ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Accelerating Multicomponent Phase-Coexistence Calculations with Physics-informed Neural Networks†

Satyen Dhamankar,†* Shengli Jiang,†* and Michael A. Webb*

† Equally contributing authors.

* Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544.

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Phase separation in multicomponent mixtures is of significant interest in both fundamental research and technology. Although the thermodynamic principles governing phase equilibria are straightforward, practical determination of equilibrium phases and constituent compositions for multicomponent systems is often laborious and computationally intensive. Here, we present a machine-learning workflow that simplifies and accelerates phase-coexistence calculations. We specifically analyze capabilities of neural networks to predict the number, composition, and relative abundance of equilibrium phases of systems described by Flory-Huggins theory. We find that incorporating physics-informed material constraints into the neural network architecture enhances the prediction of equilibrium compositions compared to standard neural networks with minor errors along the boundaries of the stable region. However, introducing additional physics-informed losses does not lead to significant further improvement. These errors can be virtually eliminated by using machine-learning predictions as a warm-start for a subsequent optimization routine. This work provides a promising pathway to efficiently characterize multicomponent phase coexistence.

Design, System, Application

Accurate phase coexistence characterization is critical for designing and optimizing systems and processes involving multiple components, yet traditional methods are often slow and computationally expensive. To overcome this, we developed a machine learning workflow grounded in physical principles to streamline and speed up these calculations. Using Flory-Huggins theory, we generated ternary phase diagrams and trained a theory-aware machine learning algorithm to predict equilibrium phases, compositions, and abundances. These predictions serve as an initial guess for numerical optimization, enabling fast and accurate determination of equilibrium states. This approach can be extended beyond ternary systems or applied to other free-energy models to describe a variety of chemical and biological processes. Ultimately, this method offers a promising way to accelerate chemical process simulations and drive innovations in multi-phase separations, as well as other system design workflows.

1 Introduction

Phase coexistence in multicomponent systems is ubiquitous in nature and technology. Examples range from diverse purification processes in the chemical industry[1–3] to the formation of membraneless organelles via liquid-liquid phase separation in biology.[4–11] Thorough characterization of multicomponent phase equilibria involves not only identifying the phases present but also determining their composition and abundance, as the distribution and composition of species across phases significantly affect system properties and functions. This information guides processing methods and underlies calculations in process-simulation software. Critically, these calculations can constitute a substantial fraction of the overall computational time dedicated to simulation.[12] Current multiphase flash calculation schemes require knowledge of the number of equilibrium phases,[12,13] are sensitive to initial guesses,[14] and can converge to spurious or trivial solutions if root-finding is not appropriately bounded and constrained.[15,16] Therefore, efficient and accurate methods for predicting equilibrium states are valuable for both industrial applications and fundamental research.

At equilibrium, species distribute across phases based on the extremization of an appropriate thermodynamic potential. For example, minimization of the Gibbs energy dictates equilibrium for a system at specified temperature T, pressure p, and global composition x_i. Equilibrium phase-coexistence arises when species partition into distinct phases with equal chemical potentials driven by the extremization, rather than forming a homogeneous mixture. For a system at fixed T and p, this yields

$$\mu_i^{\alpha}\left(T,p,\{x_j\}^{\alpha}\right)=\mu_i^{\beta}\left(T,p,\{x_j\}^{\beta}\right)\ \forall\ i \tag{1}$$

where μ_i^π is the chemical potential of species $i \in \{A,B,C,\dots\}$ in phase $\pi \in \{\alpha,\beta\}$ with composition $\{x_j\}^\pi = \{x_A^\pi, x_B^\pi, x_C^\pi, \dots\}$. Eq. (1) constrains the equilibrium state of the system, as manifest in Gibbs' phase rule (i.e. $\mathscr{F} = N - \mathscr{P} + 2$ where $\mathscr{F}$ is the number of independent intensive relationships needed to specify a system of N species and $\mathscr{P}$ phases). Provided a thermodynamic model for describing the mixture behavior as a function of intensive variables, Eq. (1), or its equivalent at other conditions, functionally comprises $N(\mathscr{P}-1)$ equations to solve for the compositions of the various phases, with others fixed. The complexity of identifying equilibrium states can vary, even while the underlying thermodynamic framework is straightforward.

Determining the conditions, expected phases, and the chemical nature of species usually depends on appropriately parameterized equations-of-state or available free-energy models. For condensed phases and binary mixtures, there are several simple free-energy models like the Margules equations[17], the van Laar model[18], or the Guggenheim-Scatchard/Redlich-Kister equation.[19] More complex models such as the Wilson models, non-random two-liquid (NRTL) models,[20] universal quasi-chemical theory (UNIQUAC),[21] UNIQUAC Functional-group Activity Coefficients (UNIFAC) models,[13,22,23] Flory-Huggins theory[24] can treat multicomponent systems. Although increasing complexity of the free-energy model or equation-of-state can facilitate more accurate representation of physical systems, the underlying calculations and theoretical principles for phase behavior remain the same for simple and complex models alike.

Given a thermodynamic model, calculating phase stability and equilibrium compositions can be approached in various ways. Simple models and binary mixtures may yield algebraic relationships that can be handled analytically or resolved using simple numerical schemes, such as self-consistent iteration or Newton's method. However, characterizing multicomponent phase coexistence typically requires dedicated software and more sophisticated numerical algorithms. Many algorithms are designed to work for only a specific set or number of phases.[25] Direct solution methods based on Newton's root-finding algorithm can be effective but are computationally intensive and sensitive to the initial seed. Jindrova et al. refined Newton's algorithm and a successive substitution strategy to locate roots. Additionally, Nichita,[26–28] Jindrova,[14,16] and Castier[29] independently performed volume stability analysis to obtain better initial guesses for the substitution strategy. There has been significant development in generating phase diagrams using constrained backmapping search algorithms.[15,30–33]

Indirect solution methods, based on thermodynamic principles and geometric criteria established via stability analysis, offer alternative approaches. Examples include Korteweg's tangent construction[34] and Binous and Bellagi's arc extension method.[35] Michelsen's multi-phase flash algorithm[36] minimizes the distance between the tangent plane and the free energy surface to identify coexisting phases. Homotopy methods have also been used to calculate critical and saturation properties of mixtures.[37,38] Additionally, Mao et al.[39] generalized phase-diagram construction to multicomponent systems using a convex-hull construction[40] applied to a discretized free-energy manifold, although accuracy and memory requirements depend on the mesh size. Overall, there is a need for simple, generalizable, and efficient methods for phase-coexistence calculations.

Machine learning (ML) techniques facilitate phase-coexistence calculations, offering prospective advantages relating to time- and memory-efficiency relative to more traditional optimization strategies.[41–50] However, many efforts only address the issue of phase stability and neglect consideration of phase composition.[42–46] Others have been restricted to binary systems with limited demonstration of more complex mixtures.[47–50] Recently, Flory-Huggins (FH) theory has been combined with ML to improve the interpretability and accuracy of mixture behavior predictions, but limitations exist in their ability to handle complex interactions and multicomponent systems beyond binary mixtures.[49,50] Nevertheless, such works highlight the potential of ML as part of a generalizable, accurate, efficient, and extensible framework for characterizing multicomponent phase behavior.

Here, we describe a data-driven workflow to characterize the phase behavior of multicomponent systems. Figure 1 illustrates the overall approach in the context of ternary systems described by Flory-Huggins (FH) theory. Using FH theory as a representative free-energy model, we construct a series of phase diagrams across the model parameter space using labor-intensive methods. This data is then used to develop an ML surrogate model, based on neural network architectures, to predict the number, composition, and relative abundance of equilibrium phases from model parameters and total system composition. Surrogate models optimized with and without physics-informed architectures and loss functions are compared. Errors are assessed for classification (number of equilibrium phases) and regression (composition and abundance of phases). Predictions from the surrogate model, which are computationally efficient and improvable, are then used to warm-start a simple optimization to precisely and accurately characterize the system's phase behavior. This procedure exemplifies an efficient, accurate, and extensible approach to phase-coexistence calculations.

2 Methods

2.1 Thermodynamic framework

For demonstration, we consider the thermodynamics of ternary systems described by FH solution theory. Systems are comprised of species A, B, and C that occupy a lattice of n sites with volume $V = nv_0$. The species can possess size disparities, reflected in their molar volumes v_i. For a polymer comprised of N_i monomers that each occupy a single lattice site, $v_i = N_i v_0$. Systems are incompressible such that $V = \sum_i n_i v_i$ where n_i is the mole number for species i. System composition is specified by the volume fractions $\phi_i = (n_i v_i)/V$ with $\sum_{i\in\{A,B,C\}} \phi_i = 1$.

