

Predicting Phase Transitions Across Temperature, Pressure, and Chemical Potential Using Exponentially Tilted Thermodynamic Maps

Corresponding Author: Professor Pratyush Tiwary

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper presents a modification of a previous paper by the same authors in PNAS, where a diffusion model is assumed to carry the joint distribution of states, temperatures and pressure to a Gaussian with a non-zero mean determined by the actual pressure of the training data. In that sense there is very little that is new here. I am not knowledgeable in any of the applications the authors describe so I will only comment on the algorithm.

1) As mentioned above the methodological modification relative to the PNAS paper is quite minimal.

2) Both in this paper and in the PNAS paper the authors fail to explain the details what kind of score function of the state space is used to approximate the gradient of the density. In fact in both models since the x variable and the β or η variables are entirely decoupled a possible solution to the problem could use a separate process for the x variable and for the η variable. In that case the system would not learn anything that can be generalized to new temperatures or pressures. So, something is forcing some kind of interaction between η and x and it is important to explain that.

3) It would also be interesting to compare the loss they get with their joint model for the score, with a model that completely decouples the two processes, i.e. one reverse process that maps the prior in x to the data distribution $p(x)$ and one reverse process (which can be a simple Gaussian process with decreasing variance) mapping the prior on η to η_0 . Or for that matter simply training a model only on the x variable but at two different temperatures. Does this last one do much worse? Maybe the model is implicitly learning to implement MCMC, which depends on temperature in a very simple way.

4) Also more detail on the number of diffusion steps and how does the diffusion computation of samples compare to MCMC computation of samples. Is there a gain in computation time?

5) Overall, papers using deep networks are expected to perform ablation experiments to show that all elements of the proposed algorithm are really necessary. So the items above are examples of that.

6) In the supplemental information they provide an pseudo code algorithm but it actually confused me. In the paper there is a clear distinction between the temperature and pressure of the training sample and the gaussian population of temperatures and pressures. This somehow disappears in the pseudo code. The derivative of the density should be in terms of all the variables x, β, α . This needs to be clarified.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The paper extends the thermodynamic maps framework to capture systems with an exponentially tilted prior. This leads to a second fluctuation variable related to the mean of the prior distribution. The fluctuation variables undergo diffusion in the forward process, and contribute to the score function. The justification for the diffusion on β is that it provides a temperature dependent correction to the score function. However, given that both α and β diffuse from their initial value to a normal distribution it is unclear what diffusing on α is adding to the model.

It is not necessary to go through essentially identical derivations for both of the considered ensembles (grand canonical and isothermal-isobaric). Similarly the sections on diffusion models, exponentially tilted diffusion models, TMs, and exponentially tilted TMs do not all need a full section of explanation and could be more consolidated. Also, there is a small typo in the text between Eqs 17 and 18 – “exponential titling factor”

The thermodynamic justification for using this approach for the CO₂ phase transition is not clearly established. It is not clear why the collective variables extracted from the MD simulations should have an exponentially tilted prior given the analysis steps separating them from the isothermal isobaric system. Further explanation could give more insight into the range of systems this approach could be applied to.

While this work does extend the existing framework to a larger range of systems, the assumptions still require equilibrium systems with harmonic Hamiltonians (or those that can be collectively described as such). The ability of the framework to generate samples from unseen points remains impressive, but a comparison to a simpler diffusion model would make the impact of the thermodynamic approach clear. The idea of adding an offset to the prior mean is novel, but it's significantly less of a step than the original thermodynamic maps framework, making the improvements described in the work largely incremental. This paper is not currently suited for this journal, but a rigorous analysis of what this approach learns above a simpler diffusion model and a stronger thermodynamic analysis of how the fluctuating variables guide the diffusion could make it suitable. I expect that doing this expanded analysis rigorously will take substantial effort but I am happy to revisit my decision and re-evaluate the work if the authors do.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

Thank you for clarifying the structural basis of the relationship between the collective variables and the thermodynamically relevant quantities. The connection to the order parameters adopted from Gimondi and Salvalaglio, and the point that the CVs track the CO₂ phases and thus the free energy, resolves my question about why the tilting term is appropriate here. The added ablation studies are what most strengthen the manuscript. They establish that the tilted prior and joint score are necessary rather than incremental additions to the original framework. Removing all thermodynamic conditioning leaves the model unable to recover even the training conditions, showing that the conditioning is necessary rather than incidental. Replacing the tilted prior with a unit Gaussian then isolates the role of the thermodynamic quantities in tilting the prior, supported by the quantitative error analysis against MCMC. Together these establish what the expTM framework contributes. The clarified score function and corrected algorithm resolve the earlier ambiguity about how the feature space couples to the thermodynamic variables, and the wall-clock comparison against MCMC shows a concrete improvement in efficiency.

