



Delft University of Technology

#### Document Version

Final published version

#### Licence

CC BY

#### Citation (APA)

Raju, D., Skartlien, R., Ramdin, M., & Vlugt, T. J. H. (2026). Machine Learning-Based Prediction of Phase Equilibria and Interfacial Properties of Multicomponent CO<sub>2</sub> Mixtures with Impurities: Application to CO<sub>2</sub> Transportation. *Industrial and Engineering Chemistry Research*, 65(35), 18863-18890. <https://doi.org/10.1021/acs.iecr.6c03388>

#### Important note

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

#### Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.  
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

#### Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

#### Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

*This work is downloaded from Delft University of Technology.*

# Machine Learning-Based Prediction of Phase Equilibria and Interfacial Properties of Multicomponent CO<sub>2</sub> Mixtures with Impurities: Application to CO<sub>2</sub> Transportation

Darshan Raju, Roar Skartlien, Mahinder Ramdin, and Thijs J. H. Vlucht\*



Cite This: *Ind. Eng. Chem. Res.* 2026, 65, 18863–18890



Read Online

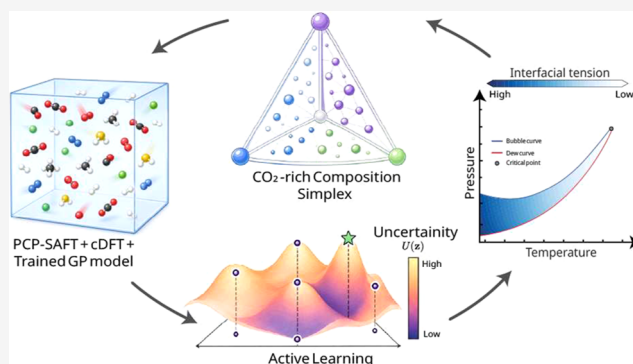
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Impurities in captured CO<sub>2</sub> broaden the two-phase envelope, decrease the interfacial tension (IFT), and affect safe dense-phase transport. These properties are difficult to measure and expensive to compute using molecular simulations. We present the first systematic machine learning (ML) surrogate framework for phase equilibria and interfacial properties of impure CO<sub>2</sub>, trained on Perturbed-Chain Polar Statistical Associating Fluid Theory (PCP-SAFT) equation of state (EoS) + cDFT data. Within this framework, the hyperparameter-optimized TabPFN most accurately predicts  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ , IFT, and interfacial thickness. For IFT, we combined the Winterfeld–Scriven–Davis correlation with an ML-based residual correction. The hybrid model with stochastic variational Gaussian process regression reduces the root mean squared error (RMSE) of IFT to 0.02 mN m<sup>-1</sup>. Symbolic regression additionally provides an interpretable expression for this correction. For sampling CO<sub>2</sub>-rich compositions, active learning is the most data-efficient strategy. Trained surrogates predict each state point in ca. 1–5 ms, compared to ca. 0.18–1.8 s for cDFT, enabling rapid property estimation for computational fluid dynamics (CFD), inverse design, and the planning of simulations and experiments.



## 1. INTRODUCTION

Carbon dioxide (CO<sub>2</sub>) is one of the major greenhouse gases that drive global warming.<sup>1,2</sup> Carbon capture has become an essential process to reduce CO<sub>2</sub> emissions and to prevent global warming.<sup>3,4</sup> Techniques related to the CO<sub>2</sub> capture, transport, storage, and utilization of CO<sub>2</sub> are collectively referred to as CCUS. The transportation of CO<sub>2</sub> is essential for multiple CCUS applications, including direct air capture,<sup>5</sup> enhanced oil recovery,<sup>6</sup> methanol production,<sup>6</sup> and dedicated reservoir storage (e.g., Northern Lights, Porthos).<sup>7,8</sup> CO<sub>2</sub> is transported efficiently from capture facilities to storage reservoirs by pipelines or ships.<sup>9</sup> CO<sub>2</sub> obtained from capture processes is typically not pure. The composition of the captured CO<sub>2</sub> depends on the capture technology and the source stream, such as flue gas in postcombustion capture, syngas in precombustion capture, and oxy-fuel flue gas in oxy-combustion capture.<sup>10</sup> The impurities found primarily in postcombustion are typically N<sub>2</sub>, O<sub>2</sub>, and Ar, whereas in oxy-fuel combustion, high concentrations of Ar, N<sub>2</sub>, and O<sub>2</sub> are found.<sup>6,7,11–13</sup> In precombustion, the impurities are mostly H<sub>2</sub>, CH<sub>4</sub>, and CO.<sup>6,7,11–13</sup> Along with the noncondensable impurities such as Ar, N<sub>2</sub>, H<sub>2</sub>, CH<sub>4</sub>, O<sub>2</sub>, and CO,<sup>6,7,11–13</sup> minor traces of corrosive impurities such as H<sub>2</sub>O, H<sub>2</sub>S, SO<sub>x</sub>, NO<sub>x</sub>, O<sub>2</sub>, and hydrocarbons (C<sub>2</sub>H<sub>6</sub> and C<sub>3</sub>H<sub>8</sub>) are also found in the captured source streams. These impurities affect the

thermophysical behavior and phase stability of the CO<sub>2</sub> stream by broadening the two-phase envelope. For efficient transportation, CO<sub>2</sub> streams must be maintained in a dense single phase.<sup>14–16</sup> This reduces the volumetric flow rate and helps prevent phase separation, which can otherwise cause flow instability and pressure losses.<sup>14,15</sup> This requirement is particularly challenging during transient operations, where pressure and temperature fluctuate.<sup>10,17</sup> While steady-state operation targets a dense phase, systems often undergo startup, shutdown, depressurization, and planned maintenance, which can lead to two-phase flow in CO<sub>2</sub> transportation systems.<sup>18–20</sup> At such conditions, accurate knowledge of the thermodynamic and transport properties of impure CO<sub>2</sub> mixtures becomes indispensable for modeling their flow and phase behavior.<sup>10,21</sup> Compared to other thermodynamic and transport properties such as density, heat capacity, speed of sound, and viscosity, interfacial tension (IFT) is one of the important properties of the fluid mixture that governs the bubble and droplet

**Received:** June 29, 2026

**Revised:** August 14, 2026

**Accepted:** August 19, 2026

**Published:** August 28, 2026



formation and stability, and interphase heat and mass transfer.<sup>22–24</sup> During depressurization of the CO<sub>2</sub> stream through a valve or nozzle, the CO<sub>2</sub> flow may experience a delayed phase change during that fast transient event.<sup>20</sup> The delayed phase change needs to be accounted for in computational fluid dynamics (CFD) simulations to correctly predict mass flow rate and pressure, thereby enabling the accurate evaluation of pipeline forces and the associated risk of running ductile fracture.<sup>23,25</sup>

Experimental measurements and molecular dynamics (MD) simulations are widely used to determine phase equilibria and interfacial properties.<sup>24–28</sup> Experimental determination of IFTs for a wide range of mixtures and compositions encountered in CO<sub>2</sub> transportation is costly and time-consuming, resulting in limited and scarce experimental data for impure CO<sub>2</sub> mixtures.<sup>10,29–32</sup> MD simulations using classical force fields have proven to be a powerful tool to predict interfacial properties.<sup>24–26,32</sup> Performing MD simulation for a wide range of compositions encompassing a broad range of two-phase conditions to compute interfacial properties can be extremely computationally intensive.<sup>25</sup> In addition to experimental measurements and MD-based computations, equation of state (EoS)-based interfacial modeling approaches, including classical Density Functional Theory (cDFT), density gradient theory (DGT), square gradient theory (SGT), linear gradient theory (LGT), and semiempirical correlations, have been used to predict interfacial properties.<sup>25,27,33,34</sup> The EoS-based cDFT, DGT, SGT, and LGT models closely reproduce experimental measurements<sup>27,35</sup> and MD simulation results.<sup>25,36–38</sup> To illustrate a few, Chow et al.<sup>27</sup> used the Statistical Associating Fluid Theory (SAFT)-variable range (VR) Mie EoS<sup>39</sup> coupled with SGT to predict the IFT of binary and ternary mixtures containing H<sub>2</sub>O, CO<sub>2</sub>, Ar, and N<sub>2</sub>. The predictions closely reproduced the experimental measurements, and the SAFT-VR Mie EoS + SGT framework was recommended for modeling the IFTs of H<sub>2</sub>O–CO<sub>2</sub> systems containing diluent gases. In our previous study,<sup>25</sup> we computed the interfacial properties of binary CO<sub>2</sub> mixtures containing Ar, N<sub>2</sub>, CH<sub>4</sub>, and H<sub>2</sub>. These properties included the IFT, interfacial thickness, and impurity enrichment at the interface. The computations were performed using MD simulations and the Perturbed-Chain Statistical Associating Fluid Theory (PC-SAFT) EoS + cDFT framework, and the results were compared with predictions from the semiempirical Parachor<sup>40</sup> and Winterfeld–Scriven–Davis (WSD)<sup>41</sup> models. The PC-SAFT EoS + cDFT with temperature-dependent binary interaction parameters (BIPs) showed a good agreement with MD simulations, and both the Parachor and the WSD semiempirical models were found to have significant deviations. A wide range of empirical and semiempirical models has been proposed in the literature to model the IFTs of mixtures, differing in their physical basis, flexibility, and number of adjustable or fitted coefficients. In a recent comparative study, Mulero et al.<sup>42</sup> assessed 10 empirical and semiempirical models for binary *n*-alkane mixtures and reported that the WSD model<sup>41</sup> yields the best overall performance for estimating IFT, provided that all components in the mixture are subcritical. When this condition is not satisfied, the Redlich–Kister model<sup>43</sup> was found to perform best. However, the WSD model requires only one fit coefficient, whereas the Redlich–Kister model requires three. Similar limitations in terms of requiring fitted coefficients and condition-dependent performance are also observed in other semiempirical models reported by Mulero et al.,<sup>42</sup> particularly

when one or a few components in the mixture are in a supercritical state.<sup>44</sup> Complementing these traditional modeling approaches, machine learning (ML) has emerged as a promising alternative by leveraging available thermophysical data.<sup>45</sup> Maginn and co-workers<sup>45</sup> used the CRC Handbook of Chemistry and Physics<sup>46</sup> as the primary experimental source to train the residuals obtained from the Joback and Reid (JR) Group Contribution (GC) method. The computationally efficient JR GC method often shows systematic biases (residuals). The systematic biases of normal boiling temperature ( $T_b$ ), enthalpy of vaporization at  $T_b$  ( $\Delta H_{\text{vap},T_b}$ ), normal melting temperature ( $T_m$ ), critical pressure ( $P_c$ ), critical molar volume ( $V_c$ ), and critical temperature ( $T_c$ ) are compared to the experimental source<sup>46</sup> of 485–5640 molecules. The computed residuals were then trained on the Gaussian Process Regression (GPR) model to improve the accuracy of thermodynamic property predictions using the JR GC method and to quantify the associated uncertainties.<sup>45</sup> The hybrid model achieved predictive performance of a coefficient of determination ( $R^2$ ) > 0.85 for  $T_b$ ,  $\Delta H_{\text{vap},T_b}$ ,  $P_c$ ,  $V_c$ , and  $T_c$ .<sup>45</sup> Focused on IFT, Mouallem et al.<sup>47</sup> developed ML models to estimate CO<sub>2</sub>-brine IFT for CO<sub>2</sub> geo-storage using a large literature data set of 2896 experimental data points covering wide ranges of pressure, temperature, salinity, salt type, and CH<sub>4</sub> and N<sub>2</sub> impurities. These authors compared Gradient Boosting (GB), eXtreme Gradient Boosting (XGB), Least Squares Boosting (LSB), GPR, Artificial Neural Networks (ANN), and Genetic Programming, and found that GB performed the best with  $R^2 = 0.964$ . When a suitable database is unavailable or nonexistent, numerous studies have used physics-based models, such as molecular simulations and EoSs, to generate representative data sets for application-specific compositions and thermodynamic conditions.<sup>47–50</sup> Albá et al.<sup>48</sup> used the polar soft-SAFT EoS to generate saturation properties data for CO<sub>2</sub>-based refrigerant mixtures and trained an ANN to predict bubble and dew-point temperatures. The ANN model achieved  $R^2 \approx 1$ , demonstrating its suitability for screening sustainable refrigerant mixtures.<sup>48</sup> It is important to note that purely data-driven ML models do not always follow the expected physical behavior. This is especially important for tree-based models, which typically produce piecewise-constant predictions and have limited ability to extrapolate beyond the training range. As a result, these models may not fully capture the smooth variation of thermophysical properties, such as IFT, with temperature and pressure. This motivates the use of physics-informed hybrid models, where a physically based model first describes the main trend, and the ML model is then used to learn only the remaining residual error. Although several ML models have been developed to predict the IFT of CO<sub>2</sub>-brine systems<sup>47,51–53</sup> and other thermophysical properties of CO<sub>2</sub>-based mixtures,<sup>54–56</sup> including saturation properties,<sup>48</sup> a comparative assessment of ML models for CO<sub>2</sub>-rich mixtures containing transportation-relevant impurities is still lacking. Such an assessment first requires a training data set, and building that data set dominates the cost because every point has to be generated by a physics-based computation, and the composition space of CO<sub>2</sub> transport is high-dimensional and thinly populated. The number of feed compositions required to reach a target accuracy, therefore, determines the computational cost of training ML models, and the most efficient sampling strategy for these compositions has not been established for CO<sub>2</sub>-rich mixtures.

In this study, we use the Perturbed-Chain Polar Statistical Associating Fluid Theory (PCP-SAFT) EoS<sup>57–59</sup> combined with cDFT<sup>60</sup> to compute the IFT, interfacial thickness, bubble-point pressure, and dew-point pressure of multicomponent CO<sub>2</sub>-rich mixtures containing transportation-relevant impurities. The CO<sub>2</sub> mole fraction ranges from 95 to 99.99%, while the total impurity concentration varies from 5 to 0.01%. This concentration range reflects practical CO<sub>2</sub> transportation conditions,<sup>7</sup> where high CO<sub>2</sub> purity is desired to maintain dense-phase operation and reduce the risk of phase splitting and flow-assurance problems.<sup>6,7,10</sup> The impurities considered are Ar, N<sub>2</sub>, CH<sub>4</sub>, H<sub>2</sub>, O<sub>2</sub>, CO, and H<sub>2</sub>S. Other impurities, such as H<sub>2</sub>O, NO<sub>*x*</sub>, and SO<sub>*x*</sub>, are typically present at parts-per-million levels and are therefore not included.<sup>6,10</sup> The inclusion of these impurities enables the model to cover the main classes of transportation-relevant contaminants, including light gases, noncondensable components, and acid–gas impurities.<sup>6,10</sup> The systems considered range from binary to 8-component mixtures. The PCP-SAFT EoS predicts the phase equilibria and interfacial properties of CO<sub>2</sub>-rich mixtures in close agreement with experimental data, yielding bubble- and dew-point pressures comparable to those obtained from the GERG-2008 EoS.<sup>61</sup> This agreement supports its use for generating the phase-equilibrium and interfacial-property training data. Using the generated PCP-SAFT EoS + cDFT<sup>57,62</sup> data set, we develop data-driven ML models for four target properties: bubble-point pressure, dew-point pressure, IFT, and interfacial thickness. We compare several ML models, including Support Vector Machine (SVM), Random Forest (RF), XGB, and Tabular Prior-data Fitted Network (TabPFN). For all target properties, we use the same input features: temperature, pressure, and mixture composition. These features correspond to the input variables required when planning experiments or molecular simulations. We found that TabPFN with Hyper-Parameter Optimization (HPO), followed by XGB, achieved the lowest Cross Validation (CV) root-mean-square error (RMSE) and mean absolute error (MAE) for all target properties. The results also showed that dew-point pressure and interfacial thickness are less sensitive to impurity concentration. We further assessed the effect of mixture size and compared random, stratified, and GPR-based active learning (AL) sampling strategies. AL was the most data-efficient sampling strategy, requiring fewer feed compositions to achieve a given accuracy. We further develop a hybrid ML model for IFT using the computationally efficient WSD model as a baseline. The ML model is trained on the deviation of the WSD model from the more rigorous PCP-SAFT EoS+cDFT predictions. The ML models used to develop the IFT hybrid models are Symbolic Regression (SR), GPR, and Stochastic Variational Gaussian Process Regression (SVGP). This hybrid strategy enables rapid estimation of IFTs for multicomponent CO<sub>2</sub>-rich mixtures relevant to CO<sub>2</sub> transportation. After training, the surrogate models can predict interfacial and saturation properties much more rapidly than direct PCP-SAFT EoS + cDFT calculations. This reduction in computational cost is the central motivation of the present study. Multiphase CFD simulations of CO<sub>2</sub> transport require repeated evaluations of interfacial and saturation properties over a wide range of compositions and thermodynamic states, which become prohibitively expensive when using direct PCP-SAFT EoS + cDFT or MD computations.<sup>20,23</sup> In sharp contrast, once trained, the ML surrogates can provide these properties at negligible computational cost, making them

suitable for repeated property evaluations in large-scale transport process simulations. To the best of our knowledge, this is the first study to combine the PCP-SAFT EoS + cDFT physics engine with two complementary ML surrogate strategies for the IFT of multicomponent CO<sub>2</sub>-rich mixtures, while also developing fully data-driven ML models for phase equilibria. The two strategies are a fully data-driven model and a physics-informed WSD-residual model. Earlier ML studies are typically limited to a single property or to simple binary or CO<sub>2</sub>–brine systems. In contrast, our framework spans binary to eight-component mixtures and includes seven CO<sub>2</sub> transportation-relevant impurities. It also provides uncertainty-aware predictions and active-learning-based composition sampling.

The paper is organized as follows: we describe the methodology used to compute the saturation properties (bubble-point and dew-point pressures) and interfacial properties using the PCP-SAFT EoS + cDFT in Section 2. The WSD semiempirical model and ML models used in the study are also explained briefly in Section 2. In Section 3, we present the predictive performance of the ML models for each target property. We also analyze the effects of composition space and mixture size and evaluate the performance of the hybrid residual ML model relative to the semiempirical WSD model. Finally, Section 4 summarizes the main findings of this work. Further methodological details and results are provided in the Supporting Information. The experimental phase equilibrium data, the training data sets, and the trained ML models used in this work are provided in the accompanying Supporting Information.

## 2. METHODOLOGY

### 2.1. Classical Density Functional Theory + PCP-SAFT

The phase equilibria (bubble and dew-point pressures) represented in the *PT* phase diagram, and the interfacial properties of multicomponent CO<sub>2</sub> mixtures with impurities are computed using the Framework for Equations of State and Classical Density Functional Theory (FeOs) v0.9.5 framework.<sup>60</sup> In FeOs, SAFT-based EoSs, such as PCP-SAFT and PC-SAFT,<sup>57–59,63</sup> are coupled with cDFT<sup>62,64–66</sup> to model inhomogeneous systems, including vapor–liquid interfaces.<sup>60,62</sup> In cDFT, the vapor–liquid interface is treated as a mesoscopic region in which the density varies smoothly from the bulk vapor phase to the bulk liquid phase. The inhomogeneous system is described using the grand potential functional,<sup>62</sup>

$$\Omega(T, \mu, [\rho(\mathbf{r})]) = F(T, [\rho(\mathbf{r})]) - \sum_i \int \rho_i(\mathbf{r})(\mu_i - V_i^{\text{ext}}(\mathbf{r}))d\mathbf{r} \quad (1)$$

where  $\mathbf{r}$  is the position vector in space,  $\rho_i(\mathbf{r})$  is the number-density profile of component  $i$ ,  $[\rho(\mathbf{r})]$  denotes the set of density profiles of all components in the mixture,  $F(T, [\rho(\mathbf{r})])$  is the intrinsic Helmholtz free-energy functional,  $\mu_i$  is the chemical potential of component  $i$ , and  $V_i^{\text{ext}}(\mathbf{r})$  is an external potential acting on all or some molecules.<sup>62</sup> The equilibrium density profiles are obtained by minimizing the grand potential with respect to the density profile at a constant temperature and chemical potential of each component,

$$\left. \frac{\delta \Omega}{\delta \rho_i(\mathbf{r})} \right|_{T, \mu_i} = 0 \quad (2)$$

Applying this equilibrium condition to the grand potential functional in eq 1 yields the Euler–Lagrange equation.<sup>60</sup>

$$\left. \frac{\delta F[\rho(\mathbf{r})]}{\delta \rho_i(\mathbf{r})} \right|_T - \mu_i + V_i^{\text{ext}}(\mathbf{r}) = 0 \quad (3)$$

For a homogeneous bulk system,  $V_i^{\text{ext}}(\mathbf{r}) = 0$ , and the corresponding bulk relation is  $F_{\rho_i}^{\text{b}} - \mu_i = 0$ , where  $F_{\rho_i}^{\text{b}}$  denotes the functional derivative of the Helmholtz free-energy functional evaluated for the bulk system. Using this bulk relation, the Euler–Lagrange equation<sup>60</sup> in the absence of an external potential can be written relative to the bulk state as

$$\left. \frac{\delta F[\rho(\mathbf{r})]}{\delta \rho_i(\mathbf{r})} \right|_T - F_{\rho_i}^{\text{b}} = 0 \quad (4)$$

The intrinsic Helmholtz free-energy functional in the Euler–Lagrange equation is commonly decomposed into ideal and residual contributions,  $F[\rho(\mathbf{r})] = F^{\text{ideal}}[\rho(\mathbf{r})] + F^{\text{res}}[\rho(\mathbf{r})]$ , where the ideal contribution is known exactly from statistical mechanics, while the residual contribution accounts for intermolecular interactions.<sup>62,67,68</sup> In PCP-SAFT-based cDFT, the residual Helmholtz free-energy functional can be decomposed into the same physical contributions as in the bulk PCP-SAFT EoS, namely, hard-sphere, hard-chain, and dispersion terms, and, when relevant, association and polar terms,<sup>62,68</sup>

$$F^{\text{res}}[\rho(\mathbf{r})] = F^{\text{hs}}[\rho(\mathbf{r})] + F^{\text{chain}}[\rho(\mathbf{r})] + F^{\text{disp}}[\rho(\mathbf{r})] + F^{\text{assoc}}[\rho(\mathbf{r})] + F^{\text{polar}}[\rho(\mathbf{r})] \quad (5)$$

For further details on the free-energy functional forms and cDFT, the reader is referred to refs 62,64,66,69. In this study, the equilibrium density profiles are obtained by solving the Euler–Lagrange equation<sup>60,64</sup> from eq 4, which combines the different contributions to the free-energy functional. Once the equilibrium density profiles are obtained, the IFTs are computed from the excess grand potential per unit interfacial area  $A$  as

$$\gamma = \frac{\Omega^{\text{Eq.}} - \Omega^{\text{Bulk}}}{A} \quad (6)$$

where  $\Omega^{\text{Eq.}}$  and  $\Omega^{\text{Bulk}}$  are the equilibrium and bulk grand potentials, respectively. The interfacial thickness ( $L_{10}^{\text{eq}}$ ), adsorption, and enrichment of components at the interface are computed from the density profiles.<sup>25,70</sup> The Euler–Lagrange equations are solved numerically in one-dimensional Cartesian coordinates, and the required functional derivatives are evaluated by automatic differentiation.<sup>33</sup> The interfacial domain is discretized with adaptive grid width and number of grid points, with interface widths ranging from 100 to 1000 Å and grid sizes from 1024 to 2048 points, where finer discretization is applied near the critical region. The density profiles are computed using the default cDFT solver settings in FeOs,<sup>60</sup> which uses two sequential Anderson mixing steps<sup>68,71</sup> with a relative density residual tolerance of  $10^{-11}$ . Although the implementation is computationally efficient in FeOs,<sup>60</sup> the computation of interfacial properties for a full  $PT$  phase diagram can still be computationally demanding for multi-

component mixtures with temperature-dependent  $k_{ij}$ . This is because the equilibrium density profiles are obtained from coupled nonlinear Euler–Lagrange equations, which must be solved iteratively over the spatial domain for all components in the mixture at each state point.<sup>68,72–74</sup> This cost becomes particularly important when repeated IFT computations are required, as in CFD and other process simulations. The pure-component parameters of the PCP-SAFT EoS<sup>57–59,63</sup> used to describe the bulk phase behavior and interfacial properties of mixtures comprising  $\text{CO}_2$ ,  $\text{H}_2$ ,  $\text{N}_2$ ,  $\text{Ar}$ ,  $\text{CH}_4$ ,  $\text{O}_2$ ,  $\text{CO}$ , and  $\text{H}_2\text{S}$  are  $m_i$ ,  $\sigma_{ii}$ ,  $\varepsilon_{ii}/k_B$ ,  $p_i$ ,  $Q_i$ ,  $n_A$ ,  $n_B$ ,  $\kappa^{\text{AB}}$ , and  $\varepsilon^{\text{AB}}/k_B$ . These parameters denote the number of spherical segments per molecule, segment size parameter, dispersion energy parameter, dipole moment, quadrupole moment, number of association sites of type A, number of association sites of type B, association volume, and association energy parameter, respectively. For the studied components, only  $\text{H}_2\text{S}$  was modeled with an explicit association contribution using a 2A – 2B association scheme, with  $n_A = 2$  and  $n_B = 2$ . Hence, cross-association interactions between different molecular species were not considered. The indices  $ii$  and  $jj$  denote interactions between segments of the same component, whereas interactions between dissimilar atom pairs were modeled using the Lorentz–Berthelot (LB) mixing rules.<sup>75,76</sup>

$$\sigma_{ij} = (1 - l_{ij}) \frac{1}{2} (\sigma_{ii} + \sigma_{jj}) \quad (7)$$

$$\varepsilon_{ij} = (1 - k_{ij}) (\varepsilon_{ii} \varepsilon_{jj})^{1/2} \quad (8)$$

where the indices  $ij$  refer to cross interactions between atoms of different types. The parameters  $k_{ij}$  and  $l_{ij}$  are the binary interaction parameters used in the mixing rules for  $\varepsilon_{ij}$  and  $\sigma_{ij}$ , respectively, for components  $i$  and  $j$ . Here,  $l_{ij}$  is not considered ( $l_{ij} = 0$ ), and only  $k_{ij}$  is considered. To accurately model the multicomponent  $\text{CO}_2$ -rich mixtures, temperature-dependent BIPs ( $k_{ij}$ ) were determined for each dissimilar  $\text{CO}_2$ -impurity pair by fitting to the available experimental  $P_{xy}$  data at a temperature where experimental data were available. Here,  $P_{xy}$  refers to isothermal vapor–liquid equilibrium data in which the equilibrium pressure,  $P$ , is reported together with the corresponding liquid-phase composition,  $x$ , and vapor-phase composition,  $y$ . The fitted BIPs, together with the pure-component PCP-SAFT EoS<sup>57–59,63</sup> parameters used in this study, are provided in Sections S1 and S2 of the Supporting Information. These parameters were used to compute the phase equilibria and interfacial properties of  $\text{CO}_2$  mixtures with impurities. In addition, Section S2 describes the method used to determine the temperature-dependent BIPs and lists the experimental  $P_{xy}$  data used for fitting. For an  $N$ -component mixture,  $N(N - 1)/2$  BIPs are required, so the 8-component mixture considered in this study requires 28 binary pairs. Experimental  $P_{xy}$  data were available for 23 pairs. For the remaining five pairs,  $\text{H}_2$ – $\text{O}_2$ ,  $\text{CH}_4$ – $\text{O}_2$ ,  $\text{O}_2$ – $\text{CO}$ ,  $\text{Ar}$ – $\text{H}_2\text{S}$ , and  $\text{O}_2$ – $\text{H}_2\text{S}$ , no experimental  $P_{xy}$  data were available in the literature. Hence,  $k_{ij}$  was set to zero for these pairs.