The dimensionless, intensive (per lattice site) Helmholtz energy of mixing follows as

$$\tilde{f} \equiv \beta \Delta f_{\text{mix}} = \sum_i \frac{\phi_i}{(v_i/v_0)} \ln \phi_i + \frac{1}{2} \sum_j \sum_k \phi_j \phi_k \chi_{jk} \tag{2}$$

where $\beta = (k_B T)^{-1}$ is the inverse temperature with k_B as Boltzmann's constant, and χ_{ij} is the Flory-Huggins interaction param-

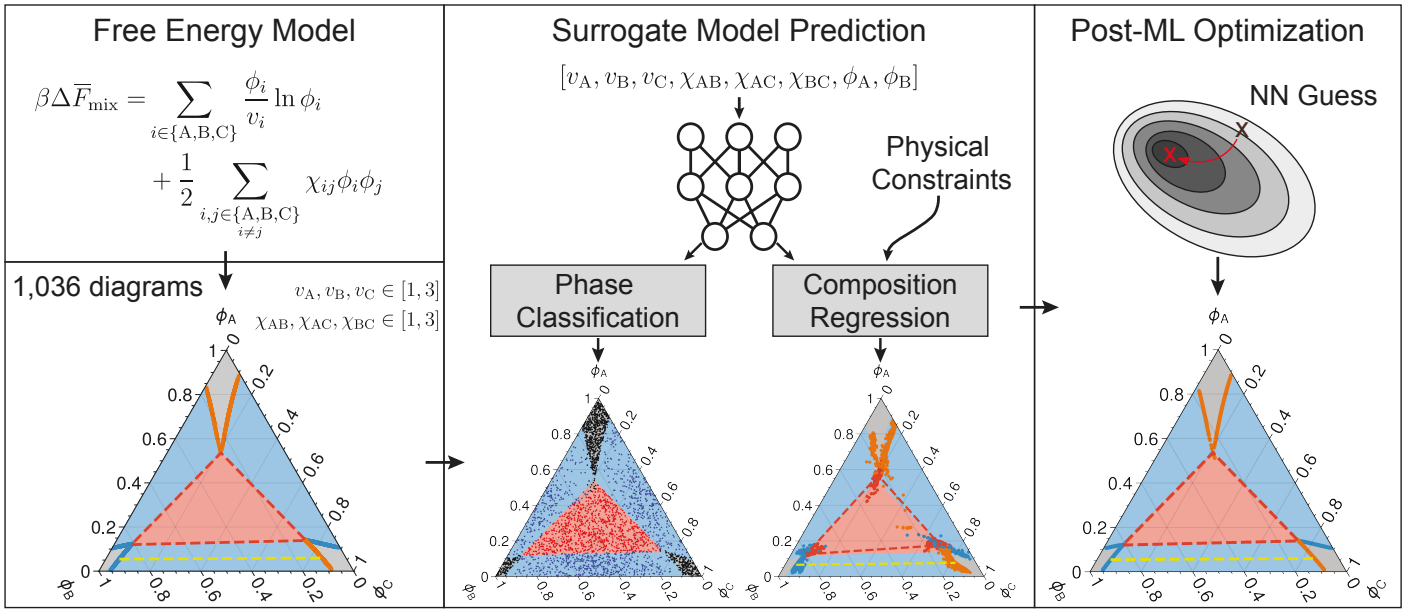

Fig. 1 **Strategy for multi-component phase-coexistence prediction using machine learning.** 1,036 ternary phase diagrams are generated using the algorithm arc continuation algorithm (Equations 6-8) and convex hull construction algorithm[39], and are used as training data for a physics-informed machine learning (ML) model to classify phase regions and predict equilibrium phase compositions. The ML predictions serve as initial guesses for the Newton-CG method to obtain equilibrium composition predictions.

eter for species i and j with $\chi_{ii} = 0$; the summations are over all components (A, B, C). Altogether, the behavior of a system is determined by the composition $\phi = (\phi_A, \phi_B, \phi_C)$, the molar volumes of the species $\mathbf{v} = (v_A, v_B, v_C)$, and the interaction parameters $\mathbf{X} = (\chi_{AB}, \chi_{AC}, \chi_{BC})$.

Up to a constant, chemical potentials are obtained by partial differentiation of the Helmholtz energy of mixing:

$$\beta\mu_i(T,V,\phi) = \frac{1}{v_0}\left(\frac{\partial\left[V\tilde{f}\right]}{\partial n_i}\right)_{T,V,n_{j\neq i}}. \tag{3}$$

Using Eq. (2) in Eq. (3), this yields

$$\beta\mu_i(T,V,\phi) = \ln(\phi_i) + \sum_{j\neq i}\phi_j\left(1-\frac{v_i}{v_j}\right) + v_i\left[\sum_{j\neq i}\sum_{k\neq i}\phi_j\phi_k\left(\chi_{ij}-\frac{1}{2}\chi_{jk}\right)\right] \tag{4}$$

where the summations exclude the species for which the chemical potential is being assessed (e.g., for μ_A, the summation for $j \neq i$ is equivalent to that for $j \in \{B,C\}$).

The thermodynamic stability of a mixture is assessed by considering the determinant of the Hessian matrix for the Helmholtz energy. For

$$\mathbf{H}_{\tilde{f}} = \begin{bmatrix} \frac{\partial^2\tilde{f}}{\partial\phi_A^2} & \frac{\partial^2\tilde{f}}{\partial\phi_A\partial\phi_B} \\ \frac{\partial^2\tilde{f}}{\partial\phi_B\partial\phi_A} & \frac{\partial^2\tilde{f}}{\partial\phi_B^2} \end{bmatrix}, \tag{5}$$

the spinodal boundary of a ternary mixture is the locus of all compositions that solve

$$|\mathbf{H}_{\tilde{f}}| = \frac{\partial^2\tilde{f}}{\partial\phi_A^2}\cdot\frac{\partial^2\tilde{f}}{\partial\phi_B^2} - \left(\frac{\partial^2\tilde{f}}{\partial\phi_A\partial\phi_B}\right)^2 = 0. \tag{6}$$

Critical points are identified by additionally considering constraints on third-order derivatives[51,52] given by

$$E_1 \equiv \begin{vmatrix} \frac{\partial|\mathbf{H}_{\tilde{f}}|}{\partial\phi_A} & \frac{\partial|\mathbf{H}_{\tilde{f}}|}{\partial\phi_B} \\ \frac{\partial^2\tilde{f}}{\partial\phi_A\partial\phi_B} & \frac{\partial^2\tilde{f}}{\partial\phi_B^2} \end{vmatrix} = 0 \tag{7}$$

and

$$E_2 \equiv \begin{vmatrix} \frac{\partial|\mathbf{H}_{\tilde{f}}|}{\partial\phi_B} & \frac{\partial|\mathbf{H}_{\tilde{f}}|}{\partial\phi_A} \\ \frac{\partial^2\tilde{f}}{\partial\phi_A\partial\phi_B} & \frac{\partial^2\tilde{f}}{\partial\phi_A^2} \end{vmatrix} = 0. \tag{8}$$

For fixed total particle density ($\rho = n/V$) and constant temperature, Gibbs' phase rules indicate there can be at most three coexisting phases. To characterize three-phase coexistence, there are 12 variables. Nine correspond to the volume fractions in each phase: ϕ^α, ϕ^β, and ϕ^γ, for which each $\phi^\pi = (\phi_A^\pi, \phi_B^\pi, \phi_C^\pi)$. Three correspond to the fractional abundances of each phase–w^α, w^β, and w^γ). Criteria for chemical equilibrium applied to each species across each phase

$$\mu_i^\alpha(T,\rho,\phi^\alpha) = \mu_i^\beta(T,\rho,\phi^\beta) = \mu_i^\gamma(T,\rho,\phi^\gamma) \tag{9}$$

provide six independent equations. For a system with a specified total composition, material balance constraints provide the remaining equations:

$$\sum_i \phi_i^\pi = 1 \text{ for } \pi \in \{\alpha,\beta\,\gamma\} \tag{10}$$

$$\sum_\pi w^\pi\phi_i^\pi = \phi_i \text{ for } i \in \{A,B,C\}. \tag{11}$$

To characterize two-phase coexistence, the variable count is reduced to eight, with commensurate reduction by three equations

from Eq. (9) and one equation from Eq. (10). Section 2.2 describes algorithmic approaches for determining equilibrium compositions.

2.2 Phase-coexistence calculations

Two different algorithms are used to characterize phase-coexistence based on the principles outlined in Section 2.1. For systems with at least one critical point and two-phase coexistence, an iterative and perturbative approach based on natural parameter continuation (NPC) is used to construct binodal curves originating from a critical point. Otherwise, the approach described by Mao et al.[39] based on convex hull construction (CHC) is used. NPC is straightforward and computationally efficient but limited, while CHC is general but computationally intensive. Nevertheless, with this combination, the equilibrium composition of phases at coexistence can be reliably determined for models described by FH solution theory. The following algorithms are thus used to provide ground-truth results and requisite training data for the development of ML models (Section 2.3).

Natural parameter continuation (NPC). For two-phase coexistence, Eq. (9) is rearranged as

$$\Delta\mu_i^{\alpha\beta}(T,\rho,\phi^\alpha,\phi^\beta) \equiv \mu_i^\beta - \mu_i^\alpha = 0 \text{ for } i \in \{\mathrm{A},\mathrm{B},\mathrm{C}\}. \tag{12}$$

Provided a point on the coexistence curve ϕ^*, a nearby point can be identified by solving a set of linear equations that enforce Eq. (12) following a small perturbation in the composition:

$$\sum_{j\in\{\mathrm{A},\mathrm{B}\}}\sum_\pi \left.\frac{\partial\Delta\mu_i^{\alpha\beta}}{\partial\phi_j^\pi}\right|_{\phi^*} \delta\phi_j^\pi = 0 \text{ for } i \in \{\mathrm{A},\mathrm{B},\mathrm{C}\} \tag{13}$$

where $\delta\phi_j^\pi$ is the small perturbation in the composition of species j in phase π. Coexistence curves (i.e., a locus of equilibrium composition tuples) can then be constructed as follows:

1. Define tolerance parameters $\delta^\varnothing$ and δ^0.
2. Identify and set the critical point to be ϕ^*.
3. Generate a random, small perturbation on the composition $\delta\phi$, yielding two new compositions: $\phi' = \phi^* + \delta\phi$ and $\phi'' = \phi^* - \delta\phi$
4. Use the compositions ϕ' and ϕ'' as initial guesses to solve Eq. (12), producing coexisting compositions ϕ_{new}^α and ϕ_{new}^β that are distinct from ϕ^*.
5. Set $\phi_{\text{old}}^\pi \leftarrow \phi_{\text{new}}^\pi$ and and use for Eq. (13). Set one of the $\delta\phi_i^\pi$ (e.g., ϕ_{B}^β) to a small perturbation and solve for the remaining $\delta\phi_i^\pi$ to produce $\delta\phi^{\alpha'}$ and $\delta\phi^{\beta'}$.
6. Set $\phi' = \phi^{\alpha(0)} + \delta\phi^{\alpha'}$ and $\phi'' = \phi^{\beta(1)} + \delta\phi^{\beta'}$ and use as initial guesses to solve Eq. (12), producing new coexisting compositions ϕ_{new}^α and ϕ_{new}^β that are those distinct from those prior.
7. Repeat steps 5 and 6 until either $||\phi_{\text{new}}^\alpha - \phi_{\text{new}}^\beta|| < \delta^\varnothing$, which indicates a closure of the coexistence curves, or when any $\phi_i^\pi < \delta^0$, which indicates termination at a composition boundary.
8. Verify validity of compositions by checking that all have $|\mathbf{H}_{\tilde{f}}| > 0$.