On balance, the revisions clarify the method and show that it is a significant enough expansion of the existing method to make it suitable for this journal. I recommend acceptance.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature

Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Revisions made to the Manuscript

First, we would like to thank the reviewers for their constructive comments and suggestions. We highly value the feedback provided, and we believe that addressing these comments has significantly improved the quality of our manuscript. The revised manuscript includes results of several ablation tests that establish how the approach in this work represents significant enhancement in speed and novelty with respect to the original Thermodynamic Maps approach from PNAS 2024. Below, we provide our response to each of the comments raised and detail the modifications we have made in the revised manuscript. All our responses are highlighted in blue and modifications are directly made in the manuscript with red for the reviewers' convenience.

Response to the reviewers

Reviewer 1

Comment 1.1 — This paper present a modification of a previous paper by the same authors in PNAS, where a diffusion model is assumed to carry the joint distribution of states, temperatures and pressure to a Gaussian with a non-zero mean determined by the actual pressure of the training data. In that sense there is very little that is new here. I am not knowledgeable in any of the applications the authors describe so I will only comment on the algorithm. As mentioned above the methodological modification relative to the PNAS paper is quite minimal.

Reply: While the modification to our previous work of Ref. [2] may look minimal, we believe that the introduction of the exponential tilting idea represents a conceptually important advancement, with non-trivial physical ramifications. This approach enables fluctuations along additional thermodynamic variables, thereby allowing the model to extend naturally beyond temperature dependence (as was done in the PNAS paper) and capture a broader range of ensembles with control parameters not limited to temperature. Furthermore, the mapping of the different isothermal-isobaric ensembles and the grand canonical ensemble to a single tunable fluctuating variable α that corresponds to the pressure and chemical potential, respectively, is the key that allows one to investigate systems under various conditions. While the implementation turned out to be straightforward, the underlying physical insight and its demonstrative ability to capture phase transitions across diverse systems underscore the significance of the contribution.

Comment 1.2 — Both in this paper and in the PNAS paper the authors fail to explain the details what kind of score function of the state space is used to approximate the gradient of the density. In fact in both models since the x variable and the β or η variables are entirely decoupled a possible solution to the problem could use a separate process for the x variable and for the η variable. In that case the system would not learn anything that can be generalized to new temperatures or pressures. So, something is forcing some kind of interaction between η and x and it is important to explain that.

Reply: We thank the reviewer for raising this point. We have revised the manuscript to clarify the exact score function used for our expTM model. To clarify, we do not learn separate scores for x and η because they are not decoupled; instead, we learn the joint score in the space of (x, η) denoted as $s_\theta(\mathbf{x}, \boldsymbol{\eta}, \tau) \approx \nabla_{\mathbf{x}, \boldsymbol{\eta}} \log p_\tau(\mathbf{x}, \boldsymbol{\eta})$. Indeed, as the reviewer intuitively feels, it is this lack of decoupling which is key to the extrapolative capability across environments. During the forward process, the data propagates stochastically and relaxes to the prior distributions. The prior is parameterized by the training conditions $[\alpha_0, \beta_0]$, which specify an exponentially tilted Gaussian prior with mean and variance explicitly controlled by $[\alpha_0, \beta_0]$, respectively. As a result, the learned joint score captures the conditional dependence of $\mathbf{x}$ on $\boldsymbol{\eta}$. In fact, this is why the joint score cannot factorize as $\nabla_{\mathbf{x}, \boldsymbol{\eta}} \log p_\tau(\mathbf{x}, \boldsymbol{\eta}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) + \nabla_{\boldsymbol{\eta}} \log p_\tau(\boldsymbol{\eta})$. Here, the feature states of $\mathbf{x}$ in the reverse SDE depend on $\boldsymbol{\eta}$ through the learned joint score function, which is what enables the extrapolation to new conditions.

