extraction temperature and organic-to-aqueous ratio was also investigated for selected solvents in experiments. This work suggested alternative solvents for 2,3-BDO LLE and proposed a versatile workflow that requires less experiments and can be applicable to a broader range of LLE studies.

KEYWORDS: 2,3-Butanediol, Liquid-Liquid Extraction, Machine Learning, Solvent Screening.

1. INTRODUCTION

Bio-based 2,3-butanediol (2,3-BDO) is gaining tremendous attention as a sustainable platform chemical across the food, agriculture, chemical, and pharmaceutical industries. 2,3-BDO can be readily converted into a variety of chemical intermediates, such as 1,3-butadiene, diacetyl, and methyl ethyl ketone, which are essential for the production of diverse products, including antifreeze agents, polymers, cosmetics, pharmaceuticals, and aviation fuels¹. In contrast to traditional chemical synthesis processes that rely on petroleum-based sources, the microbial fermentation production of 2,3-BDO utilizes renewable biomass feedstocks, offering a more cost-effective and environmentally sustainable alternative². Despite these potential advantages, the industrial-scale bioproduction of 2,3-BDO remains limited due to significant challenges in its downstream processing. Given its high boiling point (177-182°C), high viscosity, and strong affinity for water, the separation of 2,3-BDO often requires the use of evaporation and multi-stage distillation operations, which are highly energy-intensive and costly³. Moreover, the low concentration of 2,3-BDO in fermentation broth (5-10 wt%) can complicate selective recovery, especially in the presence of impurities, such as organic acids, monosaccharides, and salts⁴. These challenges have driven the search for more efficient, environmentally friendly, and cost-effective methods for the separation of bio-based 2,3-BDO.

Liquid-liquid extraction (LLE), also known as solvent extraction, is a well-established method for the efficient separation of various molecules and compounds.^{5, 6} This technique employs an organic solvent to extract target compounds from an aqueous phase, leveraging the partitioning behavior of compounds between the two phases. The extraction yield, however, is heavily dependent on several factors, such as the solvent type, solubility of the target component, phase density compatibility, extraction temperature, and solvent-to-solute ratio^{4, 7, 8}. Numerous studies have explored solvent extraction for the separation of bio-based 2,3-BDO from aqueous fermentation broth⁹⁻¹². Although these studies have reported high extraction yields, many of the solvent systems necessitate the addition of salts, sugars, polymers, or other additives, which can

complicate subsequent refining operations, thus increasing raw material and process costs¹³. Furthermore, most of these studies rely on conventional experimental methodologies for screening solvent candidates, which limits the number of distinct solvents and operating conditions tested while also increasing the time and effort required for experimentation.

Recent advances in artificial intelligence methods, such as machine learning (ML), deep learning neural networks, large language models, quantitative structure-property relationship, and hybrid ML-thermodynamic models, have revolutionized separation research and led to innovations such as computer-aided molecular design of novel green solvents and equipment operation/parameter selection for downstream process design and optimization¹⁴. Several research groups have successfully applied ML methods to optimize LLE processes, such as the extraction of pharmaceutical pollutants from wastewater¹⁵, ligand-assisted extraction of rare-earth elements¹⁶, and deep eutectic solvent extraction of furfural compounds¹⁷. Despite their potential, the predictive accuracy of ML algorithms critically depends on the quantity and quality of the input data, which may restrict their application for more complex LLE processes¹⁸. On the other hand, due to the lack of vast LLE experimental data of 2,3-BDO, the study of optimal solvents for 2,3-BDO extraction by artificial intelligence approaches has not been extensively achieved. The high-throughput experimentation to rapidly generate more comprehensive experimental datasets¹⁹⁻²¹ may fill the gap of 2,3-BDO data, but currently it is still unavailable for 2,3-BDO. Therefore, a solvent screening approach for 2,3-BDO extraction that requires only a small amount of experimental data is desirable.

Herein, we describe the development of a ML-driven workflow to perform a solvent screening for the LLE of 2,3-BDO from aqueous solutions. Specifically, a computational database of the partition coefficient of 2,3-BDO between organic solvent and aqueous phase was established based on density functional theory (DFT) calculations using implicit solvation approximation. This database was used to parameterize a ML model to evaluate a wider range of solvents. Experimental measurements of distribution coefficients for selected solvents were

performed to correlate the computational partition coefficient and experimental distribution coefficient, and to validate predictions of ML model. Then, the ML model performed a solvent screening for 2,3-BDO extraction distribution coefficient, and a list of solvent candidates was proposed and further verified by additional experimental measurements. In addition to our DFT-informed ML approach, there are alternative methods that can be applied for solvent screening. The readers can refer to several review articles for more details. Finally, selected solvent candidates were further used to investigate the impact of extraction temperature and organic-to-aqueous ratio on the extraction efficiency of 2,3-BDO. This study demonstrates an ML-driven, closed-loop screening workflow as an effective approach to optimizing the LLE of bio-based 2,3-BDO. It is also applicable to explore the promising solvents for a wider range of LLE studies.

2. MATERIALS AND METHODS

2.1. Materials

All reagents and solvents and were obtained from commercial suppliers and used without further purification. Specifically, 2,3-BDO with a purity of 98% and oleyl alcohol with a purity of 85-87% were acquired from Sigma Aldrich. All other solvents with purities ranging from 98-99.5% were sourced from Sigma Aldrich and Fisher Scientific.

2.2. Liquid-Liquid Extraction Tests

A stock solution of 2,3-BDO at 10 wt% concentration was prepared by dissolving the appropriate amount of 2,3-BDO in Millipore deionized water within a volumetric flask. For the LLE tests, mixtures comprising 2 mL of the 10 wt% 2,3-BDO solution and 2 mL of an organic solvent (1:1 v/v organic-to-aqueous phase ratio) were prepared in 8 mL glass vials, which were sealed to reduce evaporation losses. Each mixture was subjected to vortex mixing for 5 s, followed by sonication for 1 h at a controlled temperature (20°C, 40°C, or 60°C $\pm$ 3°C). After sonication, the mixtures were allowed to rest undisturbed at the specified temperature for 1.5 h to achieve phase

separation. For analytical measurement, 0.5 mL aliquots were collected from each phase of the extraction mixtures and transferred to 2 mL glass vials. All experiments were conducted in duplicate, with each sample measured twice to ensure the reproducibility of the analytical results. Standard deviations were calculated from the mean values of these measurements.

2.3. Analytical Methods

The concentration of 2,3-BDO in the aqueous phase of the extraction mixtures was determined using high-performance liquid chromatography (HPLC). The analysis was conducted using an Aminex HPX-87H column at a set temperature of 60°C, using 4 mM sulfuric acid at 0.6 mL/min. Calibration of the HPLC system was achieved by preparing a dilution series of 2,3-BDO water standards, as shown in **Figure S1**. To ensure the consistency of the HPLC measurements, a single 2,3-BDO water standard was used as an external check standard reference between all sample analyses. Additionally, a blank injection was performed after every third sample injection to purge the system of potential 2,3-BDO accumulation. To validate the accuracy of the calibration method, an alternative series of 2,3-BDO standards was prepared gravimetrically by dissolving 2,3-BDO in 1-hexanol at varying concentrations, followed by dilution with ethanol in a 1:9 v/v ratio, as shown in **Figure S2**. Upon comparing the accuracy of 2,3-BDO concentration detection between **Figures S1** and **S2**, the 2,3-BDO water calibration curve was determined to be the optimal method for 2,3-BDO measurements.