For the calculations described in this paper, $\delta^\varnothing = \delta^0 = 10^{-9}$. Initial trials for random composition perturbations are set to have a magnitude of 10^{-6}. Equations are solved numerically using *fsolve* from Python's `SciPy` module. Occasionally, the trial perturbations resulted in solutions that collapsed back to the critical point or other prior generated points, in which case new perturbations would be attempted with possibly different magnitudes.

Convex hull construction (CHC). For systems without critical points or valid coexistence curves extending from critical points, the utility of the NPC algorithm is limited. In such cases, we use the CHC to identify equilibrium compositions. On a free energy surface, compositions with equal chemical potential are cotangent, while stable compositions ($D > 0$) that are not cotangent with any other points exist as single phases. This information can be accurately reconstructed by creating a convex hull of the free energy surface and projecting it onto the composition space. We briefly remark on salient aspects of the algorithm as applied to a ternary system, but readers are referred to the work of Mao et al.[39] for a more complete description.

The composition space $(\phi_{\mathrm{A}},\phi_{\mathrm{B}})$ is discretized into a mesh of equilateral triangles, or two-dimensional simplices. Using a finer mesh results in more accurate calculations but also increases computational cost and memory requirements; this work uses a simplex edge-length of 0.0002. After generation of the mesh, the free energy surface (FES) is also discretized into points defined by the tuple $(\phi_{\mathrm{A}},\phi_{\mathrm{B}},\tilde{f}(\phi_{\mathrm{A}},\phi_{\mathrm{B}}))$. The convex hull $(\phi_{\mathrm{A}}^{\mathrm{CH}},\phi_{\mathrm{B}}^{\mathrm{CH}},\tilde{f}^{\mathrm{CH}}(\phi_{\mathrm{A}}^{\mathrm{CH}},\phi_{\mathrm{B}}^{\mathrm{CH}}))$ of the FES is calculated using the Quickhull algorithm[53]. The convex hull of a non-convex FES will necessarily deform the original simplices and facilitate the identification of cotangent points on the FES. If one of the projected simplices has three unstretched sides (maximum edge length within five times initial mesh size[39]), the system is homogeneous (no phase-separation). If two sides are stretched (side length greater than five times the initial mesh size), the two farthest vertices are cotangent, indicating two coexisting phases. If all three sides are stretched, the three vertices of the simplex are cotangent, indicating three coexisting phases. With graph theoretic techniques, the number of equilibrium phases and their compositions can be determined.

2.3 Machine learning details

We explore machine learning algorithms as computationally expedient and generalizable alternatives to more traditional approaches for characterizing phase coexistence of multicomponent systems. Neural network architectures, with and without physics-informed loss functions, are optimized using data generated by the algorithms described in Section 2.2. The performance of the ML models is evaluated based on predicting the number of coexisting phases, their compositions, and relative abundance for FH models not featured in training data.

Dataset description. The dataset in this work is comprised of 1,036 phase diagrams: 107 diagrams (10%) with no phase separation (one phase), 538 diagrams (52%) with up to two-phase coexistence, and 391 diagrams (38%) with up to three-phase coexistence. Each phase diagram is produced using the methods of Section 2.2 with a distinct parameter set: $s = (\boldsymbol{\chi}, \mathbf{v}) = (\chi_{AB}, \chi_{BC}, \chi_{AC}, v_A, v_B, v_C)$.

Parameters for the models are each selected from the range $v_i \in [1,3]$ and $\chi_{ij} \in [1,3]$ where values for both ranges are discretized with a resolution of 0.1. Let s denote a parameter set and U denote the set of all possible parameter sets. With the given discretization, the total membership of U is then $|U| = 21^6$. Initially, 750 possible parameter sets are randomly selected from U with uniform probability to form $S \subset U$; care is taken to ensure that all parameter sets from this sampling are unique. From this initial sampling, only around 6.6% (≈ 50) of the selected parameter sets yielded three-phase coexistence. Using these parameter sets to define $T \subset S$, the representation of such rare systems is augmented by generating six additional parameter sets for each parameter set $t \in T$. Each new parameter set t' is generated from t by adding a Gaussian random vector $\mathbf{X}$. In particular, we use $t' = t + \mathbf{X}$ with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where $\sigma = 0.005$. All t' that yielded three-phase coexistence are collected and added to S, resulting in a final membership of $|S| = 1,036$ parameter sets.

Input and output labels for the dataset are then generated as follows. First, the composition-space of the mixture is discretized into a uniform mesh with resolution 10^{-4}. For each parameter set, if there are more than 1,000 single-phase simplices, the centroid of 1,000 randomly chosen simplices is added to the database; otherwise, the centroid of all single-phase simplices is added. For double-phase simplices, if there are more than 1,000, a random point between the ends of 1,000 randomly chosen double-phase separations is generated; otherwise, a random point between each double-phase separation is added. For multiple three-phase separations, a uniform number of points is generated in each region, ensuring a total of 1,000 three-phase points in the database. Since the number of single and double-phase simplices are determined by the size of the discretization mesh, the number of data points per each parameter set can vary.

For each tuple (ϕ_A, ϕ_B) the number of equilibrium phases, their compositions, and their abundances are recorded. In this fashion, we define an input vector $\mathbf{x} = (\chi_{AB}, \chi_{BC}, \chi_{AC}, v_A, v_B, v_C, \phi_A, \phi_B) \in \mathbb{R}^8$ that is linked to two outputs. The first output is a one-hot encoded classification vector $\mathbf{y}_c \in \mathbb{R}^3$, for which a nonzero entry indicates the presence of one, two, or three phases at equilibrium. The second output is a vector $\mathbf{y}_r \equiv \left(\phi_A^\alpha, \phi_B^\alpha, \phi_A^\beta, \phi_B^\beta, \phi_A^\gamma, \phi_B^\gamma, w^\alpha, w^\beta, w^\gamma\right) \in \mathbb{R}^9$, which describes the composition and abundances of the equilibrium phases. The phases are ordered such that ϕ_A^α has the minimum value among all ϕ_A $\left(\phi_A^\alpha \leq \phi_A^\beta \leq \phi_A^\gamma\right)$. If two phases have the same ϕ_A, they are further ordered according to ϕ_B. Such an ordering ensures a consistent representation of the equilibrium phases.

For systems with a single phase, $\phi_A^\alpha, \phi_B^\alpha$ match the inputs ϕ_A, ϕ_B, and w^α is set to unity; the abundance entries for phases β and γ are set to zero. However, the composition abundance entries for phases β and γ are assigned a value of $1/3$. The value $1/3$ is chosen to distribute errors uniformly across species. The absolute composition of these species in equilibrium will be determined by the abundance of the respective phases. For systems with two equilibrium phases, entries for the third phase compositions (i.e., ϕ_A^γ, ϕ_B^γ) are set to $1/3$, and the abundance w^γ is set to zero.

Model architectures. Figure 2 summarizes model architectures in this study; all models are implemented using PyTorch.[54] Every model takes as input $\mathbf{x}$ and predicts two outputs: $\hat{\mathbf{y}}_c$ and $\hat{\mathbf{y}}_r$. Both $\hat{\mathbf{y}}_c$ and $\hat{\mathbf{y}}_r$ have the same number of entries and ordering as described for $\mathbf{y}_c$ and $\mathbf{y}_r$; however, $\hat{\mathbf{y}}_c$ contains the predicted probabilities that $\mathbf{x}$ yields one, two, or three phases.

The basic model architecture consists of three, fully-connected hidden layers, each with m (tunable hyperparameter) hidden units; this yields a hidden vector $\mathbf{h} \in \mathbb{R}^m$. This hidden vector is then passed through a "classification layer" with *softmax* activation to yield $\hat{\mathbf{y}}_c$. This vector is also fed into separate "regression layers" to predict the composition $(\phi^\alpha, \phi^\beta, \phi^\gamma)$ and abundance $(\mathbf{w})$. Each regression layer consists of three hidden units, representing the composition of A, B, and C for each phase, and the abundance of α, β, and γ phases. *Sigmoid* activation is applied to limit predicted values on compositions and abundances to be between zero and unity, which avoids obviously unphysical values; however, overall composition and abundance constraints are not enforced. Since the composition of C depends on A and B, only the predictions for A and B compositions are kept and combined with the abundance predictions to form $\hat{\mathbf{y}}_r$.