## 2.2. Winterfeld–Scriven–Davis (WSD) Model

A theory for estimating the IFT of multicomponent fluid mixtures was formulated by Winterfeld et al.<sup>41</sup> This theory extends the Fowler–Kirkwood–Buff framework developed<sup>77,78</sup> by simplifying the statistical mechanics expression for the calculation of IFT with the assumption of pairwise additive intermolecular potentials.<sup>41,77,78</sup> The resulting WSD model

provides a practical method for estimating IFTs of nonaqueous mixtures using the IFTs of the pure components present in the mixture and can be generalized to an  $N$ -component mixture as,

$$\gamma_{\text{mix}}(T, P, \mathbf{z}) = \sum_{\alpha=1}^N \sum_{\beta=1}^N \left[ \left( \frac{\rho_{\alpha}^{\text{L}} - \rho_{\alpha}^{\text{V}}}{\rho_{\alpha}^{0,\text{L}} - \rho_{\alpha}^{0,\text{V}}} \right)^2 \gamma_{\alpha}^0 \delta_{\alpha\beta} + \Phi_{\alpha\beta} \left( \frac{\rho_{\alpha}^{\text{L}} - \rho_{\alpha}^{\text{V}}}{\rho_{\alpha}^{0,\text{L}} - \rho_{\alpha}^{0,\text{V}}} \right) \left( \frac{\rho_{\beta}^{\text{L}} - \rho_{\beta}^{\text{V}}}{\rho_{\beta}^{0,\text{L}} - \rho_{\beta}^{0,\text{V}}} \right) \sqrt{\gamma_{\alpha}^0 \gamma_{\beta}^0} \right] \quad (9)$$

where  $\gamma_{\text{mix}}$  denotes the mixture interfacial tension, which depends on temperature  $T$ , pressure  $P$ , and mole fraction vector  $\mathbf{z} = (z_1, z_2, \dots, z_N)$ . Here,  $N$  is the number of components in the mixture, and the indices  $\alpha$  and  $\beta$  denote the components in the system. The double summation in eq 9 runs over all component pairs, with  $\alpha, \beta = 1, 2, \dots, N$ . The quantities  $\gamma_{\alpha}^0$  and  $\gamma_{\beta}^0$  represent the pure-component IFTs of components  $\alpha$  and  $\beta$ , respectively. Similarly,  $\rho_{\alpha}^{0,\text{L}}$ ,  $\rho_{\alpha}^{0,\text{V}}$ ,  $\rho_{\beta}^{0,\text{L}}$ , and  $\rho_{\beta}^{0,\text{V}}$  correspond to the equilibrium liquid and vapor molar densities, in  $\text{mol cm}^{-3}$ , of the pure components  $\alpha$  and  $\beta$ , respectively. The parameters  $\rho_{\alpha}^{\text{L}}$ ,  $\rho_{\alpha}^{\text{V}}$ ,  $\rho_{\beta}^{\text{L}}$ , and  $\rho_{\beta}^{\text{V}}$  denote the liquid and vapor-phase molar densities of components  $\alpha$  and  $\beta$  in the mixture.<sup>41</sup> The term  $\Phi_{\alpha\beta}$  is the cross-interaction, or mixing, parameter that accounts for nonideal interactions between dissimilar molecules. For nonaqueous and weakly polar fluids,  $\Phi_{\alpha\beta}$  is commonly approximated as 1 ( $\Phi_{\alpha\beta} \approx 1$ ).<sup>41</sup> The Kronecker delta  $\delta_{\alpha\beta}$  equals one when  $\alpha = \beta$ , corresponding to the pure-component contribution, and zero otherwise.<sup>41</sup> For an  $N$ -component mixture, eq 9 can also be written compactly in matrix form as  $\gamma_{\text{mix}}(T, P, \mathbf{z}) = \mathbf{a}^T \Phi \mathbf{a}$ , where the vector  $\mathbf{a}$  contains the density-scaled pure component IFT terms,  $a_{\alpha} = \left( \frac{\rho_{\alpha}^{\text{L}} - \rho_{\alpha}^{\text{V}}}{\rho_{\alpha}^{0,\text{L}} - \rho_{\alpha}^{0,\text{V}}} \right) \sqrt{\gamma_{\alpha}^0}$ , and  $\Phi$  is the symmetric matrix of cross-interaction parameters with diagonal elements  $\Phi_{\alpha\alpha} = 1$ . The WSD model has been shown to provide reliable IFT predictions for nonaqueous mixtures and therefore serves as a useful semiempirical framework when molecular-correlation data are unavailable.<sup>25,41,42</sup> The formulation becomes less reliable when one or more components in the mixture are supercritical at  $T$ . For a three-component mixture, assuming  $\Phi_{\alpha\beta} = 1$  for the cross-interaction term and treating component 3 as supercritical, eq 9 simplifies to

$$\begin{aligned} \gamma_{\text{mix}}(T, P, z_1, z_2, z_3) &= \left( \frac{\rho_1^{\text{L}} - \rho_1^{\text{V}}}{\rho_1^{0,\text{L}} - \rho_1^{0,\text{V}}} \right)^2 \gamma_1^0 + \left( \frac{\rho_2^{\text{L}} - \rho_2^{\text{V}}}{\rho_2^{0,\text{L}} - \rho_2^{0,\text{V}}} \right)^2 \gamma_2^0 \\ &+ \left( \frac{\rho_1^{\text{L}} - \rho_1^{\text{V}}}{\rho_1^{0,\text{L}} - \rho_1^{0,\text{V}}} \right) \left( \frac{\rho_2^{\text{L}} - \rho_2^{\text{V}}}{\rho_2^{0,\text{L}} - \rho_2^{0,\text{V}}} \right) \sqrt{\gamma_1^0 \gamma_2^0} \end{aligned} \quad (10)$$

In this form, component 3 does not contribute directly to the IFT because its pure-component interfacial tension is not defined at supercritical conditions. As a consequence, the WSD model neglects the direct contribution of the supercritical component, which can be physically limiting because component 3 can still influence the mixture phase behavior, density difference, and interfacial properties. In practice, component 3 is expected to lower the IFT of the subcritical mixture by weakening the cohesive forces between unlike molecules, similar to the action of a surfactant. For the  $\text{CO}_2$ -

rich mixtures with impurities considered in this study, all impurity components are supercritical at the investigated conditions, except  $\text{H}_2\text{S}$ . In future work, the Widom line<sup>79,80</sup> could be explored as a thermodynamic criterion for distinguishing liquid-like and gas-like supercritical states and for estimating the effective contribution of supercritical components to the mixture IFT. To represent this behavior, in our previous study, we introduced a correction term<sup>25</sup> for binary mixtures based on the total liquid-phase mole fraction of the supercritical components. Here, we will use the baseline WSD model eq 9 to compute IFT and then to train the residual-based hybrid ML model with respect to the PCP-SAFT EoS + cDFT.<sup>57,60,63</sup>

### 2.3. Machine Learning Models

The capabilities of ML models in predicting thermophysical properties of fluid mixtures considered in this study have been widely reported in the literature.<sup>45,81–87</sup> Hence, for brevity, we do not provide a detailed description of each model examined. Instead, we focus on the rationale for selecting the ML models, kernels, and hyperparameters used in this work. For the target properties, namely, dew-point pressure, bubble-point pressure, IFT, and interfacial thickness, we use the same set of input features to avoid overfitting and enable a consistent comparison. The selected features are physically well motivated because the feature set contains only the thermodynamic state variables  $T$ ,  $P$ , and  $\mathbf{z}$ , which are the same variables used to define phase equilibrium and IFT measurements. The supervised ML models considered in this study include Support Vector Machine (SVM),<sup>88,89</sup> Random Forest (RF),<sup>90</sup> Gradient Boosting (XGB),<sup>91,92</sup> and Tabular Prior-data Fitted Network (TabPFN).<sup>85,86,93</sup> SVM uses kernel functions to construct linear or nonlinear models in the feature space. Although it is commonly used for classification, it can also be applied to regression through Support Vector Regression (SVR), which is suitable for continuous thermophysical property prediction.<sup>89</sup> In SVR, the model fits a regression function like linear regression while allowing errors within a tolerance region called the  $\epsilon$ -insensitive tube. Data points on or outside this tube are the support vectors and mainly determine the fitted model. In this work, SVR with a Radial Basis Function (RBF) kernel was used for all target properties. The main hyperparameters are the regularization parameter  $C$ , the tube width  $\epsilon$ , and the kernel coefficient  $\gamma_{\text{RBF}}$ .<sup>89</sup>  $C$  controls the penalty for errors outside the tube,  $\epsilon$  defines the error tolerance, and  $\gamma_{\text{RBF}}$  controls the length scale of the RBF kernel.<sup>89</sup> RF is an ensemble ML method composed of multiple decision trees.<sup>90</sup> Each tree is trained on a bootstrapped sample of the training data set, where data points are selected randomly with replacement. At each node, the optimal split is selected from a random subset of input features, reducing correlation over individual trees and improving model robustness. For regression, the final prediction is obtained by averaging the outputs of all trees. The key hyperparameters of the RF model include the total number of trees,  $n_{\text{estimators}}$ , the maximum allowable depth of each tree,  $d_{\text{max}}$ , the minimum sample count required to divide an internal node,  $n_{\text{split}}$ , the minimum number of samples retained in a terminal leaf,  $n_{\text{leaf}}$ , and the number of input features randomly evaluated when determining each split,  $n_{\text{features}}$ .<sup>90</sup> XGB is a gradient-boosted decision-tree algorithm in which trees are added sequentially to correct the errors of the previous model.<sup>91,92</sup> At each iteration, a new tree is fitted to the negative gradient of the loss function,

and its contribution is scaled by a learning rate. The final prediction is obtained by summing the contributions from all trees. Compared with Adaptive Boosting (AdaBoost),<sup>94</sup> XGB uses a more general gradient-based optimization framework and includes regularization terms that help reduce overfitting. Compared with Categorical Boosting (CatBoost),<sup>95</sup> XGB is commonly used for numerical tabular data, whereas CatBoost is particularly designed for data sets with categorical features. The main hyperparameters of XGB besides  $n_{\text{estimators}}$  and  $d_{\text{max}}$  are sample fraction, feature fraction, and regularization parameters.<sup>91,92</sup> In addition to SVR and decision-tree-based models, this work uses the state-of-the-art Tabular Prior-data Fitted Network (TabPFN) deep-learning model.<sup>85</sup> TabPFN is a transformer-based tabular foundation model designed for heterogeneous tabular data with different feature scales, missing values, and mixed data types, where conventional neural networks (NNs) often struggle, and tree-based models usually perform strongly.<sup>85</sup> This makes TabPFN suitable for thermodynamic property prediction, where the inputs are structured state variables such as  $T$ ,  $P$ , and  $\mathbf{z}$ , and the targets are continuous phase-equilibrium and interfacial properties. TabPFN is pretrained on millions of synthetic tabular data sets and uses In-Context Learning (ICL), the same mechanism used in large language models, to make predictions on new data sets with little or no data set-specific training.<sup>85</sup> As such, extensive hyperparameter tuning is not strictly required, although HPO using Bayesian optimization can further adapt the model to a specific data set.<sup>85,93</sup> In this work, the TabPFN HPO was performed using  $N_{\text{trials}} = 100$ , where each trial sampled a different hyperparameter configuration, fitted the TabPFN model, and evaluated its performance using CV. Additionally, an RF-TabPFN hybrid model is also considered to combine the robustness of tree-based learning with the strong small-data prediction ability of TabPFN.<sup>93</sup> The hyperparameter tuning procedures for SVM, RF, XGB, and TabPFN, together with the corresponding optimal hyperparameter values, are provided in Section S4 of the Supporting Information. The SVM and RF models were implemented using the scikit-learn<sup>96</sup> ML library in Python, XGB was implemented using the XGBoost Python package,<sup>92</sup> and TabPFN was implemented using the TabPFN-v2.5 Python package.<sup>85,86</sup> The HPO of TabPFN and RF-TabPFN models were implemented using the TabPFN-v0.4 extensions package.<sup>93</sup> All of the ML models were trained on CPUs, except for the TabPFN-based models, which were run on an NVIDIA H100 GPU.

In addition to data-driven ML models, a residual-learning strategy was used for the IFT. The computationally inexpensive WSD model<sup>41</sup> was taken as the baseline, and ML models were trained to predict the residual between the IFT obtained from the PCP-SAFT EoS + cDFT<sup>60</sup> computation and the WSD prediction.<sup>25,41</sup> This residual represents the missing correction to the semiempirical WSD model. Three ML models were considered for this task: Symbolic Regression (SR),<sup>97</sup> Gaussian Process Regression (GPR),<sup>98</sup> and Stochastic Variational Gaussian Process Regression (SVGP).<sup>87,99,100</sup> SR<sup>97</sup> was used to examine whether the residual can be represented by a compact and physically interpretable analytical expression, especially for mixtures containing supercritical components where the WSD model may lack  $T$ ,  $P$ , and composition-dependent effects. Following the SR framework of Cranmer,<sup>97</sup> SR searches over mathematical expressions generated from selected features,

operators, and numerical constants, and identifies expressions that balance predictive accuracy and complexity. The exact GPR was chosen because it is widely used to learn smooth, unknown residual functions while also providing predictive uncertainty.<sup>45</sup> In GPR, the residual function is described by a Gaussian process (GP) with a mean function  $m(\mathbf{x})$  and a covariance kernel  $k(\mathbf{x}, \mathbf{x}')$ .<sup>45,98</sup> Here,  $\mathbf{x}$  and  $\mathbf{x}'$  denote two input feature vectors, such as thermodynamic states defined by  $T$ ,  $P$ , and composition  $\mathbf{z}$ . In this work, the Matérn-5/2 kernel was used, with the main hyperparameters being the kernel variance  $\sigma_f^2$ , the length scale  $l$ , and the noise variance  $\sigma_n^2$ .<sup>45,98</sup> These parameters control the magnitude of the residual variation, the smoothness of the residual function, and the unresolved numerical or modeling noise, respectively, and were optimized by maximizing the Log Marginal Likelihood (LML).<sup>45,98</sup> Since the standard or exact GPR requires building and factorizing on the full  $N \times N$  kernel matrix, its training time scales approximately as  $O(N^3)$  and its memory requirement as  $O(N^2)$ , where  $N$  is the number of training points.<sup>98</sup> To improve scalability, Stochastic Variational Gaussian Process Regression (SVGP)<sup>87,99,100</sup> was also used as an approximate GP model based on  $M$  inducing points, where  $M \ll N$ . The SVGP model used a constant mean function and a Matérn-5/2 kernel with Automatic Relevance Determination (ARD), allowing each input feature to have an individual length scale and thereby providing a measure of feature relevance. A Cholesky variational distribution<sup>99</sup> was used to approximate the posterior distribution over the inducing function values, and the inducing locations were optimized during training. The model was trained using the Evidence Lower Bound Objective (ELBO),<sup>101,102</sup> which balances a data-fit term with a regularization term based on the Kullback–Leibler (KL) divergence<sup>103</sup> between the variational posterior and the prior.<sup>87,100</sup> To account for the highly skewed residual distribution, 2000 inducing points were initialized through stratified sampling. This sampling covered the  $PT$  region where most residuals are concentrated, as well as the midtail and extreme-tail residual regions. The target residuals were transformed using a quantile transformation followed by standardization, and the ELBO was maximized using mini-batch Adam optimization with a batch size of 1024, tail-weighted sampling, over 500 epochs, with the learning rate cosine-annealed from  $10^{-2}$  to  $10^{-4}$ . The SVGP model was implemented using the GPyTorch package,<sup>104</sup> while the exact GPR model was implemented using the scikit-learn library in Python.<sup>96</sup> For SR, the symbolic-regression framework-v1.5 of Cranmer<sup>97</sup> was used in Julia.

## 2.4. Composition Space Sampling

To compute the saturation pressures, namely, bubble and dew pressures, IFT, and interfacial thickness for compositions relevant to  $\text{CO}_2$  transportation, a pool of mixture compositions was generated. Since transported  $\text{CO}_2$  streams typically contain high  $\text{CO}_2$  mole fractions, the composition space was mainly focused on mixtures with  $z_{\text{CO}_2}$  in the range of 95–99.99%, while the remaining fraction was distributed between the impurity components. To distribute the impurity concentration, we used Dirichlet sampling,<sup>105</sup> where the  $\text{CO}_2$  mole fraction was fixed initially, and the remaining fraction was treated as the impurity budget,  $s = 1 - z_{\text{CO}_2}$ . The Dirichlet distribution is defined on a simplex, which is the geometric space of non-negative composition vectors whose components

sum to unity. As a result, for  $k$  impurity components, the sampled normalized impurity vector lies on a  $(k - 1)$ -dimensional simplex.<sup>105</sup> In this study, with seven impurity components considered, the Dirichlet distribution samples points on a six-dimensional impurity simplex. A normalized impurity composition vector was then sampled from a Dirichlet distribution,<sup>105</sup>

$$(\tilde{z}_1, \dots, \tilde{z}_k) \sim \text{Dir}(\alpha_1, \dots, \alpha_k), \quad \tilde{z}_i \geq 0, \quad \sum_{i=1}^k \tilde{z}_i = 1 \quad (11)$$

where  $k$  is the number of impurity components. The impurity mole fractions were obtained by scaling this normalized vector with the impurity budget,

$$z_i = s\tilde{z}_i = (1 - z_{\text{CO}_2})\tilde{z}_i, \quad i = 1, \dots, k \quad (12)$$

Thus, the final composition of a mixture satisfies

$$\begin{aligned} z_{\text{CO}_2} + \sum_{i=1}^k z_i &= z_{\text{CO}_2} + (1 - z_{\text{CO}_2}) \sum_{i=1}^k \tilde{z}_i \\ &= z_{\text{CO}_2} + (1 - z_{\text{CO}_2}) = 1 \end{aligned} \quad (13)$$

This procedure ensures that all impurity mole fractions are non-negative, the impurity fractions sum exactly to the prescribed remaining fraction, and the  $\text{CO}_2$  mole fraction remains fixed at the selected level. Apart from Dirichlet sampling,<sup>105</sup> systematic grid scaling was used to ensure coverage of highly impure  $\text{CO}_2$  conditions, corresponding to  $z_{\text{CO}_2} = 95\text{--}96\%$  with an impurity combination (binary mixtures). The final composition pool was generated using 70% Dirichlet sampling, 20% systematic grid scaling, and 10% compositions from industrial specifications and reported  $\text{CO}_2$  transport streams.<sup>6</sup> The industrial compositions were taken from the review article of Simonsen et al.,<sup>6</sup> including quality specifications from transportation projects such as Northern Lights<sup>8</sup> and Porthos,<sup>7</sup> capture technologies such as post-combustion, precombustion, oxy-fuel combustion, and operational  $\text{CO}_2$  pipelines in the USA and Canada.<sup>6</sup> Trace impurities reported in parts per million, such as  $\text{SO}_x$ ,  $\text{NO}_x$ , and  $\text{NH}_3$ , were neglected, and their fractions were assigned to  $\text{CO}_2$ . Since seven impurities were considered in this study, the possible combinations of impurity types and their concentrations are numerous. To address this, three sampling strategies, namely, random sampling, stratified sampling, and active learning (AL), were used to circumvent the computational cost of evaluating every mixture in the composition space and to identify the minimum number of mixtures required for reliable ML model performance. For random sampling, a target set of  $\mathcal{S}$  mixture compositions was selected from the full composition pool of size  $\mathcal{P}$  using a fixed random seed. The pool  $\mathcal{P}$  contained compositions generated from Dirichlet sampling, systematic grid scaling, and industrially reported feed specifications. For stratified sampling, the  $\text{CO}_2$  mole fraction range was divided into five bins. Within each  $\text{CO}_2$  bin, the impurity compositions were further partitioned into five clusters using  $k$ -means clustering based on Euclidean distance.<sup>106</sup> The number of impurity clusters was selected based on the elbow method and the silhouette score.<sup>106–108</sup> This procedure resulted in 25 strata, combining 5  $\text{CO}_2$  composition bins with 5 impurity composition clusters. The stratified sampling algorithm and its detailed implementation are provided in Section S3 of the Supporting Information. In

addition to random and stratified sampling, a kernel-based AL scheme was adopted to reduce the computational cost associated with evaluating a large number of candidate mixtures in the composition pool  $\mathcal{P}$ . The AL approach used in this work is shown in Algorithm 1. This approach was adapted from the work of Xiang et al.,<sup>109</sup> where it was originally applied to chemical compound-space exploration of alkanes. In the present work, the algorithm was modified for composition-space sampling. A GP regressor with an ARD Matérn-5/2 kernel, signal variance  $\sigma_f^2$ , and additive noise variance  $\sigma_n^2$  was first fitted to the pre-existing  $P_{\text{bubble}}$  labeled set  $\mathcal{T}_{\text{lab}}$  using standardized composition vectors  $\tilde{\mathbf{x}}$ . The vector  $\tilde{\mathbf{x}}$  denotes the scaled form of the previously generated composition vector, obtained by mean-centering each composition variable and normalizing it by its standard deviation before fitting the GPR model.<sup>98</sup> Xiang et al.<sup>109</sup> similarly used the previously trained GPR model<sup>110</sup> on MD simulations and the NIST Standard Reference Database<sup>111</sup> with the marginalized graph kernel. However, in this work, we generated a composition pool of size 100 and computed the  $P_{\text{bubble}}$  curve using the PCP-SAFT EoS<sup>57,63</sup> to train the GPR<sup>98</sup> model and determine its hyperparameters. The kernel hyperparameters were obtained by maximizing the LML, giving  $\theta^* = \{l^*, \sigma_f^{2*}, \sigma_n^{2*}\}$ . The optimized component-wise length scales provide a measure of the relevance of each species to  $P_{\text{bubble}}$ . The fitted noise term was then discarded, retaining only the signal kernel  $k_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \sigma_f^{2*} M_{5/2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j; l^*)$ , which defines an ARD-weighted metric in the composition space. Starting from two randomly selected compositions, the selected set  $\mathcal{S}$  was grown by adding the candidate with the largest posterior variance (uncertainty measure), defined as  $U(\tilde{\mathbf{x}}_i) = k_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i) - \mathbf{k}_S(\tilde{\mathbf{x}}_i)^T \mathbf{K}_{SS}^{-1} \mathbf{k}_S(\tilde{\mathbf{x}}_i)$ . This uncertainty depends on the position of each candidate mixture relative to the compositions already selected in  $\mathcal{S}$  and does not require additional PCP-SAFT EoS<sup>57,63</sup> evaluations during the selection step. For the Matérn-5/2 kernel, this procedure corresponds to a sequential maximum-variance design that selects new points from regions of composition space least covered by the current selected set. After each acquisition, candidates with  $U < u_t$  were removed from the remaining pool  $\mathcal{P}'$  because they were considered informationally redundant. In contrast to Xiang et al.,<sup>109</sup> who used a fixed uncertainty threshold  $u_t$  as the stopping criterion,  $u_t$  was adjusted in this work according to the target sample size  $N$ , so that the AL loop terminates when the selected set reaches  $|\mathcal{S}| = N$ . The candidate pool  $\mathcal{P}$  consisted of 1000 multicomponent  $\text{CO}_2$  mixtures. An initial subset, denoted as  $\mathcal{P}_{\text{init}}$ , containing 100 multicomponent  $\text{CO}_2$  mixtures was generated from this pool and labeled by computing  $P_{\text{bubble}}$  using PCP-SAFT EoS.<sup>57,63</sup> These labeled compositions were then used to pretrain the GPR model used in the acquisition step. The same candidate pool  $\mathcal{P}$  was used for AL, random sampling, and stratified sampling. Each sampling strategy was repeated for 20 independent trials. The reported metrics correspond to the trial-averaged values, and the uncertainty is quantified using the 95% confidence interval for the independent trials. The temperature, pressure, and composition ranges spanned by the resulting training data set are summarized in Table 1. Predictions outside these ranges constitute extrapolation, for which the reported error metrics do not apply.

