The Journal of Chemical Physics

RESEARCH ARTICLE | OCTOBER 24 2025

Accelerating phase diagram construction through activity coefficient prediction

Special Collection: 2025 JCP Emerging Investigators Special Collection

Mohsen Farshad ; Fathya Y. M. Salih ; Dinis O. Abranches ; Yamil J. Colón

J. Chem. Phys. 163, 164118 (2025)
https://doi.org/10.1063/5.0286520

AIP Publishing

The Journal of Chemical Physics

ARTICLE

pubs.aip.org/aip/jcp

Accelerating phase diagram construction through activity coefficient prediction

Cite as: J. Chem. Phys. 163, 164118 (2025); doi: 10.1063/5.0286520
Submitted: 19 June 2025 • Accepted: 29 September 2025 •
Published Online: 24 October 2025

View Online Export Citation CrossMark

Mohsen Farshad,[1] Fathya Y. M. Salih,[1] Dinis O. Abranches,[2,a)] and Yamil J. Colón[3,b)]

AFFILIATIONS

[1] Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA
[2] CICECO – Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Aveiro 3810-193, Portugal
[3] Department of Chemical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA

Note: This paper is a part of the 2025 JCP Emerging Investigators Special Collection.
[a)] **Electronic mail:** jdinis@ua.pt
[b)] **Author to whom correspondence should be addressed:** ycolon@nd.edu

ABSTRACT

Obtaining phase diagrams from molecular simulations remains computationally demanding due to the need for extensive sampling of coexistence conditions and large system sizes. This study demonstrates a novel methodology for efficiently predicting phase behavior in Lennard-Jones mixtures by leveraging machine learning. Here, we train a Gaussian process (GP) model on Kirkwood–Buff Integrals (KBIs) to establish a predictive link between KBIs and activity coefficients—a key thermodynamic quantity encoding deviations from ideality. Through the incorporation of KBI trends, the GP model leads to the prediction of the activity coefficients of two new systems without prior knowledge of their phase behavior, eliminating the need for direct coexistence simulations and significantly reducing computational cost. This framework has broad applicability in computational thermodynamics, offering a scalable strategy for studying complex mixtures with tunable interactions.

I. INTRODUCTION

Phase diagrams are fundamental tools in thermodynamics and materials science, providing crucial insights into the stability and transitions between different phases of matter under varying conditions of temperature, pressure, and composition.[1,2] Phase diagrams help with the development of new materials, the improvement of industrial processes,[3] and the design of separation techniques,[1] such as distillation,[4] crystallization,[5] absorption,[6] solubilization,[7] and liquid–liquid extraction.[8]

Phase diagrams can be categorized based on the system under study and the phases involved.[9] Furthermore, the parameters that indicate phase transitions depend on the specific focus of the study.[10] The vapor–liquid equilibrium (VLE), which is of our particular interest in this work, encompasses liquid and vapor phases with a coexistence region that can be quantified through the construction of the phase diagram.[11] Despite their significance, the construction of phase diagrams remains expensive and time-consuming. Experimentally, studying phase behavior is challenging due to the complexities of synthesis and characterization depending on the system under investigation.[12–14] To overcome these challenges, theoretical approaches have been developed to model phase behavior, which we briefly discuss in the following paragraph.

Traditional thermodynamic modeling, such as regular solution methods[15,16] and Flory–Huggins theory,[17] has been widely successful, especially in VLEs using models such as the Non-Random Two-Liquid (NRTL) model or Margules.[10,18] These approaches rely on excess Gibbs energy models fitted to experimental data, requiring measured VLEs to compute activity coefficients before describing phase behavior.[10] While powerful for modeling and description, they are inherently limited to systems where sufficient experimental data exist.[15,19–21] Alternatively, phase diagrams can be obtained from computational simulations, but these simulations are also costly. For example, molecular dynamics (MD) simulations using direct coexistence methods require very large simulation boxes to maintain a realistic interface-to-volume ratio and typically demand wide sweeps over thermodynamic conditions.[22,23] Furthermore, MD simulations that use free energy calculations, such as umbrella sampling, require long simulations and careful setup. Although Monte Carlo (MC) methods remain the most popular for phase diagram construction, they are computationally expensive due to the need for extensive sampling across many thermodynamic states, especially in systems with complex interactions or large particle numbers.[24–28]

Instead of direct coexistence simulations, which require large biphasic systems, long equilibration to stabilize both phases and the liquid–vapor interface, and careful handling of interfacial stability, our method uses only single-phase liquid simulations of binary mixtures at fixed compositions. These homogeneous simulations involve more than two times fewer particles than comparable coexistence setups and equilibrate substantially faster, while still providing Kirkwood–Buff integrals (KBIs) as structural input to the Gaussian process (GP) model. Thus, there is a need for more efficient methods, and machine learning (ML) has the potential to bridge this gap.[29,30]

Recent ML approaches to phase diagram construction employ different methodologies depending on the type and availability of data.[31–33] The most common are deep neural networks (DNNs), which can capture highly nonlinear relationships but typically require extensive training datasets.[33] Alternatively, probabilistic models, such as Gaussian processes (GPs), have been explored for small data regimes, offering built-in uncertainty estimation and improved generalization in sparse-data settings.[34] In materials science, GPs have been applied to model atomistic properties, construct interatomic potentials,[35] and predict phase behavior by learning mappings between thermodynamic conditions and system properties.[36] In their work, Ladygin *et al.* combined GPs with thermodynamic integration data to reconstruct free energy surfaces and compute phase diagrams with quantified uncertainty. GPs also enable active learning strategies, where uncertainty guides targeted data acquisition.[33] For example, they have been used to iteratively guide molecular simulations toward phase boundaries, reducing the total number of required simulations and improving efficiency in phase diagram construction.[31] Our work builds upon this class of probabilistic models and introduces a novel approach that combines thermodynamic and structural properties with a GP model. We propose a ML-based approach using GP regression to develop an efficient method for predicting activity coefficients (γ) from Kirkwood–Buff Integrals (KBIs).[29,30,37,38]

