Supplementary Information

Supplementary Methods: Derivation of Loss Function

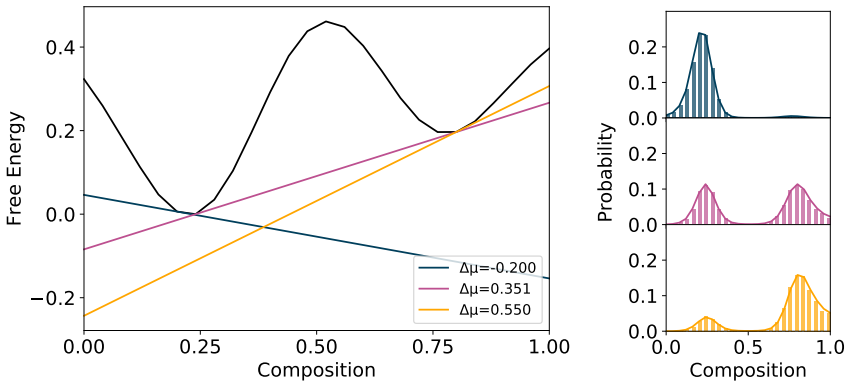
The KL divergence between the model and the true distributions can be expressed as follows:

$$\mathrm{KL}(P_{\mathrm{AR}}|P_{\mathrm{SG}}) = \mathbb{E}_{\mathrm{AR}}[\log(\frac{P_{\mathrm{AR}}}{P_{\mathrm{SG}}})] \tag{1}$$

$$\mathbb{E}_{\mathrm{AR}}[\log(\frac{P_{\mathrm{AR}}}{P_{\mathrm{SG}}})] = \mathbb{E}_{\mathrm{AR}}[\log(P_{\mathrm{AR}}) - \log(P_{\mathrm{SG}})] \tag{2}$$

$$= \mathbb{E}_{\mathrm{AR}}[\log[P_{\mathrm{AR}}(\mathbf{S})] + \frac{U(\mathbf{S})}{k_{\mathrm{b}}T} - \frac{\sum_{i\neq 0}\Delta\mu_i N_i}{k_{\mathrm{b}}T} + \log(Z_{\mathrm{SG}})] \tag{3}$$

where $\log(Z_{\mathrm{SG}})$ is a constant that can that be removed without affecting the optimization.



(a) (b)

Supplementary Figure 1: Semi-grand canonical thermodynamics of two-component system with N = 25. (a) Free energy per site with varying composition computed using enumeration of all states and tangent construction of Legendre transform at varying values of $\Delta\mu$. (b) Composition probability histograms for values of $\Delta\mu$ of -0.200 (top), 0.351 (middle), and 0.550 (bottom) obtained through enumeration of all states (lines) and self-normalized importance sampling (SNIS) (bars).