**Table 1. Temperature, Pressure, and Composition Ranges of the Data Set Used to Train the ML Models, Generated Using the PCP-SAFT EoS + cDFT Framework<sup>a</sup>**

| variable                   | minimum | maximum |
|----------------------------|---------|---------|
| $T/[K]$                    | 200     | 300     |
| $P/[bar]$                  | 2.62    | 162.88  |
| $z_{CO_2}/[mol\ mol^{-1}]$ | 0.95    | 0.99    |
| $z_{H_2}/[mol\ mol^{-1}]$  | 0       | 0.0313  |
| $z_{Ar}/[mol\ mol^{-1}]$   | 0       | 0.05    |
| $z_{N_2}/[mol\ mol^{-1}]$  | 0       | 0.04    |
| $z_{CH_4}/[mol\ mol^{-1}]$ | 0       | 0.05    |
| $z_{O_2}/[mol\ mol^{-1}]$  | 0       | 0.04    |
| $z_{CO}/[mol\ mol^{-1}]$   | 0       | 0.03    |
| $z_{H_2S}/[mol\ mol^{-1}]$ | 0       | 0.03    |

<sup>a</sup>The data set comprises 19,361 state points (ca.  $1.9 \times 10^4$ ) evaluated for 100 training feeds sampled from the composition pool with sampling fractions ( $f_{Dirichlet}, f_{systematic}, f_{industrial}$ ) = (0.70, 0.20, 0.10). For each feed, state points were generated at temperatures from 200 to 300 K in steps of 5 K, with 10 pressures spaced equally between the dew-point and bubble-point pressures at each temperature. The impurity mole fractions have a minimum of zero because not every impurity is present in every mixture. Predictions outside the reported ranges constitute extrapolation, for which the reported error metrics do not apply.

**Algorithm 1** CO<sub>2</sub>-transportation composition-space Active Learning (AL) based on Automatic Relevance Determination (ARD)-kernel uncertainty. A Gaussian Process Regression (GPR) model with an ARD Matérn-5/2 kernel is fitted to the labeled bubble-point pressure set,  $T_{lab}$ , and the optimal hyperparameters,  $\theta^*$ , are obtained by maximizing the Log Marginal Likelihood (LML).

**Input:**  $\mathcal{P}$ : candidate composition pool;  $T_{lab}$ :  $P_{bubble}$  labels;  $N$ : target size.

**Output:**  $\mathcal{S}$ : selected composition set of target size  $N$ .

```

 $\mathcal{S} \leftarrow \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}\}, \quad \mathcal{P}' \leftarrow \mathcal{P} \setminus \mathcal{S}$ 
Fit GPR using  $k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \sigma_f^2 M_{5/2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j; \ell) + \sigma_n^2 \delta_{ij}$ 
 $\theta^* \leftarrow \arg \max_{\theta} \log p(\mathbf{y} | \tilde{\mathbf{X}}, \theta), \quad \theta^* \leftarrow \{\ell^*, \sigma_f^{2*}, \sigma_n^{2*}\}$ 
 $k_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \leftarrow \sigma_f^{2*} M_{5/2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j; \ell^*)$ 

while  $|\mathcal{S}| < N$  and  $\mathcal{P}' \neq \emptyset$  do
     $U(\mathbf{x}_i) \leftarrow k_s(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i) - \mathbf{k}_S(\tilde{\mathbf{x}}_i)^T \mathbf{K}_{SS}^{-1} \mathbf{k}_S(\tilde{\mathbf{x}}_i), \quad \forall \mathbf{x}_i \in \mathcal{P}'$ 
     $\mathbf{x}^* \leftarrow \arg \max_{\mathbf{x}_i \in \mathcal{P}'} U(\mathbf{x}_i)$ 
     $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{x}^*\}, \quad \mathcal{P}' \leftarrow \mathcal{P}' \setminus \{\mathbf{x}^*\}$ 
     $\mathcal{P}' \leftarrow \mathcal{P}' \setminus \{\mathbf{x}_i \in \mathcal{P}' : U(\mathbf{x}_i) < u_t\}$ 
end while
return  $\mathcal{S}$ 

```

## 2.5. Model Evaluation

To evaluate the quantitative performance of the regression ML models, we used the following metrics: the coefficient of determination ( $R^2$ ), the root-mean-square error (RMSE), and the mean absolute error (MAE). These metrics are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (14)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (15)$$

and

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (16)$$

Here,  $N$  is the total number of data points,  $i$  indexes each data point,  $y_i$  is the reference value of the  $i$ th data point,  $\hat{y}_i$  is the

corresponding predicted value from the ML model, and  $\bar{y}$  is the mean of the reference values. The coefficient of determination ( $R^2$ ) measures how well the predicted values from an ML model reproduce the variation in the reference data. Values of  $R^2$  closer to unity indicate better predictive performance. The RMSE measures the average prediction error in the same physical units as the target property. Because the errors are squared before averaging, larger deviations are penalized more strongly. The MAE measures the average absolute prediction error and is less affected by large individual deviations than the RMSE. To compare prediction errors of target properties with different physical units and numerical ranges, the Normalized Root Mean Square Error (NRMSE) and Normalized Mean Absolute Error (NMAE) were also evaluated. These metrics are defined as

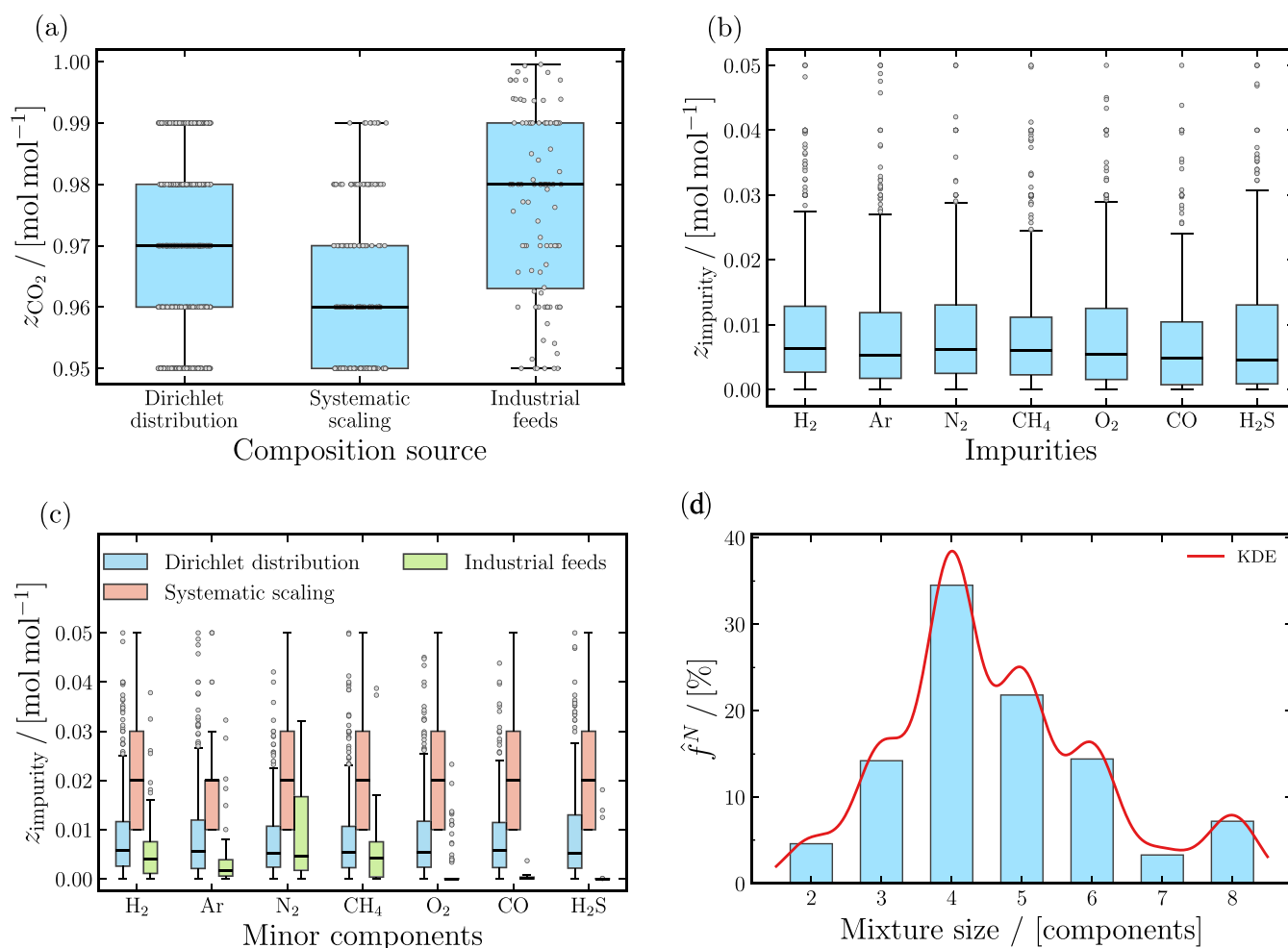
$$NRMSE = \frac{RMSE}{y_{\max} - y_{\min}} \quad (17)$$

$$NMAE = \frac{MAE}{y_{\max} - y_{\min}} \quad (18)$$

where  $y_{\max}$  and  $y_{\min}$  denote the maximum and minimum reference values of the corresponding target property, respectively. To ensure consistency, the same composition sets were used to compute  $P_{bubble}$ ,  $P_{dew}$ , IFT, interfacial thickness, and the WSD residual using PCP-SAFT EoS + cDFT. The data sets were split into training, testing, and validation sets with a 0.70:0.15:0.15 ratio, and the same split was used for all ML models for each target property. To evaluate model robustness, a 5-fold Cross Validation (CV) was performed on the training set. In a 5-fold CV, the training set was divided into five subsets. For each CV run, four subsets were used to train the model, while the remaining subset was used for validation. This process was repeated five times so that each subset served once as the validation set. The resulting fold-wise metrics were then used to assess model robustness and identify possible overfitting or underfitting of the ML model.

## 3. RESULTS AND DISCUSSION

This section evaluates the performance of machine learning (ML) models in predicting the  $PT$  phase envelope, including bubble- and dew-point pressures, and the interfacial properties ( $\gamma$  and  $L_{10}^{90}$ ) of multicomponent CO<sub>2</sub>-rich mixtures. The reference phase envelope and interfacial properties were computed using the PCP-SAFT EoS + cDFT<sup>57,59,60</sup> of the CO<sub>2</sub>-rich mixtures. Section 3.1 describes the composition pool of CO<sub>2</sub>-rich mixtures considered in this study. The effect of impurities on  $P_{bubble}$ ,  $P_{dew}$ ,  $\gamma$ , and  $L_{10}^{90}$  is presented and discussed in Section 3.2. Section 3.3 evaluates the relationships between the target properties,  $P_{bubble}$ ,  $P_{dew}$ ,  $\gamma$ , and  $L_{10}^{90}$ , and the input features,  $T$ ,  $P$ , and  $\mathbf{z}$ , using the Spearman rank correlation<sup>112</sup> and the Kolmogorov–Smirnov (KS)<sup>113</sup> statistical metrics. Two supervised ML-based modeling strategies are considered in this work. For the full data-driven ML approach, Random Forest (RF),<sup>90</sup> Support Vector Regression (SVR)<sup>88</sup> with a Radial Basis Function (RBF) kernel, Gradient Boosting (XGB),<sup>91,92</sup> and Tabular Prior-Data Fitted Network (TabPFN)<sup>86</sup> were used. The performance of these models for  $P_{bubble}$ ,  $P_{dew}$ ,  $\gamma$ , and  $L_{10}^{90}$  is presented in Section 3.4. In the full data-driven models, the thermodynamic state variables  $T$  and  $P$ , together with the feed composition vector  $\mathbf{z}$ , were used as input features. The



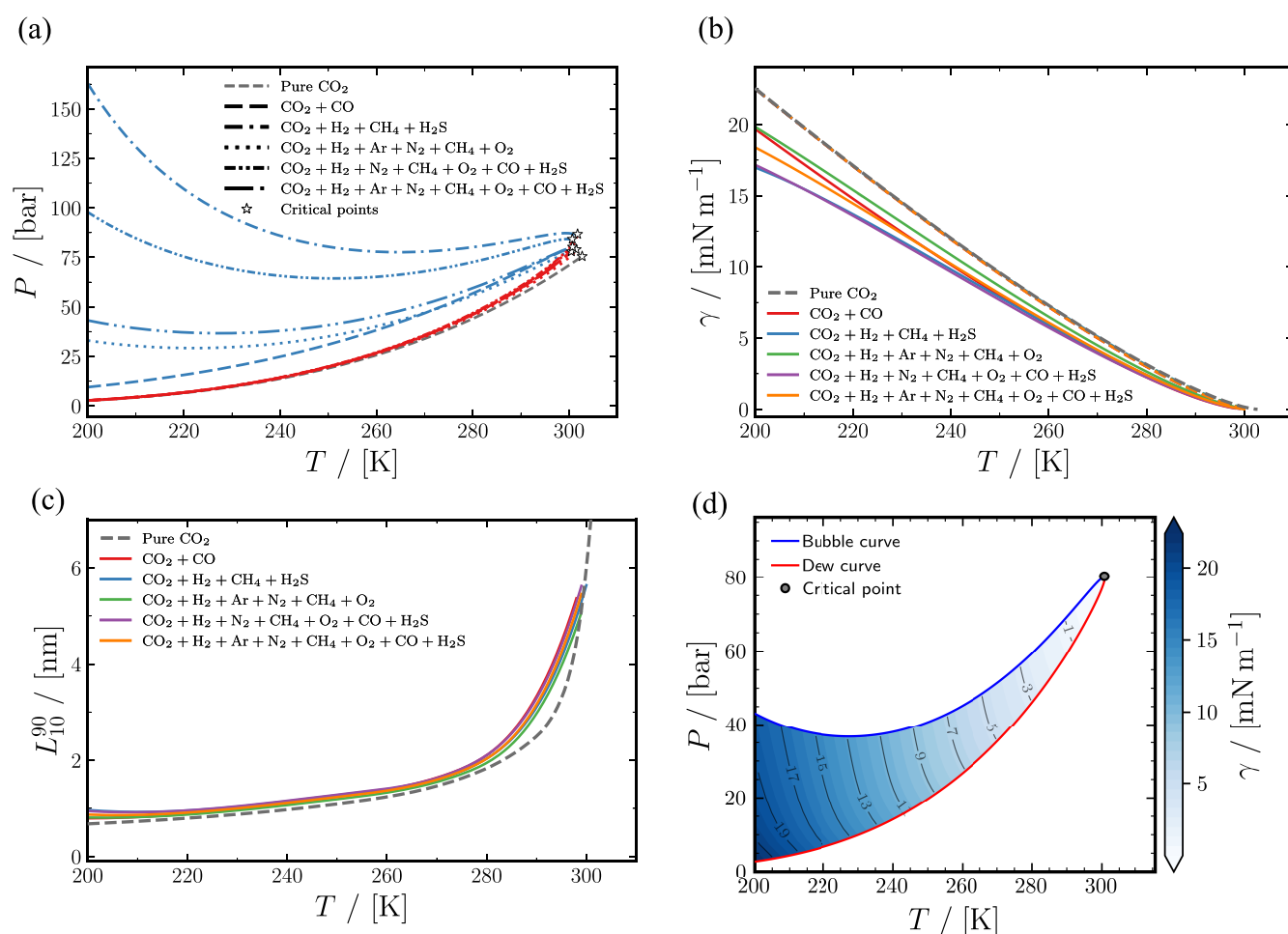
**Figure 1.** Composition-pool characteristics for the CO<sub>2</sub>-rich mixtures considered in this study for computing the phase envelope and interfacial properties. The composition pool  $\mathcal{P}$  comprises  $N = 1000$  compositions. (a)  $z_{\text{CO}_2}$  distribution grouped by each composition source. The markers represent the individual  $z_{\text{CO}_2}$  compositions from each feed obtained using the Dirichlet distribution,<sup>105</sup> systematic scaling, and industrial feeds.<sup>6</sup> (b) Box plots of the impurity mole fractions,  $z_{\text{impurity}}$  distribution for H<sub>2</sub>, Ar, N<sub>2</sub>, CH<sub>4</sub>, O<sub>2</sub>, CO, and H<sub>2</sub>S in the composition pool. (c) Comparison of  $z_{\text{impurity}}$  distributions separated by composition source, illustrating the different  $z_{\text{impurity}}$  ranges generated by the Dirichlet distribution,<sup>105</sup> systematic scaling, and industrial feeds.<sup>6</sup> (d) Normalized histogram of the mixture-size distribution. Bars give the percent-normalized relative frequency  $\hat{f}_i^N$ , and the solid red line shows the kernel density estimation (KDE). The pool spans binary to eight-component mixtures, with the upper tail contributed by the industrial feed sampling fraction.

effects of the number of compositions and the sampling strategy on the predictive accuracy of the aforementioned ML models are evaluated in Section 3.5, using random, stratified, and active learning (AL) sampling methods. In addition to the full data-driven models, Section 3.6 discusses hybrid residual ML models for improving the interfacial tension (IFT) predicted by the semiempirical Winterfeldt–Scriven–Davis (WSD) correlation.<sup>41</sup> In this residual-learning approach, Symbolic Regression (SR),<sup>97</sup> Gaussian Process Regression (GPR),<sup>98</sup> and Stochastic Variational Gaussian Process Regression (SVGP)<sup>87,100,104</sup> were trained on the difference between the WSD prediction and the PCP-SAFT EoS + cDFT<sup>57,59,60</sup> reference IFT.

### 3.1. Composition Pool

In this work, multicomponent CO<sub>2</sub>-rich mixtures were generated with the constraint  $z_{\text{CO}_2} \geq 95\%$  to represent compositions relevant to CO<sub>2</sub> transportation. The remaining impurity fraction,  $z_{\text{impurity}} \leq 5\%$ , was distributed between the impurity components using two sampling approaches: the

Dirichlet distribution and systematic scaling. For each generated mixture,  $z_{\text{CO}_2}$  was then determined from the total impurity fraction to satisfy the composition closure condition. In addition, industrial feed compositions reported by Simonsen et al.<sup>6</sup> were included to introduce realistic CO<sub>2</sub> transport mixtures into the composition pool. We considered a pool of 1000 CO<sub>2</sub>-rich compositions. Figure 1a shows the CO<sub>2</sub> mole fraction distribution for the three composition sources: Dirichlet distribution, systematic scaling, and industrial feeds. The individual  $z_{\text{CO}_2}$  compositions are plotted as markers, and the outliers are not shown separately. The middle 50% of the data, represented by the Interquartile Range (IQR), lies between approximately  $z_{\text{CO}_2} = 0.96$  and  $0.98$  for the Dirichlet distribution, between  $z_{\text{CO}_2} = 0.95$  and  $0.97$  for systematic scaling, and between  $z_{\text{CO}_2} = 0.96$  and  $1.00$  for the industrial feeds. The CO<sub>2</sub> mole fractions from the Dirichlet distribution and systematic scaling appear at fixed values, such as  $0.96$  and  $0.97$ , because they were generated with two-decimal precision.



**Figure 2.** Sensitivity analysis of CO<sub>2</sub>-rich mixtures for the feed compositions listed in Table 2. (a) Bubble and dew-point pressure curves compared with pure CO<sub>2</sub>, with critical points. (b, c) Interfacial tension (IFT),  $\gamma$ , and interfacial thickness,  $L_{10}^{90}$ , at the bubble and dew equilibrium curves for the same CO<sub>2</sub>-rich mixtures and pure CO<sub>2</sub> saturation line. (d) Two-phase envelope for the feed 4 listed in Table 2, including the bubble curve, dew curve, critical point, and the contour distribution of  $\gamma$  within the two-phase region. The phase envelope and interfacial properties were computed using the PCP-SAFT EoS coupled with cDFT using FeOs.<sup>57–60</sup>

This discretization was intentionally introduced so that the ML models could learn and interpolate thermodynamic properties for different feed-composition levels. In contrast, the industrial feed compositions have more continuous  $z_{\text{CO}_2}$  values because they come from reported feed data. Figure 1b shows the  $z_{\text{impurity}}$  distribution in a box plot with outliers (markers) for each impurity component. The majority of  $z_{\text{impurity}}$  is  $<0.03$ , since the other impurity component takes the remaining fraction, which is more likely in multicomponent systems. The distribution of  $z_{\text{impurity}}$  with respect to the composition source is presented in Figure 1c. The Dirichlet-sampling compositions mainly have their IQR in the low-impurity region, whereas the systematic scaling compositions have their IQR at higher impurity fractions, mostly for  $z_{\text{impurity}} > 0.01$ . For the industrial feeds, the major impurities, such as H<sub>2</sub>, Ar, N<sub>2</sub>, and CH<sub>4</sub>, have a wide IQR range compared with minor impurities such as O<sub>2</sub>, CO, and H<sub>2</sub>S. This combined strategy ensures that the impurity composition space is sampled efficiently while keeping the mixtures representative of CO<sub>2</sub>-rich transport streams. For the generated composition pool of size  $|P| = 1000$ , Figure 1d presents the empirical distribution of mixture sizes, where the mixture size denotes the number of active components in each composition. The bars represent

the percent-normalized relative frequency, defined as  $\hat{f}_i^N = (n_i/N) \times 100\%$ , with  $\sum \hat{f}_i^N = 100\%$ , where  $n_i$  is the number of compositions containing exactly  $i$  components and  $N = |P|$ . The solid red line represents the kernel density estimation (KDE) of the same sample, plotted on the same percent-normalized axis,  $\hat{f}^N$ . Since the mixture size is integer-valued with unit bin width, the bar heights and KDE<sup>114,115</sup> values are directly comparable. The distribution is approximately bell-shaped and centered between four and five components. Eight-component mixtures originate mainly from the industrial feed composition source, whereas the Dirichlet sampling, due to the impurity-weighting parameter  $\alpha$ , favors lower-cardinality mixtures and rarely activates all seven impurities at the same time. Consequently, the shape of  $\hat{f}_i^N$  can be controlled through the sampling fractions ( $f_{\text{Dirichlet}}, f_{\text{systematic}}, f_{\text{industrial}}$ ) = (0.70, 0.20, 0.10), allowing the composition pool to be biased toward low, medium, or high cardinality mixture sets as required. The temperature, pressure, and composition ranges of the training data set generated from this pool are summarized in Table 1.