Activity coefficients, a key thermodynamic quantity derived from excess chemical potentials, govern phase behavior and describe deviations from ideal solution behavior, providing essential information about intermolecular interactions in phase equilibria.[10] Meanwhile, KBIs quantify the spatial correlations between particles in a given system, offering insights into local structural organization.[39,40] KBIs are calculated by integrating the radial distribution functions (RDFs) obtained from simulation trajectories, thereby connecting the microscopic structure to macroscopic thermodynamics.[39–43] These two properties are connected in this work using a GP model, enabling the prediction of activity coefficients from structural correlation data. For the training dataset, we calculate activity coefficients from coexistence simulations and KBIs from single-phase simulations for systems with different Lennard-Jones interaction parameters. The trained model predicts activity coefficients for new systems based solely on KBIs calculated from single-phase simulations, eliminating the need for coexistence simulations.

By training a GP on KBI derived from radial distribution functions (RDFs),[41–44] we establish a predictive framework that allows us to map KBIs to activity coefficients. Using these predicted activity coefficients, we reconstruct phase diagrams that show reasonable agreement with those obtained from biphasic MD simulations. This workflow enables the construction of phase diagrams without prior knowledge of activity coefficients, relying only on KBIs from single-phase systems and the predictive power of the trained GP model.

II. METHODS

A. Molecular dynamics (MD)

The systems considered are binary Lennard-Jones (LJ) fluids, with a interatomic potential described by the following equation:[27]

$$U_{ij}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + U_{\circ}, \tag{1}$$

where $U_{\circ}$ is a shifting term equal to the LJ potential at the cutoff distance, ensuring that the potential smoothly converges to zero at a cutoff distance of 2.5σ. The parameters ε and σ represent the depth of the potential well and the particle size, respectively. For the first and second components, we set $\varepsilon_1 = 1$ and $\varepsilon_2 = 0.9$, while $\sigma_1 = \sigma_2 = 1$. All parameters are expressed in reduced Lennard-Jones units, where ε has units of energy and σ has units of length, and reported relative to the LJ parameters of component 1. LJ parameters and simulation conditions were selected based on the work of Stephan *et al.*[45,46]

In this study, we fixed ε and σ to reproduce established Lennard-Jones phase diagrams and varied only the cross-interaction parameter ξ to generate different levels of non-ideality. This simplifying choice allows systematic benchmarking against well-characterized systems, while still capturing a range of VLE behaviors. We note, however, that further work will be needed to test generalization across other Lennard-Jones parameters and to extend the approach to fluids with more complex intermolecular potentials.

The interactions between two different components were determined using the following combination rules:

$$\sigma_{12} = \frac{1}{2}(\sigma_1 + \sigma_2), \tag{2}$$

$$\varepsilon_{12} = \xi\sqrt{\varepsilon_1\varepsilon_2}, \tag{3}$$

where the scaling factor ξ was set to values in the range $[0.85, 1.2]$ to obtain systems with different VLE behaviors.[45,46] The scaling factor ξ is a tunable parameter for unlike the interaction energy; following Stephan *et al.*, it was selected to generate mixtures spanning different VLE behaviors, including low- and high-boiling azeotropes. This setup, in which particle sizes are fixed while interaction strength is varied, provides a controlled model for mixtures with comparable excluded volumes but differing cohesive energies, such as small non-polar fluids,[47] and facilitates comparison with prior LJ coexistence studies.

All molecular dynamics (MD) simulations were performed using LAMMPS[48] under periodic boundary conditions at a reduced temperature of $T/T_{c,1} = 0.7$ using a Nose–Hoover thermostat,[49,50] where $T_{c,1} = 1.1\varepsilon/k_B$ and represents the critical temperature of component 1.[46,51] The simulation routine includes equilibration with a production run of 4×10^6 steps for each molar fraction of each system, at a time step of 0.001 τ. A Nose–Hoover thermostat was used for all simulations, and a Nose–Hoover barostat was employed where necessary.[52,53] All initial configurations were constructed using PACKMOL.[54]

As described in the supplementary material, to construct VLEs, we performed vapor–liquid coexistence simulations. These were carried out in a single elongated box containing both liquid and vapor phases separated by an interface, from which density profiles and vapor pressures were extracted. In Fig. S1 of the supplementary material, we present exemplary plots of the vapor pressure and density profiles and resulting phase diagrams. We constructed three different VLEs, as shown in Fig. S2 of the supplementary material. Activity coefficients are then extracted from VLEs for GP output and are tabulated in Table S1 of the supplementary material. For GP input, we calculate KBIs using single-phase simulations, as described in Sec. II A 1.