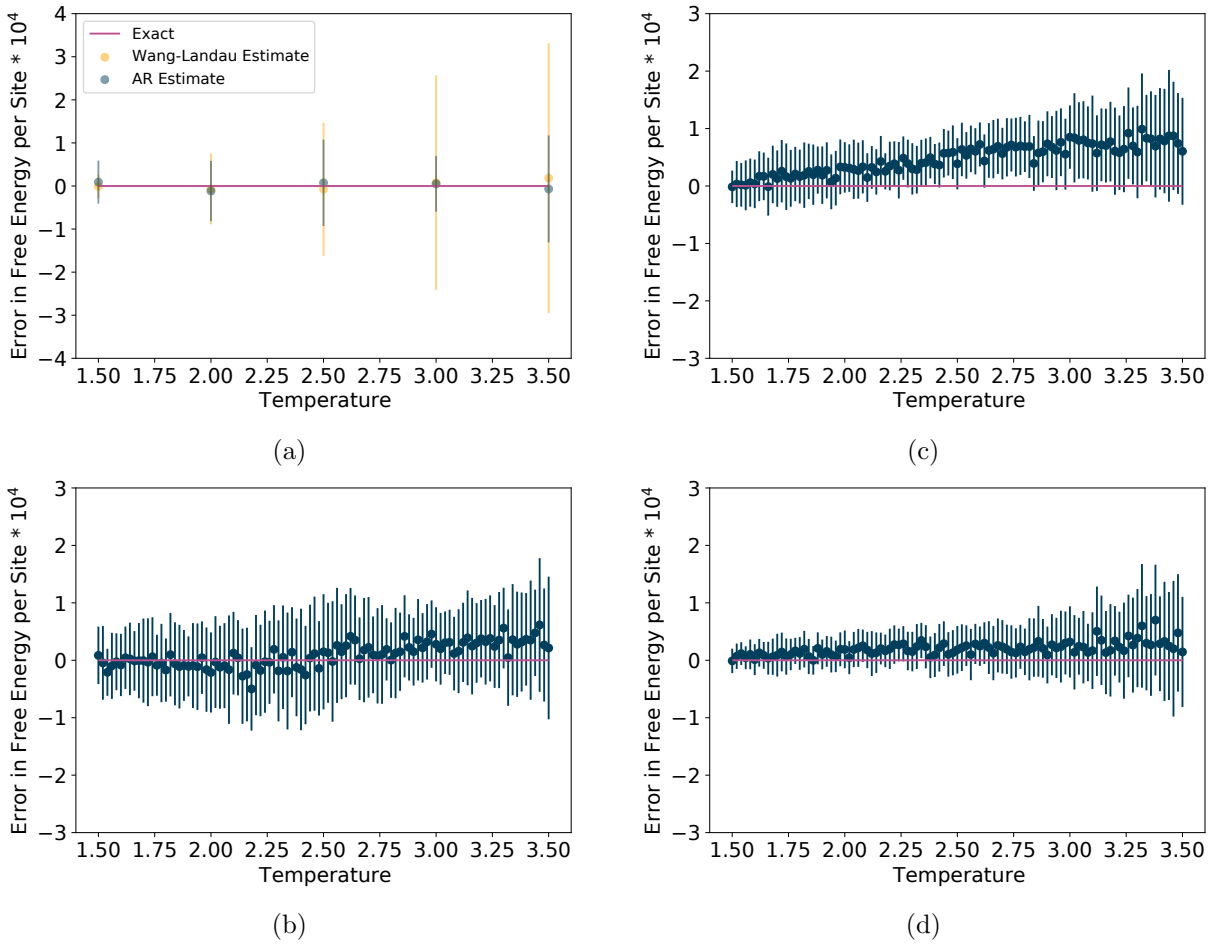
Supplementary Note 1: Architecture Details

In this section, we describe the structure of the neural network architectures used in this work. We represent each site variable $\mathbf{S}_i$ as a categorical one-hot over the number of possible components in the system. $\mathbf{S}$ is then the concatenation of $\mathbf{S}_i$. Therefore, for a 50-site binary system, $\mathbf{S}$ is 100 dimensional.

The simplest autoregressive layer is a weight matrix W with masked parameters such that the imposed dependence between the variables is preserved. With T and $\Delta\mu$ constraints appended to the start of $\mathbf{S}$, the non-zero components of a single layer W_{ij} include $j \leq 2 * \lfloor \frac{i}{2} \rfloor$. The weight matrix of the second layer can be expanded with non-zero terms W_{ij}, $j \leq 2 * (\lfloor \frac{i}{2} \rfloor + 1)$. Element-wise activation functions can be included after the application of weight matrices.

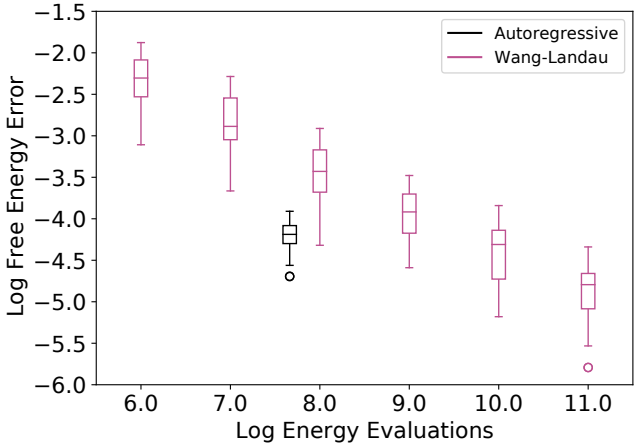
Additional layers to the autoregressive networks can be added with the same form as the second layer. The depth of all models is specified in the corresponding experimental details section.