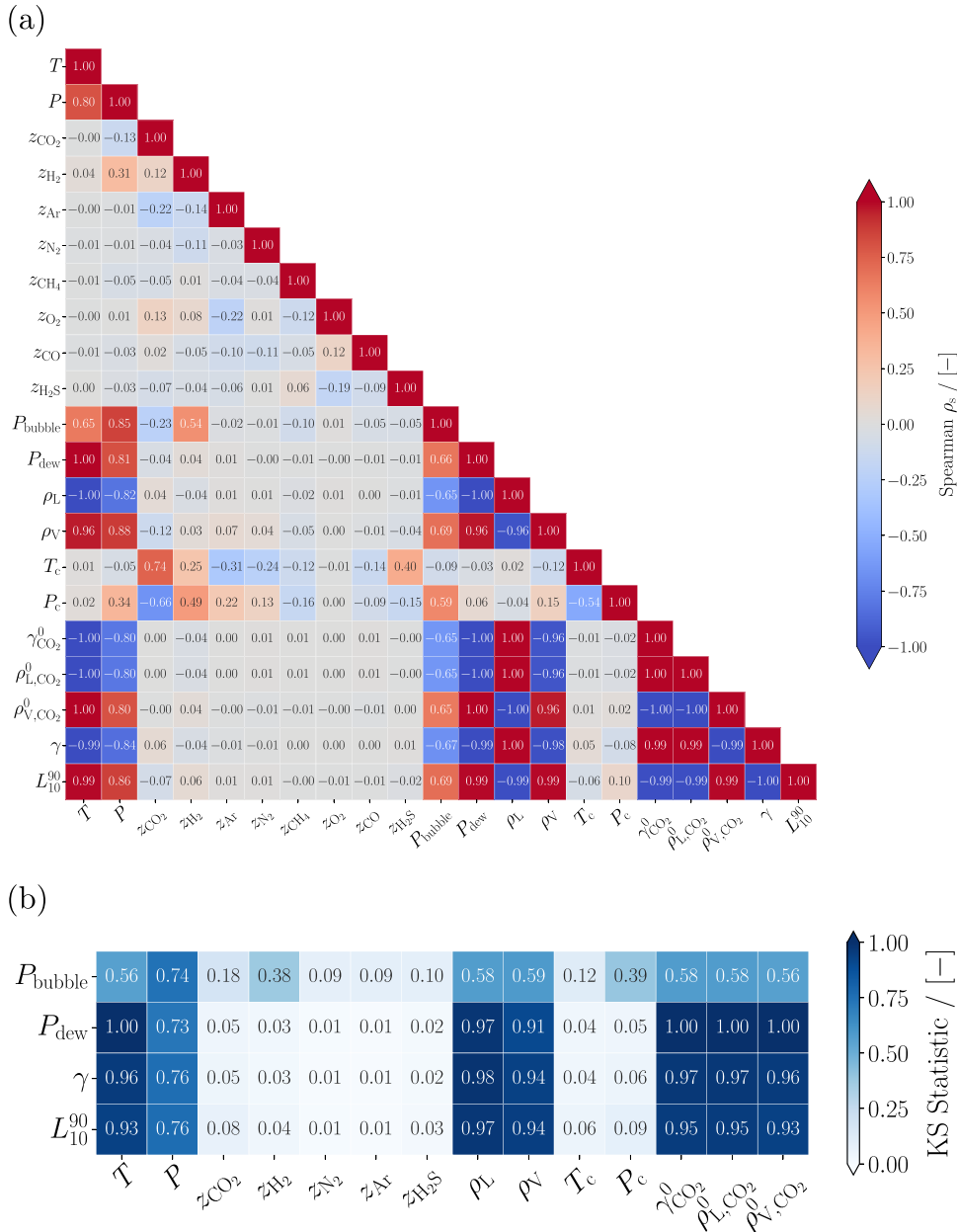
### 3.2. Effect of Impurities

Figure 2a–c presents the phase envelopes, interfacial tension (IFT), and interfacial thickness of selected CO<sub>2</sub>-rich mixtures computed using the PCP-SAFT EoS.<sup>58,60,63</sup> The corresponding

**Table 2.** Mixture Compositions  $z_i$  in Mole Fractions Used for the Sensitivity Analysis Shown in Figure 2<sup>a</sup>

| feed | $N_c$ | $z_{\text{CO}_2}$ | $z_{\text{H}_2}$ | $z_{\text{Ar}}$ | $z_{\text{N}_2}$ | $z_{\text{CH}_4}$ | $z_{\text{O}_2}$      | $z_{\text{CO}}$       | $z_{\text{H}_2\text{S}}$ |
|------|-------|-------------------|------------------|-----------------|------------------|-------------------|-----------------------|-----------------------|--------------------------|
| 1    | 2     | 0.9700            |                  |                 |                  |                   |                       | 0.0300                |                          |
| 2    | 4     | 0.9500            | 0.0313           |                 |                  | 0.0039            |                       |                       | 0.0148                   |
| 3    | 6     | 0.9800            | 0.0058           | 0.0091          | 0.0011           | 0.0029            | 0.0011                |                       |                          |
| 4    | 7     | 0.9600            | 0.0200           |                 | 0.0080           | 0.0040            | 0.0040                | 0.0020                | 0.0020                   |
| 5    | 8     | 0.9700            | 0.0075           | 0.0040          | 0.0079           | 0.0099            | $3.79 \times 10^{-5}$ | $6.99 \times 10^{-4}$ | $3.90 \times 10^{-7}$    |

<sup>a</sup>The feeds considered for the sensitivity analysis were selected from a composition pool of size 1000. The selected feeds span mixture sizes from 2 to 8 components.



**Figure 3.** Feature-relevance analysis for the CO<sub>2</sub>-rich composition pool  $\mathcal{P} = 100$ . (a) Spearman rank-correlation matrix<sup>112</sup> between the input features ( $T$ ,  $P$ ,  $z$ ,  $\rho_L$ ,  $\rho_V$ ,  $T_c$ ,  $P_c$ ,  $\gamma_{\text{CO}_2}^0$ ,  $\rho_{L,\text{CO}_2}^0$ , and  $\rho_{V,\text{CO}_2}^0$ ) and the target properties  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^{90}$ . (b) Kolmogorov–Smirnov (KS) statistic<sup>113</sup> between each input feature and target property. The Spearman coefficient,  $\rho_s$ , identifies monotonic feature–target relationships, while the KS statistic quantifies distributional differences and provides an additional measure of feature sensitivity.

feed compositions are listed in Table 2. These five feeds were selected from the composition pool,  $\mathcal{P}$ , to include different numbers of active components present in CO<sub>2</sub>-rich mixtures and to capture distinct  $P_{\text{bubble}}$  behavior. The phase envelopes in

Figure 2a are compared with the pure-CO<sub>2</sub> saturation line. The dew-point curves of the CO<sub>2</sub>-rich mixtures are nearly identical. Compared to the pure CO<sub>2</sub> saturation line, noticeable differences near the critical region were identified. Compared

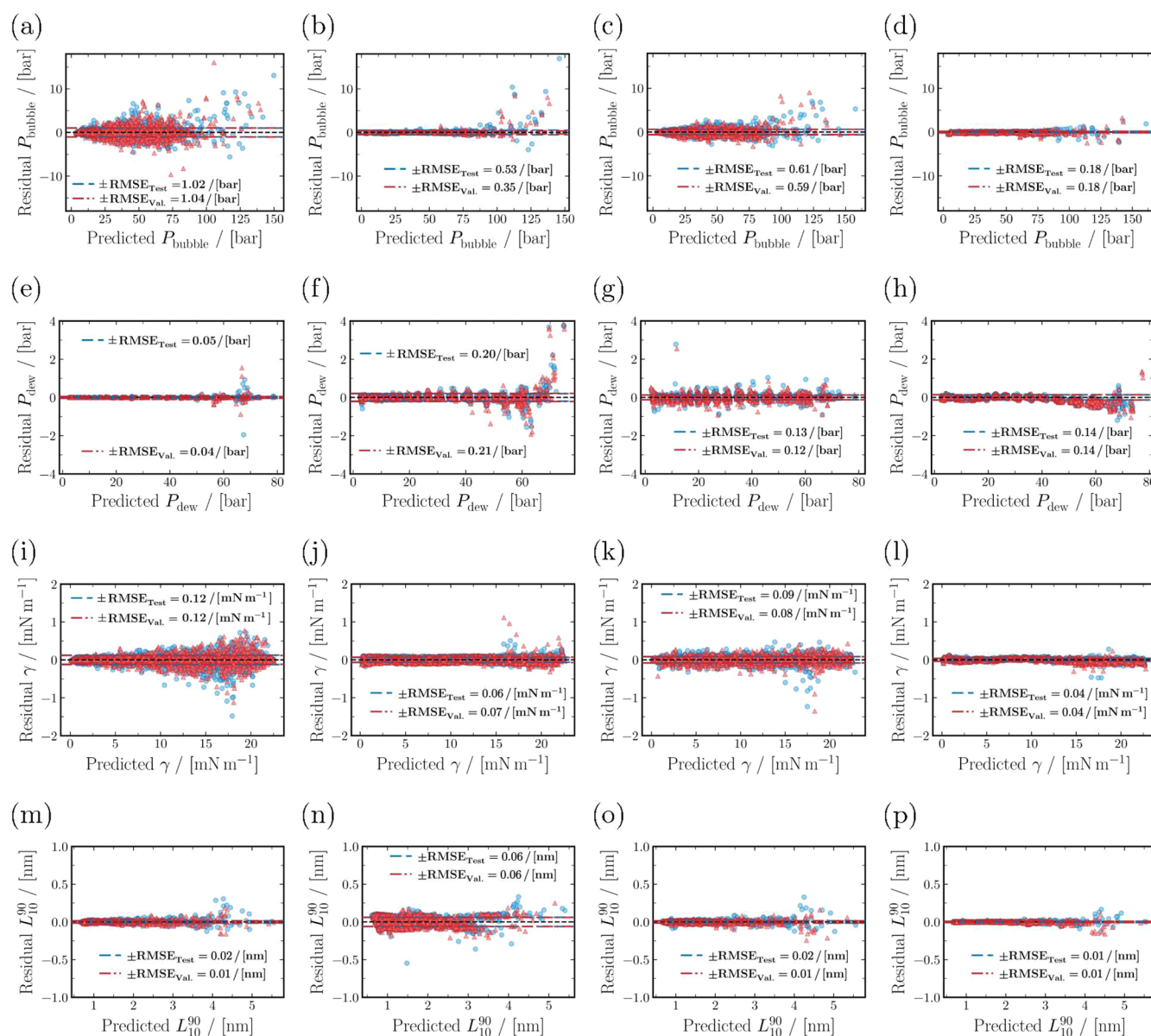
to dew-point curves, bubble-point curves are much more sensitive to impurity type and its composition. For the investigated composition range, most impurities broaden the phase envelope of CO<sub>2</sub>-rich mixtures, while H<sub>2</sub>S tends to narrow it. For CO<sub>2</sub> transportation, this broadening increases the range of temperatures and pressures over which vapor–liquid phase splitting can occur, so higher operating pressures may be required to maintain single-phase dense CO<sub>2</sub> transport. This behavior is mainly related to the impurity critical temperature and volatility relative to CO<sub>2</sub>. Low critical temperature impurities such as H<sub>2</sub>, N<sub>2</sub>, Ar, and CH<sub>4</sub> are more volatile than CO<sub>2</sub> and increase the separation between the bubble- and dew-point curves. In contrast, H<sub>2</sub>S has a higher critical temperature and stronger condensability than CO<sub>2</sub>, and therefore behaves more like a heavy component. The broadening effect is especially clear for mixtures containing H<sub>2</sub>. For example, feed 1, a binary mixture without H<sub>2</sub>, has a narrow phase envelope, whereas feed 2, a quaternary mixture containing H<sub>2</sub>, has a much wider phase envelope. For feeds 2 and 4,  $P_{\text{bubble}}$  is higher at lower temperatures than at higher temperatures, which can be attributed to the negative Joule–Thomson behavior of H<sub>2</sub>.<sup>116</sup> Figure 2b presents IFTs of the same selected mixtures computed at the  $P_{\text{bubble}}$  and  $P_{\text{dew}}$  curves and compared with pure CO<sub>2</sub>. The IFT values along the dew curves remain nearly identical to the pure-CO<sub>2</sub> interfacial tension (IFT), almost independent of impurity type and its composition. IFT values at bubble curves show a stronger dependence on the feed composition and impurity. Compared with pure CO<sub>2</sub>, all selected CO<sub>2</sub>-rich mixtures have lower IFT on the bubble branch, with mean relative deviation (MRD) values ranging from ca. –14 to –26%. Feed 3 shows the lowest average deviation of ca. –14%, whereas feed 4 exhibits the largest average decrease of ca. –26%. The largest deviations are noticed at the high temperatures, around  $T = 298\text{--}300\text{ K}$ . Figure 2c presents the corresponding interfacial thickness,  $L_{10}^0$ , computed at the  $P_{\text{bubble}}$  and  $P_{\text{dew}}$  curves. Similar to the IFT, the  $L_{10}^0$  values of the CO<sub>2</sub>-rich mixtures along the dew branch are almost identical to those of pure CO<sub>2</sub>, indicating that the impurities have only a minor effect on the dew branch interfacial properties. In sharp contrast, the mixtures have larger  $L_{10}^0$  values than pure CO<sub>2</sub> along the bubble branch, with MRD values ranging from approximately 10 to 20%. Feed 3 shows the smallest deviation, with an MRD of ca. 10%, whereas feed 4 exhibits the largest deviation, with an MRD of ca. 20%. However, the differences in  $L_{10}^0$  between the CO<sub>2</sub>-rich mixtures listed in Table 2 are small compared with their deviations from pure CO<sub>2</sub>. Overall, impurities lower the IFT and increase the interfacial thickness, indicating a more diffuse interface than in pure CO<sub>2</sub>. From a transportation perspective, these changes can affect two-phase flow behavior once phase splitting occurs, because lower IFT facilitates phase dispersion and increases the interfacial surface area available for heat and mass transfer. The phase envelopes and interfacial properties were computed for several CO<sub>2</sub>-rich mixtures from the composition pool,  $\mathcal{P}$ , similar to the five representative feeds listed in Table 2. The phase envelopes and interfacial properties data computed using the PCP-SAFT EoS + cDFT were then used for training ML models. For each feed composition, a two-phase contour map was generated to check the consistency of the PCP-SAFT EoS + cDFT computations, as illustrated in Figure 2d for feed 4. The contour map indicates that the IFT is strongly a function of  $T$ . At constant  $T$ , the IFT changes almost linearly with pressure, indicating

that its pressure dependence along an isotherm can be approximated using a simple linear relation.

### 3.3. Correlation Analysis of Features

To guide the input feature selection for the ML models, the feature–target relationships were evaluated using the Spearman correlation matrix<sup>112</sup> and the Kolmogorov–Smirnov (KS) statistic,<sup>113</sup> as shown in Figure 3. These analyses were performed on a composition pool with  $|\mathcal{P}| = 100$ , resulting in 19,361 thermodynamic state points generated using the PCP-SAFT EoS + cDFT framework.<sup>57,59,60</sup> The Spearman correlation matrix in Figure 3a was used to identify monotonic nonlinear feature–target relationships and was preferred over Pearson correlation because the investigated phase-equilibrium and interfacial properties show nonlinear behavior, as illustrated in Figure 2a–c. However, a weak Spearman correlation does not necessarily imply feature irrelevance because nonmonotonic relationships can still give a  $\rho_s$  value close to zero. For this reason, the KS statistic in Figure 3b was used as a complementary direction-free relevance measure by comparing the target–property distributions between low- and high-feature-value groups. A high KS value with low  $\rho_s$  therefore indicates an important but potentially nonmonotonic feature–target relationship. This combination is useful here because all target properties show nonlinear behavior. For example,  $P_{\text{bubble}}$  can show a strongly nonlinear dependence on  $T$  for mixtures with higher H<sub>2</sub> content, where the curve shape is affected by Joule–Thomson inversion behavior.<sup>116</sup> The  $P_{\text{dew}}$  curves, which closely follow the pure-CO<sub>2</sub> saturation line, can be related to an Antoine-type relation,  $\ln P \sim -A/T + B$ .<sup>117</sup> Similarly, the IFT decreases nonlinearly with temperature, consistent with semiempirical correlations of the form  $\gamma = \gamma_0(1 - T/T_c)^n$ ,<sup>118</sup> where  $n \neq 1$ . The interfacial thickness also shows critical behavior and can be represented qualitatively as  $L \sim (1 - T/T_c)^{-\nu}$ ,<sup>119,120</sup> where the positive exponent  $\nu$  causes  $L$  to increase sharply as  $T$  approaches  $T_c$ . Overall, using both Spearman correlation and the KS statistic provides a more useful feature-relevance analysis than relying only on a linear correlation measure.

The KS analysis confirms the importance of the thermodynamic state variables, particularly  $T$  and  $P$ . Pressure shows relatively high KS values of ca. 0.7 for  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^0$ , while temperature reaches values close to one for  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^0$ . The phase densities,  $\rho^L$  and  $\rho^V$ , also show high KS values of ca. 0.9 for  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^0$ . The Spearman matrix supports the same physical interpretation, where  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^0$  are strongly related to  $T$ ,  $\rho^L$ , and  $\rho^V$ , whereas  $P_{\text{bubble}}$  is more sensitive to the feed composition, particularly  $z_{\text{H}_2}$  and  $z_{\text{CO}_2}$ . Based on this feature-relevance analysis, the full data-driven models were trained using  $T$ ,  $P$ , and the feed composition vector  $\mathbf{z}$  as the final input features, defined as  $\mathbf{X} = \{T, P, \mathbf{z}\}$ . Since the feed mole fractions satisfy  $\sum_i z_i = 1$ , only  $N_c - 1$  composition variables are independent, giving a scalar feature-set cardinality of  $|\mathbf{X}_{\text{scalar}}| = 2 + (N_c - 1)$ . The composition vector was kept explicit to ensure that the models remain direct functions of mixture composition and to allow the effects of individual impurities to be evaluated. The phase mole fraction vectors,  $\mathbf{x}$  and  $\mathbf{y}$ , pure-CO<sub>2</sub> properties, and mixture critical properties were excluded because they largely repeat information already contained in  $T$ ,  $P$ , and  $\mathbf{z}$ . This avoids multicollinearity and unnecessary input dimensions while keeping the model physically interpretable and composition-dependent.



**Figure 4.** Residual analysis for the four target properties using the selected Machine Learning (ML) models. For each target property, the models are arranged from left to right as Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (XGB), and Tabular Prior-data Fitted Network (TabPFN). Panels (a–d), (e–h), (i–l), and (m–p) correspond to  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and interfacial thickness, respectively. The test and validation sets are presented separately to evaluate the residual distribution for both fitted and untrained data. The root-mean-square error (RMSE) values of each target property for the test and validation sets are reported in each subplot to quantify the corresponding prediction error. The test set is shown as blue circles with dashed  $\pm$  RMSE bands, and the validation set is shown as red triangles with dash-dotted  $\pm$  RMSE bands. The residuals are plotted against the predicted values to identify possible model bias, heteroscedasticity, and systematic deviations throughout the prediction range.<sup>85,89–92</sup>

### 3.4. Machine Learning (ML) Model Performance and Outlier Analysis

The data set for the machine learning (ML) regression models was generated using the PCP-SAFT EoS + cDFT framework.<sup>59,60,63</sup> To ensure broad coverage of the composition space, a pool of 100 feed compositions was constructed, resulting in approximately  $1.9 \times 10^4$  data points. The feed pools combined Dirichlet sampling, systematic scaling, and industrial feed compositions, with relative contributions of 0.7, 0.2, and 0.1, respectively. The predictive performance of the full data-driven models was evaluated using parity plots and prediction-error plots. The parity plots are presented in Figures S12–S15 of the Supporting Information for clarity, whereas

the prediction-error plots are presented in Figure 4. These plots compare the predicted and reference values for  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^{90}$  using Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (XGB), and Tabular Prior-data Fitted Network (TabPFN). The parity plots report the  $R^2$  values, while Figure 4 reports the corresponding test and validation RMSE values using consistent axis ranges for each target property. The 5-fold Cross Validation (CV) metrics, including the mean and uncertainty of  $R^2$ , RMSE, and mean absolute error (MAE), are reported in Table 3. Although all models achieved high  $R^2$  values in the single train-test-validation split, their error magnitudes, outlier behavior, and CV robustness varied widely. Outliers were identified using a

**Table 3. Five-Fold Cross Validation (CV) Performance of the Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (XGB), Tabular Prior-Data Fitted Network (TabPFN), and Random Forest (RF) + Tabular Prior-Data Fitted Network (TabPFN) Machine Learning (ML) Models for All Target Properties<sup>a</sup>**

| target property      | ML model                 | $R^2$ ( $\uparrow$ ) | RMSE ( $\downarrow$ ) | MAE ( $\downarrow$ ) |
|----------------------|--------------------------|----------------------|-----------------------|----------------------|
| $P_{\text{bubble}}$  | RF                       | $0.9691 \pm 0.0114$  | $3.657 \pm 0.849$     | $2.456 \pm 0.529$    |
|                      | SVR                      | $0.9402 \pm 0.0438$  | $4.658 \pm 2.368$     | $2.436 \pm 1.041$    |
|                      | XGB                      | $0.9726 \pm 0.0161$  | $3.406 \pm 1.181$     | $2.300 \pm 0.717$    |
|                      | TabPFN <sup>b</sup>      | $0.9828 \pm 0.0062$  | $2.732 \pm 0.587$     | $1.841 \pm 0.254$    |
|                      | TabPFN <sup>c</sup>      | $0.9960 \pm 0.0016$  | $1.313 \pm 0.302$     | $0.766 \pm 0.168$    |
|                      | RF + TabPFN <sup>b</sup> | $0.9721 \pm 0.0176$  | $3.390 \pm 1.308$     | $2.034 \pm 0.513$    |
| $P_{\text{dew}}$     | RF                       | $0.9995 \pm 0.0004$  | $0.392 \pm 0.139$     | $0.120 \pm 0.032$    |
|                      | SVR                      | $0.9730 \pm 0.0213$  | $2.708 \pm 1.314$     | $1.336 \pm 0.500$    |
|                      | XGB                      | $0.9993 \pm 0.0006$  | $0.419 \pm 0.219$     | $0.170 \pm 0.060$    |
|                      | TabPFN <sup>b</sup>      | $0.9997 \pm 0.0002$  | $0.298 \pm 0.077$     | $0.167 \pm 0.010$    |
|                      | TabPFN <sup>c</sup>      | $0.9998 \pm 0.0004$  | $0.201 \pm 0.185$     | $0.041 \pm 0.012$    |
|                      | RF + TabPFN <sup>b</sup> | $0.9997 \pm 0.0003$  | $0.296 \pm 0.138$     | $0.124 \pm 0.027$    |
| $\gamma$             | RF                       | $0.9991 \pm 0.0003$  | $0.186 \pm 0.030$     | $0.101 \pm 0.011$    |
|                      | SVR                      | $0.9765 \pm 0.0184$  | $0.856 \pm 0.391$     | $0.449 \pm 0.168$    |
|                      | XGB                      | $0.9992 \pm 0.0004$  | $0.174 \pm 0.042$     | $0.109 \pm 0.024$    |
|                      | TabPFN <sup>b</sup>      | $0.9993 \pm 0.0002$  | $0.163 \pm 0.017$     | $0.122 \pm 0.010$    |
|                      | TabPFN <sup>c</sup>      | $1.0000 \pm 0.0000$  | $0.040 \pm 0.009$     | $0.020 \pm 0.004$    |
|                      | RF + TabPFN <sup>b</sup> | $0.9993 \pm 0.0004$  | $0.157 \pm 0.043$     | $0.091 \pm 0.019$    |
| $L_{10}^{\text{90}}$ | RF                       | $0.9964 \pm 0.0046$  | $0.0353 \pm 0.0219$   | $0.0116 \pm 0.0019$  |
|                      | SVR                      | $0.7786 \pm 0.1583$  | $0.2981 \pm 0.1240$   | $0.1726 \pm 0.0526$  |
|                      | XGB                      | $0.9965 \pm 0.0048$  | $0.0339 \pm 0.0230$   | $0.0118 \pm 0.0015$  |
|                      | TabPFN <sup>b</sup>      | $0.9980 \pm 0.0021$  | $0.0272 \pm 0.0143$   | $0.0111 \pm 0.0017$  |
|                      | TabPFN <sup>c</sup>      | $0.9982 \pm 0.0010$  | $0.0277 \pm 0.0081$   | $0.0074 \pm 0.0026$  |
|                      | RF + TabPFN <sup>b</sup> | $0.9970 \pm 0.0038$  | $0.0320 \pm 0.0205$   | $0.0096 \pm 0.0022$  |

<sup>a</sup>The reported values correspond to the mean  $\pm$  standard deviation for the five validation folds. The arrows ( $\uparrow$  and  $\downarrow$ ) indicate the desired direction of model performance. Higher values are better for  $R^2$  ( $\uparrow$ ), whereas lower values are better for root-mean-square error (RMSE) and mean absolute error (MAE) ( $\downarrow$ ). The RMSE and MAE are reported in bar for  $P_{\text{bubble}}$  and  $P_{\text{dew}}$ , mN m<sup>-1</sup> for  $\gamma$ , and nm for the interfacial thickness ( $L_{10}^{\text{90}}$ ). Superscripts indicate the TabPFN configuration, where <sup>b</sup> denotes no HyperParameter Optimization (HPO) and <sup>c</sup> denotes HPO. <sup>b</sup>TabPFN without HPO. <sup>c</sup>TabPFN with HPO.

3-standard-deviation criterion ( $3\sigma$ ), where a data point was flagged when its absolute residual exceeded 3 times the residual standard deviation. Since the  $3\sigma$  threshold was computed separately for each model, a more accurate model can have a smaller threshold and still contain flagged outliers while maintaining better overall predictive performance.