1. Kirkwood–Buff integrals (KBIs)

To compute running KBIs (G_{ij}^R), we first calculated RDFs $[g_{ij}(r)]$ for component pairs (1–1, 1–2, and 2–2) from the generated single-phase trajectories (Figs. S5–S7 of the supplementary material).[41–43] The RDFs were obtained by histogramming interparticle separations into 1200 uniformly spaced radial bins up to $r_{\max} = 5\sigma$. The KBI values were then obtained from the following integral, evaluated numerically as a discrete sum (rectangle rule) for the running KBIs, with final values obtained using trapezoidal integration over the RDF grid:

$$G_{ij}^{\infty} = \int_0^{\infty} [g_{ij}(r) - 1] 4\pi r^2 \, dr. \tag{4}$$

The Kirkwood–Buff integral (KBI) in Eq. (4) is defined in the thermodynamic limit, that is, for an infinite system. However, molecular simulations are inherently finite, which introduces size-dependent errors in the direct evaluation of the KBI. To address this, Krüger *et al.*[41] proposed a finite-size correction that modifies the integrand to account for the finite spherical volume of diameter (L). The corrected form is shown in the following equation and can be used to compute $G_{ij}^V(L)$ for different system sizes:[55]

$$G_{ij}^V(L) = \int_0^L 4\pi r^2 [g_{ij}(r) - 1]\left(1 - \frac{3}{2}\left(\frac{r}{L}\right) + \frac{1}{2}\left(\frac{r}{L}\right)^3\right) dr. \tag{5}$$

By plotting the computed KBI values against $1/L$, $1/V$, or $1/L^3$, and extrapolating to zero, one can recover the infinite-volume limit G_{ij}^{∞}, since $L \to \infty$ implies $1/L \to 0$ (Figs. S8–S10 of the supplementary material). This approach significantly improves convergence and accuracy in estimating KBIs from finite-size simulations. The corrected KBI values in Tables S2–S4 of the supplementary material that are obtained from this procedure were subsequently used to input the GP model for training.

2. Single-phase MD simulations

For these systems, a cubic box with dimensions $40\sigma \times 40\sigma \times 40\sigma$ was packed with 16 000 particles, resulting in a starting density of $\sim 0.25\sigma^{-3}$. This liquid-phase density is allowed to equilibrate in an NPT ensemble at a reduced pressure of 1. After equilibration, we perform the production run in the NVT ensemble for $4 \times 10^6 \tau$.

B. Gaussian process (GP)

The GP model[34] was implemented using the Python package `GPflow`[56] to predict activity coefficients (γ) from Kirkwood–Buff integrals (KBIs) extracted from single-phase MD simulations. This means that the ML task can be defined as

$$\gamma_i(x_2) = GP([G_{11}(x_2), G_{12}(x_2), G_{22}(x_2)]). \tag{6}$$

Here, the notation $GP([\cdot])$ is shorthand to indicate that the activity coefficients are modeled as the outputs of a GP, with G_{11}, G_{12}, and G_{22} at a given composition x_2 as input features. In practice, this means that the GP learns a distribution over possible functions mapping KBIs to activity coefficients, rather than a fixed parametric form, with a mean prediction and associated predictive variance.

As a non-parametric stochastic approach,[57] GPs capture uncertainty in predictions and are particularly effective for learning complex relationships between thermodynamic properties with small and scarce datasets.

We tested multiple kernel combinations and normalization strategies and found that selecting an appropriate kernel–normalization pairing was critical for improving prediction accuracy across systems with varying interaction strengths. Given the very limited dataset (five systems in total), we employed a two-way split (training/testing) without a separate validation set, as further subdivision would leave too few samples for meaningful training. Model selection was, therefore, guided by performance on held-out test systems, observing whether predictions reproduced the corresponding MD results, together with the uncertainty estimates provided by the GP model. This limitation will be addressed in future work by expanding the dataset to enable a full train/validation/test protocol. Toward this end, a composite kernel, formed by the product of a Matern12 with a polynomial and a linear kernel, was employed to model correlations in the data, with a white noise kernel added to account for observational noise.[58,59] Feature and label normalization was performed using standardization, as it yielded the best predictive performance. The model was optimized using the SLSQP algorithm.[60]

Certain systems with available phase diagrams in the literature[46] were used as training data for the GP. Meanwhile, the testing set consists of systems that exhibit different levels of non-ideality to those available in the literature. Table I lists the systems considered in this work, along with their status as training or testing data for the GP.

TABLE I. List of VLE systems studied in this work.

System ξ	System usage
1.2	Training
1.1	Testing
1.0	Training
0.9	Testing
0.85	Training

C. Activity coefficient and VLE relationship in thermodynamics

To obtain ground-truth activity coefficient values for both the training and testing sets, activity coefficients are extracted from the phase diagram using the following relation:

$$\gamma_i = \frac{x_i^v\, p}{x_i^l\, p_i^{\mathrm{sat}}}, \tag{7}$$

where γ_i is the activity coefficient of component i, x_i^v is the vapor-phase molar fraction, x_i^l is the liquid-phase molar fraction, p is the total pressure, and p_i^{sat} is the saturation vapor pressure of pure component i.

Finally, to use the GP model in inference mode—for systems where the phase diagram and, hence, partial pressures are not known—the predicted activity coefficients $\hat{\gamma}_i$ and the known saturation pressures p_i^{sat} of each pure component extracted from interface simulations are used to reconstruct the phase diagrams of new systems. The partial pressure of component i in the vapor phase satisfies

$$p_i = x_i^v\, p = x_i^l\, p_i^{\mathrm{sat}} \hat{\gamma}_i. \tag{8}$$

Consequently, the total vapor pressure is given by

$$p = \sum_i x_i^l\, p_i^{\mathrm{sat}} \hat{\gamma}_i. \tag{9}$$

This formalism allows us to go from VLEs to activity coefficients and also from activity coefficients to VLEs, given the saturated pressure of pure components.