Supplementary Note 2: Ising Alloy Experimental Details



Supplementary Figure 2: Errors in free energy per site computed by SEGAL through self-normalized importance sampling (SNIS). (a) SEGAL and Wang-Landau (run for 10^9 evaluations) free energies compared with exact values for $B = 0.0$ case. (b-d) SEGAL estimates compared with Wang-Landau (run for 10^{11} evaluations) values for (b) $B = 0.0$, (c) $B = 0.2$, and (d) $B = 0.4$ cases. The benchmark Wang-Landau approach was an implementation of an algorithm by Beladinelli et al.[1,2] using portions of code developed by Kristjan Haule[3]. In (a), the mean (points) and standard deviation (bars) are shown over 50 trials of SEGAL sampling and 50 trials of the Wang-Landau algorithm. In (b-d), the mean (points) and standard deviation (bars) are shown over 50 trials of SEGAL sampling.

We determined the accuracy of SEGAL's free energy estimates when compared to the exact values at $B = 0$ by computing the mean of the absolute error over temperatures in [1.5,2.0,2.5,3.0,3.5]. We also determined the accuracy of the benchmark Wang-Landau method, computed in the same manner, when it was run at $B = 0$ for varying numbers of energy evaluations.



Supplementary Figure 3: Accuracy of SEGAL's estimated free energies compared with Wang-Landau benchmark at $B = 0$. The energy evaluation cost of training SEGAL is $2.8 * 10^7$. The cost of one series of 20,000 samples at each set of constraints is $1.8 * 10^7$ energy evaluations. The box plots are composed of 50 trials of SEGAL sampling or 50 trials of the relevant Wang-Landau algorithm. The box plots are defined as follows: the center-line is the median, the box edges are first and third quartiles, the whiskers stretch 1.5*(Interquartile Range) beyond the quartiles, and outliers outside of the whiskers are shown as points.

Usually, thermodynamic quantities obtained through derivatives such as the heat capacity and magnetic susceptibility are computed with Monte Carlo methods using the fluctuations of the samples.

$$C_{\mathrm{B}} = [\langle U^2 \rangle - \langle U \rangle^2]/k_{\mathrm{b}}T \tag{4}$$

$$\chi = [\langle M^2 \rangle - \langle M \rangle^2]/k_{\mathrm{b}}T \tag{5}$$

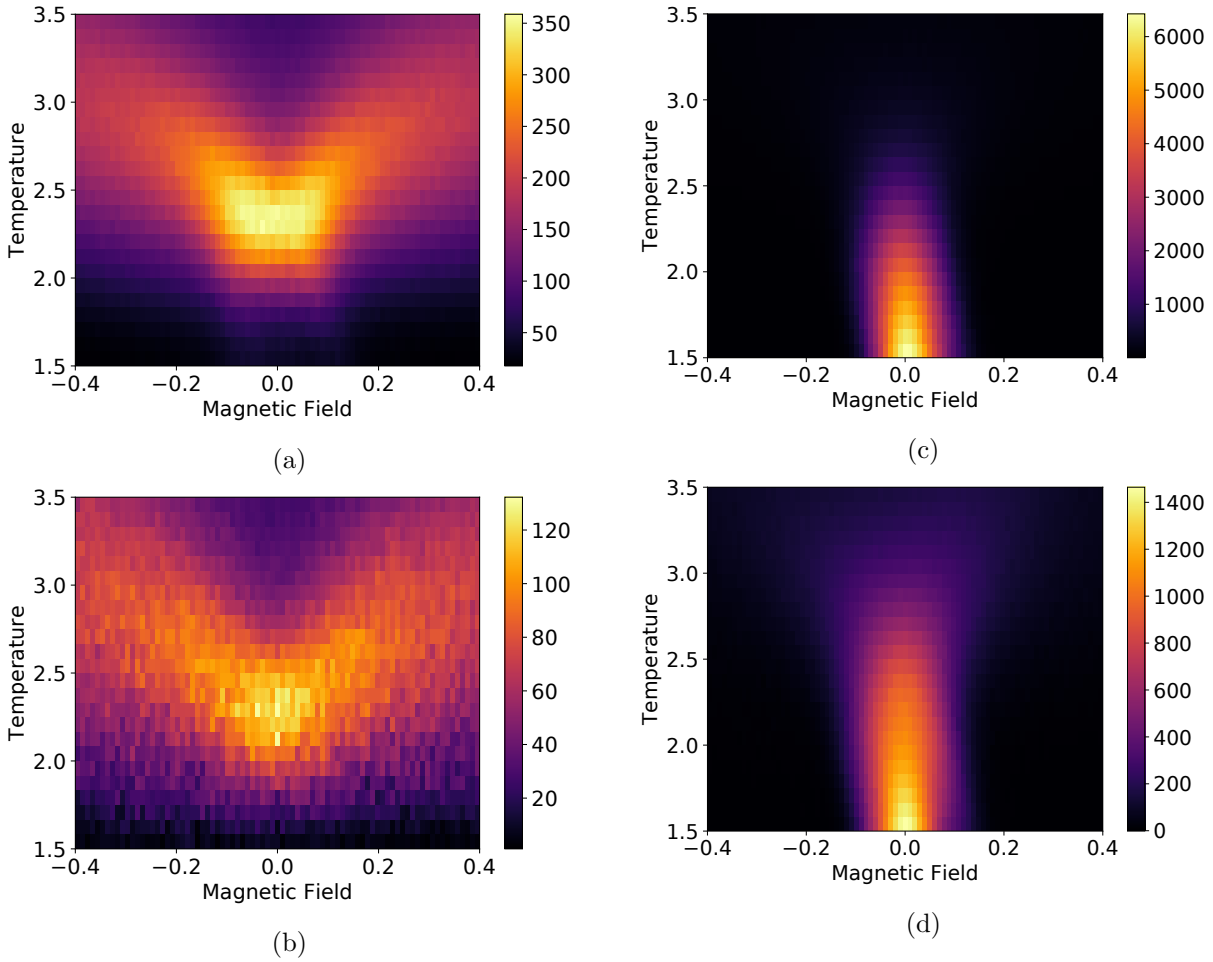
However, the differentiable structure of the function $P(\mathbf{S})$ allows for another strategy to obtain these derivatives. Treating χ as the example:

$$\chi = \frac{d}{dB} M(B,T) = \frac{d}{dB} \mathbb{E}_{\mathrm{AR}}[M(\mathbf{S})] \tag{6}$$

Using the same log-derivative trick found the methods section:

$$\frac{d}{dB} \mathbb{E}_{\mathrm{AR}}[M(\mathbf{S})] = \mathbb{E}_{\mathrm{AR}}[M(\mathbf{S}) \frac{d}{dB} \log(P_{\mathrm{AR}}(\mathbf{S}))] \tag{7}$$

where $\frac{d}{dB}\log(P_{\mathrm{AR}}(\mathbf{S}))$ can be computed with automatic differentiation implemented in PyTorch. We report thermodynamic quantites obtained using this method below. We note that when heat capacities and susceptibilities are computed using automatic differentiation, they exhibit similar general trends to the fluctuation approach, but estimates are noisier and reduced in magnitude.



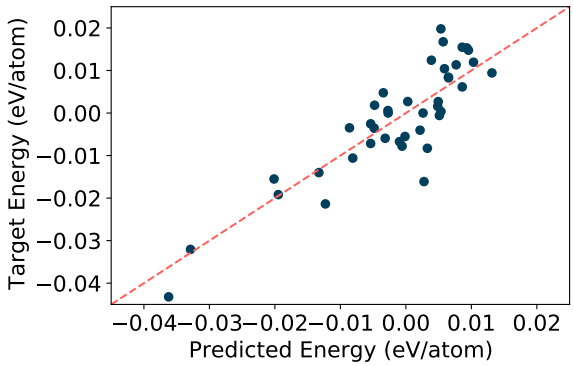
(a) (c)

(b) (d)

Supplementary Figure 4: Thermodynamic quantities computed using fluctuations of observables and neural network differentiation. All calculations use 10,000 samples at each set of constraints. (a) Heat capacity computed using fluctuations. (b) Heat capacity computed using neural network differentiation. (c) Magnetic susceptibility computed using fluctuations. (d) Magnetic susceptibility computed using neural network differentiation.

The Ising alloy models were composed of three layers as described above with tanh activation functions after the first two layers. Networks were trained for 22762 epochs with the rmsprop optimizer and a learning rate of $10^{-2.92}$. Each iteration contained 1250 total samples that were divided among 25 sets of conditions (50 samples per condition). The 25 conditions were chosen from all possible combinations of fields $B \in [-0.4, -0.2, 0.0, 0.2, 0.4]$ and 5 randomly chosen temperatures in range $1.5 < T_i < 3.5$. Hyperparameters were optimized using SigOpt[4] and all neural networks implementations were built with PyTorch[5].