2.4. Experimental Distribution Coefficient Calculations

The extraction performance of each solvent was evaluated by calculating two key parameters: the Distribution Coefficient (K_d) and Extraction Efficiency (EE). The K_d is defined as the ratio of the mass fraction of 2,3-BDO in the organic phase (w_{org}) to the mass fraction of 2,3-BDO in the aqueous phase (w_{aq}), as shown in **Equation 1**. The w_{org} is determined from the mass of 2,3-BDO in the organic phase upon reaching equilibrium (m_{org}), which is calculated as the difference

between the initial mass of 2,3-BDO in the aqueous feedstock solution ($m_{aq\ feed}$) and the final mass of 2,3-BDO in the aqueous phase post-extraction ($m_{aq\ final}$), as detailed in **Equations 2-3**.

$$K_d = \frac{w_{org}}{w_{aq}} \quad (1)$$

$$w_{org} = \frac{m_{org}}{m_{org\ total}} \quad (2)$$

$$m_{org} = m_{aq\ feed} - m_{aq\ final} \quad (3)$$

The *EE* is defined as the mass of 2,3-BDO in the organic phase upon reaching equilibrium (m_{org}) to the initial mass of 2,3-BDO in the aqueous feedstock solution ($m_{aq\ feed}$), as expressed by **Equation 4**. To convert the HPLC-detected concentration of 2,3-BDO to mass, the volumes of both the aqueous and organic phases were measured before and after phase separation, with the masses of each phase calculated based on the known phase densities.

$$EE = \frac{m_{org}}{m_{aq\ feed}} \times 100\% \quad (4)$$

2.5. Density Functional Theory Calculations

The density functional theory (DFT) calculations were performed to compute the partition coefficient of 2,3-BDO between organic and aqueous solvents. All DFT calculations were carried out with the ORCA software²² using B3LYP functional^{23, 24} and Grimme-D3²⁵ dispersion correction. The def2-TZVP basis set²⁶ were used for all the atoms and SMD implicit solvent model²⁷ as used to approximate the solvation environment. Only organic solvents containing carbon, oxygen, nitrogen or hydrogen atoms were considered in this work, leading to 130 organic solvents in total from the original SMD model. For every solvent, a single 2,3-BDO molecule was first geometrically optimized using BFGS algorithm implemented in ORCA, followed by the vibrational frequency calculation in ORCA to estimate the entropy (TS) and enthalpy (H) terms for Gibbs free energy (H - TS) evaluation. Readers can refer to ORCA's user manual for detail. A single calculation

typically costs 12 minutes using a 64-core CPU. The partition coefficient ($P_{\text{organic/water}}$) of 2,3-BDO was computed by **Equation 5** below, where $\Delta G_{\text{organic}}^o$ and $\Delta G_{\text{water}}^o$ indicate the Gibb's free energy of 2,3-BDO solvation in the organic solvent and aqueous phase, respectively, at the standard condition (25°C, 1 atm), T was temperature, and R was the gas constant. For simplicity, we labeled ΔG as the difference of the two Gibb's free energy of 2,3-BDO solvation in organic and aqueous phase (i.e. $\Delta G = \Delta G_{\text{organic}}^o - \Delta G_{\text{water}}^o$).

$$\log P_{\text{organic/water}} = \frac{-(\Delta G_{\text{organic}}^o - \Delta G_{\text{water}}^o)}{2.303RT} \quad (5)$$

2.6. Machine Learning modeling

The earlier work of Smith et al focused on 2,3-BDO solubility instead of distribution coefficient or separation factor as measured in our experiments. Therefore, in this work, we aim to develop a ML model that can directly evaluate the K_d values of 2,3-BDO between solvent and aqueous phase, and our experimental results can be numerically compared to our ML predictions. The computed ΔG values for 130 organic solvents were used to parameterize ML models to evaluate the ΔG of new organic solvents, which can be converted to logP. The logP values were further converted to K_d values by fitting a linear correlation to experimental results. Due to the small size of our data, the ΔG dataset was randomly split into training and testing sets by the ratio 60%:40% only during the fast examination of ML algorithms. 38 different ML algorithms were first examined to train predictive models to estimate ΔG values of DFT calculations, and their detail is provided in the supporting material **Table S1**. The ML algorithm showing potentially the highest accuracy was further fine-tuned by a five-fold cross validation through random search, and the prediction results from the fine-tuned model were used to compare to experimental results.

The RDKit²⁸ molecular property descriptor was used to represent the solvent molecules based on their Simplified Molecular Input Line Entry System (SMILES) strings. For every solvent in the dataset, a 131-length feature descriptor was first generated from its SMILES string. The

features showing no variation (i.e. constant value for all data) were removed. Then these molecular features are compared across all solvents to analyze correlations between each other. The features showing high correlation coefficients ($R^2 > 0.75$) were condensed by removing those with larger overlaps. This eventually reduced the descriptor from 131-length to 64-length. The resulting RDKit descriptor was used as the input to ML models for the initial ML algorithm examination after the 60%:40% data split. Once the most promising ML algorithm was identified, these RDKit descriptors of solvents were further used as input for ML hyperparameter tuning using the five-fold cross validation. The fine-tuned model was saved as our final ML model for solvent screening. For a given new solvent, its SMILES string was first collected or generated from literature or public source, and then its 64-length feature descriptor was generated by RDKit from the SMILES string and passed into the final ML model to compute the ΔG values, which was later converted to distribution coefficients and compared to experimental results. The ML training was performed within the framework of Python package Scikit-learn²⁹. Structural similarity analysis was performed using cosine similarity based on the molecular features, and 1-butanol was always used as the reference structure. All scripts and data used in this work is freely available on the Github repository (https://github.com/zhangdf07/BDO_ML).

3. RESULTS AND DISCUSSION

3.1. Experimental Distribution Coefficients of An Initial List of Solvents

We first gathered an initial list of organic solvents that has been suggested in the previous computational work of Smith et al for LLE of 2,3-BDO in the aqueous solution. The authors used open-access organic and aqueous solubility databases and developed ML models to estimate 2,3-BDO solubility of several types of organic solvents. They suggested a list of solvents containing various alcohols, esters, and hydrocarbons, for potential experimental exploration.

Therefore, in this work, we started with these solvents by experimentally measuring their performance in extracting 2,3-BDO from aqueous solution.

To quantify the effectiveness of these initial organic solvents, the experimental distribution coefficients for 24 selected solvents were measured at standard conditions (1:1 v/v organic-to-aqueous ratio, 20°C). As shown in **Figure 1**, these solvents were highlighted in three classes: alcohols, esters, and hydrocarbons. Their experimental K_d values at standard conditions are visualized as a function of their solvent molecular weights. The numeric data of our measurement is provided in **Table S2** and their experimental K_d error bar values are highlighted in **Figure S3**.