III. RESULTS AND DISCUSSION

A. ML-generated phase diagrams

1. Data preparation

For input data to the GP model, we calculated KBIs using single-phase simulations following the procedure outlined in Secs. II A 1 and II A 2. The smoothness and consistency of the KBI profiles in Fig. 1 indicate sufficient statistical sampling in the single-phase MD simulation, ensuring reliable structural correlations. Figure 1 shows that KBIs follow a systematic and unique trend with respect to molar fractions for different interactions and systems. Moreover, the trends of G_{11}, G_{12}, and G_{22} exhibit systematic variations across different interaction strengths ξ, which quantifies the strength of cross-interactions relative to like-species interactions. For lower ξ, where 1–1 and 2–2 interactions are more favorable than 1–2, G_{11} increases with increasing x_2 as shown by the brown and green lines, whereas for higher ξ, where 1–2 interactions dominate, G_{11} decreases with x_2 as shown by the red and blue lines in Fig. 1(a). In contrast, G_{22} generally decreases for lower ξ as shown by the green and red lines, while it increases with increasing x_2 for higher ξ as shown by the red and blue lines in Fig. 1(c). The purple lines in Figs. 1(a)–1(c) represent KBIs for the intermediate case of $\xi = 1$ for which G_{11}, G_{12}, and G_{22} remain nearly constant.

G_{12} presented in Fig. 1(b) captures cross correlations between components 1 and 2, showing a maximum or minimum around $x_2 \approx 0.5$. A peak appears in G_{12} for higher ξ shown with a blue line, while a minimum is observed at lower interaction parameters shown with a brown line. As an indication, Fig. 1 suggests that G_{12} decreases with a decrease in interaction strength of ξ. The shift in G_{12} trends aligns with the phase boundaries shown in Fig. S2 of the supplementary material, emphasizing the role of cross-interactions in phase behavior. The variations in G_{11} and G_{22} at different ξ further correlate with coexistence regions, demonstrating how interaction strengths influence phase stability. As described in Sec. II, KBIs and activity coefficients are used as the input and target data for the GP model, respectively. As described in Sec. II, the activity coefficients are calculated from the constructed phase diagrams in Fig. S2 of the supplementary material using Eq. (7).

Using the procedure described in the supplementary material using coexistence simulations, VLE phase diagrams that are shown in Fig. S2 of the supplementary material were constructed, confirming that the VLEs generated in this work consistently align with the MD results in the literature[45,46] for a variety of VLE non-idealities. We were able to reproduce results for VLE systems exhibiting both low- and high-boiling azeotropes, as shown in Figs. S2(a) and S2(c) of the supplementary material, respectively, and a system with no azeotropic behavior, as shown in Fig. S2(b) of the supplementary material. As described in Sec. II C, the activity coefficients as output for GP were extracted from VLE phase diagrams. We present the KBIs and activity coefficient values used for training of GP in Tables S1–S4 of the supplementary material .

2. Fitting the GP model (predicting γ_i)

Leveraging these extracted KBIs, the GP model is trained and is shown to successfully map KBIs to activity coefficients, as shown in Fig. 2. The model was then used to predict the activity coefficients of two new systems using only their KBIs, eliminating the need for direct phase equilibrium simulations and significantly reducing the computational costs. Figure 2 shows the predicted activity coefficients γ_1 and γ_2 as a function of the molar fraction of the second component (x_2), comparing values obtained from MD simulations and GP predictions.

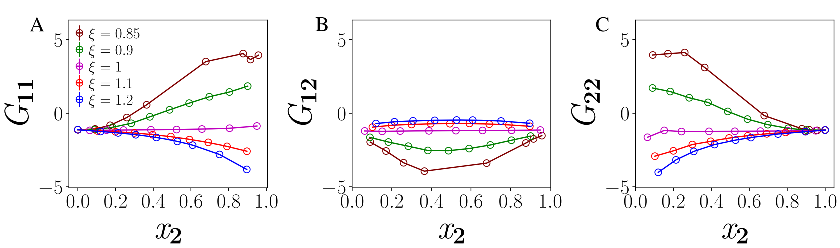

FIG. 1. KBI plots from single-phase cubic systems used to train the GP model for three different activity coefficients (a)–(c). The brown, green, purple, red, and blue lines represent results associated with ξ = 0.85, 0.9, 1.0, 1.1, and 1.2, respectively.

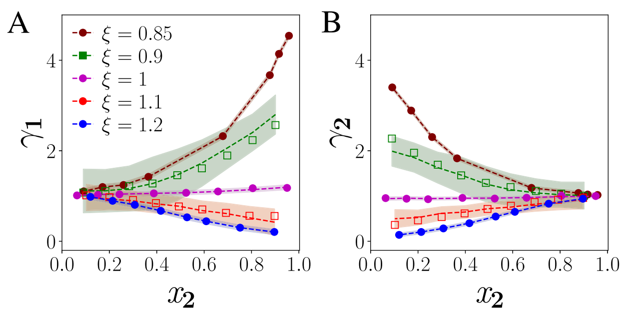

FIG. 2. Activity coefficients γ_1 (a) and γ_2 (b) as a function of molar fraction x_2, obtained from MD simulations (markers) and GP predictions (dashed lines). The solid circles represent training data, while the hollow squares denote test data. The shaded regions around the GP predictions indicate the 95% confidence intervals derived from the GP predictive variance. The close agreement between MD and predicted values, together with bounded uncertainties, demonstrates the model's ability to generalize across systems with different interaction strengths.

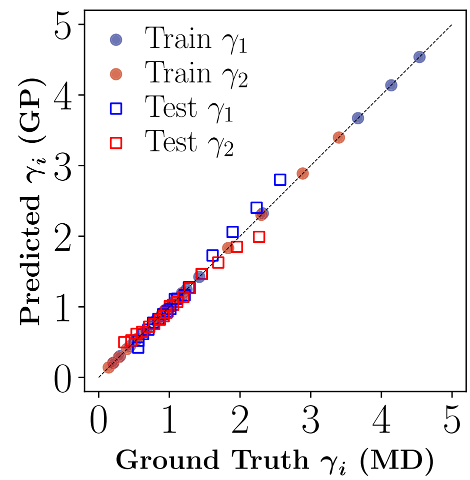

FIG. 3. Parity plot comparing predicted and MD-derived activity coefficients for both training and test systems. The solid circles represent training data (γ_1 in blue and γ_2 in red); the hollow squares indicate GP predictions on the test systems. The dashed line is a parity line for perfect agreement.