Supplementary Note 3: CuAu Alloy Experimental Details

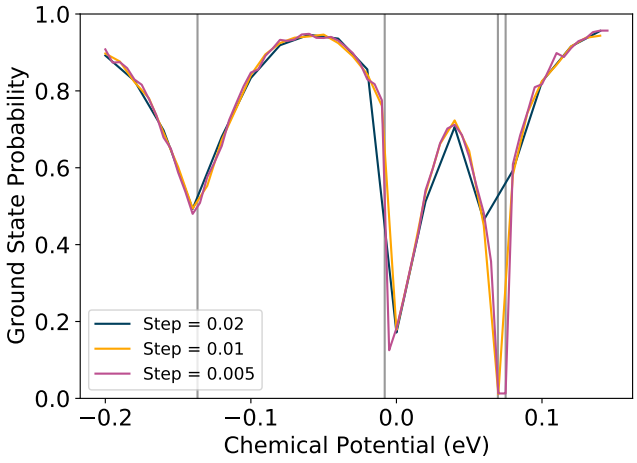


(a)

Supplementary Figure 5: Fit of CuAu cluster expansion to formation energy.

The CuAu cluster expansion was obtained using CLEASE[6] on an fcc structure with cell parameter 3.8 Å. If multiple symmetrically equivalent structures were in the database, the lowest energy structure was included in the training of the cluster expansion. Max cluster diameters for two, three, and four body clusters were 6.0 Å, 4.5 Å, and 4.5 Å respectively. The expansion was fit with LOOCV and L2 regularization.

The autoregressive models were composed of three layers with sigmoid activation. Networks were trained for 5000 epochs with the Adam optimizer and a learning rate of 10^{-3}. Each iteration contained 200 total samples that were divided among 4 sets of conditions (50 samples per condition). The 4 conditions were chosen using random chemical potentials $\Delta\mu$ in range [-0.24 eV, 0.24 eV] and an annealed temperature schedule $T = 3000 * (0.999)^{\text{iter}}$ K.



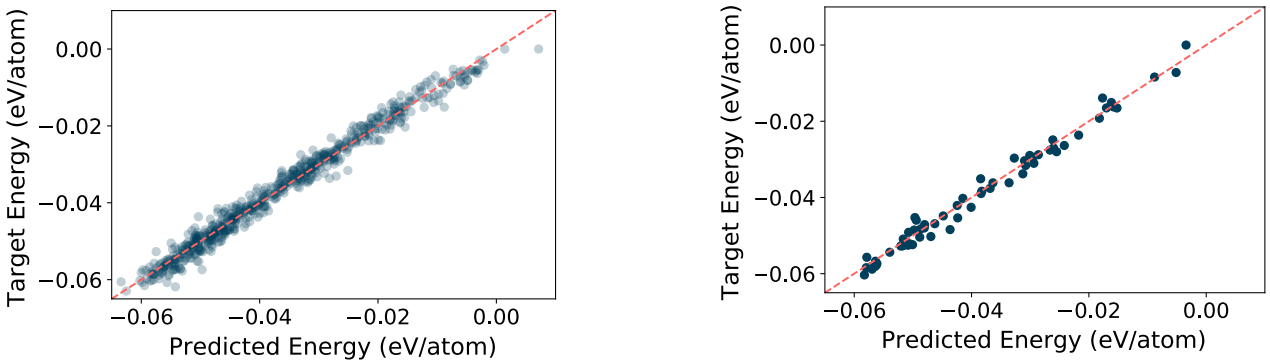
Supplementary Figure 6: Probability of sampling the grand potential minima as a function of chemical potential. Each $\Delta\mu$ is evaluated using a batch of 2500 samples. Varying step sizes are plotted to show the effects of approaching the phase transitions with increasing resolution. The minimum probability over all resolutions is 27/2500 and occurs at a $\Delta\mu$ value of 0.07 eV.

Supplementary Note 4: AgPd Alloy Experimental Details

CE: The AgPd cluster expansion was obtained using CLEASE[6] on an fcc structure with cell parameter 4.09 Å. Max cluster diameters for two, three, and four body clusters were 8.0 Å, 6.5 Å, and 5.5 Å respectively. Of the 625 structures in the ICET example, 613 were correctly added to the CLEASE database. The expansion was fit with k-fold cross validation and L2 regularization.