We also consider a variation on the basic model architecture that enforces consistency between $\hat{\mathbf{y}}_r$ and the majority class featured in $\hat{\mathbf{y}}_c$. This is achieved using a mask-layer that sets abundance entries in $\hat{\mathbf{y}}_r$ to zero based on the plurality class indicated in $\hat{\mathbf{y}}_c$ (see dashed box in Figure 2). For example, if one equilibrium phase is predicted, then abundance entries associated with the β and γ phase are set to zero. If two equilibrium phases are predicted, then entries associated with the γ phase are set to zero. If three equilibrium phases are predicted, then $\hat{\mathbf{y}}_r$ is preserved from the regression layer. Compositions of species for nonexistent phases are set to $1/3$, as described earlier. To enforce constraints on overall composition and abundance, the physics-informed (PI) model incorporates *softmax* activation functions, ensuring that predicted phase compositions and abundances sum to unity. As an alternative approach, following the masking, *softmax* normalization is applied to the abundances to ensure their sum equals unity.

Loss functions. Models are optimized using loss functions that target raw numerical accuracy as well as physical sensibility. Both the simple baseline and PI models optimize a composite loss function

$$\mathscr{L}_{\text{base}} = \lambda_{\text{CE}}\mathscr{L}_{\text{CE}} + \mathscr{L}_{\text{MAE}} \tag{14}$$

that combines losses for classification cross-entropy (CE), $\mathscr{L}_{\text{CE}}$, and regression mean absolute error (MAE) loss, $\mathscr{L}_{\text{MAE}}$. The

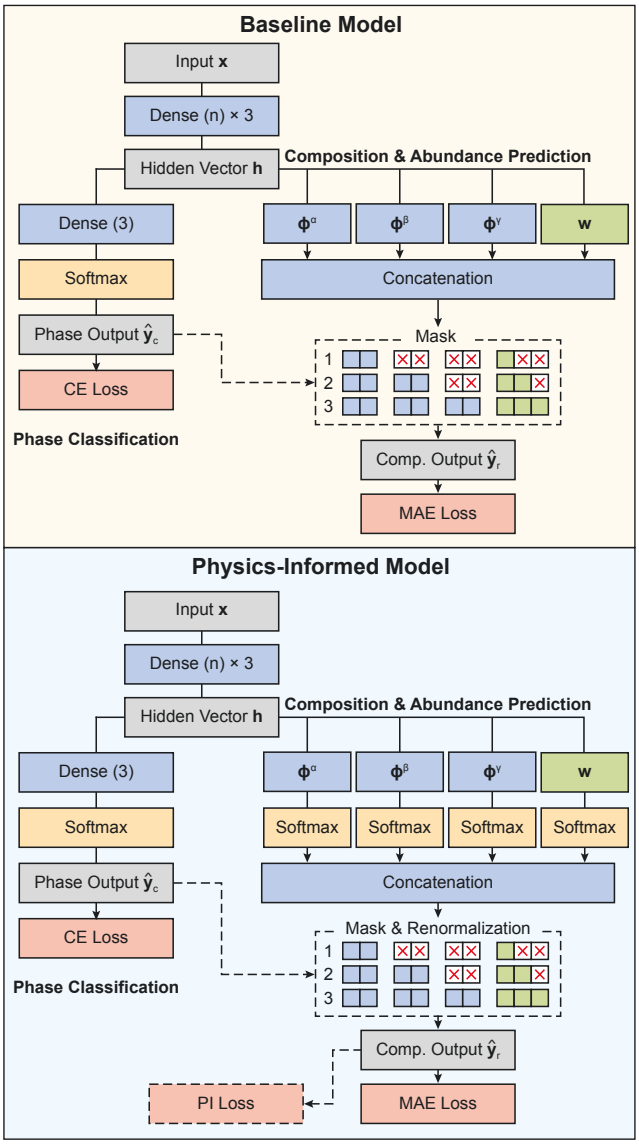

Fig. 2 **Architectures of the machine learning models**. In both the baseline and physics-informed (PI) models, the parameter vector $\mathbf{x}$ is fed into the model to produce an intermediate hidden vector $\mathbf{h}$. The hidden vector $\mathbf{h}$ produces two outputs: (1) a phase classification probability vector $\hat{\mathbf{y}}_c$, trained with cross-entropy (CE) loss, and (2) an equilibrium composition and abundance vector $\hat{\mathbf{y}}_r$, trained with mean absolute error (MAE) loss. Softmax activation is applied in PI models to ensure that the equilibrium composition and abundance vectors sum to unity. Optional functionalities (indicated by dashed lines and boxes) include a "mask", activated based on $\hat{\mathbf{y}}_c$, which sets corresponding elements in $\hat{\mathbf{y}}_r$ to zero if an input is classified as one- or two-phase. For PI models, softmax renormalization is applied to the masked abundance to ensure the sum equals unity. Additionally, in PI models, a PI composite loss can be incorporated alongside the MAE loss during training for $\hat{\mathbf{y}}_r$ prediction.

weighting parameter ($\lambda_{CE} = 0.1$) is determined empirically to balance loss magnitudes throughout training. While a perfect and physically meaningful model would necessarily minimize $\mathscr{L}_{base}$, with data limitations, simply minimizing the baseline loss function may not strictly satisfy all the criteria prescribed for thermodynamic systems at equilibrium. We therefore also consider augmented PI models (referred to as PI+) optimized with a composite loss function that includes additional regression targets

$$\mathscr{L}_{PI} = \mathscr{L}_{base} + \lambda_{split}\mathscr{L}_{split} + \lambda_{\Delta\mu}\mathscr{L}_{\Delta\mu} + \lambda_f\mathscr{L}_f \tag{15}$$

where $\lambda_{split} = 0.01$, $\lambda_{\Delta\mu} = 0.01$, and $\lambda_f = 0.001$ (identical across all models) are weighting parameters chosen through grid search over {0.001, 0.01, 0.1, 1} using the same architecture (detailed in Supplementary Information Tables S4 and S5). The chosen values yielded statistically comparable equilibrium composition regression R^2 and equilibrium phase classification F_1 scores to other parameter combinations, with performance deteriorating only when $\lambda_f \geq 0.01$. These values balance the influence of the PI losses to focus on minimizing $\mathscr{L}_{base}$ while incorporating physical constraints. The specific functional forms for these PI losses are described next.

In Eq. (15), the additional loss terms aim to satisfy different constraints on the thermodynamics of physical systems. In particular, $\mathscr{L}_{split}$ relates to constraints on the total composition of a given species distributed across phases:

$$\mathscr{L}_{split} = \sum_{i\in\{A,B\}} \left(\phi_i - \sum_{\pi\in\{\alpha,\beta,\gamma\}} w^\pi \phi_i^\pi\right)^2. \tag{16}$$

The second term $\mathscr{L}_{\Delta\mu}$ relates to the condition of equal chemical potentials for species across coexisting equilibrium phases. This loss is calculated as

$$\mathscr{L}_{\Delta\mu} = \frac{1}{2}\sum_\pi\sum_{\pi'}\sum_{i\in\{A,B,C\}} \log\left(1 + (\Delta\mu_i^{\pi\pi'})^2\right) \tag{17}$$

where $\Delta\mu_i^{\pi\pi'}$ is as defined in Eq. (12), and the first two summations are over the ground-truth equilibrium coexisting phases (i.e., $\pi, \pi' \in \{\alpha,\beta\}$ for two-phase coexistence and $\pi, \pi' \in \{\alpha,\beta,\gamma\}$ for three-phase coexistence). The additional term $\mathscr{L}_f$ promotes the minimization of the free energy of the equilibrium system:

$$\mathscr{L}_f = \sum_\pi \sum_{i\in\{A,B,C\}} w^\pi \phi_i^\pi \mu_i^\pi. \tag{18}$$

We acknowledge there are various reasonable ways to constraint losses for physical constraints; the current work examines the overall strategy of incorporating physical information into the ML workflow rather than identifying optimal implementations.

Model training and assessment. To assess model generalizability and mitigate selection bias on test data, a nested five-fold cross-validation (CV) procedure is used. Stratified sampling is employed to evenly distribute diagrams featuring one, two, and three phases across the five folds. Then, five iterations are performed in a process referred to as the *outer CV*.

Each iteration uses a unique fold as the test set and the remaining four folds as the overall training set to provide a more robust assessment of model performance.

The overall training set is further divided into training and validation sets, using a similar five-fold CV approach (*inner CV*) to the outer CV process. Each fold of the inner CV is trained with 10% of the training data for efficient hyperparameter optimization. Tunable hyperparameters include batch sizes of $\{5000, 10000, 20000\}$, learning rates of $\{0.001, 0.005, 0.01\}$, the presence (or absence) of a mask, and the number of neurons selected from $\{64, 128, 256\}$ for each hidden layer. The optimal hyperparameter setting for each fold is identified by the highest average validation composite score across five sub-folds, calculated as the sum of the F_1 score for classification and the average R^2 score.

Each fold of the outer CV uses the optimal hyperparameter settings identified from its corresponding five-fold inner CV and retrains the model for up to 500 epochs. The retraining involves selecting 80% of the overall training diagrams as the training set and 20% as the validation set, using the same stratified splitting. During the retraining process, the impact of training data sizes on model performance is assessed by using 1%, 5%, 10%, 20%, 30%, and up to 100% of the training data. Because each diagram contains a different number of data points, the number of training, validation, and test set data points ranges from 1,447,511 to 1,458,241, from 358,987 to 366,470, and from 452,308 to 460,074.