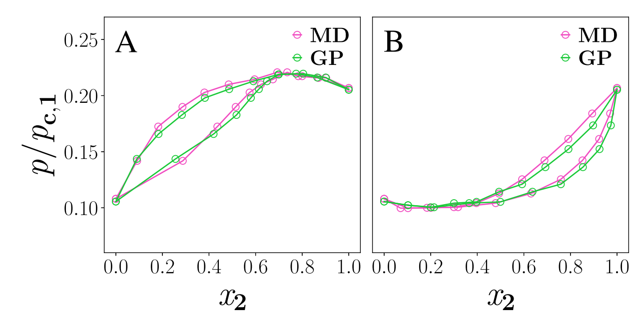

FIG. 4. Phase diagrams for test data constructed from GP-predicted activity coefficients (green) and compared with our MD results (pink). (a) $\xi = 0.9$ and (b) $\xi = 1.1$. Mean GP predictions are shown for clarity; the full propagated uncertainty is provided in the supplementary material (Fig. S15).

As described in Sec. II, we selected a multiplication of Matern12, polynomial, and linear kernels. Furthermore, labels and features were standardized using z-score normalization. It was observed that a single GP model trained on G_{11}, G_{12}, and G_{22} to predict both γ_1 and γ_2 gave accurate and consistent results across all systems tested. While γ_1 and γ_2 are predicted independently within the model, they share the same input features. Using Gaussian processes together with thermodynamic consistency checks allows us to capture their interdependence and improves generalization compared to training two separate models.[10,61,62]

Table S7 of supplementary material summarizes the GP-predicted and MD-generated activity coefficients for test data. The two subplots in Fig. 2 illustrate strong agreement between the predicted γ_1 and γ_2 values and MD data, particularly for the new systems where the GP model had no prior knowledge of their activity coefficients.

Importantly, the model accurately captures both positive and negative deviations from ideality across a wide range of interaction strengths, as evidenced by the close alignment of predicted values with molecular dynamics (MD) data. To further assess model performance, a parity plot is shown in Fig. 3, where the predicted activity coefficients are plotted against their MD counterparts. Training data for γ_1 and γ_2 are shown as solid blue and red circles, respectively, while test data are displayed as hollow squares. The proximity of all points to the parity line confirms the model's robustness and generalization across systems with varying ξ values.

3. Model inference (constructing predicted phase diagrams)

To show whether this level of agreement in γ_i translates to accurate prediction of phase diagrams, the predicted activity coefficients together with the known saturation pressures p_i^{sat} of each pure component were used to reconstruct the phase diagrams of the two new systems using Eq. (9). Figure 4 displays the predicted phase diagrams for the test data, constructed using GP-predicted activity coefficients. While the GP provides a predictive distribution for the activity coefficients, these are visualized as $\pm 1.96\sigma$ bands in Fig. 2, propagating this uncertainty through Eqs. (8) and (9) to obtain pressure bounds results in very wide confidence intervals, as shown in Figure S15, due to the high sensitivity of the vapor pressure to γ. For clarity, we, therefore, present only the mean GP-predicted phase diagrams in Fig. 4, while the propagated uncertainty is shown in the supplementary material (Fig. S15) at the 1σ level for visibility.

A direct comparison with MD-generated phase diagrams demonstrates reasonable agreement, validating the predictive capability of the GP model across systems with different interaction strengths. To assess the thermodynamic consistency of the predicted activity coefficients, we examined their compliance with the Gibbs–Duhem relation and computed the excess Gibbs free energy, as shown in Figs. S13 and S14 of the supplementary material . These analyses provide local and global validation of thermodynamic consistency, respectively. Furthermore, the ability to reconstruct phase behavior solely from structural data represents a significant advancement in reducing the computational cost associated with phase equilibrium calculations.

The overall agreement across all figures demonstrates the effectiveness of integrating MD simulations with machine learning for phase behavior predictions. The strong correlation between KBIs and thermodynamic properties enables a substantial reduction in computational cost while maintaining high accuracy using GP. By successfully predicting phase diagrams for test systems without prior knowledge of their phase behavior, this approach establishes a promising framework for efficiently modeling complex fluid mixtures.

The training set considered here is limited to nearly symmetric Lennard-Jones mixtures with equal particle sizes, similar well depths, and variation only in ξ, chosen to benchmark the framework against reference phase diagrams. Within this restricted setting, KBIs are formally derived from pair correlations, but they implicitly encode the influence of all interactions present in the simulation, since $g_{ij}(r)$ reflects the collective outcome of many-body forces. Nevertheless, KBIs do not capture higher-order correlations explicitly, and this limitation may become important in systems with strong asymmetry or near critical points owing to collective fluctuations. For the Lennard-Jones mixtures considered here, the three KBIs (G_{11}, G_{12}, and G_{22}) provided sufficient information to reproduce activity coefficients, but more complex systems may require additional structural descriptors beyond KBIs. This study focuses exclusively on binary mixtures; extension to multicomponent systems, while conceptually straightforward within the KBI-GP framework, will require addressing the increased number of KBIs and the need for larger training datasets.

Although the GP model predicts γ_1 and γ_2 independently, we verified that the results broadly satisfy thermodynamic consistency through checks of the Gibbs–Duhem relation and the excess Gibbs free energy (Figs. S13 and S14 of the supplementary material). Embedding these constraints directly into the learning model, for example, using physics-informed Gaussian processes or related constrained formulations,[63,64] would provide a systematic way to ensure consistency but may also reduce predictive flexibility in small-data regimes. Physics-informed neural networks (PINNs) and related differentiable models have also been proposed for activity coefficient prediction,[65–68] but these typically require large datasets, whereas Gaussian processes are more effective in the small-data regime considered here. Finally, investigating the fidelity of KBIs against analytic or semi-analytic benchmarks (e.g., Widom–Rowlinson,[69] Percus–Yevick,[70] or lattice regular solution models) would provide complementary validation, which we identify as an interesting direction for future work.