Among the three solvent classes tested under standard conditions, alcohols generally displayed the highest K_d values, followed by esters and hydrocarbons. The superior extraction yields of alcohol solvents can be attributed to the presence of hydroxyl groups in their molecular structures, which facilitate strong hydrogen bonding with the hydroxyl groups present in the 2,3-BDO molecules³¹. Among all these solvents under standard conditions, 1-butanol exhibited the highest K_d value of 0.92, indicating its strong affinity for extracting 2,3-BDO from aqueous solution, likely due to the molecular structure of 1-butanol where the hydroxyl group exists on the short four-carbon chain. Furthermore, water molecules can easily distribute to the 1-butanol organic phase, forming polar molecular clusters that can serve as local aqueous environments for the co-extraction of additional 2,3-BDO molecules³².

In addition, the longer carbon-chain backbone of fatty alcohols can lower polarity and reduce the intermolecular interactions with 2,3-BDO molecules in the aqueous phase³¹. As a result, the shorter-chain alcohols such as 1-hexanol, 2-hexanol, and 4-methyl-2-pentanol, display relatively higher K_d values (0.37, 0.35, and 0.30, respectively) than those longer-chain alcohols such as 1-octanol, 1-heptanol, 1-nonanol, and oleyl alcohol (0.22, 0.22, 0.19, 0.14, and 0.07, respectively). As depicted in **Figure 1 (b)**, the alcohol solvents with lower molecular weights (74-

116 g/mol) exhibit elevated K_d values compared to those of higher molecular weights (>116 g/mol). Interestingly, despite previous studies reporting oleyl alcohol as an effective solvent for extracting 2,3-BDO from aqueous fermentation broths^{33, 34}, our repeated measurements show a low K_d value of 0.07 and its long carbon-chain suggests a poor solvent candidate for 2,3-BDO extraction.

A similar trend is observed among ester solvents, where those with shorter carbon-chains and lower molecular weights (88-116 g/mol) generally display higher K_d values than those with longer chains and higher molecular weights (>116 g/mol), as shown in **Figure 1 (c)**. The ethyl acetate and isopropyl acetate exhibit sufficient distribution behavior with K_d values of 0.19 and 0.10, respectively. The other esters such as isobutyl acetate, butyl butyrate, and isopropyl myristate, show low K_d values ranging from 0.05 to 0.01, indicating the reduced hydrogen bonding and dipole-dipole interactions with hydrophilic 2,3-BDO due to the increasing length of the hydrophobic carbon-backbone of these ester molecules.

Not surprisingly, the hydrocarbon solvents exhibit the lowest K_d values of 2,3-BDO among the three solvent classes, typically ranging from 0.03 to 0.005, since they mostly interact with 2,3-BDO via weaker van der Waals forces. As illustrated in **Figure 1 (d)**, the variation of K_d values do not appear to strongly correlate with solvent molecular length (or weight), but more likely be influenced by subtle changes in molecular geometries among linear alkane, cycloalkane, and aromatic solvent structures. Overall, our measurements indicate that alcohol generally outperforms ester and hydrocarbon in the extraction of 2,3-BDO, which agrees with the findings of previous works^{30, 31, 35}.

3.2. Screening of New Solvents via Machine Learning

Although we have obtained the experimental K_d values for 24 solvents, it remains to be insufficient for a dataset of ML development. Moreover, there is currently no other existing database for a

significant amount of experimental K_d values of 2,3-BDO in the literature. As a result, we pursue an alternative pathway where DFT calculations with SMD implicit solvation generate a dataset of computational ΔG of 2,3-BDO between organic solvents and water, and partition coefficient (P) can be easily computed from by ΔG using **Equation 5**. To overcome the limited availability of SMD parameters, the DFT dataset allows for the development of ML models to explore a wider range of solvents. Our 24 experimental K_d values were only used to find and verify the correlation between experimental K_d and computational P results, not as training data.

In the first step, we quantified how DFT-computed P values correlate to the experimentally measured K_d values for 2,3-BDO. Among the 24 solvents in our initial list, only 11 solvents are available in the SMD parameters, and their results are illustrated in **Figure 2(a)**. The experimental K_d and DFT P values demonstrate a clear correlation, which allows us to fit a linear equation to convert a given P value into the K_d value, with a R^2 of 0.946 and a Root Mean Squared Error (RMSE) of 0.082. We assume that such a correlation equation can be extended to other solvents. Therefore, we expect that the experimental K_d value of 2,3-BDO for a given solvent can be estimated by linearly transforming a computational P value from DFT calculation, as illustrated in **Figure 2(b)**. However, as noted previously, the DFT calculations using SMD solvation model is limited by the availability of its fitted parameters, and it is not cost-efficient to develop such parameters for all kinds of solvents.

To overcome this difficulty, we used our DFT calculations as a small database and parameterized ML models to evaluate the computational ΔG , thus computational P, for more solvents. Considering the relatively small size of our training data, we examined the appropriateness of several ML algorithms to identify the potential ML models with the highest accuracy. As shown in **Figure 3(a)**, the testing performance of 38 ML algorithms are compared using the same training/testing data. Their numerical values of R^2 and RMSE are provided in Table **S1**. Overall, the Extra-Trees ensemble algorithm shows the lowest RMSE error and highest

R^2 coefficient. Therefore, we further refined the parameterization of the Extra-Trees model using a five-fold cross-validation via a random search method. The resulting model demonstrates a decent accuracy (RMSE=0.19, R^2 =0.98), as illustrated in **Figure 3(b)**, to evaluate the computational ΔG values in agreement with DFT calculations. The solvents showing the largest predictive errors (ML vs. DFT) are highlighted in **Figure 3(c)**. The propanoic acid and aniline show the largest ΔG errors, but the error is still below 0.55 kcal/mol. Due to the small size of our dataset, it can be challenging to further significantly minimize the prediction error for all types of organic solvents. Since the predictive errors for most solvents are acceptable (<0.38), we believe the accuracy of this model is reasonable to explore other solvents.

Feature importance analysis was performed to understand the most impactful structural features of solvents for 2,3-BDO extraction, as shown in **Figure 3(d)**. We found that three features: the minimum atomic partial charge (MinPartialCharge), the topological polar surface area (TPSA), and the number of hydrogen bond donors (NumHDonors), play more significant roles than the other features. These three features are all influenced by polar segments in the molecule, for instance, hydroxyl groups, which indicate the importance of primarily oxygen and nitrogen, as well as their attached hydrogen atoms.

The fidelity of our developed ML model was further interrogated by comparing the ML prediction to the experimental K_d values for the 13 solvents in the initial list without available SMD parameters. The ML-predicted K_d values for these solvents, provided in **Figure 4(a)**, suggest an excellent agreement with experimental values (R^2 of 0.931 and RMSE of 0.07). The accuracy of the two types of data (solvents included in DFT and solvents not included in DFT) is comparable, suggesting no significance of overfitting in our ML model. Note that these 13 solvents were not used in either the ML training/validation evaluation or in the fitting of the linear equation to convert P to K_d values, demonstrating the predictive performance of the ML model for unseen solvents. Furthermore, as shown in **Figure 4(b)**, the structural similarity between these 13 solvents and

those involved in the DFT dataset indicates that the experimental solvents generally fall within the solvent space defined by the DFT dataset. This is likely the reason for the high predictive accuracy of our model because the interpolation accuracy of ML is far better than its extrapolation capability. Therefore, we believe that our ML model is trustworthy to estimate the K_d values of organic solvents, and it is rational to conduct a solvent screening of 2,3-BDO extraction.