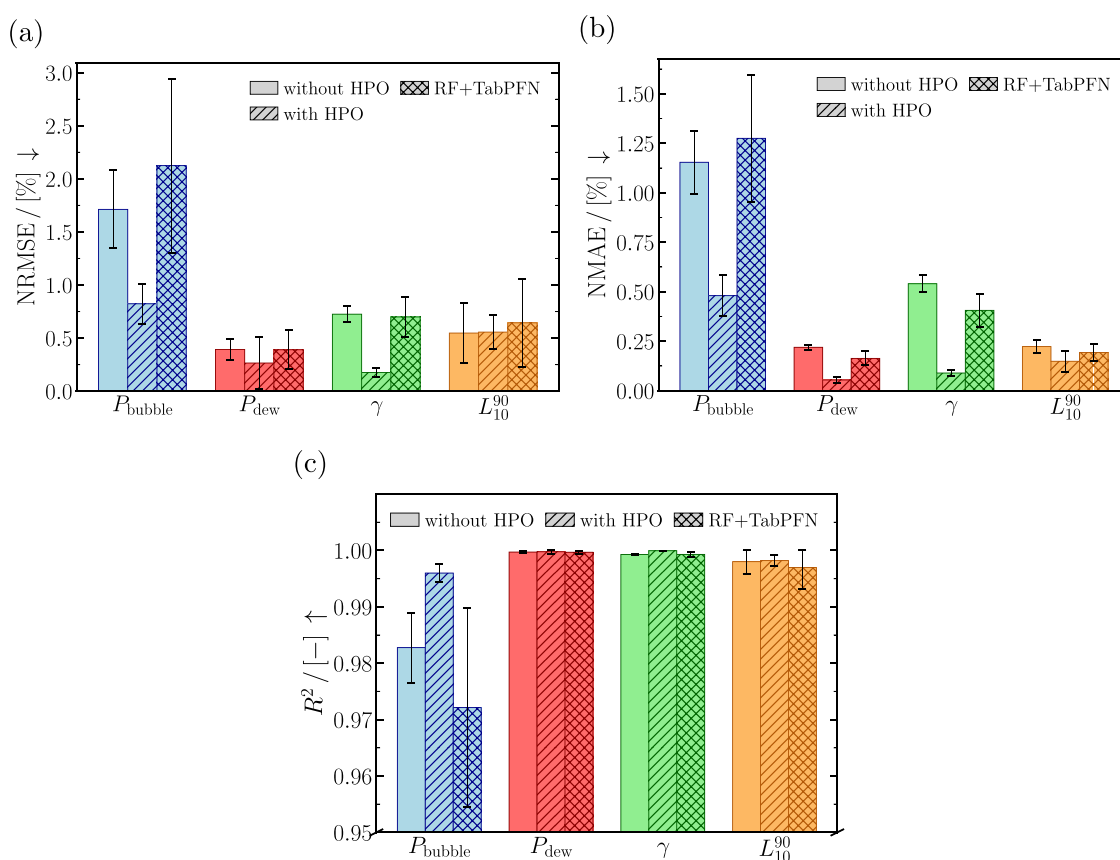
**3.4.1. Bubble-Point Pressure ( $P_{\text{bubble}}$ ).** For  $P_{\text{bubble}}$ , the prediction-error plots are shown in Figure 4a–d. The maximum residual is ca. 3 bar for TabPFN, ca. 9 bar for XGB, and ca. 16 bar for both RF and SVR. These large residuals occur mainly at low  $T$ , especially in the 200–210 K range, suggesting that this region of the thermodynamic space is more difficult to learn. This behavior is associated with mixtures containing higher H<sub>2</sub> concentrations, where  $P_{\text{bubble}}$  changes strongly with the composition. The single-split test and validation RMSE values show that TabPFN achieves the lowest error, ca. 0.18 bar, followed by SVR, XGB, and RF, with values of ca. 0.53, 0.60, and 1.0 bar, respectively. However, the 5-fold CV metrics in Table 3 indicate that XGB generalizes better than SVR. Overall, TabPFN provides the best balance between accuracy and robustness, followed by XGB, SVR, and RF.

**3.4.2. Dew-Point Pressure ( $P_{\text{dew}}$ ).** For  $P_{\text{dew}}$ , the prediction-error plots are shown in Figure 4e–h. Unlike  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$  is less affected by impurities and remains close to pure CO<sub>2</sub> saturation behavior. For the single train-test-validation split, RF yielded the lowest RMSE of ca. 0.05 bar, followed by TabPFN and XGB with ca. 0.1 bar and SVR with

ca. 0.2 bar. However, TabPFN showed the best 5-fold RMSE, indicating more consistent fold-to-fold generalization. The maximum residuals were ca. 1.4 bar for TabPFN and ca. 2 bar for RF,  $\sim 3$  bar for XGB, and  $\sim 4$  bar for SVR. Most outliers were found in the near-critical region, mainly at  $T > 280$  K. Overall, RF, TabPFN, and XGB perform similarly well for  $P_{\text{dew}}$ , while SVR is the weakest model.

**3.4.3. Interfacial Tension ( $\gamma$ ).** For  $\gamma$ , the prediction-error plots are shown in Figure 4i–l. Most outliers occur in the low- $T$  region, particularly in the range of 200–210 K, similar to  $P_{\text{bubble}}$ . The maximum residual is ca. 0.5 mN m<sup>-1</sup> for TabPFN, ca. 1.1 mN m<sup>-1</sup> for SVR, ca. 1.3 mN m<sup>-1</sup> for XGB, and ca. 1.5 mN m<sup>-1</sup> for RF. The single-split RMSE values also show that TabPFN performs best, followed by SVR, XGB, and RF. This suggests that the tree-based models struggle more in the low- $T$  region compared to the transformer-based and SVM models. The 5-fold CV RMSE values in Table 3 show that the tree-based models generalize better than SVR, while TabPFN outperforms all models. Overall, TabPFN provides the best  $\gamma$  prediction, combining the lowest RMSE, the lowest worst-case residuals, and the best CV performance, while XGB serves as the closest conventional baseline.

**3.4.4. Interfacial Thickness ( $L_{10}^{\text{90}}$ ).** For  $L_{10}^{\text{90}}$ , the prediction-error plots are shown in Figure 4m–p. Similar to  $P_{\text{dew}}$ ,  $L_{10}^{\text{90}}$  is less sensitive to impurities, leading to low RMSE values for all investigated models. In contrast, SVR yielded the highest RMSE and the lowest  $R^2$ , with a 5-fold CV  $R^2$  of ca. 0.77, indicating weak generalization. This may be attributed to the

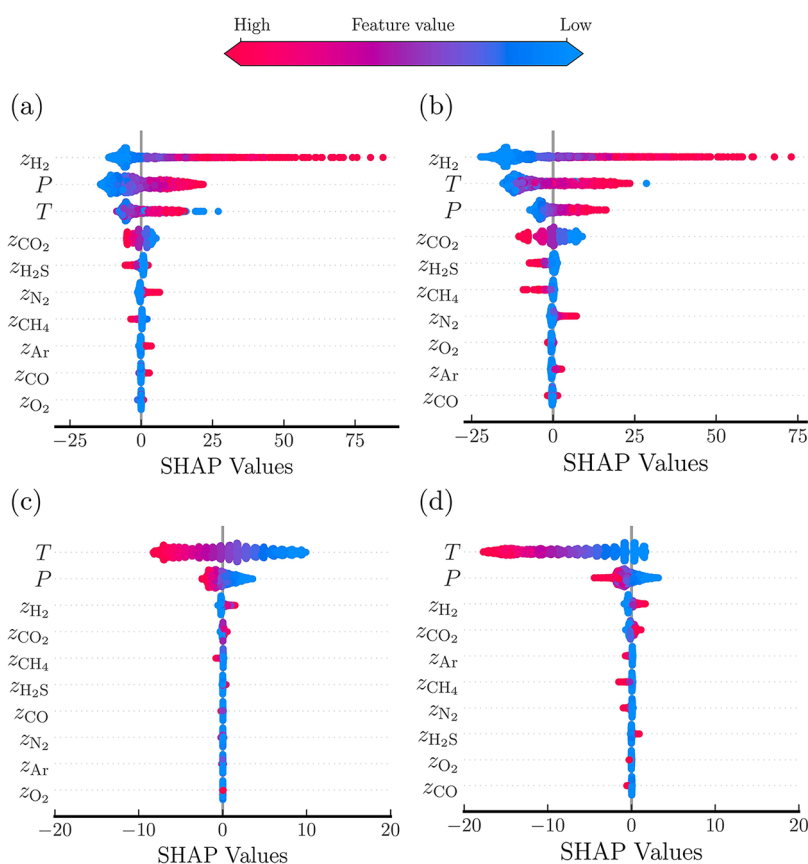


**Figure 5.** 5-fold Cross Validation (CV) performance comparison of the Tabular Prior-data Fitted Network (TabPFN)-based models for the four target properties. (a–c) Comparison of the normalized root mean square error (NRMSE), normalized mean absolute error (NMAE), and  $R^2$ , respectively, for  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^{90}$ . The three model configurations correspond to TabPFN without HyperParameter Optimization (HPO), TabPFN with HPO, and the hybrid Random Forest (RF) + TabPFN models. The bar heights represent the mean values from the five CV folds, while the error bars indicate the corresponding standard deviations. The downward arrows in panels (a) and (b) indicate that lower error values are preferred, whereas the upward arrow in panel (c) indicates that higher  $R^2$  values are preferred.<sup>85,93</sup>

difficulty of the  $\epsilon$ -insensitive loss in capturing the strongly nonlinear increase of  $L_{10}^{90}$  near the critical point. TabPFN, XGB, and RF performed similarly on the single train-test-validation split, with RMSE values below ca. 0.02 nm, while TabPFN achieved the lowest test RMSE and the best 5-fold CV RMSE. Based on the outlier analysis, the maximum residuals for TabPFN, XGB, and RF were below ca. 0.3 nm, whereas for SVR, the maximum residual was about 0.5 nm. Overall, TabPFN is the best model for interfacial thickness prediction, with XGB and RF as close competitors.

**3.4.5. Selection and Evaluation of TabPFN-Based Models.** Based on the prediction-error, outlier, and 5-fold CV analyses, TabPFN was selected as the most reliable full data-driven model for predicting the bubble- and dew-point pressures and interfacial properties of  $\text{CO}_2$ -rich mixtures. For the other models, XGB showed the most consistent overall performance, RF performed well for impurity-insensitive properties, and SVR showed weaker fold-to-fold robustness during CV. Based on this result, different transformer-based TabPFN configurations<sup>93</sup> were further investigated for  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ ,  $\gamma$ , and  $L_{10}^{90}$ . These configurations include the standard TabPFN model without HPO, TabPFN with HPO, and a hybrid RF + TabPFN model. In the hybrid model, RF partitions the input feature space, while TabPFN improves local predictions within the resulting partitions, thereby combining the structured partitioning ability of tree-based models with the predictive capability of the transformer-based

TabPFN model.<sup>93</sup> The same data set used for the RF, SVR, and XGB models was used for this comparison. Figure 5a–c compares the NRMSE, NMAE, and  $R^2$  values, respectively, for all target properties. To enable direct comparison between properties with different units and magnitudes, the error metrics were normalized with respect to each target property, as shown in Figure 5a,b. All three TabPFN-based models resulted in low normalized errors for the four target properties. Overall, the tuned TabPFN model with HPO showed the best performance, giving the lowest NRMSE and NMAE for  $P_{\text{bubble}}$ ,  $P_{\text{dew}}$ , and  $\gamma$ , as well as the highest  $R^2$  values for all four properties. The improvement from tuning was most pronounced for  $\gamma$ , where the NRMSE decreased from ca. 0.7% for the default model to ca. 0.2%, corresponding to nearly a 4-fold reduction. For  $P_{\text{bubble}}$ , the NMAE also decreased from ca. 1.2 to 0.5% after tuning. The default TabPFN and RF-TabPFN models showed similar performance, although both were slightly less accurate than the tuned model. The RF-TabPFN model yielded the largest errors for  $P_{\text{bubble}}$ , with an NRMSE of ca. 2.1% and an NMAE of ca. 1.3%, but performed comparably to the other models for  $P_{\text{dew}}$  and  $\gamma$ . For  $L_{10}^{90}$ , the three models showed very similar results: the default TabPFN had the lowest NRMSE of ca. 0.6%, whereas the tuned model had the lowest NMAE of ca. 0.2% and the highest  $R^2$  of 0.9982. These results show that HPO improved TabPFN for most target properties, with the strongest effect observed for  $\gamma$ . In contrast, the gain was limited for  $L_{10}^{90}$ , where all three models

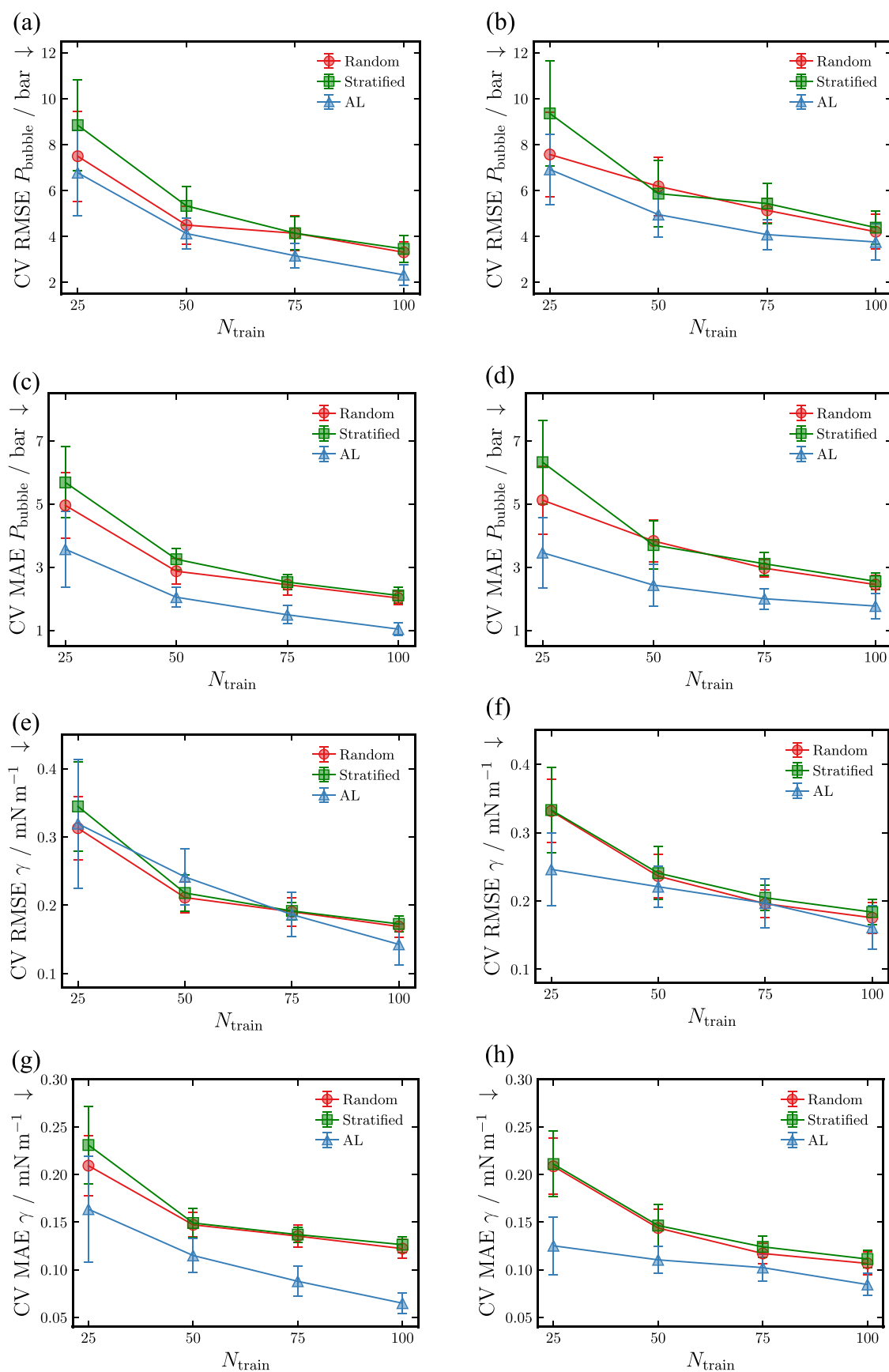


**Figure 6.** SHapley Additive exPlanation (SHAP) value distribution for the Gradient Boosting (XGB) and Tabular Prior-Data Fitted Network (TabPFN) models with HyperParameter Optimization (HPO). For  $P_{\text{bubble}}$ , panel (a) corresponds to Gradient Boosting (XGB) and panel (b) corresponds to Tabular Prior-data Fitted Network (TabPFN). For  $\gamma$ , panel (c) corresponds to XGB and panel (d) corresponds to TabPFN. The features are ordered by overall importance, and each point represents the SHAP value of a feature for a single sample. The color scale indicates the feature value, with high values in red and low values in blue.<sup>85,91,92</sup>

already performed similarly well. For  $\text{CO}_2$  transportation, these accuracies show that the tuned TabPFN model can rapidly screen the effects of impurities on phase boundaries and interfacial properties. This can help identify operating conditions near the two-phase region and estimate the compression pressure required to maintain single-phase dense  $\text{CO}_2$  transport without repeatedly performing expensive PCP-SAFT EoS + cDFT computations. The trained TabPFN surrogate for  $P_{\text{bubble}}$  and  $P_{\text{dew}}$  can therefore replace the rigorous flash calculations required in CFD and process simulations. The same holds for  $\gamma$  and  $L_{10}^0$  in two-phase flow conditions. A surrogate is applicable only where it can be trusted. For a given temperature, pressure, and feed composition, TabPFN outputs a distribution of possible values for the target property rather than a single predicted value. Each prediction therefore carries a prediction interval (PI) at no additional cost.<sup>85</sup> For example, Section S4.3 of the Supporting Information reports the PI for the HPO-tuned  $P_{\text{bubble}}$  model of feed 2 in Table 2, together with the PCP-SAFT EoS reference values.

**3.4.6. SHAP Analysis.** To interpret the feature contributions for the best-performing ML models, SHAP analysis was performed for  $P_{\text{bubble}}$  and  $\gamma$ .<sup>93,121</sup> These properties were selected because they are more sensitive to impurity type and concentration than  $P_{\text{dew}}$  and  $L_{10}^0$ , and are therefore more critical for  $\text{CO}_2$  transportation. The SHAP distributions for  $P_{\text{bubble}}$  predicted using XGB and TabPFN with HPO are shown in Figure 6a and b, respectively, and the corresponding

distributions for  $\gamma$  are shown in Figure 6c and d. In each beeswarm plot, features are ranked by mean absolute SHAP value, and the sign indicates whether a feature increases or decreases the prediction relative to the model base value. For  $P_{\text{bubble}}$ , both models identify  $z_{\text{H}_2}$  as the dominant feature, with high  $z_{\text{H}_2}$  values giving positive SHAP values and therefore increasing the predicted bubble-point pressure. This agrees with the expected behavior of  $\text{H}_2$  as a light impurity, which increases the pressure required to maintain the liquid phase.  $T$  and  $P$  are the next most important variables, while high  $z_{\text{CO}_2}$  values shift the prediction toward lower  $P_{\text{bubble}}$ , consistent with more  $\text{CO}_2$ -rich mixtures having lower bubble-point pressures. The remaining impurities have smaller contributions, although TabPFN shows some bidirectional trace-impurity effects, suggesting composition- and state-dependent interactions. For IFT, the dominant features shift from composition to thermodynamic state variables, with  $T$  ranked highest in both models, followed by  $P$ , while the composition features remain close to zero. High-temperature points mainly appear on the negative SHAP side, confirming that increasing temperature lowers the predicted interfacial tension. Overall, the agreement between XGB and TabPFN on the dominant features and their direction of influence supports the physical consistency of the learned trends, while the broader impurity response in TabPFN suggests that the transformer-based model captures more composition-dependent interactions than XGB.



**Figure 7.** Comparison of random, stratified, and active learning (AL) sampling strategies for different training composition sizes,  $N_{\text{train}}$ . For  $P_{\text{bubble}}$ , panels (a) and (b) correspond to root-mean-square error (RMSE) for Tabular Prior-data Fitted Network (TabPFN) and Gradient Boosting (XGB), respectively, while panels (c) and (d) correspond to mean absolute error (MAE) for TabPFN and XGB, respectively. For  $\gamma$ , panels (e) and (f) correspond to RMSE and (g, h) to MAE, for TabPFN and XGB, respectively. The RMSE and MAE values were obtained from 5-fold Cross

Figure 7. continued

Validation (CV). The markers represent the mean values over 20 independent trials, and the error bars represent the corresponding 95% confidence intervals. Low RMSE and MAE values indicate better predictive performance, as indicated by the arrow ↓.<sup>85,91,92,109</sup>

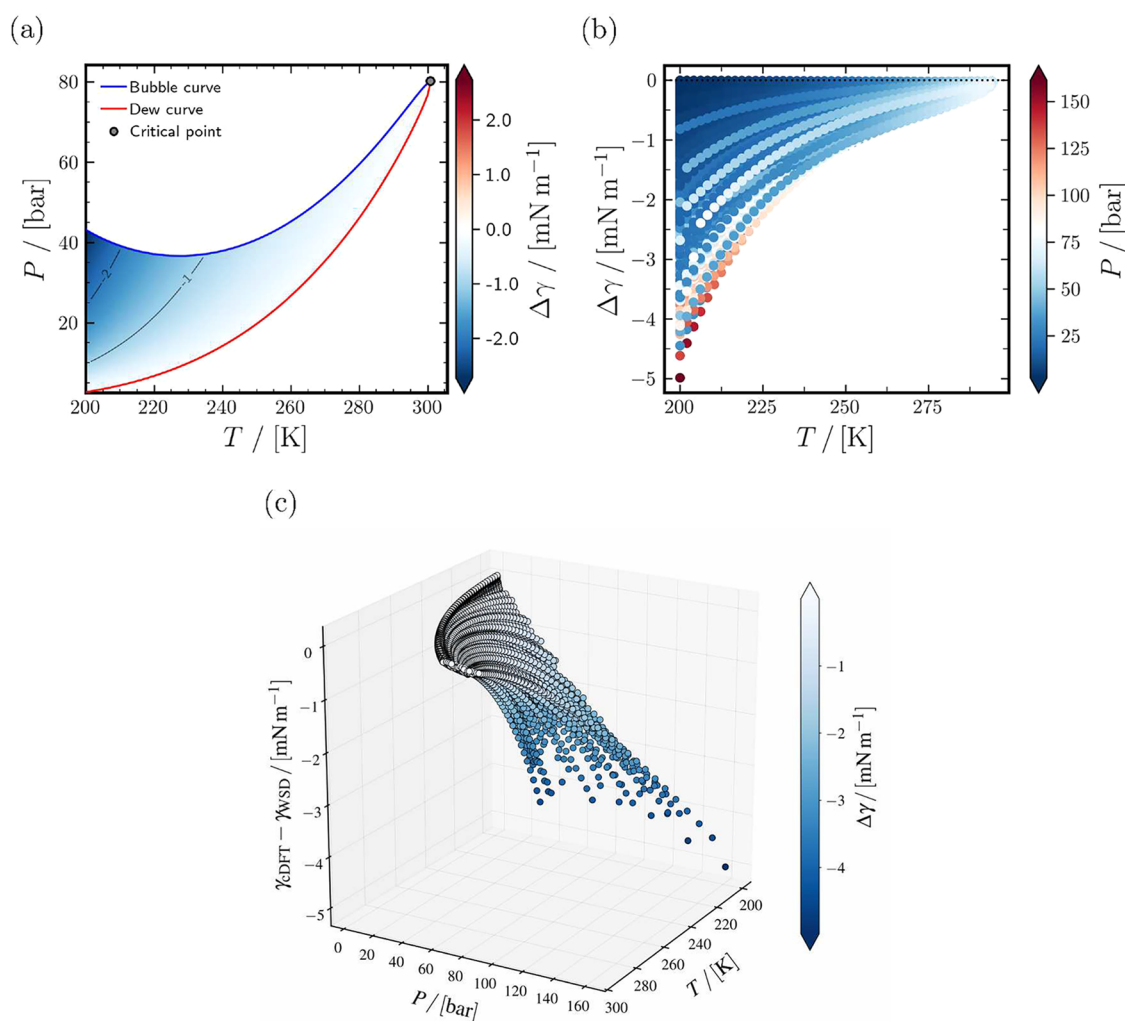
To summarize, the model performance and interpretability analyses show that TabPFN provides the most reliable full data-driven model for predicting phase-equilibrium and interfacial properties of CO<sub>2</sub>-rich mixtures. Compared with the tested models, the HPO-tuned TabPFN model achieves the best overall performance, especially for  $P_{\text{bubble}}$  and  $\gamma$ . These two properties are particularly important because they are more sensitive to impurity type and concentration and therefore cannot be easily generalized for different CO<sub>2</sub>-rich mixture compositions. The SHAP analysis further confirms that  $P_{\text{bubble}}$  predictions are governed mainly by the  $z_{\text{H}_2}$  concentration, whereas  $\gamma$  predictions are governed primarily by  $T$  and  $P$ , supporting the physical consistency of the learned trends. Since this study aims not only to screen suitable ML models but also to evaluate the number of feed compositions required to achieve a target prediction accuracy, a composition-space sampling analysis was performed using the two best-performing models, TabPFN and XGB, for  $P_{\text{bubble}}$  and  $\gamma$ , as described in the next section.