IV. CONCLUSIONS

This study introduces a novel methodology for efficiently predicting phase behavior in Lennard-Jones mixtures by integrating molecular dynamics (MD) simulations with Gaussian process (GP) regression. This approach enables a cost-effective strategy for determining activity coefficients and phase diagrams using only single-phase simulations, significantly reducing the computational expense associated with extensive phase equilibrium calculations. By mapping Kirkwood–Buff integrals (KBIs) to activity coefficients via GP, we developed a predictive model capable of accurately estimating activity coefficients for new datasets based on KBIs extracted from single-phase simulations. This framework is broadly useful in computational thermodynamics—for quickly estimating phase diagrams in model systems, exploring mixture behavior across different interaction strengths, and supporting the development of force fields or coarse-grained models that capture non-ideal mixing. Its effectiveness relies on using sufficiently large systems for accurate sampling of KBIs and selecting training data that represent the range of conditions relevant to the systems of interest. While demonstrated here for Lennard-Jones mixtures, where deviations from ideality arise from generic size and energetic disparities rather than chemically specific interactions, the GP–KBI framework can, in principle, be extended to chemically realistic systems, since KBIs also encode more specific contributions, such as hydrogen bonding and electrostatics. Extending the methodology to such systems is an important direction for future work.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional plots: reference phase diagrams extracted from coexistence simulations, RDF $[g_{ij}(r)]$, KBI integrand $[G_{ij}(r)]$, KBI running integrals $[G_{ij}^{V}(r)]$, and a table of activity coefficients of test systems.

ACKNOWLEDGMENTS

The authors thank Esteban Cea-Klapp for insightful discussions on Kirkwood–Buff integral (KBI) calculations and phase diagram extraction. The authors acknowledge the Center for Research Computing (CRC) at the University of Notre Dame for providing computational resources. This research was partially supported by the National Science Foundation under Award No. ERC-2330175 for the Engineering Research Center EARTH and partially supported by Breakthrough Electrolytes for Energy Storage (BEES), an Energy Frontier Research Center funded by the DOE Office of Science, Basic Energy Sciences (BES), under Award No. DE-SC0019409. This work was also developed within the scope of the project CICECO-Aveiro Institute of Materials, UID/50011 and LA/P/0006/2020 (DOI 10.54499/LA/P/0006/2020), financed by national funds through the FCT/MCTES (PIDDAC).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Mohsen Farshad: Data curation (equal); Formal analysis (lead); Investigation (lead); Methodology (lead); Writing – original draft (equal); Writing – review & editing (equal). **Fathya Y. M. Salih**: Formal analysis (supporting); Investigation (equal); Methodology (supporting); Writing – original draft (supporting); Writing – review & editing (equal). **Dinis O. Abranches**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (supporting); Investigation (equal); Methodology (equal); Project administration (equal); Writing – original draft (equal); Writing – review & editing (equal). **Yamil J. Colón**: Conceptualization (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The GitHub repository of this project contains example MD input files, activity coefficients and KBI data in Excel format, and Python scripts—including the GP model—to reproduce all figures in the main text and the supplementary material: https://github.com/mfarshad1/KBI-GAMMA-GP.

REFERENCES

[1] P. Y. Chew and A. Reinhardt, "Phase diagrams—Why they matter and how to predict them," J. Chem. Phys. **158**, 030902 (2023).

[2] A. Kaul, "The phase diagram," in *Aqueous Two-Phase Systems* (Humana Press, NJ, 2000), Vol. 11, pp. 11–21.

[3] F. Zipoli, V. Viterbo, O. Schilter, L. Kahle, and T. Laino, "Prediction of phase diagrams and associated phase structural properties," Ind. Eng. Chem. Res. **61**, 8378–8389 (2022).

[4] A. Fredenslund, J. Gmehling, M. L. Michelsen, P. Rasmussen, and J. M. Prausnitz, "Computerized design of multicomponent distillation columns using the UNIFAC group contribution method for calculation of activity coefficients," Ind. Eng. Chem. Process Des. Dev. **16**, 450–462 (1977).

[5] A. Öncül, K. Sundmacher, and D. Thévenin, "Numerical investigation of the influence of the activity coefficient on barium sulphate crystallization," Chem. Eng. Sci. **60**, 5395–5405 (2005).

[6] M. Krannich, F. Heym, and A. Jess, "Characterization of six hygroscopic ionic liquids with regard to their suitability for gas dehydration: Density, viscosity, thermal and oxidative stability, vapor pressure, diffusion coefficient, and activity coefficient of water," J. Chem. Eng. Data **61**, 1162–1176 (2016).

[7] A. Diedrichs and J. Gmehling, "Solubility calculation of active pharmaceutical ingredients in alkanes, alcohols, water and their mixtures using various activity coefficient models," Ind. Eng. Chem. Res. **50**, 1757–1769 (2011).

[8] H. Suleman, A. S. Maulud, S. N. Shah, Z. Man, and M. I. A. Mutalib, "Thermodynamic modelling of liquid-liquid extraction systems involving ionic liquids: A new approach," J. Mol. Liq. **252**, 18–23 (2018).

[9] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, New York, 1985).

[10] J. M. Prausnitz, R. Lichtenthaler, and E. G. d. Azevedo, *Molecular Thermodynamics of Fluid-Phase Equilibria*, 2nd ed., *Prentice-Hall International Series in the Physical and Chemical Engineering Sciences* (Prentice-Hall, Englewood Cliffs, NJ, 1986).