A solvent screening using our ML model was then performed for 6717 organic solvents gathered from the literature and open sources. The full results for all 6717 solvents are provided in our repository. A structural similarity analysis was also performed, as shown in **Figure 4(b)**, for these new solvents. They mostly fall within the space defined by the DFT data, with several exceptions that lies slightly beyond the edge of the DFT-space. Their K_d values were evaluated and ranked as illustrated in **Figure 4(c)**. Since 1-hexanol (predicted $K_d = 0.405$) displays an intermediate performance for 2,3-BDO LLE³⁶, we used its K_d value as a lower cutoff threshold and looked for the solvents with larger K_d . On the other hand, we note that our current screening only considers the partitioning of 2,3-BDO between the organic and aqueous phases, but the solubility of these two phases themselves has been ignored. As a result, several solvents showing extremely high 2,3-BDO K_d values, such as methanol (1.05) and ethanol (0.96), are not feasible in practice because they are highly soluble in water. Therefore, we selected 1-butanol (predicted $K_d = 0.754$), which was successful for 2,3-BDO LLE in our experiments and show the highest K_d value, as an upper cutoff to screen out solvents that are miscible with water. In total, we obtained a second list of 50 solvent candidates, and their results are provided in **Table S3**. We believe these solvents have higher priorities for future experimental explorations.

To verify the accuracy of solvents in our second list, we selected four solvents from the potential 50 candidates and three other solvents with lower K_d values outside the potential candidates. The experimental K_d values of these seven solvents were measured under standard conditions and compared to ML predictions, presented in **Figure 4(d)**. Their raw data is provided

in **Table S4**. The ML prediction demonstrates decent agreements with experimental results ($\Delta K_d < 0.12$) for six solvents, including 2-methyl-1-butanol, 3-methyl-1-butanol, 3-pentanol, 2-ethyl-1-butanol, 3-hexanol and 2-cyclohexylethanol. Only 2-ethyl-1,3-hexanediol shows a larger prediction error ($\Delta K_d = 0.32$). These errors could be from several sources: The computed K_d values were converted from the linear fit of P values, introducing flaw from the linear approximation as well as the uncertainty of using P values. The parameterization of the Extra-Tree model using DFT data also introduced errors in the fitting. Moreover, the relatively small size of the DFT dataset is challenging to cover the entire solvent space, thus less representative to capture solvent molecules beyond the scope of dataset. This is likely the major source of the large error for 2-ethyl-1,3-hexanediol, since diol solvent molecules are not extensively included in the SMD parameters of DFT calculations. In summary, we observed an overall good agreement between the experimental and predicted K_d values for the new seven solvents ($R^2 = 0.913$, RMSE = 0.096) as shown in **Figure S4(a)**. Further reducing the discrepancy between experiment and ML prediction, especially for solvents with complex structures such as cyclic rings or multiple functional groups, will require additional data reinforcement.

3.3. Impact of Extraction Temperature and Organic-to-Aqueous Ratio

Beyond the type of solvents, extraction temperature and organic-to-aqueous ratio are two other important factors that may significantly influence the solvent extraction efficiency, energy usage, and operational costs of an extraction process^{7, 37}. These properties cannot be efficiently evaluated using traditional molecular modeling such as DFT calculations. The lack of extensive experimental 2,3-BDO results for different solvents under various temperatures and organic-to-aqueous ratios has also made the development and validation of ML models challenging. Therefore, we only used our experiments to assess the impact of extraction temperature and organic-to-aqueous ratios on 2,3-BDO extraction in this work.

The experimental K_d values and extraction efficiencies of the solvents from the initial list were measured at three different extraction temperatures (20°C, 40°C, or 60°C), with the complete numeric data provided in **Table S2**. As depicted in **Figure 5(a)**, most solvents, irrespective of solvent class, demonstrated an increase in extraction efficiency with rising temperature. Among all tested solvents, 1-butanol exhibited the highest extraction efficiency of 32.2% at 40°C and 34.0% at 60°C. Although the exact mechanism underlying the impact of temperature on the LLE of 2,3-BDO is beyond the scope of this study, the observed increase in extraction efficiency may be attributed to the increased solvation of 1-butanol in water at elevated temperatures, which could facilitate more efficient mass transfer of 2,3-BDO molecules between the aqueous and 1-butanol phases^{7, 38}. This positive trend is also observed in all other alcohol solvents, shown in **Table S2**. For instance, 1-hexanol exhibited an increase in extraction efficiency from 15.0% at 20°C to 18.5% at 40°C and 19.0% at 60°C. In addition, 2-hexanol showed an increase in 2,3-BDO extraction efficiency from 14.5% at 20°C to 16.7% at 40°C and 18.2% at 60°C. Similarly, all ester solvents demonstrated an increase in 2,3-BDO extraction efficiency with increasing extraction temperature. Ethyl acetate remained the better solvent among the ester class for 2,3-BDO extraction with extraction efficiency values of 9.0%, 9.6%, and 10.6% at 20°C, 40°C, and 60°C, respectively. Sec-butyl acetate showed a relatively significant increase in 2,3-BDO extraction efficiency, with values of 2.1%, 4.2%, and 7.8% at 20°C, 40°C, and 60°C, respectively. Isopropyl myristate also exhibited a relatively significant increase in efficiency, compared to the other ester solvents, with values of 1.2%, 2.6%, and 6.8% at 20°C, 40°C, and 60°C, respectively. The hydrocarbon solvents followed a similar trend of increasing extraction efficiency of 2,3-BDO with increasing extraction temperature. Notably, tert-Butylcyclohexane showed a marked increase in extraction efficiency with increasing temperature, with measured values of 0.7%, 1.1%, and 5.5% at 20°C, 40°C, and 60°C, respectively. However, the overall extraction efficiencies of hydrocarbon solvents were still lower compared to those of the alcohol and ester solvents for 2,3-BDO extraction. It is important to note that while increasing the

extraction temperature may improve the extraction efficiency of 2,3-BDO from aqueous solution, such experimental conditions could be cost-prohibitive at larger scales due to high energy consumption. Therefore, a careful balance between solvent selection and extraction temperature should be considered.

The impact of varying organic-to-aqueous ratios on 2,3-BDO extraction was evaluated by measuring the experimental extraction efficiencies for three solvents (1-butanol, 2-methyl-1-butanol, and 3-methyl-1-butanol) at three different ratios (0.5, 1, and 2), as depicted in **Figure 5(b)**. It is not a surprise to observe that the extraction efficiency increases with increasing organic-to-aqueous ratios for all three solvents. Among these solvents, 1-butanol showed the highest extraction efficiency across all three organic-to-aqueous ratios, with an initial extraction efficiency of 27.6% at a ratio of 0.5, to 44.0% at a ratio of 1, and reached 76.2% for a ratio of 2. 2-methyl-1-butanol exhibited an extraction efficiency of 12.9% at a ratio of 0.5, 34.1% at a ratio of 1, and 46.0% at a ratio of 2. 3-methyl-1-butanol displayed an extraction efficiency of 16.4% at a ratio of 0.5, 33.2% at a ratio of 1, and 44.9% at a ratio of 2. Overall, the increase of extraction efficiency for 1-butanol is more significant with increasing organic-to-aqueous ratio than the other two solvents. At low organic-to-aqueous ratios, a more favorable condition in practical applications due to lower solvent usage, the performance among 1-butanol, 2-methyl-1-butanol, and 3-methyl-1-butanol is more similar. Therefore, we suggest 2-methyl-1-butanol, and 3-methyl-1-butanol as two possible alternatives to 1-butanol for 2,3-BDO LLE with low solvent usage.