Thus, in a fully decoupled solution, where the reverse processes are trained for x and η , it would indeed fail to generalize across temperatures or pressures, and as the reviewer says, this point should be emphasized better which we do now. During the reverse process, the main driving force is the learned score function, which enables mapping between the tilted Gaussian prior distribution and the target distribution. This is what allows the feature space $\mathbf{x}$ to arrive at the desired η condition of the given conditional target distribution. This indicates two main aspects: the prior is directly mapped to a tilted Gaussian, and the score is learned jointly during the reverse process. To clarify, we have added the following in the main manuscript:

Another technical but important point worth emphasizing is that we do not learn separate scores for x and η because they are not decoupled; instead, we learn the joint score in the space of (x, η) denoted as $s_\theta(\mathbf{x}, \boldsymbol{\eta}, \tau) \approx \nabla_{\mathbf{x}, \boldsymbol{\eta}} \log p_\tau(\mathbf{x}, \boldsymbol{\eta})$. It is this lack of decoupling that is key to the extrapolative capability across environments in this work, as well as in the original TM [2]. Furthermore, this joint architecture is fundamentally different from conditional generation approaches where $\boldsymbol{\eta}$ might be treated as a fixed label. Here, $\boldsymbol{\eta}$ plays a significant role, allowing it to be directly mapped to a tilted Gaussian, and the score is learned jointly during the reverse process.

Comment 1.3 — It would also be interesting to compare the loss they get with their joint model for the score, with a model that completely decouples the two processes, i.e. one reverse process that maps the prior in $\mathbf{x}$ to the data distribution $p(\mathbf{x})$ and one reverse process (which can be a simple Gaussian process with decreasing variance) mapping the prior on η to η_0 . Or for that matter simply training a model only on the $\mathbf{x}$ variable but at two different temperatures. Does this last one do much worse? Maybe the model is implicitly learning to implement MCMC, which depends on temperature in a very simple way.

Reply: We thank the reviewer for this excellent suggestion. The reviewer suggests two ablation choices: i) Decoupling the model into two separate diffusion processes. One that maps prior to the data distribution for x alone, and another diffusion process with the η variable alone. Or ii) train a model only on lattice configuration x at two fixed thermodynamic conditions, such that the model can interpolate at various thermodynamic variables without explicitly modeling the joint (x, η) space.

In response, we conducted the second ablation experiment suggested by the reviewer. Specifically, we trained diffusion models only on the lattice configurations x at two fixed thermodynamic conditions (i.e., chemical potentials μ), without embedding any conditional thermodynamic information. This model failed to generate samples, even for the conditions in the training data. This was because the training data contained mixed information about the thermodynamic conditions. At that point, extrapolating or interpolating to different conditions became meaningless, as the model lost its ability to generate