[11] P. M. Mathias, "Guidelines for the analysis of vapor–liquid equilibrium data," J. Chem. Eng. Data **62**, 2231–2233 (2017).

[12] J. Chen, S. R. Cross, L. J. Miara, J.-J. Cho, Y. Wang, and W. Sun, "Navigating phase diagram complexity to guide robotic inorganic materials synthesis," Nat. Synth. **3**, 606–614 (2024).

[13] M. S. Alam, M. A. Chowdhury, M. A. Kowser, M. S. Islam, M. M. Islam, and T. Khandaker, "Advances of MAX phases: Synthesis, characterizations and challenges," Eng. Rep. **6**, e12911 (2024).

[14] C. G. Salzmann, "Advances in the experimental exploration of water's phase diagram," J. Chem. Phys. **150**, 060901 (2019).

[15] Y. A. Chang, S. Chen, F. Zhang, X. Yan, F. Xie, R. Schmid-Fetzer, and W. A. Oates, "Phase diagram calculation: Past, present and future," Prog. Mater. Sci. **49**, 313–345 (2004).

[16] B. Sundman and J. Ågren, "A regular solution model for phases with several components and sublattices, suitable for computer applications," J. Phys. Chem. Solids **42**, 297–301 (1981).

[17] Y. Zhao, P. Inbar, H. P. Chokshi, A. W. Malick, and D. S. Choi, "Prediction of the thermal phase diagram of amorphous solid dispersions by Flory–Huggins theory," J. Pharm. Sci. **100**, 3196–3207 (2011).

[18] H. Renon and J. M. Prausnitz, "Local compositions in thermodynamic excess functions for liquid mixtures," AIChE J. **14**, 135–144 (1968).

[19] Ø. Wilhelmsen, A. Aasen, G. Skaugen, P. Aursand, A. Austegard, E. Aursand, M. A. Gjennestad, H. Lund, G. Linga, and M. Hammer, "Thermodynamic modeling with equations of state: Present challenges with established methods," Ind. Eng. Chem. Res. **56**, 3503–3515 (2017).

[20] U. R. Kattner, "The thermodynamic modeling of multicomponent phase equilibria," JOM **49**, 14–19 (1997).

[21] H. Okamoto and T. B. Massalski, "Guidelines for binary phase diagram assessment," J. Phase Equilib. **14**, 316–335 (1993).

[22] C. Vega, E. Sanz, J. L. F. Abascal, and E. G. Noya, "Determination of phase diagrams via computer simulation: Methodology and applications to water, electrolytes and proteins," J. Phys.: Condens. Matter **20**, 153101 (2008).

[23] J. R. Espinosa, J. M. Young, H. Jiang, D. Gupta, C. Vega, E. Sanz, P. G. Debenedetti, and A. Z. Panagiotopoulos, "On the calculation of solubilities via direct coexistence simulations: Investigation of NaCl aqueous solutions and Lennard-Jones binary mixtures," J. Chem. Phys. **145**, 154111 (2016).

[24] K. Binder, J. L. Lebowitz, M. K. Phani, and M. H. Kalos, "Monte Carlo study of the phase diagrams of binary alloys with face centered cubic lattice structure," Acta Metall. **29**, 1655–1665 (1981).

[25] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, 4th ed. (Cambridge University Press, 2014).

[26] D. Frenkel, B. Smit, J. Tobochnik, S. R. McKay, and W. Christian, "Understanding molecular simulation," Comput. Phys. **11**, 351–354 (1997).

[27] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, 2nd ed. (Oxford University Press, Oxford, 2017).

[28] A. Z. Panagiotopoulos, "Monte Carlo methods for phase equilibria of fluids," J. Phys.: Condens. Matter **12**, R25–R52 (2000).

[29] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).

[30] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys. **145**, 170901 (2016).

[31] C. Dai and S. C. Glotzer, "Efficient phase diagram sampling by active learning," J. Phys. Chem. B **124**, 1275–1284 (2020).

[32] G. Deffrennes, K. Terayama, T. Abe, and R. Tamura, "A machine learning–based classification approach for phase diagram prediction," Mater. Des. **215**, 110497 (2022).

[33] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," npj Comput. Mater. **5**, 21 (2019).

[34] C. E. Rasmussen, *Gaussian Processes for Machine Learning*, *Adaptive Computation and Machine Learning* (MIT Press, Cambridge, Massachusetts, 2006).

[35] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," Chem. Rev. **121**, 10073–10141 (2021).

[36] V. Ladygin, I. Beniya, E. Makarov, and A. Shapeev, "Bayesian learning of thermodynamic integration and numerical convergence for accurate phase diagrams," Phys. Rev. B **104**, 104102 (2021).

[37] D. O. Abranches, E. J. Maginn, and Y. J. Colón, "Activity coefficient acquisition with thermodynamics-informed active learning for phase diagram construction," AIChE J. **69**, e18141 (2023).

[38] M. DelloStritto and M. L. Klein, "Understanding strain and failure of a knot in polyethylene using molecular dynamics with machine-learned potentials," J. Phys. Chem. Lett. **15**, 9070–9077 (2024).

[39] J. G. Kirkwood and F. P. Buff, "The statistical mechanical theory of solutions. I," J. Chem. Phys. **19**, 774–777 (1951).

[40] A. Ben-Naim, "Inversion of the Kirkwood–Buff theory of solutions: Application to the water–ethanol system," J. Chem. Phys. **67**, 4884–4890 (1977).

[41] P. Krüger, S. K. Schnell, D. Bedeaux, S. Kjelstrup, T. J. H. Vlugt, and J.-M. Simon, "Kirkwood–Buff integrals for finite volumes," J. Phys. Chem. Lett. **4**, 235–238 (2013).