4. CONCLUSIONS

This study developed a ML-driven workflow to enhance the LLE of bio-based 2,3-BDO, successfully combining molecular modeling, computational screening, and experimental validation, to identify alternative solvents. The DFT calculations using SMD method established not only a computational database to enable the development of a more versatile ML model

(Extra-Trees ensemble) but also a quantitative correlation to convert computational results (P) to experimentally measurable term (K_d). As a result, the ML model was developed to overcome the limitation of DFT calculations, validated by 24 experimental measurements, and used to suggest new solvent candidates via a solvent screening. Additional experiments were carried out to verify suggested solvents and to investigate the impact of temperature and organic-to-aqueous ratio on extraction efficiency. We suggested alternative solvents such as 2-methyl-1-butanol and 3-methyl-1-butanol for 2,3-BDO LLE. This workflow overcomes the shortage of extensive experimental data of 2,3-BDO extraction, and it requires only a small number of experimental measurements (on the order of tens) to correlate and validate computational results to experimental results. This work delivers a targeted list of solvent candidates for 2,3-BDO extraction, and also establishes a versatile approach to broader applications in solvent screening for biomass-derived chemical separations and potentially a wider range of LLE studies.

5. AUTHOR CONTRIBUTIONS

J.E. and D.Z. contributed equally to this work. J.E. performed the experimental validation, and D.Z. performed DFT calculations and machine learning development. J.E. and D.Z. co-drafted the manuscript. T.L. made substantial contributions to sample analysis. D.L. and D.Z. conceptualized the study and defined its scope. D.L., N.S., and V.G. provided critical insights on manuscript content and contributed significant revisions. All authors reviewed and approved the final manuscript.

6. ACKNOWLEDGEMENTS

This work is financially sponsored by the U.S. Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies Office (BETO), the

Bioprocessing Separations Consortium. Computational resources were provided by the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility, using NERSC award BES ERCAP0031452. The authors from the ABPDU acknowledge support from the U.S. Department of Energy's Bioenergy Technologies Office (BETO), which is part of the Office of Energy Efficiency and Renewable Energy (EERE), and funding from the American Recovery and Reinvestment Act. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represents that its use would not infringe privately owned rights. We extend our gratitude to Stephen Tan and Randy Louie from the Joint BioEnergy Institute (JBEI) for their help with high-throughput experimental tests, and to Dr. Ramkrishna Singh from ABPDU for his insightful discussions on sample analysis and other assistance.

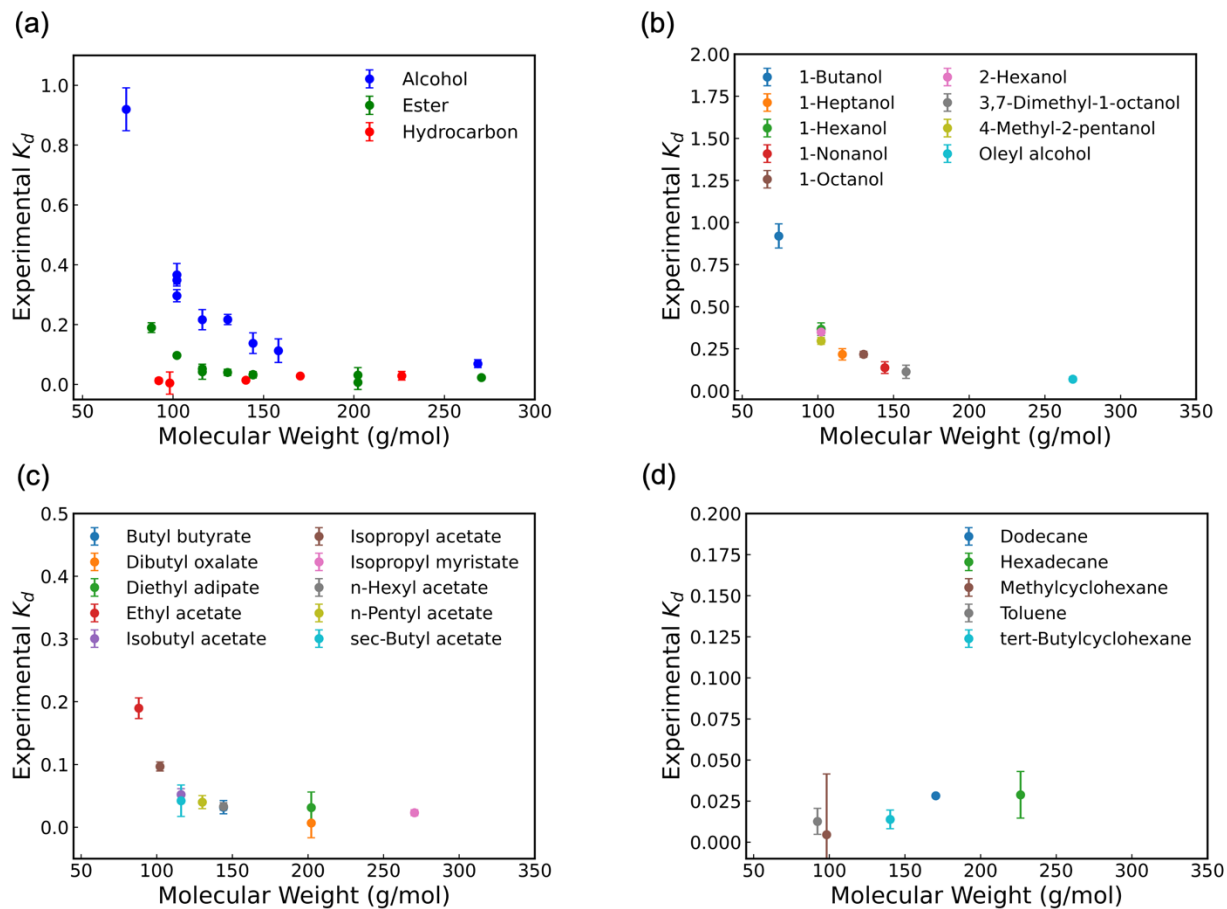


Figure 1. (a) The experimental K_d as a function of molecular weight of initial solvents measured at 20°C and 1:1 v/v organic-to-aqueous ratio, classified as (b) alcohol, (c) ester, or (d) hydrocarbon.