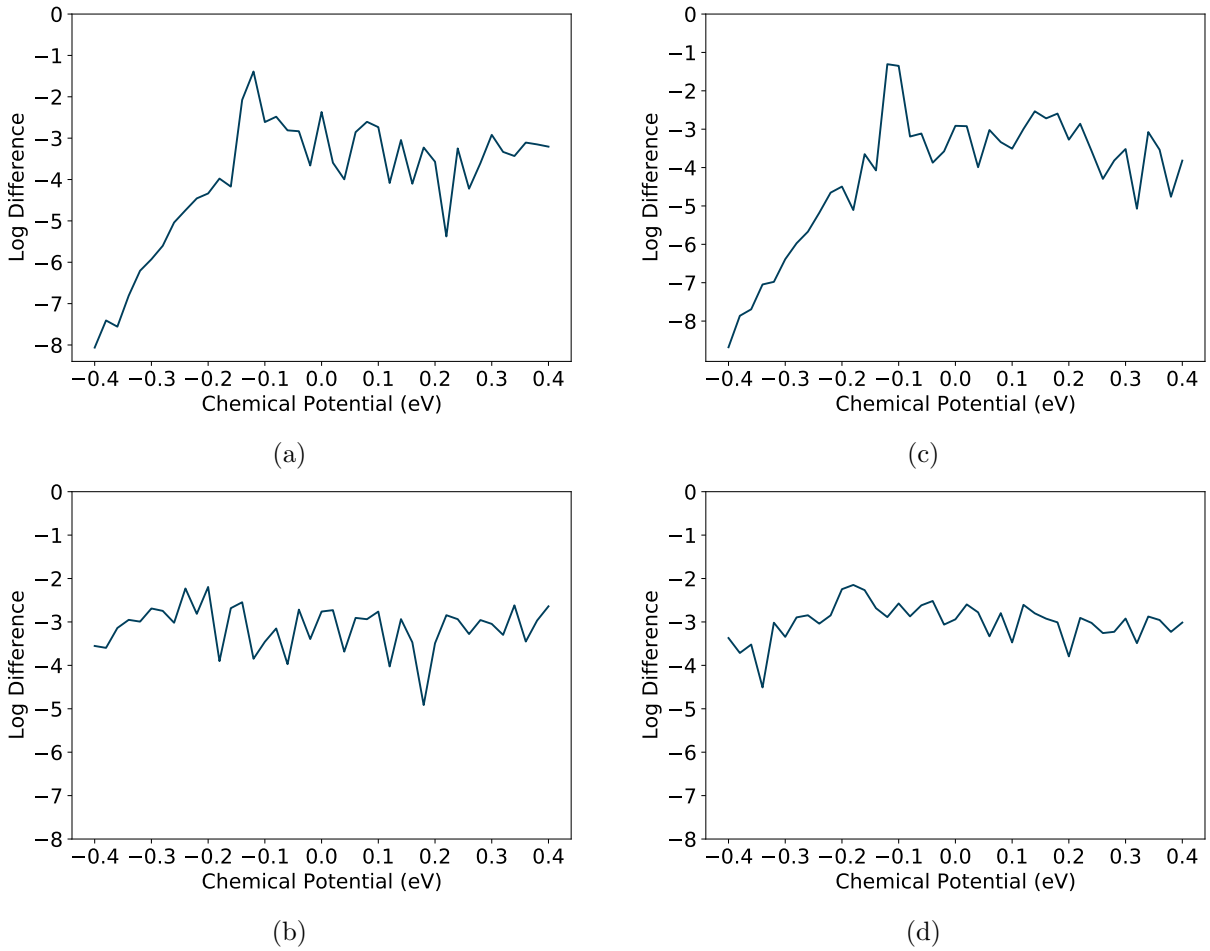
CGC: Crystal graph convolutions were trained with a (80,10,10) (training, validation, test) split. Training continued for 300 epochs with a batch size of 100. The objective was optimized with Adam. The learning rate was initially set to 0.01 and reduced to 0.001 after 100 epochs. Various hyperparameters of the convolutional structure are as follows: hidden atom features in convolution layers = 64, hidden features after pooling = 64, number of convolution layers = 3, and number of hidden layers after pooling = 2.