### 3.5. Composition Space Sampling

To evaluate the number of feed compositions required to train full data-driven ML models to a given accuracy, a composition-space sampling analysis was performed. The feed compositions were sampled from the composition pool  $\mathcal{P}$ , which represents CO<sub>2</sub> transportation-relevant mixtures. The pool contained 1000 feed compositions, and the pool characteristics are shown in Figure 1. The three sampling methods investigated in this study were random sampling, stratified sampling, and AL sampling with four feed composition sizes ( $N_{\text{train}}$ ) of 25, 50, 75, and 100. The minimum size was set to 25 because the stratified sampling method used 25 strata. For each selected feed composition, the target properties were generated using the PCP-SAFT EoS + cDFT framework.<sup>59,60,63</sup> For each sampling method and  $N_{\text{train}}$  size, evaluation was performed over 20 independent trials. In each trial, a different random seed was used for both  $N_{\text{train}}$  composition selection and 5-fold CV partitioning so that the reported metrics reflect the effect of the selected mixture compositions rather than a particular split. For TabPFN, the default model without HPO was used, because performing HPO for 20 independent trials and 4  $N_{\text{train}}$  sizes was computationally demanding. Figure 7a,b shows the 5-fold CV RMSE of  $P_{\text{bubble}}$ , averaged over the 20 trials for the TabPFN and XGB models, respectively. For both models, the RMSE decreases as the number of feed compositions increases, with the largest improvement observed when increasing the training size from 25 to 50 compositions. For AL sampling, the TabPFN RMSE decreases from 6.8 to 4.1 bar, while the MAE decreases from 3.6 to 2.0 bar, as shown in Figure 7c, corresponding to an error reduction of about 40%. For XGB, the RMSE decreases from 6.9 to 5.0 bar, while the MAE decreases from 3.5 to 2.4 bar, as shown in Figure 7d, corresponding to an error reduction of about 28%. For  $N_{\text{train}}$  larger than 50, the error continues to decrease, but the improvement becomes smaller, reaching RMSE values of 2.3 bar for TabPFN and 3.8 bar for XGB at 100 compositions. This indicates that the models are approaching a performance plateau where further training compositions yield diminishing

improvements. The same trend is observed in the CV MAE results shown in Figure 7c,d for TabPFN and XGB, respectively. Compared with random and stratified sampling, AL achieves the lowest RMSE and MAE for all  $N_{\text{train}}$  sizes, indicating that adaptive selection of informative feed compositions is more data-efficient than nonadaptive sampling. Random and stratified sampling show similar performance at larger training sizes, namely,  $N_{\text{train}} = 50, 75$ , and 100, but at the smallest size,  $N_{\text{train}} = 25$ , random sampling performs better than stratified sampling, with RMSE values of 7.5 and 8.8 bar, respectively, for TabPFN. For this reason, random sampling is preferred as the nonadaptive baseline because it matches stratified sampling at larger training sizes while outperforming it in the low-data regime. Overall, these results indicate that approximately 50 feed compositions provide a favorable balance between the accuracy and computational cost for predicting  $P_{\text{bubble}}$ . Moreover, AL provides the most data-efficient sampling strategy by reducing the number of feed compositions required to reach a given level of prediction accuracy.

Figure 7e and f shows the 5-fold CV RMSE of  $\gamma$  for the TabPFN and XGB models, respectively. The values are averaged over the 20 independent trials and are shown as a function of  $N_{\text{train}}$ . The corresponding 5-fold CV MAE values are shown in Figure 7g,h. For TabPFN, all sampling strategies show similar RMSE values for the investigated  $N_{\text{train}}$  sizes, indicating that the model is less sensitive to the sampling strategy for  $\gamma$  than for  $P_{\text{bubble}}$ . Similar to  $P_{\text{bubble}}$ , the largest improvement is observed when  $N_{\text{train}}$  is increased from 25 to 50. For AL sampling, the TabPFN RMSE decreases from 0.32 to 0.24 mN m<sup>-1</sup>, while the MAE decreases from 0.16 to 0.12 mN m<sup>-1</sup>. For XGB, the RMSE decreases from 0.25 to 0.22 mN m<sup>-1</sup>, while the MAE decreases from 0.13 to 0.11 mN m<sup>-1</sup>. When the  $N_{\text{train}}$  is increased to 100, the RMSE values of 0.14 mN m<sup>-1</sup> for TabPFN and 0.16 mN m<sup>-1</sup> for XGB are noticed, showing that the improvement becomes smaller beyond 50 compositions. Overall, the  $\gamma$  predictions are highly accurate, with errors below 0.5 mN m<sup>-1</sup> for all sampling methods and  $N_{\text{train}}$ . For XGB, AL sampling produces the lowest RMSE for all  $N_{\text{train}}$ , whereas random and stratified sampling perform similarly. The MAE results show the benefit of AL more clearly, as AL produces the lowest mean error for both TabPFN and XGB for the investigated feed composition sizes. This indicates that AL selects more informative feed compositions for predicting  $\gamma$ , although for TabPFN, the RMSE values of AL and random sampling remain comparable at  $N_{\text{train}} < 100$ . Random and stratified sampling again show similar performance, with random sampling slightly better at  $N_{\text{train}} = 25$ , making it the preferred nonadaptive baseline. Together with the  $P_{\text{bubble}}$  results, these findings indicate that approximately 50 feed compositions provide a good balance between accuracy and computational cost for training full data-driven ML models, while AL offers the most data-efficient sampling strategy, especially when the target is to reduce the number of training compositions required to reach a given accuracy. Based on these results, we recommend GPR-based AL as the sampling strategy for extending the present data set, for example, to additional impurity types.

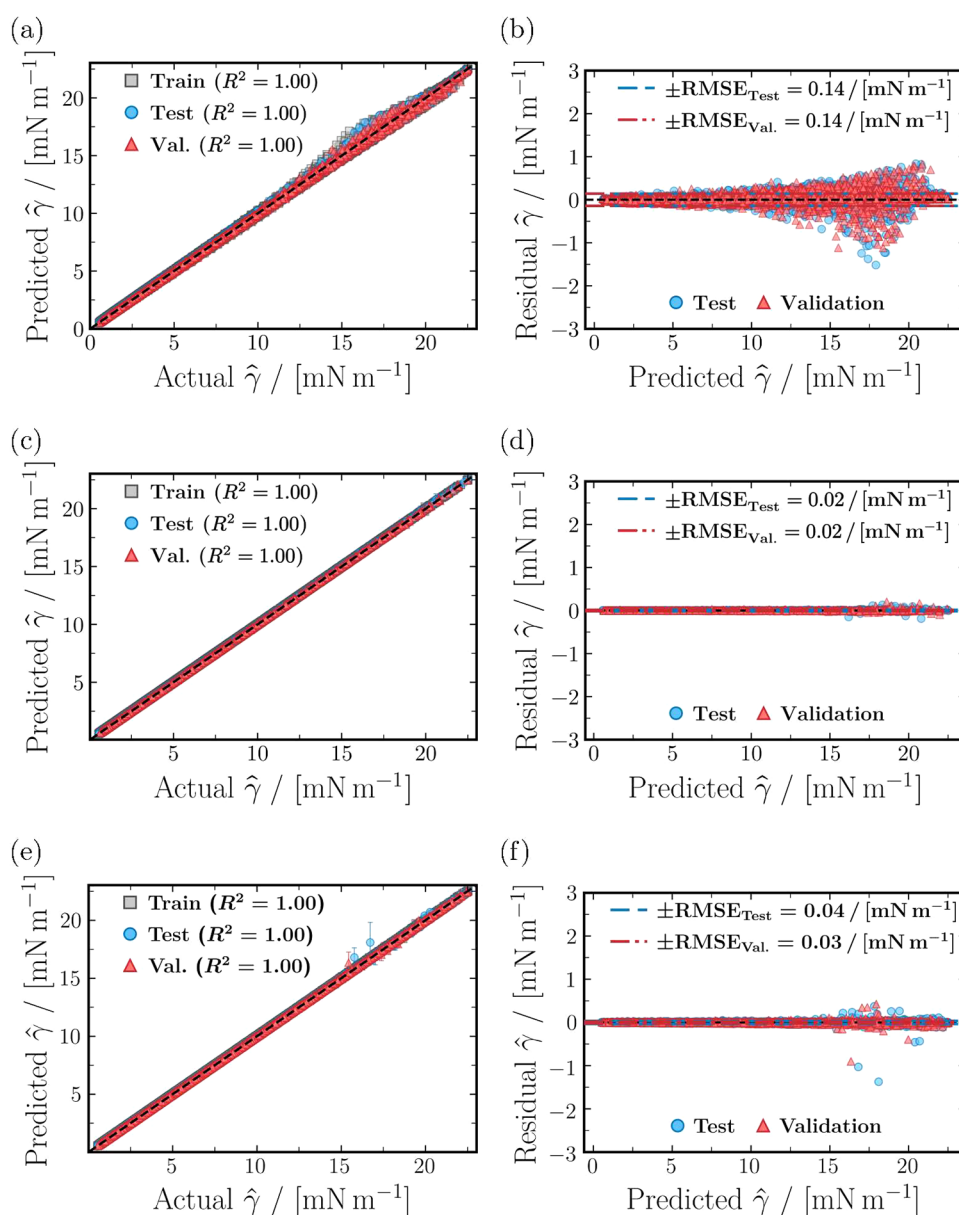


**Figure 8.** Residual analysis of the Winterfeld–Scriven–Davis (WSD) model for interfacial tension (IFT), using the Perturbed-Chain Polar Statistical Associating Fluid Theory (PCP-SAFT) Equation of State (EoS) + classical Density Functional Theory (cDFT) results as reference. (a)  $PT$  two-phase envelope with the bubble curve, dew curve, critical point, and contour distribution of the IFT residual,  $\Delta\gamma$ . (b) Variation of  $\Delta\gamma$  with temperature inside the two-phase region, with the color scale indicating pressure. (c) Three-dimensional residual surface of  $\Delta\gamma$  as a function of temperature and pressure. The residual maps show the deviation of the WSD model from the PCP-SAFT + cDFT reference calculations throughout the two-phase region.<sup>41</sup>

### 3.6. Hybrid Machine Learning (ML) Model for Interfacial Tension (IFT)

We further evaluated  $\gamma$  using a hybrid residual-learning approach, in which the simple semiempirical WSD model<sup>41</sup> was combined with an ML model trained to learn the systematic residuals of the WSD predictions. This approach tests whether incorporating physical prior knowledge improves the predictive accuracy of IFT. The residual,  $\Delta\gamma$ , was defined as the difference between the IFT computed using the PCP-SAFT EoS + cDFT framework<sup>57,59,60</sup> and that predicted by the WSD correlation, i.e.,  $\Delta\gamma = \gamma_{\text{cDFT}} - \gamma_{\text{WSD}}$ . The residuals were computed for the same 100 feed compositions constructed in Section 3.4. Figure 8a shows the contour map of  $\Delta\gamma$  within the two-phase region, bounded by the dew- and bubble-point curves, for feed 4 listed in Table 2. The residuals are mostly negative, indicating that the WSD correlation overpredicts IFT compared to the PCP-SAFT EoS + cDFT framework.<sup>57,59,60</sup> This overprediction is more pronounced at lower temperatures and higher pressures, where the magnitude of  $\Delta\gamma$  increases. This behavior can be attributed to the larger density difference between the coexisting phases at these conditions. At lower

temperatures, the liquid phase becomes more cohesive and denser, whereas volatile impurities preferentially partition into the vapor phase. As a result, the WSD model tends to overestimate the cohesive contribution of the subcritical mixture and does not fully capture the impurity-induced reduction in IFT. The residual distribution also shows a smooth, structured pattern within the two-phase region rather than random scatter. In particular,  $\Delta\gamma$  remains close to zero near the critical point and becomes increasingly negative toward the low-temperature and high-pressure region. This indicates that the error of the WSD correlation has a systematic dependence on the thermodynamic state. Similar contour maps were generated for all 100 feed compositions, and the residuals were found to range from 0 to ca.  $-5 \text{ mN m}^{-1}$ , as shown in Figure 8b. The pressure-colored scatter plot shows that the largest negative residuals are mainly associated with high-pressure states, particularly above ca. 75 bar. This trend is also evident in Figure 8c, where  $\Delta\gamma$  is plotted as a function of both  $T$  and  $P$ , confirming that the largest deviations are concentrated in the low-temperature and high-pressure region of the two-phase envelope. The structured residual field shown



**Figure 9.** Parity and prediction-error analysis for the hybrid interfacial tension (IFT) residual models applied to  $\gamma$ . (a, c, e) Comparison of the predicted values with the reference values and report the corresponding  $R^2$  values. The right panels (b, d, f) present the residuals as a function of the predicted values with the corresponding root-mean-square error (RMSE) values for the test and validation data sets. (a, b), (c, d), and (e, f) correspond to Symbolic Regression (SR), Gaussian Process Regression (GPR), and Stochastic Variational Gaussian Process Regression (SVGP), respectively. For GPR and SVGP, the parity plots include the model-predicted uncertainties.<sup>87,97–100</sup>

in Figure 8c supports the residual-learning strategy because the ML model only needs to learn a smooth state-dependent ( $T$  and  $P$ ) correction to the semiempirical WSD prediction rather than learning the complete IFT response. For this purpose, we used Symbolic Regression (SR), Gaussian Process Regression (GPR), and Stochastic Variational Gaussian Process Regression (SVGP).

**3.6.1. Physically Interpretable Symbolic Regression (SR) Correction for the Winterfeld–Scriven–Davis (WSD) Semiempirical Correlation.** Figure 9a and b shows the parity plot and prediction-error analysis for the Symbolic Regression (SR)-based hybrid model for IFT, in which the baseline WSD semiempirical correlation was combined with a learned correction for the relative residual. Instead of training the SR model directly on the dimensional residual  $\Delta\gamma$  (mN

$\text{m}^{-1}$ ), the target was defined as the dimensionless relative residual  $\epsilon_\gamma = (\gamma_{\text{cDFT}} - \gamma_{\text{base}})/\gamma_{\text{base}}$ , where  $\gamma_{\text{cDFT}}$  is the PCP-SAFT EoS + cDFT reference value and  $\gamma_{\text{base}}$  is the baseline WSD prediction. The predicted residual was then mapped back to the IFT using  $\hat{\gamma} = \gamma_{\text{base}}(1 + \epsilon_\gamma)$ . Hence, the parity plot and prediction-error analysis for the reconstructed  $\hat{\gamma}_{\text{cDFT}}$  are shown in Figure 9a and b, respectively. The input feature set  $\mathcal{F}$  for training the  $\epsilon_\gamma$  was constructed from two groups of physically motivated variables. The first group contained reduced state and interfacial quantities, including  $T_r^* = T/T_c$ ,  $1 - T_r^*$ ,  $P_r^* = P/P_c$ ,  $\Pi^* = P/P_{\text{sat}}^0$ ,  $\Delta\rho = \rho_l - \rho_v$ ,  $\Delta\rho^* = (\rho_l - \rho_v)/(\rho_l^0 - \rho_v^0)$ , and  $\gamma_r^* = \gamma_{\text{base}}/\gamma_{\text{CO}_2}^0$ . The dimensionless input features used are represented by \*. The  $T_c$  and  $P_c$  of pure  $\text{CO}_2$  are used in  $T_r^*$  and  $P_r^*$ . The second group contained composition-based impurity aggregates such as the liquid-

vapor-, and feed-phase impurity sums, which are defined as  $x_{\text{imp}} = \sum_{i \neq \text{CO}_2} x_{iV}$ ,  $y_{\text{imp}} = \sum_{i \neq \text{CO}_2} y_{iV}$  and  $z_{\text{imp}} = \sum_{i \neq \text{CO}_2} z_{iV}$  respectively. In addition, the second group was also combined with the phase-partitioning difference  $\Delta_{\text{imp}} = x_{\text{imp}} - y_{\text{imp}}$  and the impurity partition coefficient  $l_K = \ln[(y_{\text{imp}} + \xi)/(x_{\text{imp}} + \xi)]$ , with  $\xi = 10^{-12}$  to avoid division by zero. Combining both groups resulted in a 12-feature set, i.e.,  $|\mathcal{F}| = 12$ , which provides the SR model with physically informed variables to correct the limitations of the baseline WSD model for  $\text{CO}_2$ -rich mixtures with impurities. Because the baseline WSD model already contains the dominant temperature- and pressure-dependent contribution to IFT, the SR model was also tested using only the composition-based impurity aggregates  $\{z_{\text{imp}}, x_{\text{imp}}, y_{\text{imp}}, \Delta_{\text{imp}}, l_K\}$  to examine whether the missing correction could be represented only through impurity partitioning features. For this purpose, the SR search was performed over 200 iterations using simple unary and binary operators, yielding several candidate equations with varying complexities and corresponding RMSE values. This reduced feature set SR search recovered the previously suggested correction  $\hat{\epsilon}_\gamma \approx -2.56x_{\text{imp}}$ ,<sup>25</sup> where the prefactor 2.56 is an additional empirical constant obtained compared to the previous correction factor suggested by us.<sup>25</sup> This shows that the correction is approximately proportional to the liquid-phase impurity fraction, although it remains less accurate. From the full 12-feature set, the selected correction from the Pareto frontier was

$$\hat{\epsilon}_\gamma \approx \ln(\Delta\rho^*)/[\sqrt{P_r^*} + (\gamma_r^* - \Delta\rho^*)] \quad (19)$$

which achieved a test  $R^2$  of 0.94 in residual space. After reconstruction to IFT, this correction yielded a  $R^2 \approx 1$  (as shown in Figure 9a), with an RMSE of 0.14  $\text{mN m}^{-1}$  and an MAE of 0.074  $\text{mN m}^{-1}$ . In comparison, the correction ( $\hat{\epsilon}_\gamma \approx -2.56x_{\text{imp}}$ ) obtained using only the composition-based feature set resulted in a residual-space test  $R^2$  of 0.618 and a reconstructed-IFT RMSE of 0.29  $\text{mN m}^{-1}$ . The robustness of both the full 12-feature and composition-only corrections was further confirmed by 5-fold CV. The corresponding scores are summarized in Table 4.

Although both corrections were obtained from the SR search, the selected input features are physically motivated and allow a clear interpretation. The Pareto fronts obtained from the physics-informed and composition-only feature sets are shown in Tables S9 and S10, respectively. The Pareto-frontier equations shown in Table S9 are expressed in terms of the density difference ( $\Delta\rho$ ) between the liquid and vapor phases. Similarly, the equations listed in Table S10 are expressed in terms of the total liquid mole fraction of impurities ( $x_{\text{imp}}$ ). The  $x_{\text{imp}}$ -based equations have a direct physical analogy to surfactant behavior because the impurities decrease the IFT of the  $\text{CO}_2$ -rich mixture much like a surfactant, consistent with our previous findings.<sup>25</sup> The negative prefactor ( $-2.56$ ) reflects this monotonic reduction in IFT with increasing liquid-phase impurity content. However, the WSD residual is not governed by composition alone. It also varies with temperature and pressure, with the largest deviations observed at low temperature and high pressure, as shown in Figure 8. The physics-informed Pareto-frontier equations account for these dependencies, even though the baseline WSD model already contains strong  $T$ - and  $P$ -dependent terms. For the chosen correction (eq 19), the numerator  $\ln(\Delta\rho^*)$  reflects the

**Table 4. Five-Fold Cross Validation (CV) Performance of the Hybrid ML Residual Model, Evaluated for the Reconstructed IFT ( $\hat{\gamma}$ ): Symbolic Regression (SR), Gaussian Process Regression (GPR), and Stochastic Variational Gaussian Process Regression (SVGP)<sup>a,b</sup>**

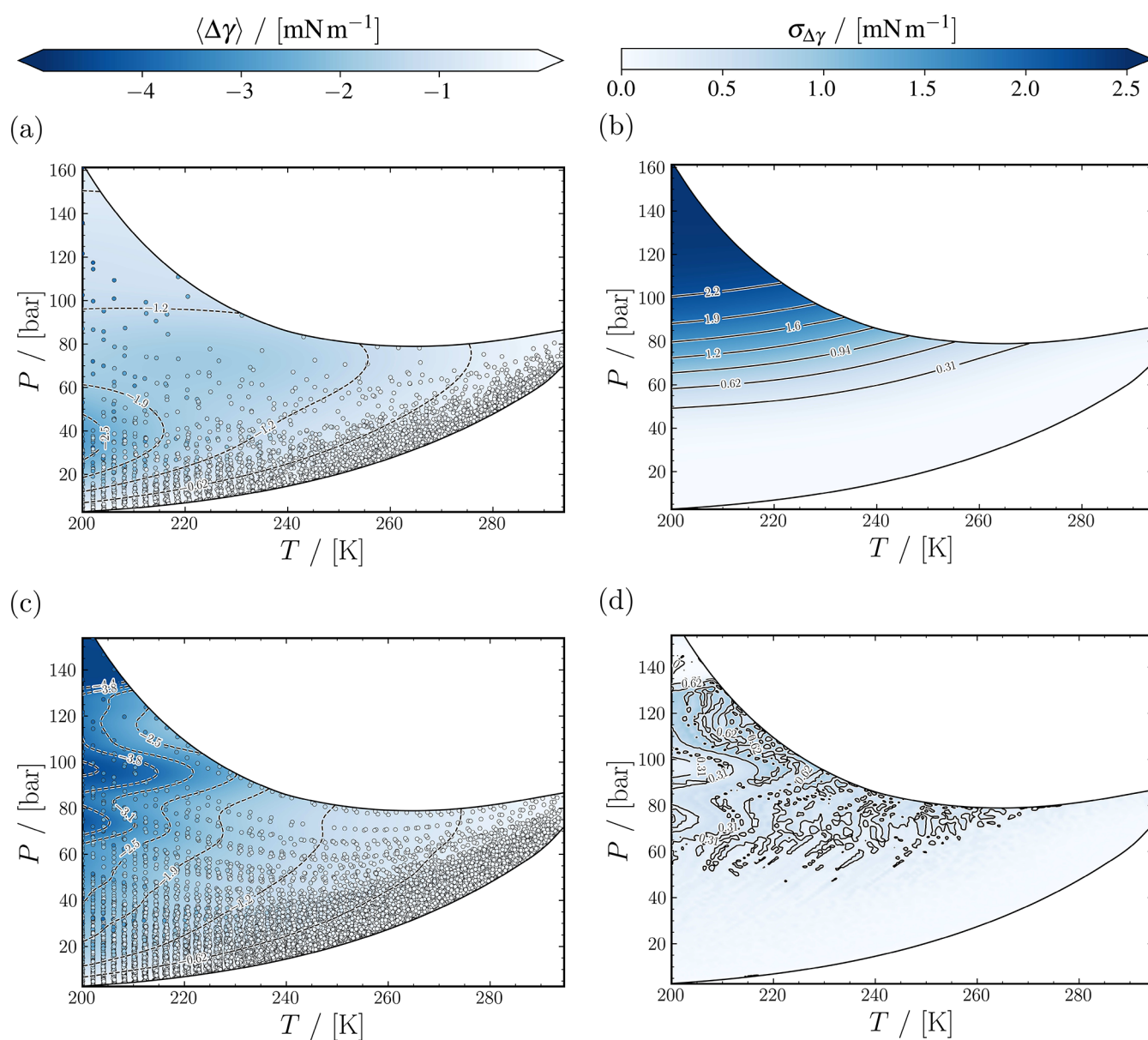
| model                                  | $R^2$ ( $\uparrow$ )  | RMSE/[ $\text{mN m}^{-1}$ ]<br>( $\downarrow$ ) | MAE/[ $\text{mN m}^{-1}$ ]<br>( $\downarrow$ ) |
|----------------------------------------|-----------------------|-------------------------------------------------|------------------------------------------------|
| SR<br>( $\mathcal{F}_{\text{Comp.}}$ ) | $0.9983 \pm 0.0002$   | $0.2475 \pm 0.0110$                             | $0.1434 \pm 0.0073$                            |
| SR ( $\mathcal{F}$ )                   | $0.9996 \pm 0.0000$   | $0.1130 \pm 0.0073$                             | $0.0603 \pm 0.0031$                            |
| GPR                                    | $0.9999 \pm 0.0001^b$ | $0.0379 \pm 0.0262$                             | $0.0051 \pm 0.0011$                            |
| SVGP                                   | $1.0000 \pm 0.0000^b$ | $0.0233 \pm 0.0030$                             | $0.0119 \pm 0.0014$                            |

<sup>a</sup>The reported  $R^2$ , root-mean-square error (RMSE), and mean absolute error (MAE) values are the mean  $\pm$  standard deviation over the five folds. The arrows indicate the desired direction of model performance. The SR model was trained on the dimensionless relative residual  $\epsilon_\gamma$ , whereas the GPR and SVGP models were trained on the dimensional residual  $\Delta\gamma$ . For the GPR and SVGP, the reconstructed IFT RMSE and MAE are therefore identical to the residual-space values. The two SR entries correspond to the composition feature set  $\mathcal{F}_{\text{Comp.}} = \{z_{\text{imp}}, x_{\text{imp}}, y_{\text{imp}}, \Delta_{\text{imp}}, l_K\}$  and the full physics-informed feature set  $\mathcal{F} = \mathcal{F}_{\text{Comp.}} + \{T_r^*, 1 - T_r^*, P_r^*, \Pi^*, \Delta\rho, \Delta\rho^*, \gamma_r^*\}$ .