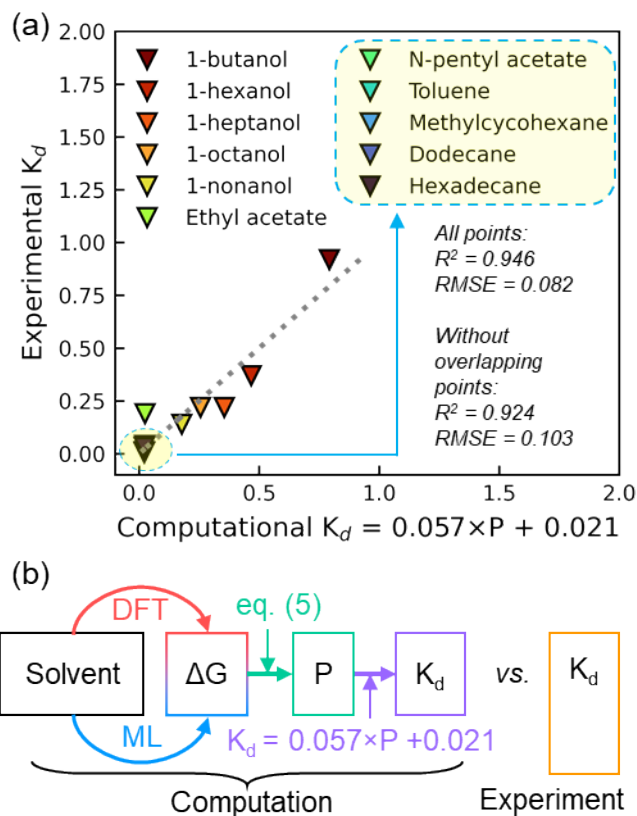


Figure 2. (a) The fitted linear correlation between the experimental K_d values measured at 20°C and the computational K_d values computed through a linear transformation of computational partition coefficient from DFT calculations with SMD implicit solvation. The label of the horizontal axis indicates the linear equation to convert P to K_d , which is fitted using all 11 solvents involved in the plot. The shaded area highlights the overlapping points of 5 solvents. (b) The schematic flow for evaluating the computational K_d value of a given solvent. ΔG , P , and K_d on the left side indicate the free energy difference, partition coefficient, and distribution coefficient, respectively, from computation.

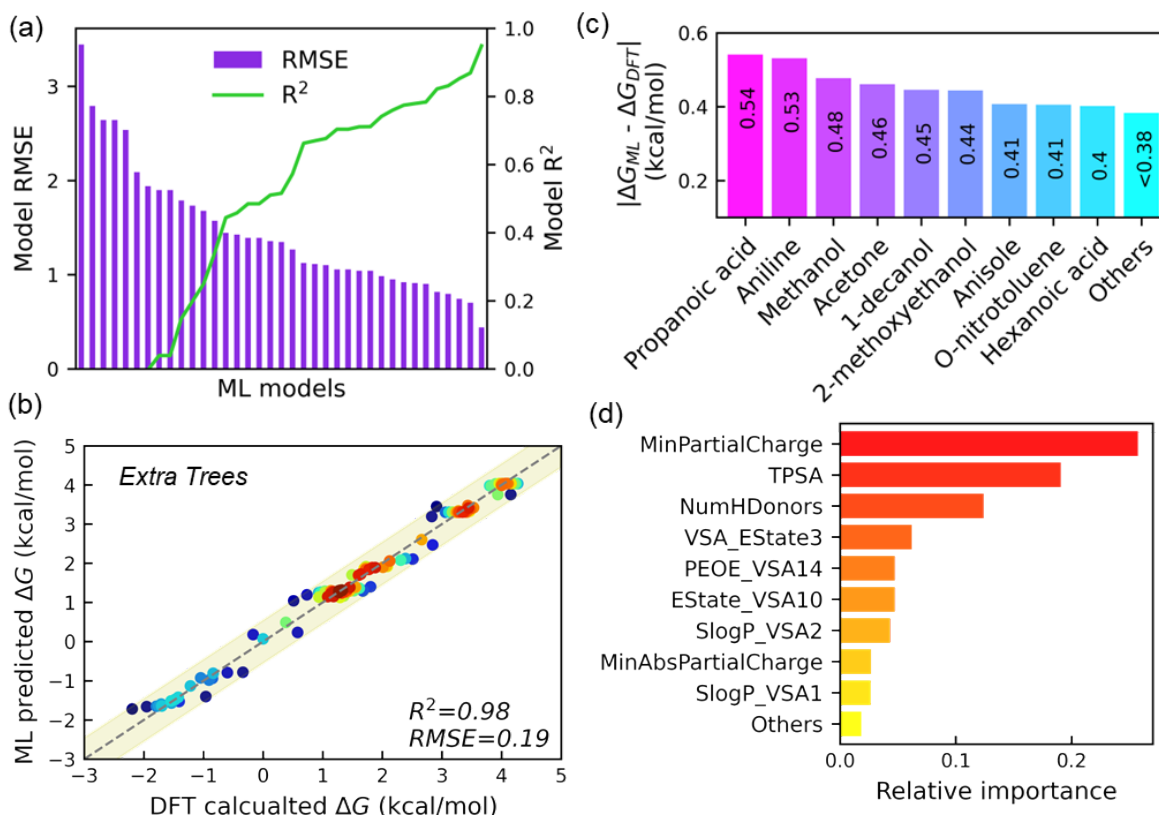


Figure 3. (a) The predictive performance (testing set) of different ML algorithms using the same DFT training dataset of DFT ΔG values. (b) The accuracy of the fine-tuned Extra-Trees model by five-fold cross validation for predicted ΔG compared to DFT results. Data points are colored by their local distribution density where warmer colors indicate dense distribution and colder colors indicate thin distribution. The shaded region highlights the confidence band within a numerical error of 0.55 kcal/mol. (c) The solvents showing the largest predicted errors compared to DFT values. Numerical values are labeled by the text in the bars. (d) The feature importance for the top 10 features impacting the predicted extraction performance by the Extra-Trees model.