Experimental Details: Both 27-site AgPd models were composed of three layers as described above with sigmoid activation functions after the first two layers. Networks were trained for 17229 epochs with the Adam optimizer and a learning rate of $10^{-2.52}$. Each iteration contained 1250 total samples that were divided among 25 sets of conditions (50 samples per condition). The 25 conditions were chosen from all possible combinations of chemical potentials $\Delta\mu \in$ [-0.4 eV, -0.2 eV, 0.0 eV, 0.2 eV, 0.4 eV] and 5 randomly chosen temperatures in range 200 K $< T_i <$ 900 K. Hyperparameters were optimized using SigOpt[4] and all neural networks implementations were built with PyTorch[5]. Benchmark SGC MCMC simulations were run for 1000 sweeps (27 MC steps) in each chain with one sample recorded for each sweep. At each temperature, we performed self-normalized importance sampling from SEGAL and ran a benchmark MCMC simulation for 41 values of $\Delta\mu$ ranging from -0.4 eV to 0.4 eV. All SEGAL estimates of NESS and composition are made using 10,000 samples.



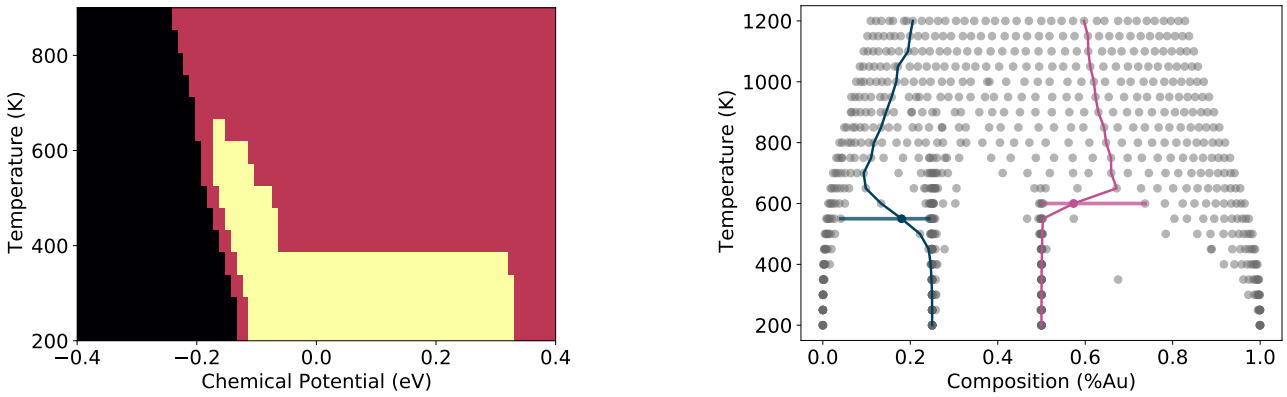
(a) (b)

Supplementary Figure 7: AgPd models for U. (a) Fit of AgPd cluster expansion to formation energy. (b) Performance of crystal graph convolutional fit of formation energy on test set.



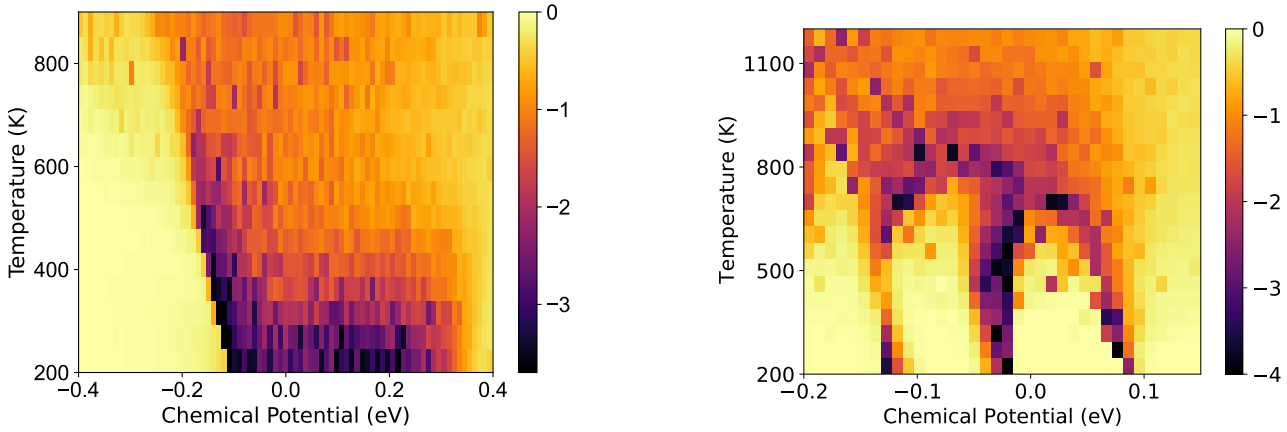
Supplementary Figure 8: Deviations in composition between SEGAL estimates and SGC MCMC simulations. (a) cluster expansion at 250 K. (b) cluster expansion at 750 K. (c) crystal graph convolution at 250 K. (d) crystal graph convolution at 750 K. Below the peak of miscibility gap, deviations peak at the discontinuity in composition due to SEGAL's performance near the phase transition.