The nested CV strategy yields a mean and standard deviation of F_1 and R^2 scores as determined from the five-fold outer CV test sets. Given the imbalanced phase distribution in the dataset, the F_1 score evaluates classification performance, while the R^2 score assesses regression accuracy for the variables in $\mathbf{y}_\mathrm{r}$.

2.4 Post-inference optimization

We implement a post-inference optimization procedure to correct some deficiencies in ML model predictions. This procedure uses the predictions from the ML model as a warm-start on initial values for more traditional optimization algorithms (e.g., truncated Newton method). The objective function for minimization is

$$\mathscr{L}_\mathrm{post} = \left(\frac{1}{2}\sum_{\pi}\sum_{\pi'}\sum_{i\in\{\mathrm{A,B,C}\}} \log\left(1+\left(\Delta\mu_i^{\pi\pi'}\right)^2\right)\right)\ldots \qquad (19)$$

$$+\mathbb{1}_{w^\alpha<1}(1-\mathbb{1}_{w^\gamma>0})\mathscr{L}_\mathrm{col}$$

where $\mathbb{1}_c$ is an indicator function equal to unity when the condition c is satisfied and zero otherwise and

$$\mathscr{L}_\mathrm{col} = \frac{1}{2}\sum_{\pi}\sum_{\pi'}\frac{\phi^\pi-\phi}{|\phi^\pi-\phi|}\cdot\frac{\phi^{\pi'}-\phi}{|\phi^{\pi'}-\phi|}. \qquad (20)$$

For both Eq. (19) and (20), the first two summations are over predicted equilibrium coexisting phases (i.e., π, $\pi' \in \{\alpha\}$ for a single equilibrium phase, π, $\pi' \in \{\alpha,\beta\}$ for two-phase coexistence, and π, $\pi' \in \{\alpha,\beta,\gamma\}$ for three-phase coexistence). Eq. (20) is specifically relevant for two-phase coexistence and is minimized when the tie-line composition vectors are collinear and oriented in opposite directions. The indicator function $\mathbb{1}_{w^\alpha<1}$ excludes computation of $\mathscr{L}_\mathrm{col}$ when there is only a single predicted phase, as this term would otherwise diverge. In fact, this procedure has no effect when only a single equilibrium phase is predicted. Relatedly, we note that this algorithm is asymmetrically robust against erroneous misclassification of the system phase behavior. If the predicted number of phases exceeds the true number of phases, then converged solutions will "collapse" compositions onto those of the true equilibrium phases. However, this procedure will not identify the true solution if the predicted number of phases is fewer than the ground-truth number.

The final optimization employs the `Newton-CG` optimizer in `scipy` module in Python. The Jacobian and Hessian matrix for the objective function are computed using the `autograd` package through automatic differentiation. The maximum number of iterations for optimization is limited to 10,000. If newly optimized compositions are within a tolerance of 10^{-7} of the ideal value of the objective function, these values replace the predictions proffered by the ML model. Optimizations are only considered successful if they satisfy the stability criterion $|\mathbf{H}_{\tilde{f}}| > 0$ (see Eq. (6)).

3 Results

3.1 Performance with a basic architecture

With the standard loss functions (i.e. $\mathscr{L}_\mathrm{CE}$ and $\mathscr{L}_\mathrm{MAE}$) and a basic architecture, the ML model predicts phase separation and equilibrium compositions reasonably well. Figures 3a and 3b qualitatively depict performance in both classification and regression for some representative phase diagrams. Figures 3c and 3d quantitatively summarize results across all phase diagrams.

The ML model capably predicts the number of phases at equilibrium with an overall accuracy rate of about 97% (Figure 3c). The primary source of error (4.6%) stems from misclassifying three-phase points as two-phase, which is attributed to the relative paucity of three-phase splits (only 17% of the total data). Additionally, a portion of two-phase points (2.2%) are misclassified as one-phase. A closer inspection of predicted phase diagrams suggests that misclassifications mostly occur near binodals.

The model also performs well in predicting phase abundances and their compositions (Figure 3d). By inspection, there are vertical error regions in the predicted abundance for all phases at true abundances of 0 and 1. These errors stem from inaccurate predictions of equilibrium abundance for non-existent phases, such as phases β and γ in a one-phase region and phase γ in a two-phase region. This misprediction also leads to similar error regions for equilibrium compositions around 1/3.

Table 1 provides the baseline expectations for a standard ML model. It highlights nuances in regression performance across different phase regions. The single-phase region has the lowest average MAE (0.006), followed by the two-phase (0.023) and three-phase (0.037) regions. This trend suggests increasing difficulty in predicting compositions as the number of coexisting phases increases. Notably, all R^2 values remain high across all phases, with the three-phase region exhibiting a value above 0.88.

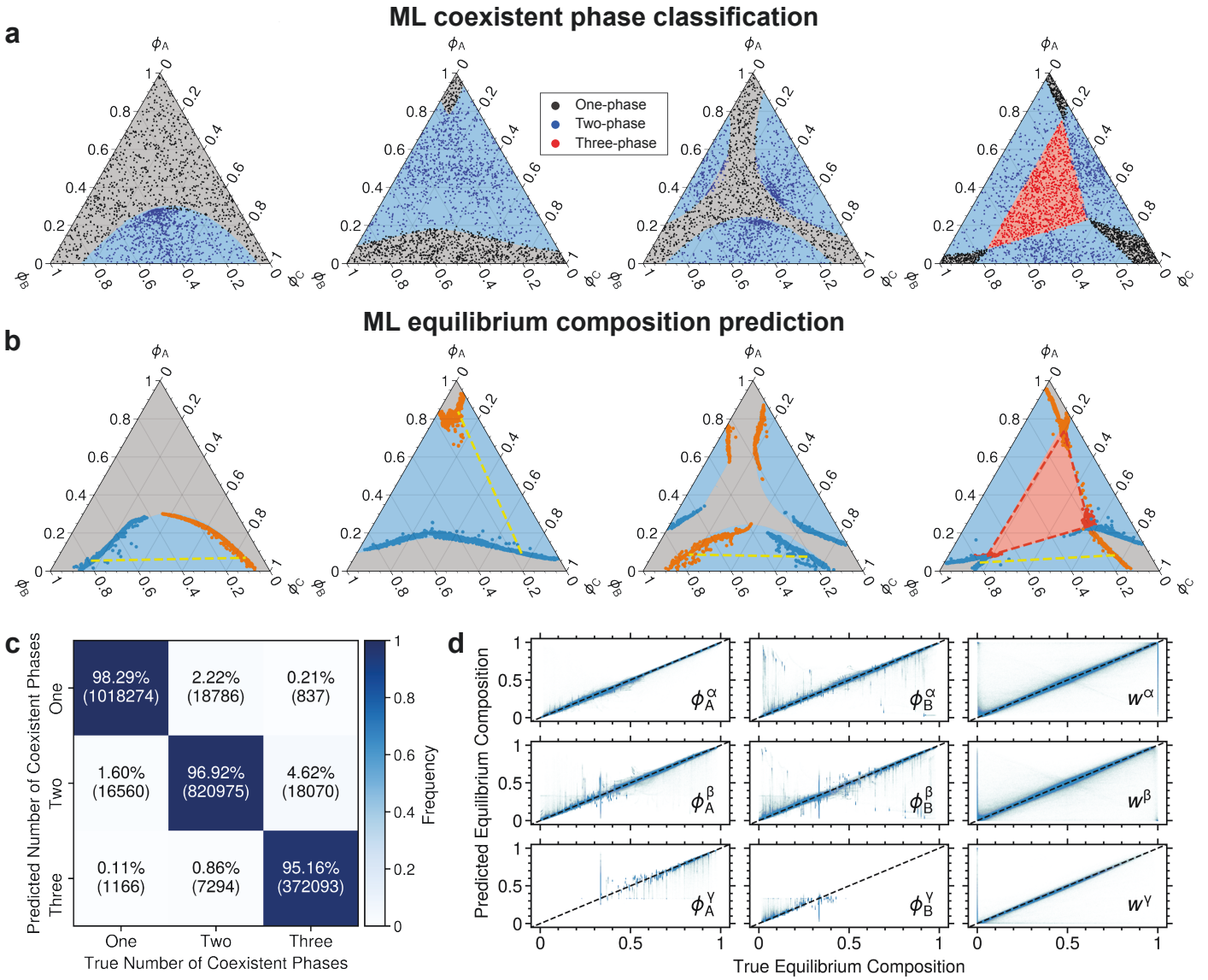

Fig. 3 **Performance summary of the baseline model**. **a)** Classification of the number of coexisting phases. The background color in all phase diagrams denotes the ground truth phase: gray (one-phase), blue (two-phase), and red (three-phase). Scatter points represent the predicted phase splits for a given initial composition, and the legend colors indicate the types of predicted splits. The parameters of the phase diagrams are detailed in Supplementary Information Tables S1 and S2. **b)** Predicted equilibrium compositions. Blue and orange scatter points represent two-phase equilibrium compositions. The yellow dashed line is a tie line for the two-phase split. Red scatter points depict composition that split into three phases. The red dashed triangle connects the three compositions at equilibrium. **c)** Confusion matrix for the predicted number of equilibrium phases. Diagonal entries represent correctly classified instances, while off-diagonal entries represent misclassifications. **d)** Parity plot for predicted equilibrium compositions. The diagonal dashed line represents perfect performance.

Table 1 **Performance of representative models for equilibrium composition prediction on the test set across different phase regions.** Mean values are reported with standard deviation in parentheses. The bold and underscored number indicates the best result.