<sup>b</sup>The reconstructed-IFT  $R^2$  is obtained per fold as  $1 - \text{RMSE}^2/\text{Var}(\gamma_{\text{CDFT}})$ .

sharpness of the interface. A large density difference corresponds to a steep density gradient and a highly cohesive, sharp interface. In sharp contrast, a smaller density difference corresponds to a more diffuse interface, with  $\Delta\rho^*$  vanishing smoothly toward the critical point. The impurity effect enters implicitly through the bulk densities because impurities lower both the liquid and vapor densities relative to pure  $\text{CO}_2$ . The denominator,  $\sqrt{P_r^*} + (\gamma_r^* - \Delta\rho^*)$ , combines a pressure-dependent term with a mismatch between the reduced baseline IFT and the reduced density difference. The term  $\sqrt{P_r^*}$  increases with pressure and therefore damps the correction at higher pressures. The difference  $(\gamma_r^* - \Delta\rho^*)$  accounts for deviations between the interfacial response and the bulk density contrast, which can arise from impurity partitioning and interfacial enrichment. Together, these terms make the denominator a state-dependent damping factor. The equations, feature-engineering details, and SR configuration are provided in Section S5.2 and Tables S8–S12 of the Supporting Information.

**3.6.2. GP-Based Models.** Figure 9c and d shows the parity plot and prediction-error analysis for the GPR hybrid model, respectively. GPR is well-suited for learning the residual function shown in Figure 8c because it defines a posterior distribution over functions consistent with the training data. This allows the GPR to provide both a posterior mean prediction and a calibrated predictive uncertainty for each test and validation data point. However, the main limitation of GPR is its poor scalability with data set size, as the training cost of exact GPR grows as  $O(n^3)$  with the number of training points. To keep the training tractable, a subset of ca. 5000 points was selected from the full  $\sim 1.9 \times 10^4$ -point data set by stratified sampling over  $T$  and  $P$  using 10 bins for each variable, ensuring adequate coverage of the state-space variables. With a 70/15/15 split, this left ca. 3500 data points for training. The GPR model was trained on the dimensional residual  $\Delta\gamma = \gamma_{\text{CDFT}} - \gamma_{\text{base}}$  using the feature set  $\mathcal{X} = \{T, P, \mathbf{z}\}$ , which is the



**Figure 10.** Two-dimensional response surfaces of the hybrid residual Machine Learning (ML) models over the  $PT$  phase envelope for (a, b) Gaussian Process Regression (GPR) and (c, d) Stochastic Variational Gaussian Process Regression (SVGP). (a, c): Posterior-mean interfacial tension (IFT) residual  $\langle \Delta\gamma \rangle$ , with the training data overlaid as points colored on the same scale. (b, d) Predictive standard deviation  $\sigma_{\Delta\gamma}$ . Each column shares the color bar at the top of the column. The black curves denote the bubble and dew boundaries.<sup>87,98–100</sup>

same feature set used for the full data-driven ML models. The model reproduced the residual field with a test  $R^2$  of 0.99, an RMSE of  $0.016 \text{ mN m}^{-1}$  for both the test and validation data sets, and an MAE of  $0.006 \text{ mN m}^{-1}$ . After reconstructing the IFT using  $\hat{\gamma} = \gamma_{\text{base}} + \Delta\hat{\gamma}$ , the model achieved  $R^2 \approx 1$  for both the test and validation sets, with a reconstructed IFT RMSE of  $0.02 \text{ mN m}^{-1}$  for both data sets, as shown in Figure 9c,d. Because  $\gamma_{\text{base}}$  is a fixed quantity, the reconstruction leaves the absolute errors unchanged, so the reconstructed-IFT RMSE and MAE equal their residual-space values, whereas the reconstructed-IFT  $R^2$  is rescaled to the larger variance of  $\gamma_{\text{cDFT}}$  as  $1 - \text{RMSE}^2/\text{Var}(\gamma_{\text{cDFT}})$ . A 5-fold CV was performed to further confirm the robustness of the model. The corresponding reconstructed-IFT CV scores are summarized in Table 4. From Table 4, it is evident that the GPR correction outperforms the SR correction in terms of absolute accuracy

while additionally providing a per-point uncertainty estimate, although this improvement comes with a strict limitation on the trainable data set size. To overcome this scalability limit and use the full data set, a Stochastic Variational Gaussian Process Regression (SVGP) model was used. The SVGP model uses inducing points and a variational approximation to retain the probabilistic, uncertainty-aware character of GPR, while reducing the training cost so that it scales linearly with the number of data points. Figure 9e and f shows the parity plot and prediction-error analysis for the SVGP hybrid model, respectively. In contrast to the exact GPR, the SVGP model approximates the full posterior using 2000 inducing points. This strategy allowed the model to be trained on the complete  $\sim 1.9 \times 10^4$  data set while retaining per-point predictive uncertainty. Using the same residual target,  $\Delta\gamma$ , the SVGP model reproduced the residual field with a test  $R^2$  of 0.996, an

RMSE of  $0.037 \text{ mN m}^{-1}$ , and an MAE of  $0.010 \text{ mN m}^{-1}$ . For the validation set, the model produced an  $R^2 = 0.998$  and an RMSE of  $0.026 \text{ mN m}^{-1}$ . After reconstructing the IFT, the model achieved  $R^2 \approx 1$  on both the test and validation sets, as shown in Figure 9e. The reconstructed IFT RMSE values shown in Figure 9f were  $0.04$  and  $0.03 \text{ mN m}^{-1}$  for the test and validation sets, respectively, while the corresponding MAE values were  $0.01 \text{ mN m}^{-1}$  for both data sets. 5-fold CV also confirmed the robustness of the SVGP model (Table 4), with a very small spread for all of the folds, indicating robust performance.

Figure 10a,b shows the posterior mean  $\langle \Delta\gamma \rangle$  and its predictive uncertainty  $\sigma_{\Delta\gamma}$  within the two-phase *PT* envelope considered in this study for the exact GPR model. The markers in Figure 10a represent the training data overlaid, sharing the same color map as the posterior mean. The posterior mean is negative throughout the envelope, ranging from ca.  $-0.02$  to  $-2.84 \text{ mN m}^{-1}$ . This is consistent with the actual training residuals spanning a range of  $0$  to ca.  $-5 \text{ mN m}^{-1}$  (Figure 8b). In Figure 10a, the underestimation of IFT is most pronounced at low  $T$  and high  $P$ , where the overlaid training markers show the largest negative residuals, reaching ca.  $-5 \text{ mN m}^{-1}$  at  $T \approx 200 \text{ K}$  and  $P \approx 160 \text{ bar}$ . In this region, the posterior-mean surface reaches only ca.  $-2.8 \text{ mN m}^{-1}$ , indicating sparse training data. The exact GPR nevertheless reproduces these markers accurately (test  $R^2 = 0.99$ , RMSE =  $0.016 \text{ mN m}^{-1}$ ), so the apparent outlier reflects that more training data at high pressures need to be included for training to improve the posterior mean predictions. From Figure 10b, the predictive uncertainty rises sharply in this same high-pressure region, where  $\sigma_{\Delta\gamma}$  exceeds  $2.4 \text{ mN m}^{-1}$  above  $100 \text{ bar}$ , also reflecting the sparse training coverage in the high-pressure zone. This also reflects that more  $\text{CO}_2$ – $\text{H}_2$ -based mixtures need to be added to generate the training data set for GPR. Figure 10c,d show the  $\langle \Delta\gamma \rangle$  and  $\sigma_{\Delta\gamma}$  of the SVGP model. Compared to the exact GPR, the overlaid training data are well reproduced by the SVGP, especially in the low-temperature and high-pressure zones, where the posterior-mean surface reaches ca.  $-4.6 \text{ mN m}^{-1}$  and closely matches the dark markers, in contrast to the exact GPR surface that flattened to ca.  $-2.8 \text{ mN m}^{-1}$ . As observed in Figure 10c,  $\langle \Delta\gamma \rangle$  exceeds  $3.8 \text{ mN m}^{-1}$  in magnitude for pressures larger than ca.  $100 \text{ bar}$ , showing that training the SVGP on the full data set yields better predictions of  $\Delta\gamma$ . This improved performance is also reflected in the  $\sigma_{\Delta\gamma}$  contour shown in Figure 10d, which reveals that  $\sigma_{\Delta\gamma}$  is well below ca.  $1.2 \text{ mN m}^{-1}$ . This is roughly half the peak uncertainty of the exact GPR (ca.  $2.4 \text{ mN m}^{-1}$ ), with the largest values again confined to the high-pressure zones. Although the residual-space accuracy of the SVGP model (Table 4) is marginally lower than that of the exact GPR trained on the 3500-point subset, the SVGP attains better performance while training on the full data set and scaling linearly with the number of data points. This makes the SVGP a more practical uncertainty-aware correction for large data sets.

In summary, the SVGP learns the residual function in Figure 8c more accurately than the exact GPR, whereas the corrected WSD expression obtained from SR reproduces the same residual only approximately, with a reconstructed-IFT CV RMSE of ca.  $0.1 \text{ mN m}^{-1}$  for the full physics-informed feature set (Table 4). The SR correction is nonetheless the most computationally efficient route to compute IFT, since its WSD expression carries essentially no inference cost, illustrating the

trade-off between accuracy and computational efficiency. To compare the two best ML surrogate strategies investigated in this study, namely, the full data-driven model (TabPFN with HPO) and the residual-correction model (SVGP), we examine their CV RMSE and MAE in Tables 3 and 4. Both models predict IFT accurately, with CV RMSE values of  $0.040 \text{ mN m}^{-1}$  (TabPFN) and  $0.023 \text{ mN m}^{-1}$  (SVGP). Likewise, the CV MAE of SVGP ( $0.012 \text{ mN m}^{-1}$ ) is about 1.7 times lower than that of TabPFN ( $0.020 \text{ mN m}^{-1}$ ). We further compared the per-state-point inference time, TabPFN with HPO infers in ca.  $1.7 \text{ ms}$  on an NVIDIA H100 GPU, whereas the SVGP requires ca.  $5.2 \text{ ms}$  on an Intel Xeon Silver 4214R CPU. On the corresponding hardware, the SVGP is about three times slower, reflecting the same accuracy versus cost trade-off. Regardless of the surrogate, both TabPFN and SVGP reproduce the IFT of the PCP-SAFT EoS + cDFT reference with high accuracy. To quantify the cost, we benchmarked the wall time to compute the IFT map (similar to Figure 2d) of the multicomponent  $\text{CO}_2$  mixtures listed in Table 2 with temperature-dependent  $k_{ij}$  on the AMD EPYC 9754 node and reduced it to a per-state-point cost. This cost ranges from ca.  $184 \text{ ms}$  for the binary system to ca.  $1791 \text{ ms}$  for the 8-component system. The cDFT wall time generally increases with the number of components, roughly by 1 order of magnitude from the 2-component to the 8-component feeds. However, the trend is not strictly monotonic because the cDFT cost also depends on the convergence behavior of the coupled density profiles, which varies between mixtures. Consequently, for each state point, the cDFT reference computation is approximately 2–3 orders of magnitude slower than TabPFN and 1–2 orders of magnitude slower than SVGP. This confirms that the ML surrogates reduce the physics-engine cost by 2–3 orders of magnitude with negligible loss of accuracy.

## 4. CONCLUSIONS

In this study, we used the PCP-SAFT EoS + cDFT framework to compute the phase equilibria and interfacial properties of multicomponent  $\text{CO}_2$ -rich mixtures. The phase-equilibrium properties include the bubble- and dew-point pressures in the *PT* phase envelope, while the interfacial properties include the IFT and interfacial thickness. The mixtures contain seven  $\text{CO}_2$  transportation-relevant impurities, namely,  $\text{CH}_4$ ,  $\text{H}_2$ ,  $\text{N}_2$ ,  $\text{Ar}$ ,  $\text{O}_2$ ,  $\text{CO}$ , and  $\text{H}_2\text{S}$ . To obtain reliable mixture predictions, the values of  $k_{ij}$  were first optimized for the 23 binary  $\text{CO}_2$ -containing pairs using available experimental  $P_{xy}$  data. For binary systems with sufficient data at different temperatures, temperature-dependent  $k_{ij}$  correlations were fitted to capture the variation in phase behavior with temperature. Using the optimized BIP, we generated a data set of ca.  $1.9 \times 10^4$  points for 100 multicomponent  $\text{CO}_2$ -rich mixtures. The mixture compositions were sampled by combining Dirichlet sampling, systematic scaling, and  $\text{CO}_2$  feeds from standards and operating pipelines in the ratio 0.7:0.2:0.1. A sensitivity analysis was performed for five multicomponent  $\text{CO}_2$ -rich mixtures with varying compositions and numbers of components from the 100 multicomponent  $\text{CO}_2$ -rich mixtures. The results showed that the dew-point pressure and interfacial thickness are less sensitive to impurity concentration than the bubble-point pressure and the IFT at the bubble pressure. The bubble-point pressure and bubble-point IFT both respond strongly to the impurity content. Screening of the full data-driven ML models was performed using the same input

variables used to plan an experiment or molecular simulation, namely temperature ( $T$ ), pressure ( $P$ ), and mixture composition ( $z$ ). Compared to the tested models, TabPFN with HyperParameter Optimization (HPO) provided the best balance between accuracy and robustness for all four properties, especially for the impurity-sensitive  $P_{\text{bubble}}$  and  $\gamma$ , for which HPO was most beneficial, reaching 5-fold CV RMSE values of 1.31 bar for  $P_{\text{bubble}}$  and  $0.040 \text{ mN m}^{-1}$  for  $\gamma$ . The next-best models were XGB, RF, and SVR, except for  $P_{\text{dew}}$ , where RF and XGB showed comparable performance. The SHAP analysis further confirmed that the prediction of  $P_{\text{bubble}}$  is influenced primarily by the  $\text{H}_2$  concentration, followed by  $T$  and  $P$ , whereas  $\gamma$  is governed mainly by  $T$  and  $P$ . It is also shown that TabPFN captures more composition-dependent impurity interactions than XGB. We also evaluated the number of feed compositions required to train the full data-driven ML models to a given accuracy by comparing random, stratified, and GPR-based active-learning sampling strategies. The results showed that active learning is the most data-efficient strategy and that  $\sim 50$  feed compositions already provide a favorable accuracy-cost balance for the  $\text{CO}_2$ -rich multicomponent mixtures. Active learning is therefore the strategy we recommend for extending the data set to new impurity types. In addition, we developed a physics-informed hybrid model for IFT, in which the semiempirical WSD correlation provides the baseline  $\gamma_{\text{base}}$ , and an ML model learns its residual obtained from the PCP-SAFT EoS + cDFT. The IFT is then reconstructed as  $\hat{\gamma} = \gamma_{\text{base}}(1 + \hat{\epsilon}_{\gamma})$ . Using SR, we obtained the interpretable correction  $\hat{\epsilon}_{\gamma} = \ln(\Delta\rho^*)/[\sqrt{P_r^*} + (\gamma_r^* - \Delta\rho^*)]$ , with a CV RMSE of  $0.11 \text{ mN m}^{-1}$ . The SVGP, trained on the full data set, achieved the lowest CV RMSE of  $0.023 \text{ mN m}^{-1}$ , outperforming the exact GPR, which was limited to a  $\sim 3500$ -point subset and yielded  $0.038 \text{ mN m}^{-1}$ . Notably, the best residual-correction model (SVGP,  $0.023 \text{ mN m}^{-1}$ ) outperformed the best full data-driven model for IFT (TabPFN with HPO,  $0.040 \text{ mN m}^{-1}$ ), confirming that embedding the WSD physical baseline improves the prediction of IFT. Finally, benchmarking the computational cost showed that a single cDFT evaluation of  $\gamma$  ranges from ca. 184 to 1791 ms as the system size increases from the binary to the 8-component system. In sharp contrast, the trained surrogates predict a state point in ca. 2 ms for TabPFN and ca. 5 ms for SVGP, resulting in substantial speed-ups with negligible loss of accuracy. This rapid inference is particularly relevant for dynamic  $\text{CO}_2$  transportation simulations, including CFD-based models and transient multiphase flow simulators such as Olga<sup>122</sup> and LedaFlow,<sup>123,124</sup> where efficient pipeline-scale computations require fast property evaluations or precomputed lookup tables. In such applications, the surrogate models act as a composition-aware replacement for precomputed lookup tables, taking  $T$ ,  $P$ , and  $z$  as direct inputs so that no new table is required per feed. To the best of our knowledge, this is the first systematic comparison of the physics engine, PCP-SAFT + cDFT, with two complementary ML surrogate strategies for the interfacial and phase-equilibrium properties of multicomponent  $\text{CO}_2$ -rich mixtures relevant to  $\text{CO}_2$  transportation. These strategies include a full data-driven model and a physics-informed residual-correction model. This framework enables rapid and uncertainty-aware estimation of IFT and saturation properties for multiphase process simulations and inverse design. This study supports the design of experiments and molecular simulations by prescreening the

most informative feed compositions for measurement or simulation and by identifying the most suitable ML model for each target property.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

All raw and processed data supporting this work are available in the Supporting Information and are archived in an open Zenodo repository (DOI: 10.5281/zenodo.21825307).

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.iecr.6c03388>.

Repository contains the raw and processed data, the trained machine-learning models, and their predictions supporting this manuscript (ZIP)

PCP-SAFT EoS pure-component parameters for  $\text{CO}_2$ ,  $\text{H}_2$ , Ar,  $\text{N}_2$ ,  $\text{CH}_4$ ,  $\text{O}_2$ , CO, and  $\text{H}_2\text{S}$ ; the methodologies used to determine the temperature-dependent binary interaction parameter (BIP) ( $k_{ij}$ ) by fitting to experimental  $P_{xy}$  data, the experimental  $P_{xy}$  data for the binary  $\text{CO}_2$  pairs, and the resulting  $k_{ij}(T)$  correlations; the stratified composition-space sampling algorithm with  $k$ -means clustering and elbow/silhouette cluster selection; the kernel variance active-learning sampling algorithm with ARD feature importance and learning curves; the full data-driven ML parity plots for bubble-point pressure, dew-point pressure, IFT, and interfacial thickness using RF, SVR, XGB, and TabPFN, together with the model hyperparameter optimization details; the SR pure-component IFT correlations (Guggenheim–Katayama form); and the hybrid WSD + SR residual-correction model for the IFT. The supporting material also provides the experimental  $P_{xy}$  data (in Excel format), the data sets used to train the ML models, which contain the raw and processed phase-equilibrium and interfacial-property data of the multicomponent  $\text{CO}_2$ -rich mixtures, and the trained TabPFN and SVGP models, together with their test and validation set predictions (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

Thijs J. H. Vlught – Engineering Thermodynamics, Process & Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Delft 2628 CB, The Netherlands; [orcid.org/0000-0003-3059-8712](https://orcid.org/0000-0003-3059-8712); Email: [T.J.H.Vlught@tudelft.nl](mailto:T.J.H.Vlught@tudelft.nl)

### Authors

Darshan Raju – Engineering Thermodynamics, Process & Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Delft 2628 CB, The Netherlands  
Roar Skartlien – Institute for Energy Technology, Kjeller N-2027, Norway  
Mahinder Ramdin – Engineering Thermodynamics, Process & Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Delft 2628 CB, The Netherlands; [orcid.org/0000-0002-8476-7035](https://orcid.org/0000-0002-8476-7035)

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.iecr.6c03388>

## Author Contributions

D.R.: Simulations, data analysis, and writing of original and revised manuscripts. R.S.: Conceptualization, data analysis, co-supervision of project, and proofreading of original and revised manuscript. M.R.: Funding acquisition, conceptualization, data analysis, co-supervision of the project, and proofreading of the original and revised manuscripts. T.J.H.V.: Funding acquisition, conceptualization, data curation, main supervision of the project, and writing of the original and revised manuscripts.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The work presented herein is part of the ENCASE project (A European Network of Research Infrastructures for CO<sub>2</sub> Transport and Injection). ENCASE has received funding from the European Union's Horizon Europe Research and Innovation program under grant Number 101094664. This work was also sponsored by NWO domain Science for the use of supercomputer facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (The Netherlands Organization for Scientific Research, NWO). The authors acknowledge the use of computational resources of the DelftBlue supercomputer, provided by the Delft High Performance Computing Center (<https://www.tudelft.nl/dhpc>).