Supplementary Note 5: Phase Diagrams Experimental Details



(a) (b)

Supplementary Figure 9: (a) Image segmentation for 125-site AgPd SEGAL model obtained with $k = 1400$ and $\sigma = 1.1$[7,8]. $\log_{10}$NESS is computed using 5,000 samples at 15 temperatures $\in$ [200 K, 900 K] and 81 values of $\Delta\mu \in$ [-0.4 eV, 0.4 eV] (b) Procedure for identifying order-disorder two-phase region for CuAu system. Vertical paths are at constant $\Delta\mu$ (blue: $\Delta\mu$ = -0.13 eV, purple: $\Delta\mu = 0.04$ eV), and points denote the critical temperature T_C. Horizontal lines show the approximated bounds of the two-phase region. Figure shows self-normalized importance sampling estimates of composition for 36 values of $\Delta\mu \in$ [-0.2 eV, 0.15 eV] and 21 temperatures $\in$ [200 K, 1200 K]. Each estimate is computed using 5,000 SEGAL samples.



(a) (b)

Supplementary Figure 10: Normalized Effective Sample Size of (a) 125-site AgPd model and (b) 128-site CuAu model. Estimates for AgPd (CuAu) were made using 5,000 (10,000) samples from SEGAL at each set of constraints. All values are reported in log10 scale.

Experimental Details: The 125-site AgPd model had 2 layers and tanh activation functions. Networks were trained for 13286 epochs with a learning rate of $10^{-2.68}$ and the Adam optimizer. Each batch contained 500 total samples. The samples were divided among 25 sets of conditions that were chosen in the same manner as the 27-site model.

The 128-site CuAu SEGAL model was trained with the same settings as the 27-site AgPd model, except each batch contained 1500 total samples. The temperature training range was [200 K, 1200 K], and the $\Delta\mu$ values were chosen as [-0.16832 eV, -0.08416 eV, 0.0 eV, 0.05625 eV, 0.1125 eV].

Metadynamics simulations were run using the CLEASE[6] package. For AgPd, simulations were run for 15 temperatures from 200 K to 900 K. The flatness criteria was set to 0.8. The modification factor was initialized at 1000, and the simulation continued until the modification factor was less than 0.0005, where the modification factor was reduced by a factor of 2.0 after the flatness criteria had been reached. For CuAu, simulations were run for 41 temperatures from 200 K to 1200 K. The flatness criteria was set to 0.8. The modification factor was initialized at 10000, and the simulation continued until the modification factor was less than 0.0001.

Supplementary References

1. Belardinelli, R. E. & Pereyra, V. D. Wang-Landau algorithm: A theoretical analysis of the saturation of the error. *J. Chem. Phys.* **127**, 184105 (2007).

2. Belardinelli, R. E. & Pereyra, V. D. Fast algorithm to calculate density of states. *Phys. Rev. E* **75**, 046701 (2007).

3. Haule, K. Wang-Landau algorithm for 2D Ising model, http://www.physics.rutgers.edu/∼haule/681/src_MC/python_codes/wangLand.py, (2010).

4. Clark, S. & Hayes, P. SigOpt, https://sigopt.com, (2019).

5. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Proc. of the 33rd Conf. on Neural Inf. Process. Syst.*, 8024–8035 (2019).

6. Chang, J. H. *et al.* CLEASE: A versatile and user-friendly implementation of cluster expansion method. *J. Phys. Condens. Matter* **31**, 325901 (2019).

7. Felzenszwalb, P. F. & Huttenlocher, D. P. Efficient graph-based image segmentation. *Int. J. Comput. Vision* **59**, 167–181 (2004).

8. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).