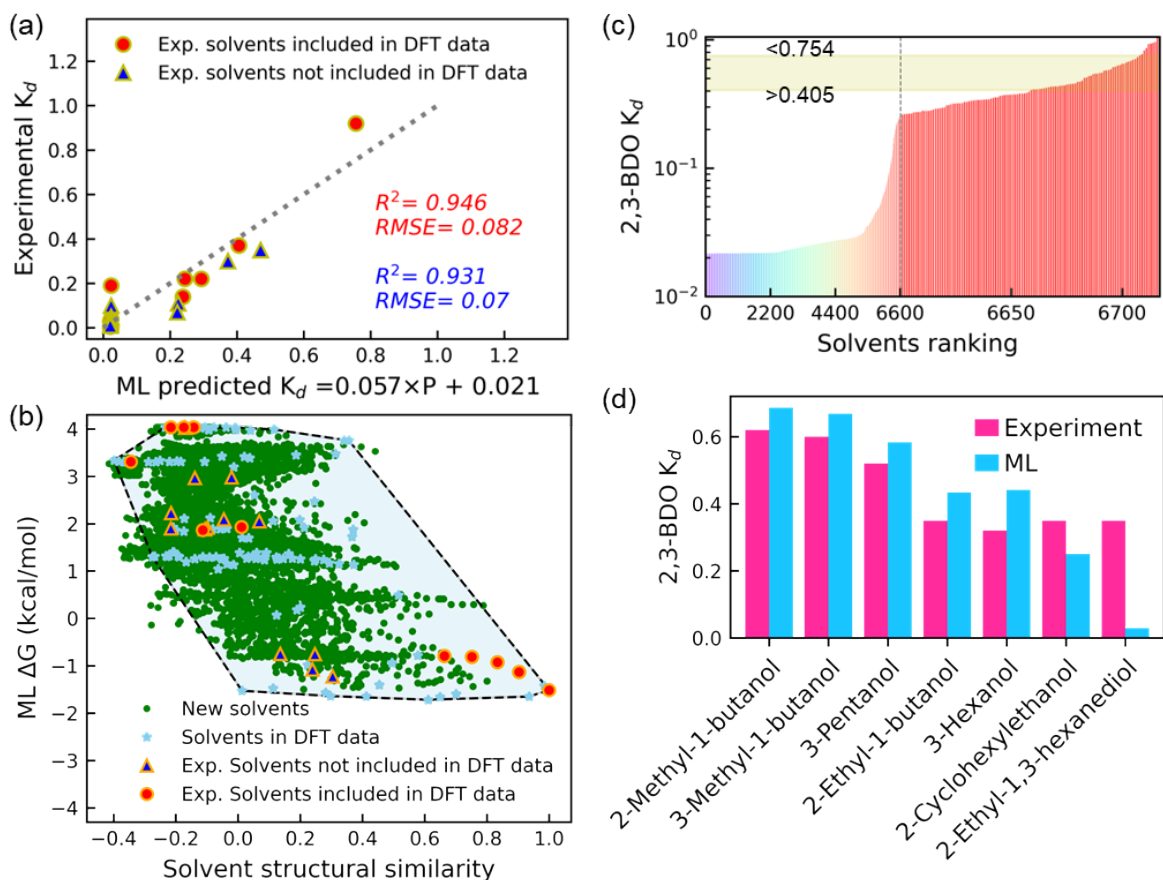


Figure 4. (a) The K_d predictive performance of our Extra-Trees model for the 24 solvents in our initial list compared to experimental results. The metrics of two types of data (included in the DFT calculations or not) are highlighted in the text. (b) The structural similarity of solvents from DFT data, experimental measurement and ML screening. The shaded area highlights the space (convex hull) defined by the DFT data. (c) The ML solvent screening results for the K_d values of 2,3-BDO in 6717 organic solvents. Solvents are ranked based on their K_d values. The inhomogeneous horizontal axis highlights the solvents with higher K_d values on the right side of the vertical dash line. The yellow shaded region highlights the K_d from 0.405 to 0.754. (d) Selected solvent candidates from our new list of solvents, and their experimental and ML-predicted K_d values.

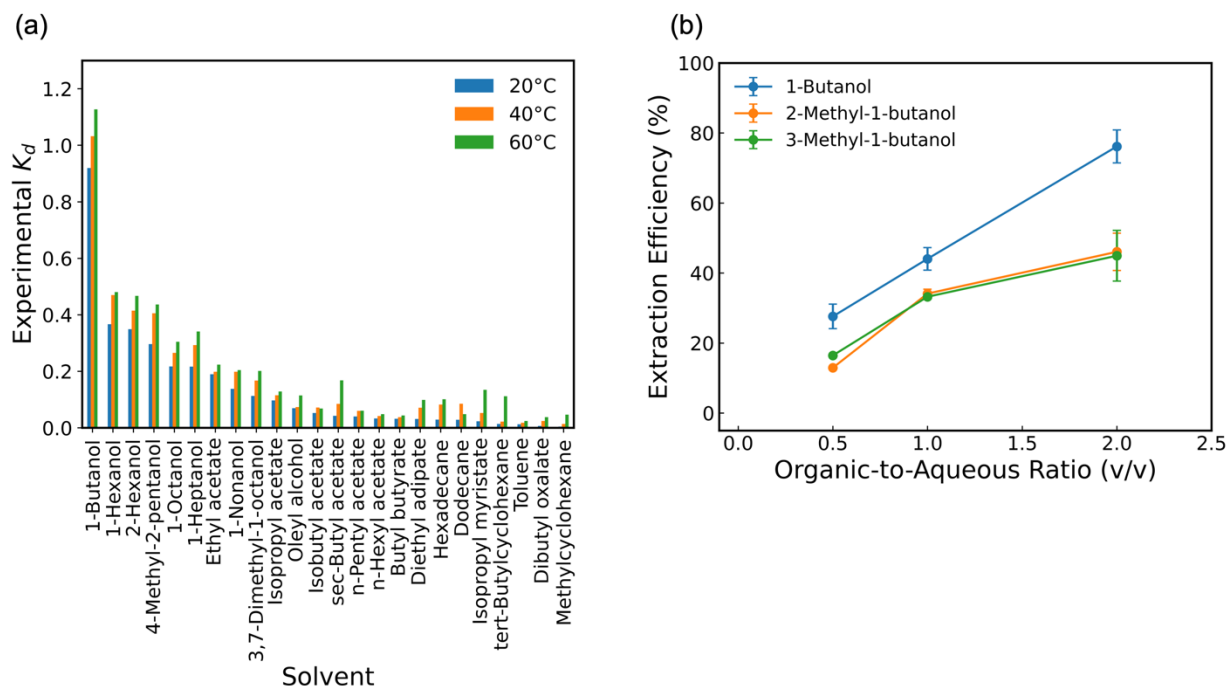


Figure 5. (a) The extraction efficiency of initial solvents measured at different extraction temperatures and 1:1 v/v organic-to-aqueous ratio. (b) The extraction efficiency of 1-butanol, 2-methyl-1-butanol, and 3-methyl-1-butanol measured at 20°C and increasing organic-to-aqueous ratio.