## REFERENCES

- (1) Lee, H.; Calvin, K.; Dasgupta, D.; Krinner, G.; Mukherji, A.; Thorne, P.; Trisos, C.; Romero, J.; Aldunce, P.; Barrett, K.; Blanco, G.; Cheung, W. W. L.; Connors, S. L.; Denton, F.; Diongue-Niang, A.; Dodman, D.; Garschagen, M.; Geden, O.; Hayward, B.; Jones, C.; Jotzo, F.; Lasco, R.; Lee, J.-Y.; Masson-Delmotte, V.; Meinshausen, M.; Mintenbeck, K.; Mokssit, A.; Otto, F. E. L.; Pathak, M.; Pirani, A.; Poloczanska, E.; Pörtner, H.-O.; Revi, A.; Roberts, D. C.; Romero, J.; Roy, J.; Ruane, A. C.; Skea, J.; Shukla, P. R.; Slade, R.; Slangen, A.; Sokona, Y.; Sörensson, A. A.; Tignor, M.; van Vuuren, D.; Wei, Y.-M.; Winkler, H.; Zhai, P.; Zommers, Z. *Climate Change: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 2023. [https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC\\_AR6\\_SYR\\_FullVolume.pdf](https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf) (accessed Sept 26, 2025).
- (2) Jones, M. W.; Peters, G. P.; Gasser, T.; Andrew, R. M.; Schwingshackl, C.; Gütschow, J.; Houghton, R. A.; Friedlingstein, P.; Pongratz, J.; Le Quéré, C. National contributions to climate change due to historical emissions of carbon dioxide, methane, and nitrous oxide since 1850. *Sci. Data* **2023**, *10*, No. 155.
- (3) Boot-Handford, M. E.; Abanades, J. C.; Anthony, E. J.; Blunt, M. J.; Brandani, S.; Mac Dowell, N.; Fernández, J. R.; Ferrari, M.-C.; Gross, R.; Hallett, J. P.; Haszeldine, S. R.; Heptonstall, P.; Lyngfelt, A.; Makuch, Z.; Mangano, E.; Porter, R. T.; Pourkashanian, M.; Rochelle, G. T.; Shah, N.; Yao, J. G.; Fennell, P. S. Carbon capture and storage update. *Energy Environ. Sci.* **2014**, *7*, 130–189.
- (4) Wilberforce, T.; Olabi, A.; Sayed, E. T.; Elsaid, K.; Abdelkareem, M. A. Progress in carbon capture technologies. *Sci. Total Environ.* **2021**, *761*, No. 143203.
- (5) Yafiee, O. A.; Mumtaz, F.; Kumari, P.; Karanikolos, G. N.; Decarlis, A.; Dumée, L. F. Direct air capture (DAC) vs. Direct ocean capture (DOC)-A perspective on scale-up demonstrations and environmental relevance to sustain decarbonization. *Chem. Eng. J.* **2024**, *497*, No. 154421.
- (6) Simonsen, K. R.; Hansen, D. S.; Pedersen, S. Challenges in CO<sub>2</sub> transportation: Trends and perspectives. *Renewable Sustainable Energy Rev.* **2024**, *191*, No. 114149.
- (7) Porthos. Porthos CO<sub>2</sub> Specifications, 2021. <https://www.porthosco2.nl/wp-content/uploads/2021/09/CO2-specifications.pdf> (accessed Aug 25, 2025).
- (8) Northern Lights About the Longship Project, 2026. <https://norlights.com/about-the-longship-project/> (accessed May 12, 2026).
- (9) Onyebuchi, V. E.; Kolios, A.; Hanak, D. P.; Biliyok, C.; Manovic, V. A systematic review of key challenges of CO<sub>2</sub> transport via pipelines. *Renewable Sustainable Energy Rev.* **2018**, *81*, 2563–2583.
- (10) Munkejord, S. T.; Hammer, M.; Løvseth, S. W. CO<sub>2</sub> transport: Data and models-A review. *Appl. Energy* **2016**, *169*, 499–523.
- (11) Shirley, P.; Myles, P. Quality Guidelines for Energy System Studies: CO<sub>2</sub> Impurity Design Parameters, 2019. <https://www.osti.gov/servlets/purl/1566771> (accessed Sept 26, 2025).
- (12) De Visser, E.; Hendriks, C.; Barrio, M.; Mølnvik, M. J.; de Koeijer, G.; Liljemark, S.; Le Gallo, Y. Dynamis CO<sub>2</sub> Quality Recommendations. *Int. J. Greenhouse Gas Control* **2008**, *2*, 478–484.
- (13) Brownsort, P. A. 1st Report of the Thematic Working Group on: CO<sub>2</sub> Transport, Storage and Networks 2019 [https://www.ccusnetwork.eu/sites/default/files/TG3\\_Briefing-CO2-Specifications-for-Transport.pdf](https://www.ccusnetwork.eu/sites/default/files/TG3_Briefing-CO2-Specifications-for-Transport.pdf) (accessed August 25, 2025).
- (14) Nazeri, M.; Haghighi, H.; McKay, C.; Erickson, D.; Zhai, S. In *Impact of CO<sub>2</sub> Specifications on Design and Operation Challenges of CO<sub>2</sub> Transport and Storage Systems in CCUS*, SPE Offshore Europe Conference and Exhibition, 2021; PE-205472-MS. DOI: 10.2118/205472-MS (accessed August 25, 2025).
- (15) Chapoy, A.; Nazeri, M.; Kapateh, M.; Burgass, R.; Coquelet, C.; Tohidi, B. Effect of impurities on thermophysical properties and phase behaviour of a CO<sub>2</sub>-rich system in CCS. *Int. J. Greenhouse Gas Control* **2013**, *19*, 92–100.
- (16) Morland, B. H.; Dugstad, A.; Svenningsen, G. Corrosion of Carbon Steel in Dense Phase CO<sub>2</sub> with Water Above and Below the Solubility Limit. *Energy Proc.* **2017**, *114*, 6752–6765.
- (17) Klinkby, L.; Nielsen, C. M.; Krogh, E.; Smith, I. E.; Palm, B.; Bernstone, C. Simulating rapidly fluctuating CO<sub>2</sub> flow into the Vedsted CO<sub>2</sub> pipeline, injection well and reservoir. *Energy Proc.* **2011**, *4*, 4291–4298.
- (18) Munkejord, S. T.; Hammer, M. Depressurization of CO<sub>2</sub>-rich mixtures in pipes: Two-phase flow modelling and comparison with experiments. *Int. J. Greenhouse Gas Control* **2015**, *37*, 398–411.
- (19) Coutinho, R. P.; Hadjicharalambous, M. In *Demystifying Vapor Collapse in CO<sub>2</sub> Injection Wells*, Proceedings of the 17th International Conference on Greenhouse Gas Control Technologies (GHGT-17), Calgary, Canada, 2024. <https://ssrn.com/abstract/5010518> (accessed Sept 26, 2025).
- (20) Henkes, R. A. W. M.; Ellepola, J. E.; Rao, A. In *Simulations of Fast Transients for Two-Phase CO<sub>2</sub> Flow in Pipes*, ISAVFT International Conference on Multiphase Production Technology, 2025; ISAVFT-2025-471. <https://onepetro.org/isavftmpt/proceedings/MPT25/MPT25/ISAVFT-2025-471/795645> (accessed June 9, 2026).
- (21) Wilhelmsen, Ø.; Aasen, A.; Skaugen, G.; Aursand, P.; Austegard, A.; Aursand, E.; Gjennestad, M. A.; Lund, H.; Linga, G.; Hammer, M. Thermodynamic Modeling with Equations of State: Present Challenges with Established Methods. *Ind. Eng. Chem. Res.* **2017**, *56*, 3503–3515.
- (22) Log, A. M.; Diky, V.; Huber, M. L. Assessment of a Parachor Model for the Surface Tension of Binary Mixtures. *Int. J. Thermophys.* **2023**, *44*, No. 110.
- (23) Skarsvåg, H. L.; Hammer, M.; Munkejord, S. T.; Log, A. M.; Dumoulin, S.; Gruben, G. Towards an engineering tool for the prediction of running ductile fractures in CO<sub>2</sub> pipelines. *Process Saf. Environ. Prot.* **2023**, *171*, 667–679.
- (24) Stephan, S.; Bråten, V.; Hasse, H. Mass transfer at vapor-liquid interfaces of H<sub>2</sub>O + CO<sub>2</sub> mixtures studied by molecular dynamics simulation. *J. Non-Equilib. Thermodyn.* **2024**, *49*, 441–461.
- (25) Raju, D.; Skartlien, R.; Ramdin, M.; Vlught, T. J. H. Vapor-Liquid Interfacial Properties of CO<sub>2</sub> Mixtures for Sequestration Applications: Molecular Simulations, Classical Density Functional

Theory, and Equations of State. *Ind. Eng. Chem. Res.* **2026**, *65*, 5213–5238.

(26) Stephan, S. Vapor-Liquid Interfaces: Molecular Simulation, Density Gradient Theory, and Experiments. Ph.D. Thesis; Technische Universität: Kaiserslautern, 2020. <https://kluedo.ub.rptu.de/frontdoor/deliver/index/docId/6089/file/Dissertation=Simon=Stephan.pdf> (accessed June 9, 2026).

(27) Chow, Y. F.; Eriksen, D. K.; Galindo, A.; Haslam, A. J.; Jackson, G.; Maitland, G. C.; Trusler, J. M. Interfacial tensions of systems comprising water, carbon dioxide and diluent gases at high pressures: Experimental measurements and modelling with SAFT-VR Mie and square-gradient theory. *Fluid Phase Equilib.* **2016**, *407*, 159–176.

(28) Jiao, F.; Ghafri, S. Z. A.; Seneviratne, K. N.; Akhfar, M.; Hughes, T. J.; Johns, M. L.; May, E. F. Interfacial tension measurements of methane + propane binary and methane + propane + n-heptane ternary mixtures at low temperatures. *J. Chem. Thermodyn.* **2022**, *171*, No. 106786.

(29) Kontogeorgis, G. M.; Dohrn, R.; Economou, I. G.; de Hemptinne, J.-C.; Ten Kate, A.; Kuitunen, S.; Mooijer, M.; Žilnik, L. F.; Vesović, V. Industrial requirements for thermodynamic and transport properties: 2020. *Ind. Eng. Chem. Res.* **2021**, *60*, 4987–5013.

(30) Hendriks, E.; Kontogeorgis, G. M.; Dohrn, R.; de Hemptinne, J.-C.; Economou, I. G.; Žilnik, L. F.; Vesovic, V. Industrial requirements for thermodynamics and transport properties. *Ind. Eng. Chem. Res.* **2010**, *49*, 11131–11141.

(31) Li, H.; Jakobsen, J. P.; Wilhelmsen, Ø.; Yan, J. PVTxy properties of CO<sub>2</sub> mixtures relevant for CO<sub>2</sub> capture, transport and storage: Review of available experimental data and theoretical models. *Appl. Energy* **2011**, *88*, 3567–3579.

(32) Schenk, M. R.; Köddermann, T.; Kirschner, K. N.; Knauer, S.; Reith, D. Molecular dynamics in the energy sector: Experiment and modeling of the CO<sub>2</sub>/CH<sub>4</sub> mixture. *J. Chem. Eng. Data* **2020**, *65*, 1117–1123.

(33) Rehner, P.; Bauer, G. Application of Generalized (Hyper-) Dual Numbers in Equation of State Modeling. *Front. Chem. Eng.* **2021**, *3*, No. 758090.

(34) Miqueu, C.; Mendiboure, B.; Graciaa, A.; Lachaise, J. Modelling of the surface tension of pure components with the gradient theory of fluid interfaces: a simple and accurate expression for the influence parameters. *Fluid Phase Equilib.* **2003**, *207*, 225–246.

(35) Pan, Z.; Trusler, J. M. Measurement and modelling of the interfacial tensions of CO<sub>2</sub>+ decane-iododecane mixtures at high pressures and temperatures. *Fluid Phase Equilib.* **2023**, *566*, No. 113700.

(36) Llovel, F.; Galindo, A.; Blas, F. J.; Jackson, G. Classical density functional theory for the prediction of the surface tension and interfacial properties of fluids mixtures of chain molecules based on the statistical associating fluid theory for potentials of variable range. *J. Chem. Phys.* **2010**, *133*, No. 024704.

(37) Algaba, J.; Miguez, J. M.; Mendiboure, B.; Blas, F. J. An accurate density functional theory for the vapor-liquid interface of chain molecules based on the statistical associating fluid theory for potentials of variable range for Mie chainlike fluids. *Phys. Chem. Chem. Phys.* **2019**, *21*, 11937–11948.

(38) Algaba, J.; Mendiboure, B.; Gomez-Alvarez, P.; Blas, F. J. Density functional theory for the prediction of interfacial properties of molecular fluids within the SAFT- $\gamma$  coarse-grained approach. *RSC Adv.* **2022**, *12*, 18821–18833.

(39) Lafitte, T.; Apostolou, A.; Avendaño, C.; Galindo, A.; Adjiman, C. S.; Müller, E. A.; Jackson, G. Accurate statistical associating fluid theory for chain molecules formed from Mie segments. *J. Chem. Phys.* **2013**, *139*, No. 154504.

(40) Ferguson, A. The Parachor and Molecular Volume. *Nature* **1930**, *125*, 597.

(41) Winterfeld, P. H.; Scriven, L.; Davis, H. An approximate theory of interfacial tensions of multicomponent systems: Applications to binary liquid-vapor tensions. *AIChE J.* **1978**, *24*, 1010–1014.

(42) Mulero, A.; Hernández, A.; Vadillo-Rodríguez, V.; Cachadiña, I. Correlations of surface tension for mixtures of n-alkanes as a function

of the composition: applicability and performance analysis of existing models. *Phys. Chem. Chem. Phys.* **2025**, *27*, 12812–12836.

(43) Redlich, O.; Kister, A. T. Algebraic Representation of Thermodynamic Properties and the Classification of Solutions. *Ind. Eng. Chem.* **1948**, *40*, 341–345.

(44) Gupta, S.; Elliott, J. R.; Anderko, A.; Crosthwaite, J.; Chapman, W. G.; Lira, C. T. Current practices and continuing needs in thermophysical properties for the chemical industry. *Ind. Eng. Chem. Res.* **2023**, *62*, 3394–3427.

(45) Agbodekhe, B. P.; Carlozo, M. N.; Abranches, D. O.; Jones, K. D.; Dowling, A. W.; Maginn, E. J. Enhanced thermophysical property prediction with uncertainty quantification using group contribution-Gaussian process regression. *Mol. Syst. Des. Eng.* **2026**, *11*, 85–106.

(46) Weast, R. C. *CRC Handbook of Chemistry and Physics*, 105th ed.; CRC Press Inc.: Boca Raton, FL, USA, 1985.

(47) Mouallem, J.; Raza, A.; Glatz, G.; Mahmoud, M.; Arif, M. Estimation of CO<sub>2</sub>-brine interfacial tension using machine learning: implications for CO<sub>2</sub> geo-storage. *J. Mol. Liq.* **2024**, *393*, No. 123672.

(48) Albà, C. G.; Alkhatib, I. I.; Vega, L. F.; Llovel, F. Machine learning integration in thermodynamics: Predicting CO<sub>2</sub> mixture saturation properties for sustainable refrigeration applications. *J. CO<sub>2</sub> Util.* **2025**, *95*, No. 103072.

(49) Sharma, S.; Yang, P.; Liu, Y.; Rossi, K. R.; Bai, P.; Rigutto, M. S.; Zuidema, E.; Agarwal, U.; Baur, R.; Calero, S.; Dubbeldam, D.; Vlucht, T. J. H. Machine Learning-Based Predictions of Henry Coefficients for Long-Chain Alkanes in One-Dimensional Zeolites: Application to Hydroisomerization. *J. Phys. Chem. C* **2025**, *129*, 18234–18249.

(50) Johnsson, E.; Sharma, S.; Rao, A. G.; Dubbeldam, D.; Calero, S.; Vlucht, T. J. H. Predicting the Maximum Loading in Zeolites for Hydroisomerization Applications: A Machine Learning Approach. *J. Phys. Chem. C* **2026**, *130*, 4299–4314.

(51) Zhang, J.; Feng, Q.; Wang, S.; Zhang, X.; Wang, S. Estimation of CO<sub>2</sub>-brine interfacial tension using an artificial neural network. *J. Supercrit. Fluids* **2016**, *107*, 31–37.

(52) Kamari, A.; Pournik, M.; Rostami, A.; Amirlatif, A.; Mohammadi, A. H. Characterizing the CO<sub>2</sub>-brine interfacial tension (IFT) using robust modeling approaches: A comparative study. *J. Mol. Liq.* **2017**, *246*, 32–38.

(53) Song, T.; Zhu, W.; Emami-Meybodi, H.; Jiang, Y.; Chen, S.; Yue, M.; Mahani, H.; Liao, Q.; Iglaier, S.; Pan, B. An explicit machine learning model for brine-gas interfacial tension prediction: Implications for H<sub>2</sub>, CH<sub>4</sub>, and CO<sub>2</sub> geo-storage. *Fuel* **2026**, *405*, No. 136502.

(54) Makke, M. H.; Ghorayeb, K. Integrating physics and machine learning: Viscosity estimation in impure CO<sub>2</sub> transport systems. *Fuel* **2026**, *406*, No. 137078.

(55) Makke, M. H.; Ghorayeb, K. In *Machine Learning-Based Assessment of Impurity Effects on Thermophysical Properties in CO<sub>2</sub>-Rich Mixtures for CCUS Applications*, D021S015R003, International Petroleum Technology Conference, 2026.

(56) Zhao, B.; Ding, L.; Su, Y. A Bayesian network-based robust framework for determining density, viscosity, and thermal conductivity of near-critical-state CO<sub>2</sub> applied in CCUS. *Chemom. Intell. Lab. Syst.* **2026**, *269*, No. 105636.

(57) Gross, J. An Equation-of-State Contribution for Polar Components: Quadrupolar Molecules. *AIChE J.* **2005**, *51*, 2556–2568.

(58) Gross, J.; Sadowski, G. Application of the perturbed-chain SAFT equation of state to associating systems. *Ind. Eng. Chem. Res.* **2002**, *41*, 5510–5515.

(59) Gross, J.; Sadowski, G. Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules. *Ind. Eng. Chem. Res.* **2001**, *40*, 1244–1260.

(60) Rehner, P.; Bauer, G.; Gross, J. FeO<sub>3</sub>: An Open-Source Framework for Equations of State and Classical Density Functional Theory. *Ind. Eng. Chem. Res.* **2023**, *62*, 5347–5357.

(61) Kunz, O.; Wagner, W. The GERG-2008 wide-range equation of state for natural gases and other mixtures: an expansion of GERG-2004. *J. Chem. Eng. Data* **2012**, *57*, 3032–3091.

- (62) Rehner, P. Interfacial Properties using Classical Density Functional Theory: Curved Interfaces and Surfactants. Ph.D. Thesis, University of Stuttgart, Institute of Thermodynamics and Thermal Process Engineering, 2021. <https://elib.uni-stuttgart.de/server/api/core/bitstreams/2de1098b-4778-4245-8fe6-cb8d0a77deaf/content> (accessed June 9, 2026).
- (63) Gross, J.; Vrabec, J. An Equation-of-State Contribution for Polar Components: Dipolar Molecules. *AIChE J.* **2006**, *52*, 1194–1204.
- (64) Evans, R. The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. *Adv. Phys.* **1979**, *28*, 143–200.
- (65) Telo da Gama, M.; Evans, R. The structure and surface tension of the liquid-vapour interface near the upper critical end point of a binary mixture of Lennard-Jones fluids: I. The two phase region. *Mol. Phys.* **1983**, *48*, 229–250.
- (66) Lutsko, J. F. *Advances in Chemical Physics*; Rice, S. A., Ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2010; Vol. 144; Chapter 1, pp 1–92.
- (67) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*, 3rd ed.; Elsevier: Amsterdam, 2006.
- (68) Mairhofer, J.; Gross, J. Numerical aspects of classical density functional theory for one-dimensional vapor-liquid interfaces. *Fluid Phase Equilib.* **2017**, *444*, 1–12.
- (69) Tarazona, P. Free-energy density functional for hard spheres. *Mol. Phys.* **1985**, *55*, 557–572.
- (70) Fleckenstein, F.; Becker, S.; Hasse, H.; Stephan, S. Vapor-liquid interfacial properties of the system acetone+ CO<sub>2</sub>: Experiments, molecular simulation, density gradient theory, and density functional theory. *Fluid Phase Equilib.* **2025**, *596*, No. 114436.
- (71) Anderson, D. G. Iterative procedures for nonlinear integral equations. *JACM* **1965**, *12*, 547–560.
- (72) Mairhofer, J.; Gross, J. Modeling of interfacial properties of multicomponent systems using density gradient theory and PCP-SAFT. *Fluid Phase Equilib.* **2017**, *439*, 31–42.
- (73) Shen, G.; Sun, Y.; Zhang, X.; Gao, X.; Qian, Y.; Lu, X.; Ji, X. Partition and selectivity of electrolytes in cylindrical nanopores with heterogeneous surface charge. *J. Mol. Liq.* **2021**, *340*, No. 116839.
- (74) Maidl, P.; Langenbach, K.; Frank, F. Achieving Convergence in Multiphase Multicomponent Density Gradient Theory Calculations through Regularization. *Ind. Eng. Chem. Res.* **2024**, *63*, 14367–14380.
- (75) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*, 3rd ed.; Academic Press: Elsevier, Oxford, UK, 2023.
- (76) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, 2nd ed.; Oxford University Press: Oxford, UK, 2017.
- (77) Fowler, R. H. A tentative statistical theory of Macleod's equation for surface tension, and the parachor. *Proc. R. Soc. London, Ser. A* **1937**, *159*, 229–246.
- (78) Kirkwood, J. G.; Buff, F. P. The statistical mechanical theory of surface tension. *J. Chem. Phys.* **1949**, *17*, 338–343.
- (79) Raju, M.; Banuti, D. T.; Ma, P. C.; Ihme, M. Widom Lines in Binary Mixtures of Supercritical Fluids. *Sci. Rep.* **2017**, *7*, No. 3027.
- (80) Simeoni, G. G.; Bryk, T.; Gorelli, F. A.; Krisch, M.; Ruocco, G.; Santoro, M.; Scopigno, T. The Widom Line as the Crossover between Liquid-Like and Gas-Like Behaviour in Supercritical Fluids. *Nat. Phys.* **2010**, *6*, 503–507.
- (81) Vo Thanh, H.; Rahimi, M.; Tangparitkul, S.; Promsuk, N. Modeling the thermal transport properties of hydrogen and its mixtures with greenhouse gas impurities: A data-driven machine learning approach. *Int. J. Hydrogen Energy* **2024**, *83*, 1–12.
- (82) Liu, Y.; Hong, W.; Cao, B. Machine learning for predicting thermodynamic properties of pure fluids and their mixtures. *Energy* **2019**, *188*, No. 116091.
- (83) Jirasek, F.; Hasse, H. Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures. *Annu. Rev. Chem. Biomol. Eng.* **2023**, *14*, 31–51.
- (84) Succetti, F.; Panella, M.; Giannitrapani, P.; Rigo, J.-C.; Colonnese, S. Machine Learning for Fluid Thermodynamics: State of the Art and Open Challenges. *IEEE Access* **2025**, *13*, 171073–171092.
- (85) Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S. B.; Schirrmeister, R. T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326.
- (86) Grinsztajn, L.; Flöge, K.; Key, O.; Birkel, F.; Jund, P.; Roof, B.; Jäger, B.; Safaric, D.; Alessi, S.; Hayler, A.; Manium, M.; Yu, R.; Jablonski, F.; Hoo, S. B.; Garg, A.; Robertson, J.; Bühler, M.; Moroshan, V.; Purucker, L.; Cornu, C.; Wehrhahn, L. C.; Bonetto, A.; Schölkopf, B.; Gambhir, S.; Hollmann, N.; Hutter, F. TabPFN-2.5: Advancing the State of the Art in Tabular Foundation Models. 2026, arXiv:2511.08667. arXiv.org e-Print archive. <https://arxiv.org/abs/2511.08667> (accessed March 5, 2026).
- (87) Hensman, J.; Matthews, A. G. d.; Ghahramani, Z. In *Scalable Variational Gaussian Process Classification*, Artificial Intelligence and Statistics (AISTATS), 2015; pp 351–360.
- (88) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- (89) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. In *Support Vector Regression Machines*, Advances in Neural Information Processing Systems, 1997; pp 155–161.
- (90) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (91) Chen, T.; Guestrin, C. In *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; pp 785–794 (accessed March 5th, 2026).
- (92) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; Li, M.; Xie, J.; Lin, M.; Geng, Y.; Li, Y.; Yuan, J.; Cortes, D. *xgboost: Extreme Gradient Boosting*, Python package version 3.2.1, 2026. <https://github.com/dmlc/xgboost>.
- (93) Prior Labs TabPFN Extensions: Community extensions for TabPFN, 2025. <https://github.com/PriorLabs/tabpfm-extensions> (accessed May 16, 2026).
- (94) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
- (95) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. In *CatBoost: Unbiased Boosting with Categorical Features*, Advances in Neural Information Processing Systems 31, 2018; pp 6639–6649.
- (96) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (97) Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. 2023, arXiv:2305.01582. arXiv.org e-Print archive. <https://arxiv.org/abs/2305.01582> (accessed May 9, 2026).
- (98) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*, 1st ed.; MIT Press: Cambridge, MA, USA, 2006.
- (99) Titsias, M. K. In *Variational Learning of Inducing Variables in Sparse Gaussian Processes*, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), 2009; pp 567–574 (accessed March 12, 2026).
- (100) Hensman, J.; Fusi, N.; Lawrence, N. D. In *Gaussian Processes for Big Data*, Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI), 2013; pp 282–290.
- (101) Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; Saul, L. K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233.
- (102) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877.
- (103) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

- (104) Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; Wilson, A. G. In *GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration*, Advances in Neural Information Processing Systems, 2018; pp 7576–7586.
- (105) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*, 1st ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 1995.
- (106) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
- (107) Bishop, C. M. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, 2006.
- (108) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
- (109) Xiang, Y.; Tang, Y.-H.; Gong, Z.; Liu, H.; Wu, L.; Lin, G.; Sun, H. Efficient exploration of chemical compound space using active learning for prediction of thermodynamic properties of alkane molecules. *J. Chem. Inf. Model.* **2023**, *63*, 6515–6524.
- (110) Xiang, Y.; Tang, Y.-H.; Liu, H.; Lin, G.; Sun, H. Predicting single-substance phase diagrams: A kernel approach on graph representations of molecules. *J. Phys. Chem. A* **2021**, *125*, 4488–4497.
- (111) Diky, V.; Muzny, C. D.; Smolyanitsky, A. Y.; Bazyleva, A.; Chirico, R. D.; Magee, J. W.; Paulechka, Y.; Kazakov, A. F.; Townsend, S. A.; Lemmon, E. W. *ThermoData Engine (TDE) Version 10: Pure Compounds, Binary Mixtures, Ternary Mixtures, and Chemical Reactions: NIST Standard Reference Database 103b*, 2015.
- (112) Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1987**, *100*, 441–471.
- (113) Massey, F. J., Jr The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78.
- (114) Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, *27*, 832–837.
- (115) Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
- (116) Okoro, F.; Owuna, F. J.; Chapoy, A.; Ahmadi, P.; Burgass, R. Bubble points and densities of H<sub>2</sub> (up to 5%) in CO<sub>2</sub>-rich binary systems. *Fluid Phase Equilib.* **2025**, *592*, 114321.
- (117) Antoine, M. C. Nouvelle relation entre les tensions et les temperatures. *C. R. Hebd. Seances Acad. Sci.* **1888**, *107*, 681–684.
- (118) Guggenheim, E. A. The principle of corresponding states. *J. Chem. Phys.* **1945**, *13*, 253–261.
- (119) Rowlinson, J. S.; Widom, B. *Molecular Theory of Capillarity*, 1st ed.; Courier Corporation: New York: US, 2013.
- (120) Rowlinson, J. S.; Swinton, F. *Liquids and Liquid Mixtures: Butterworths Monographs in Chemistry*, 3rd ed.; Butterworth-Heinemann, Elsevier: Oxford, UK, 2013.
- (121) Lundberg, S. M.; Lee, S.-I. In *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems 30, 2017; pp 4765–4774.
- (122) SLB OLGA Dynamic Multiphase Flow Simulator, SLB, 2026. <https://www.slb.com/products-and-services/delivering-digital-at-scale/software/olga> (accessed June 25, 2026).
- (123) Kongsberg Digital LedaFlow: Transient Multiphase Flow Simulation Software, Kongsberg Digital, 2026. <https://kongsbergdigital.com/products/ledaflow> (accessed June 25, 2026).
- (124) SINTEF LedaFlow, SINTEF, 2026. <https://www.sintef.no/en/software/ledaflow/> (accessed June 25, 2026).