	MAE			R^2		
	Base	PI	PI+	Base	PI	PI+
One-phase	0.006 (0.001)	**0.005 (0.001)**	**0.005 (0.001)**	0.982 (0.005)	0.987 (0.004)	**0.988 (0.003)**
Two-phase	0.023 (0.003)	**0.022 (0.001)**	0.023 (0.003)	0.912 (0.015)	**0.915 (0.009)**	0.913 (0.015)
Three-phase	**0.037 (0.006)**	0.038 (0.003)	0.038 (0.008)	0.884 (0.038)	0.883 (0.023)	**0.889 (0.041)**

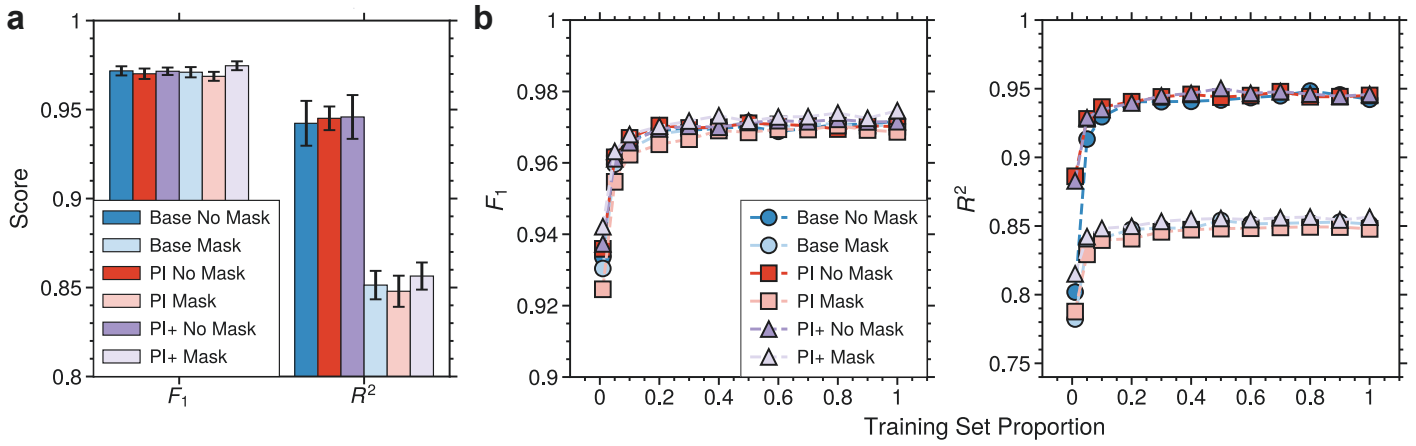

Fig. 4 **Impact of physical constraints and data size on phase-coexistence prediction**. **a)** Comparison of test set phase classification F_1 and equilibrium composition prediction R^2 across baseline (base), physics-informed (PI), and augmented PI (PI+) models, with and without classification masks, using five-fold cross-validation (CV). Bars represent mean values, and error bars indicate standard deviations. **b)** The impact of training data size on model performance. Each dot represents the average score calculated across the five-fold CV.

3.2 Performance with physics-informed losses and consistency constraints

To build on the prior model, we evaluate the potential of incorporating additional physical information on prediction accuracy (Figure 4a). In particular, physical constraints on overall composition and abundance, along with several physics-informed losses (detailed in Section 2.3) are implemented, and classification masks are used to zero the abundances of non-existent phases. While the baseline, PI, and PI+ models without classification masks achieve comparable F_1 and R^2 scores, models with masks significantly underperform in equilibrium composition prediction (Figure 4b).

The baseline model exhibits lower accuracy compared to the PI and PI+ models in one-phase and two-phase regions, while its performance is comparable to other models in three-phase scenarios (Table 1). The PI and PI+ models show similar performance across all scenarios under these metrics. The coexistence curve predictions of both the PI and PI+ models are similar (Figures 5a and S6), producing smooth and physically sensible two- and three-phase coexistence curves. In contrast, the baseline model generates erratic two-phase coexistence curves that significantly deviate from the true curves. This is also evident from the distribution of the MAE loss in Figure 5b, where the baseline model (red) has a higher average MAE than the PI (blue) and PI+ (green) models, which perform similarly. The unphysical coexistence curves of the baseline model highlight the limitations of using broad performance metrics to assess improvements in predictive accuracy. Errors are better resolved by examining deviations in chemical equilibrium potential ($\Delta\mu$) and split loss ($\mathscr{L}_{\text{split}}$), where the baseline model shows significantly larger errors than both the PI and PI+ models.

To further examine the impact on composition and abundance constraints, we analyze two additional metrics: $\mathscr{L}_{\text{unity}}$, which relates to the volume fractions of each species within a given phase, and $\mathscr{L}_{\text{weight}}$, which measures overall material conservation. These metrics are defined as:

$$\mathscr{L}_{\text{unity}} = \sum_{\pi\in\{\alpha,\beta,\gamma\}} \text{ReLU}\left(\phi_{\text{A}}^{\pi} + \phi_{\text{B}}^{\pi} - 1\right), \tag{21}$$

and

$$\mathscr{L}_{\text{weight}} = \left(1 - \sum_{\pi\in\{\alpha,\beta,\gamma\}} w^{\pi}\right)^2. \tag{22}$$

Since the PI and PI+ models enforce unity in composition and abundance through the *softmax* activation function, $\mathscr{L}_{\text{unity}}$ and $\mathscr{L}_{\text{weight}}$ remain zero for these models, whereas the baseline model violates these constraints (Figure 5b). The PI+ model, trained with additional constraints, demonstrates smaller deviations in chemical equilibrium potential ($\Delta\mu$) and marginally improves composition prediction, split loss, and free energy minimization loss compared to the PI model. Overall, designing a physics-informed model architecture to enforce material constraints is essential; however, the addition of extra losses or masks complicates training without yielding significant improvements in phase classification or equilibrium composition prediction. Therefore, the PI architecture, without additional losses, emerges as the best practical choice for implementation.

3.3 Performance with limited data

Having achieved accurate phase-coexistence predictions with a dataset of over 1.4 million data points, we investigated whether PI models would be more data efficient and achieve comparable performance with less data. Figure 4b demonstrates that even with only 10% of the data, the PI and PI+ models maintain high accuracy in phase classification (F_1: 0.967 and 0.967, respectively) and equilibrium composition prediction (R^2: 0.937 and 0.935, respectively). In contrast, the baseline model performs worse in equilibrium composition prediction (R^2: 0.930) but achieves comparable accuracy in phase classification (F_1: 0.967). The advantages of incorporating composition and abundance constraints are particularly evident in low-data scenarios (training with 1% and 5% of the data), where the PI and PI+