[42] N. Dawass, P. Krüger, S. K. Schnell, J.-M. Simon, and T. J. H. Vlugt, "Kirkwood-Buff integrals from molecular simulation," Fluid Phase Equilib. **486**, 21–36 (2019).

[43] J.-M. Simon, P. Krüger, S. K. Schnell, T. J. H. Vlugt, S. Kjelstrup, and D. Bedeaux, "Kirkwood–Buff integrals: From fluctuations in finite volumes to the thermodynamic limit," J. Chem. Phys. **157**, 130901 (2022).

[44] K. E. Newman, "Kirkwood–Buff solution theory: Derivation and applications," Chem. Soc. Rev. **23**, 31–40 (1994).

[45]S. Stephan, K. Langenbach, and H. Hasse, "Interfacial properties of binary Lennard-Jones mixtures by molecular simulation and density gradient theory," J. Chem. Phys. **150**, 174704 (2019).

[46]S. Stephan, J. Liu, K. Langenbach, W. G. Chapman, and H. Hasse, "Vapor–liquid interface of the Lennard-Jones truncated and shifted fluid: Comparison of molecular simulation, density gradient theory, and density functional theory," J. Phys. Chem. C **122**, 24705–24715 (2018).

[47]E. A. Ploetz, N. Bentenitis, and P. E. Smith, "Kirkwood–Buff integrals for ideal solutions," J. Chem. Phys. **132**, 164501 (2010).

[48]S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys. **117**, 1–19 (1995).

[49]S. Nose, "Constant-temperature molecular dynamics," J. Phys.: Condens. Matter **2**, SA115–SA119 (1990).

[50]S. Nosé, "A molecular dynamics method for simulations in the canonical ensemble," Mol. Phys. **52**, 255–268 (1984).

[51]M. Thol, G. Rutkai, R. Span, J. Vrabec, and R. Lustig, "Equation of state for the Lennard-Jones truncated and shifted model fluid," Int. J. Thermophys. **36**, 25–43 (2015).

[52]M. E. Tuckerman, J. Alejandre, R. López-Rendón, A. L. Jochim, and G. J. Martyna, "A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal–isobaric ensemble," J. Phys. A: Math. Gen. **39**, 5629–5651 (2006).

[53]G. J. Martyna, D. J. Tobias, and M. L. Klein, "Constant pressure molecular dynamics algorithms," J. Chem. Phys. **101**, 4177–4189 (1994).

[54]L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, "PACKMOL: A package for building initial configurations for molecular dynamics simulations," J. Comput. Chem. **30**, 2157–2164 (2009).

[55]E. Cea-Klapp, D. O. Abranches, E. Marin-Rimoldi, N. F. Gajardo-Parra, R. I. Canales, J. M. Garrido, and E. J. Maginn, "Understanding solute-hydrotrope aggregation in aqueous solutions: A molecular dynamics approach," J. Phys. Chem. B **129**, 5016 (2025).

[56]A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman, "Gpflow: A Gaussian process library using tensorflow," J. Mach. Learn. Res. **18**, 1–6 (2017).

[57]C. K. I. Williams and D. Barber, "Bayesian classification with Gaussian processes," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1342–1351 (1998).

[58]M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," J. Mach. Learn. Res. **2**, 299–312 (2002).

[59]C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning* (Springer, 2003), pp. 63–71.

[60]M. van der Wilk, V. Dutordoir, S. T. John, A. Artemev, V. Adam, and J. Hensman, "A framework for interdomain and multioutput Gaussian processes," arXiv:2003.01115 (2020).

[61]M. A. Álvarez, L. Rosasco, N. D. Lawrence *et al.*, "Kernels for vector-valued functions: A review," Found. Trends Mach. Learn. **4**, 195–266 (2012).

[62]E. V. Bonilla, K. Chai, and C. Williams, "Multi-task Gaussian process prediction," Adv. Neural Inf. Process. Syst. **20**, 153–160 (2007).

[63]B. Winter, C. Winter, T. Esper, J. Schilling, and A. Bardow, "SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients," Fluid Phase Equilib. **568**, 113731 (2023).

[64]H. W. Sullivan, B. L. Shanks, M. Cervenka, and M. P. Hoepfner, "Physics-informed Gaussian process inference of liquid structure from scattering data," arXiv:2507.07948 (2025).

[65]J. G. Rittig, K. C. Felton, A. A. Lapkin, and A. Mitsos, "Gibbs–Duhem-informed neural networks for binary activity coefficient prediction," Digit. Discov. **2**, 1752–1767 (2023).

[66]E. I. Sanchez Medina, S. Kunchapu, and K. Sundmacher, "Gibbs–Helmholtz graph neural network for the prediction of activity coefficients of polymer solutions at infinite dilution," J. Phys. Chem. A **127**, 9863–9873 (2023).

[67]E. I. Sanchez Medina, S. Linke, M. Stoll, and K. Sundmacher, "Graph neural networks for the prediction of infinite dilution activity coefficients," Digit. Discov. **1**, 216–225 (2022).

[68]T. Specht, M. Nagda, S. Fellenz, S. Mandt, H. Hasse, and F. Jirasek, "Hanna: Hard-constraint neural network for consistent activity coefficient prediction," Chem. Sci. **15**, 19777–19786 (2024).

[69]B. Widom and J. S. Rowlinson, "New model for the study of liquid–vapor phase transitions," J. Chem. Phys. **52**, 1670–1684 (1970).

[70]J. L. Lebowitz, "Exact solution of generalized Percus-Yevick equation for a mixture of hard spheres," Phys. Rev. **133**, A895 (1964).