REFERENCES

- (1) Xie, S.; Li, Z.; Zhu, G.; Song, W.; Yi, C. Cleaner production and downstream processing of bio-based 2, 3-butanediol: A review. *Journal of Cleaner Production* **2022**, *343*, 131033.
- (2) Bai, Y.; Feng, H.; Liu, N.; Zhao, X. Biomass-derived 2, 3-butanediol and its application in biofuels production. *Energies* **2023**, *16* (15), 5802.
- (3) Gawal, P. M.; Subudhi, S. Advances and challenges in bio-based 2, 3-BD downstream purification: A comprehensive review. *Bioresource Technology Reports* **2023**, 101638.
- (4) Xiu, Z.-L.; Zeng, A.-P. Present state and perspective of downstream processing of biologically produced 1, 3-propanediol and 2, 3-butanediol. *Applied microbiology and biotechnology* **2008**, *78*, 917-926.
- (5) Bokhary, A.; Leitch, M.; Liao, B. Q. Liquid-liquid extraction technology for resource recovery: Applications, potential, and perspectives. *Journal of Water Process Engineering* **2021**, *40*, 101762. DOI: <https://doi.org/10.1016/j.jwpe.2020.101762>.
- (6) Mazzola, P. G.; Lopes, A. M.; Hasmann, F. A.; Jozala, A. F.; Penna, T. C. V.; Magalhaes, P. O.; Rangel-Yagui, C. O.; Pessoa Jr, A. Liquid-liquid extraction of biomolecules: an overview and update of the main techniques. *Journal of Chemical Technology & Biotechnology* **2008**, *83* (2), 143-157. DOI: <https://doi.org/10.1002/jctb.1794> (accessed 2025/02/04).
- (7) Zhang, Q.-W.; Lin, L.-G.; Ye, W.-C. Techniques for extraction and isolation of natural products: A comprehensive review. *Chinese medicine* **2018**, *13*, 1-26.
- (8) Ewunie, G. A.; Morken, J.; Lekang, O. I.; Yigezu, Z. D. Factors affecting the potential of *Jatropha curcas* for sustainable biodiesel production: A critical review. *Renewable and Sustainable Energy Reviews* **2021**, *137*, 110500.
- (9) Narisetty, V.; Amraoui, Y.; Abdullah, A.; Ahmad, E.; Agrawal, D.; Parameswaran, B.; Pandey, A.; Goel, S.; Kumar, V. High yield recovery of 2, 3-butanediol from fermented broth accumulated on xylose rich sugarcane bagasse hydrolysate using aqueous two-phase extraction system. *Bioresource Technology* **2021**, *337*, 125463.
- (10) Wan, F.; Kang, T.; Liu, A.; Zhou, C.; Liu, S.; Xu, Y.; Si, S. Salt induced phase separation extraction of 2, 3-Butanediol from aqueous solutions: Recovery and recycling of potassium triphosphate. *Process Biochemistry* **2023**, *125*, 222-231.
- (11) Dai, J.; Wang, H.; Li, Y.; Xiu, Z.-L. Imidazolium ionic liquids-based salting-out extraction of 2, 3-butanediol from fermentation broths. *Process Biochemistry* **2018**, *71*, 175-181.
- (12) Birajdar, S. D.; Rajagopalan, S.; Sawant, J. S.; Padmanabhan, S. Continuous predispersed solvent extraction process for the downstream separation of 2, 3-butanediol from fermentation broth. *Separation and Purification Technology* **2015**, *151*, 115-123.
- (13) Tinoco, D.; Borschiver, S.; Coutinho, P. L.; Freire, D. M. Technological development of the bio-based 2, 3-butanediol process. *Biofuels, Bioproducts and Biorefining* **2021**, *15* (2), 357-376.
- (14) Liu, D.; Sun, N. Prospects of artificial intelligence in the development of sustainable separation processes. *Frontiers in Sustainability* **2023**, *4*, 1210209.
- (15) Casas, A.; Rodríguez-Llorente, D.; Rodríguez-Llorente, G.; García, J.; Larriba, M. Machine learning screening tools for the prediction of extraction yields of pharmaceutical compounds from wastewaters. *Journal of Water Process Engineering* **2024**, *62*, 105379.
- (16) Liu, T.; Johnson, K. R.; Jansone-Popova, S.; Jiang, D.-e. Advancing rare-earth separation by machine learning. *JACS Au* **2022**, *2* (6), 1428-1434.
- (17) Darwish, A. S.; Lemaoui, T.; AlYammahi, J.; Taher, H.; Benguerba, Y.; Banat, F.; AlNashef, I. M. Molecular insights into potential hydrophobic deep eutectic solvents for furfural extraction guided by COSMO-RS and machine learning. *Journal of Molecular Liquids* **2023**, *379*, 121631.
- (18) Choi, Y.; Park, H.; Roh, J.; Lim, J.; Han, I.-s.; Kim, D.-K.; Jeon, S.; Nam, H.-G.; Moon, I.; Cho, H. A Data-based Predictive Model for Distillation Column of Bio-based 2, 3-Butanediol. In *Computer Aided Chemical Engineering*, Vol. 50; Elsevier, 2021; pp 1005-1011.
- (19) Liang, Y.; Job, H.; Feng, R.; Parks, F.; Hollas, A.; Zhang, X.; Bowden, M.; Noh, J.; Murugesan, V.; Wang, W. High-throughput solubility determination for data-driven materials design and discovery in redox flow battery research. *Cell Reports Physical Science* **2023**, *4* (10).
- (20) Qiu, J.; Patel, A.; Stevens, J. M. High-throughput salt screening of synthetic intermediates: effects of solvents, counterions, and counterion solubility. *Organic process research & development* **2020**, *24* (7), 1262-1270.
- (21) Pomberger, A.; Jose, N.; Walz, D.; Meissner, J.; Holze, C.; Kopczynski, M.; Müller-Bischof, P.; Lapkin, A. Automated pH adjustment driven by robotic workflows and active machine learning. *Chemical Engineering Journal* **2023**, *451*, 139099.
- (22) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics* **2020**, *152* (22). DOI: 10.1063/5.0004608 (accessed 10/25/2024).
- (23) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38* (6), 3098-3100. DOI: 10.1103/PhysRevA.38.3098.
- (24) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37* (2), 785-789. DOI: 10.1103/PhysRevB.37.785.

- (25) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132* (15), 154104. DOI: 10.1063/1.3382344 (accessed 2022/12/13).
- (26) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7* (18), 3297-3305, 10.1039/B508541A. DOI: 10.1039/B508541A.
- (27) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009**, *113* (18), 6378-6396. DOI: 10.1021/jp810292n.
- (28) RDKit: Open-source cheminformatics. <https://www.rdkit.org>. (accessed).
- (29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **2011**, *12*, 2825-2830.
- (30) Brett Smith, H. W., Difan Zhang, Vissiliki-Alexandra Glezakou. Examining Solvents for 2,3-Butanediol Extraction From Water Using Molecular Modeling and Machine Learning. *ChemRx* **2024**.
- (31) Birajdar, S. D.; Padmanabhan, S.; Rajagopalan, S. Rapid solvent screening using thermodynamic models for recovery of 2, 3-butanediol from fermentation by liquid-liquid extraction. *Journal of Chemical & Engineering Data* **2014**, *59* (8), 2456-2463.
- (32) König, G.; Reetz, M. T.; Thiel, W. 1-Butanol as a solvent for efficient extraction of polar compounds from aqueous medium: theoretical and practical aspects. *The Journal of Physical Chemistry B* **2018**, *122* (27), 6975-6988.
- (33) Anvari, M.; Khayati, G. In situ recovery of 2, 3-butanediol from fermentation by liquid-liquid extraction. *Journal of industrial microbiology and biotechnology* **2009**, *36* (2), 313-317.
- (34) Khayati, G.; Pahlavanzadeh, H.; Vasheghani-Farahani, E.; Ghaemi, N. (Liquid+ liquid) phase equilibria for (water+ 2, 3-butanediol+ oleyl alcohol) at T=(300.2, 307.2, and 314.2) K. *The Journal of Chemical Thermodynamics* **2009**, *41* (2), 150-154.
- (35) Shao, P.; Kumar, A. Recovery of 2,3-butanediol from water by a solvent extraction and pervaporation separation scheme. *Journal of Membrane Science* **2009**, *329* (1), 160-168. DOI: <https://doi.org/10.1016/j.memsci.2008.12.033>.
- (36) Wu, Y.; Chen, W.; Li, R.; Wu, B.; Ji, L.; Chen, K. Measurement and thermodynamic modeling of ternary (liquid + liquid) equilibria for extraction of 2,3-butanediol from aqueous solution with different solvents at T = 298.2 K, T = 308.2 K, and T = 318.2 K. *Chemical Engineering Communications* **2022**, *209* (5), 636-647. DOI: 10.1080/00986445.2021.1884553.
- (37) Gawal, P. M.; Mudgil, D.; Subudhi, S. Hybrid extraction distillation for downstream purification of 2, 3-butanediol, and ethanol from fermentation broth. *Bioresource Technology Reports* **2023**, *24*, 101649.
- (38) Yu, T.; Yang, L.; Shang, X.; Bian, S. Recovery of Cembratrien-Diols from Waste Tobacco (*Nicotiana tabacum* L.) Flowers by Microwave-Assisted Deep Eutectic Solvent Extraction: Optimization, Separation, and In Vitro Bioactivity. *Molecules* **2024**, *29* (7), 1563.