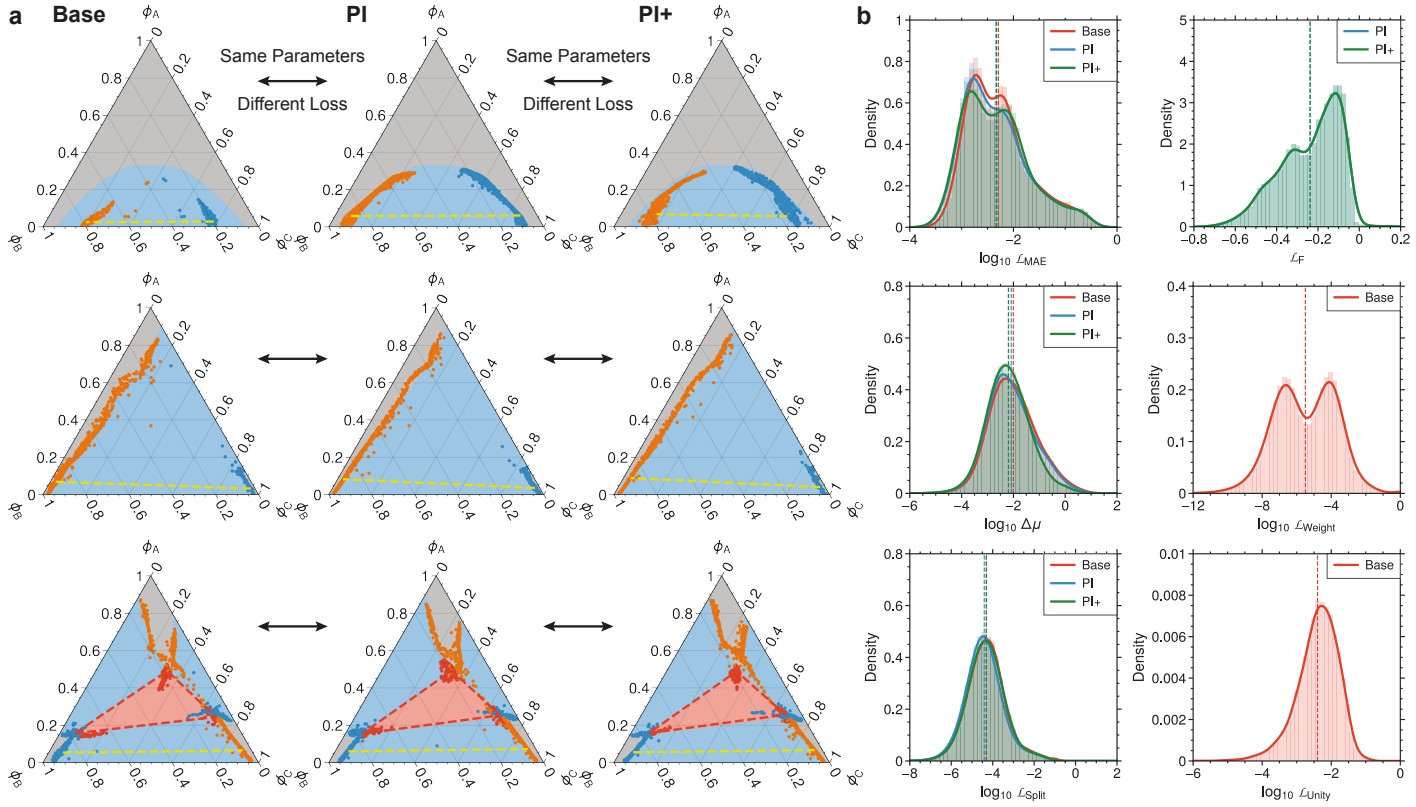

Fig. 5 **Comparison of performance metrics and physical constraints among baseline, PI, and PI+ models. a)** Predicted phase coexistence curves for the baseline (left), PI (middle), and PI+ (right) models. Arrows indicate that both predictions are for the same phase diagram. The background color in all phase diagrams denotes the ground truth phase: gray (one-phase), blue (two-phase), and red (three-phase). Blue and orange scatter points represent two-phase equilibrium compositions. The yellow dashed line is a tie line for the two-phase split. Red scatter points depict composition that split into three phases. The red dashed triangle connects the three compositions at equilibrium. The parameters of the phase diagrams are detailed in Supplementary Information Tables S1 and S2. **b)** Data distribution (shaded bars) and kernel density estimation fits (lines) for performance metrics and physical constraints. Vertical dashed lines indicate mean values.

models significantly outperform the baseline. Although the PI+ model slightly outperforms the others with the full dataset (F_1: 0.972, R^2: 0.946), the improvement over using 10% of the data is marginal. The predicted phase diagrams with coexistence curves (Figures S2, S4, S5) using 10% of the data are qualitatively accurate across the baseline, PI, and PI+ models. These findings underscore the critical role of physical constraints in enhancing model generalization under limited data conditions.

3.4 Post-ML optimization

Seeded with ML predictions, a Newton-CG method can efficiently converge to arbitrarily accurate and precise equilibrium compositions (Figure 6). This is demonstrated for the PI model, trained on the full dataset, where initial errors (Figures 6a,b) can be virtually eliminated after the optimization procedure (Figures 6c and S1). The baseline and PI+ models also achieve comparable performance after post-ML optimization (Figures S3-S6), even when trained on only 10% of the data (Figures S2, S4, S5). This combination of efficiency and accuracy could enable the handling of more complex systems and scaling to resource-intensive measurements, where data may be sparse or scarce.

Errors for models trained on fold 1 data were analysed across a random sample of 187 two-phase and 76 three-phase equilibrium phase diagrams to better characterize post-ML optimization errors (Tables 2 and S1). The results indicate that Newton-CG optimization, initialized with predictions from ML models, achieves near-perfect success rates and significantly reduces deviations from true equilibrium compositions compared to individual ML model predictions. After post-ML optimization, the PI model trained on the full dataset outperforms both the baseline and PI+ models in predicting two-phase and three-phase coexistence. The relative ranking of performances post-optimization aligns well with the relative ranking of the ML predictions alone, underscoring the importance of initial ML prediction accuracy in determining the effectiveness of the post-ML optimization process. With limited training data, all models perform similarly, with PI+ showing slightly better overall performance. Errors remain comparable to those observed with the full dataset, although three-phase coexistence errors consistently exceed those for two-phase coexistence. This disparity likely stems from the relative scarcity of three-phase coexistence in the training set, which increases complexity and complicates precise prediction.

The post-ML optimization process is also efficient and parallelizable – taking less than 1 second to converge to the optimal solution (Table S1). The ML model training requires less than 1 MB for 140,000 parameters, a substantial reduction in mem-

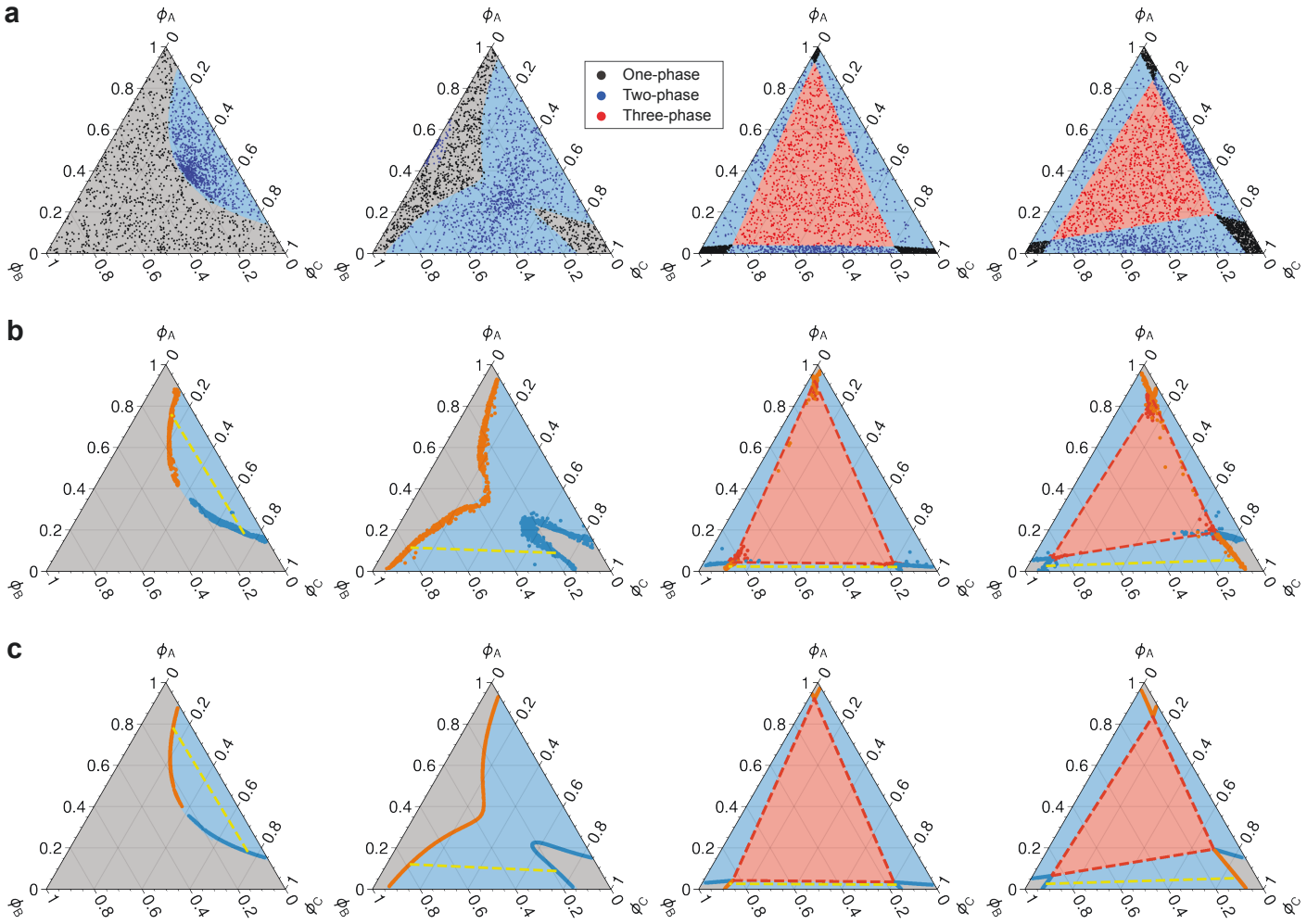

Fig. 6 **Refinement of PI model predictions. a)** Classification of the number of coexisting phases. The background color in all phase diagrams denotes the true phase: gray (one-phase), blue (two-phase), and red (three-phase). The scatter points indicate the predicted phase splits for a given initial composition. Colors in the legend denote the types of predicted splits. The parameters of the phase diagrams are detailed in Supplementary Information Tables S1 and S2. **b)** Predicted coexistence curves. Blue and orange scatter points indicate two-phase coexistence curves, with the yellow dashed line denoting an example tie line. The vertices of the red triangle indicate three-phase coexistence points. **c)** Coexistence curves produced with the post-ML optimization strategy. The results are obtained using ML inference to warm-start Newton-CG optimization.

Table 2 **Performance of equilibrium composition prediction with machine learning (ML) and post-optimization prediction.** Mean absolute errors (MAE) for the composition of species A and B from the fold 1 model are reported, with standard error of the mean in parentheses. The best result is bold and underlined.

	Data size	Two-phase ($\times 10^{-2}$)		Three-phase ($\times 10^{-2}$)	
		ML Prediction	Post-optimization	ML Prediction	Post-optimization
Base	100%	2.40 (0.01)	0.88 (0.02)	3.66 (0.03)	2.76 (0.03)
PI	100%	2.39 (0.01)	0.87 (0.02)	**3.19 (0.02)**	**2.45 (0.03)**
PI+	100%	2.94 (0.01)	0.95 (0.02)	4.26 (0.03)	3.32 (0.03)
Base	10%	2.92 (0.01)	0.96 (0.02)	3.93 (0.02)	2.94 (0.06)
PI	10%	2.95 (0.01)	1.03 (0.02)	4.58 (0.03)	3.56 (0.03)
PI+	10%	**2.36 (0.01)**	**0.83 (0.02)**	3.98 (0.03)	3.00 (0.03)

ory usage compared to arc continuation or convex hull methods, which demand approximately 1 GB of storage and 50 GB of memory per run. Overall, the accurate ML predictions of equilibrium compositions enable rapid convergence to highly accurate solutions, offering significant advantages in both memory- and time-efficiency.

4 Conclusions

In this work, we presented an efficient and extensible machine learning-based approach for calculating phase coexistence in

ternary systems. A neural network trained on phase coexistence data was able to predict the number and compositions of equilibrium phases for a solution prepared at a given composition under a specific mixing potential. Incorporating physical constraints into the neural network architecture enhanced prediction accuracy, while additional physics-informed losses offered no significant improvement. The physics-constrained architecture produced higher-quality models with less data, offering advantages in scenarios where data acquisition is labor- or resource-intensive. However, the resulting models still exhibit errors that may be unacceptable for certain applications, such as process simulation software. To achieve precise results, a Newton conjugate gradient method was used, with machine-learning predictions serving as a warm start for optimization to determine final equilibrium phase compositions. This integration of neural networks with numerical refinement enabled rapid and accurate predictions of coexisting phases, their compositions, and abundances.

This work motivates several areas of future inquiry. Extensions to systems with more components would increase utility for complex industrial and biological processes. Expanding beyond the Flory-Huggins theory by incorporating other free energy models or data from molecular simulation, perhaps in a single framework, would further enhance its generalizability across diverse chemical systems. Additionally, exploring more advanced physics-informed learning strategies, incorporating uncertainty quantification, and refining neural network architectures could boost prediction efficiency and reliability. Collectively, these directions could enhance both the theoretical and practical impact of leveraging ML for phase coexistence calculations.

Acknowledgement

S.D. and M.A.W. acknowledge support from the National Science Foundation. The development of certain machine learning approaches and analyses is based on work supported by the National Science Foundation under Grant No. 2118861. The investigation of phase behavior in polymer solutions is supported by the National Science Foundation under Grant No. 2237470. S.J. acknowledges support from the Princeton Catalysis Initiative. Calculations were performed using resources from Princeton Research Computing at Princeton University, which is a consortium led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing. The authors also acknowledge Andrej Košmrlj and Qiwei Yu for helpful discussions regarding the convex hull algorithm.

Supplementary Information

Additional optimized phase diagrams; phase classification confusion matrices; equilibrium composition prediction parity plots; post-ML optimization performance.

Data availability

The data associated with this study are publicly accessible at `https://doi.org/10.5281/zenodo.13776946`.

Code availability

The code associated with this study is publicly accessible at `https://github.com/webbtheosim/ml-ternary-phase`.

Notes and references

1 G. R. Guillen, Y. Pan, M. Li and E. M. V. Hoek, *Industrial and Engineering Chemistry Research*, 2011, **50**, 3798–3817.
2 D. R. Lloyd, S. S. Kim and K. E. Kinzer, *Journal of Membrane Science*, 1991, **64**, 1–11.
3 Y. Fily and M. C. Marchetti, *Physical Review Letters*, 2012, **108**, 235702.
4 J. Runnström, *Developmental Biology*, 1963, **7**, 38–50.
5 H. Walter and D. E. Brooks, *FEBS Letters*, 1995, **361**, 135–139.
6 M. Muschol and F. Rosenberger, *The Journal of Chemical Physics*, 1997, **107**, 1953–1962.
7 F. J. Iborra, *Theoretical Biology and Medical Modelling*, 2007, **4**,.
8 S. Weber and C. Brangwynne, *Current Biology*, 2015, **25**, 641–646.
9 A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. Kanagaraj, H. Kim, T. Mittag and J. Taylor, *Cell*, 2015, **163**, 123–133.
10 C. P. Brangwynne, T. J. Mitchison and A. A. Hyman, *Proceedings of the National Academy of Sciences*, 2011, **108**, 4334–4339.
11 C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher and A. A. Hyman, *Science*, 2009, **324**, 1729–1732.
12 J. M. Michael Michelsen, *Thermodynamic Modelling: Fundamentals and Computational Aspects*, Tie-Line Publications, 2004.
13 D.-Y. Peng and D. B. Robinson, *Industrial and Engineering Chemistry Fundamentals*, 1976, **15**, 59–64.
14 T. Jindrová and J. Mikyška, *Fluid Phase Equilibria*, 2013, **353**, 101–114.
15 Y. Zhan, Z. Hu, J. Kou and Q. Su, *Physics of Fluids*, 2024, **36**,.
16 T. Jindrová and J. Mikyška, *Fluid Phase Equilibria*, 2015, **393**, 7–25.
17 N. A. Gokcen, *Journal of Phase Equilibria*, 1996, **17**, 50–51.
18 J. J. v. Laar, *Zeitschrift für Physikalische Chemie*, 1910, **72U**, 723–751.
19 M. L. McGlashan, *Journal of Chemical Education*, 1963, **40**, 516.
20 H. Renon and J. M. Prausnitz, *AIChE Journal*, 1968, **14**, 135–144.
21 G. Maurer and J. Prausnitz, *Fluid Phase Equilibria*, 1978, **2**, 91–99.
22 A. Fredenslund, R. L. Jones and J. M. Prausnitz, *AIChE Journal*, 1975, **21**, 1086–1099.
23 X. Feng, M.-H. Chen, Y. Wu and S. Sun, *Journal of Computational Physics*, 2022, **463**, 111275.
24 P. J. Flory, *The Journal of Chemical Physics*, 1941, **9**, 660–660.
25 J. Kou, S. Sun and X. Wang, *Computational Geosciences*, 2016, **20**, 283–295.

26 D. V. Nichita, *Fluid Phase Equilibria*, 2018, **466**, 31–47.

27 D. V. Nichita, *Fluid Phase Equilibria*, 2017, **447**, 107–124.

28 D. V. Nichita, J.-C. de Hemptinne and S. Gomez, *Petroleum Science and Technology*, 2009, **27**, 2177–2191.

29 M. Castier, *Fluid Phase Equilibria*, 2014, **379**, 104–111.

30 T. Zhang, Y. Li, Y. Li, S. Sun and X. Gao, *Computer Methods in Applied Mechanics and Engineering*, 2020, **369**, 113207.

31 P. Civicioglu, *Applied Mathematics and Computation*, 2013, **219**, 8121–8144.

32 H. Wang, Z. Hu, Y. Sun, Q. Su and X. Xia, *Computational Intelligence and Neuroscience*, 2018, **2018**, 1–27.

33 S. Wang, X. Da, M. Li and T. Han, *Journal of Systems Engineering and Electronics*, 2016, **27**, 395–406.

34 J. L. Sengers, *How Fluids Unmix: Discoveries by the School of Van der Waals and Kamerlingh Onnes*, Edita-the Publishing House of the Royal; 1st edition, 2002.

35 H. Binous and A. Bellagi, *Engineering Reports*, 2020, **3**,.

36 M. L. Michelsen, *Fluid Phase Equilibria*, 1982, **9**, 1–19.

37 H. Sidky, J. K. Whitmer and D. Mehta, *AIChE Journal*, 2016, **62**, 4497–4507.

38 H. Sidky, A. C. Liddell, D. Mehta, J. D. Hauenstein and J. K. Whitmer, *Industrial & Engineering Chemistry Research*, 2016, **55**, 11363–11370.

39 S. Mao, D. Kuldinow, M. P. Haataja and A. Košmrlj, *Soft Matter*, 2019, **15**, 1297–1311.

40 C. B. Barber, D. P. Dobkin and H. Huhdanpaa, *ACM Transactions on Mathematical Software*, 1996, **22**, 469–483.

41 Y. An, M. A. Webb and W. M. Jacobs, *Science Advances*, 2024, **10**, eadj2448.

42 K. Terayama, K. Han, R. Katsube, I. Ohnuma, T. Abe, Y. Nose and R. Tamura, *Scripta Materialia*, 2022, **208**, 114335.

43 G. Deffrennes, K. Terayama, T. Abe and R. Tamura, *Materials and Design*, 2022, **215**, 110497.

44 J. C. R. Thacker, D. J. Bray, P. B. Warren and R. L. Anderson, *The journal of physical chemistry. B*, 2023, **127**, 3711–3727.

45 R. Tamura, G. Deffrennes, K. Han, T. Abe, H. Morito, Y. Nakamura, M. Naito, R. Katsube, Y. Nose and K. Terayama, *Science and Technology of Advanced Materials: Methods*, 2022, **2**, 153–161.

46 J. G. Ethier, R. K. Casukhela, J. J. Latimer, M. D. Jacobsen, B. Rasin, M. K. Gupta, L. A. Baldwin and R. A. Vaia, *Macromolecules*, 2022, **55**, 2691–2702.

47 J. Lund, H. Wang, R. D. Braatz and R. E. García, *Materials Advances*, 2022, **3**, 8485–8497.

48 Y. Li, T. Zhang and S. Sun, *Industrial and Engineering Chemistry Research*, 2019, **58**, 12312–12322.

49 Y. Aoki, S. Wu, T. Tsurimoto, Y. Hayashi, S. Minami, O. Tadamichi, K. Shiratori and R. Yoshida, *Macromolecules*, 2023, **56**, 5446–5456.

50 J. G. Ethier, D. J. Audus, D. C. Ryan and R. A. Vaia, *Giant*, 2023, **15**, 100171.

51 H. Tompa, *Transactions of the Faraday Society*, 1949, **45**, 1142.

52 R. L. Scott, *The Journal of Chemical Physics*, 1949, **17**, 268–279.

53 C. B. Barber, D. P. Dobkin and H. Huhdanpaa, *ACM Transactions on Mathematical Software (TOMS)*, 1996, **22**, 469–483.

54 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, *Advances in neural information processing systems*, 2019, **32**,.