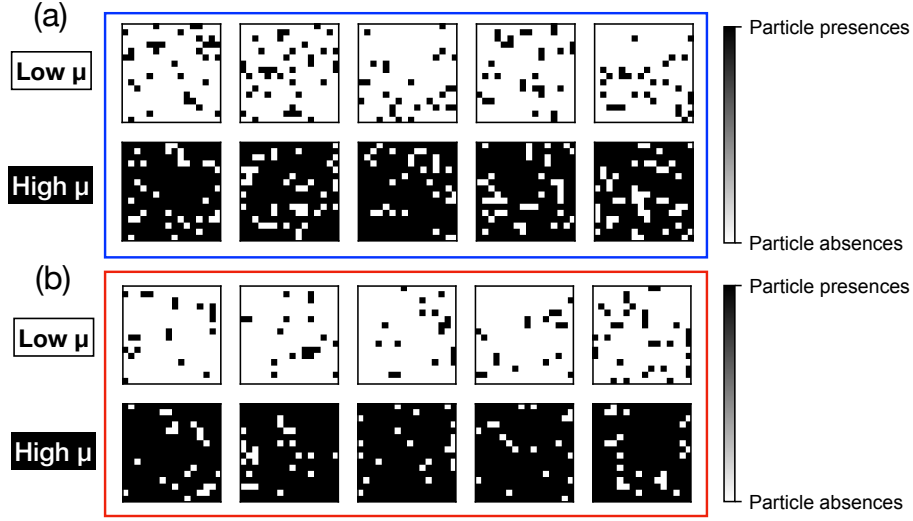


Figure 1: Ablation test on the lattice gas model. A standard conditional diffusion model was trained using configurations at two chemical potentials (low μ and high μ). (a) Representative lattice configurations from the training data at low and high chemical potentials. (b) Lattice gas configuration generated by the conditional diffusion model at the same two chemical potentials as in the training data.

samples relevant to specific thermodynamic conditions. Thus, the test clearly indicates that when the models are trained only on two fixed lattice configurations, they fail to generate meaningful results.

To further validate this point, we tested a conditional diffusion model where we treated the thermodynamic conditions by adding corresponding labels. In this case, the model embeds information regarding the chemical potential labels. As a result, the model can successfully recover the original training data at the two chemical potentials. However, it fails to generalize across thermodynamic conditions, particularly for unseen conditions during diffusion. The model cannot generate samples for unseen conditions because the corresponding labels do not exist. The results are shown in Fig. 1 and can be further found in Section IV. A. in the SI.

This highlights the strength of our expTM framework, which conditions not only on a single thermodynamic variable but on two variables simultaneously, learning their non-trivial underlying coupling. The model is not implicitly learning temperature-dependent MCMC during the diffusion dynamics. Instead, the thermodynamic variables enter through the exponentially tilted prior and the joint score, allowing the model to correctly steer toward the target distribution and capture the nontrivial behavior of phase transitions.

Comment 1.4 — Also more detail on the number of diffusion steps and how does the diffusion computation of samples compare to MCMC computation of samples. Is there a gain in computation time?

Reply: We thank the reviewer for raising this very important point, which was not discussed in the previous version of the manuscript. To address this, we directly compared the computational cost of MCMC sampling with the training and sampling costs of our expTM model. For a fair comparison, all

computations were carried out on a single NVIDIA A100 GPU throughout the test.

For MCMC, a single independent lattice configuration of a 20×20 system was generated by performing 250 equilibration steps followed by 200 production steps. For each $[T, \mu]$ pair, we generated 50 independent lattice configurations and computed the average density across them, which is the value for each pixel of the heatmap shown in the main text Fig.2 (a) and (b). The total wall-clock time required to generate these 50 configurations and compute the averaged density for one $[T, \mu]$ point was approximately 79 seconds. Repeating this procedure for 625 distinct pair of $[T, \mu]$ combinations resulted in a total wall-clock time of 49,597 seconds (approximately 13.7 hours).

In contrast, the diffusion model was trained within the SDE framework using an L_1 -smooth loss function for 10 epochs, which completed in approximately 183 seconds on the same GPU using data from only two training $[T, \mu]$ pairs. We used a VP-SDE (Variance preserving) discretization with $N_s = 100$ diffusion timesteps. Using this fixed discretization, we generated a single independent lattice configuration by running the reverse diffusion process for N_s steps. For each $[T, \mu]$ point, we generated 100 independent lattice configurations and computed the average density across them, which is the value reported for each $[T, \mu]$ point. The average sampling time required to generate these 100 configurations for one $[T, \mu]$ point was approximately 20 seconds, resulting in a total sampling time of 12,316 seconds for all 625 $[T, \mu]$ combinations. Since the training cost is negligible compared to the total sampling cost, the total computation time (training + sampling) is approximately 12,499 seconds (about 3.4 hours). A summary of the computational costs is provided in Table 1.

Method	MCMC	expTM model
Training time	N/A	183 s (10 epochs)
Sampling time per $[T, \mu]$ point	79 s	20 s
Total sampling time (625 points)	49,597 s	12,316 s
Total wall-clock time	49,597 s (~ 13.7 h)	12,499 s (~ 3.4 h)

Table 1: Wall-clock time comparison between MCMC sampling and expTM training and sampling, performed on a single NVIDIA A100 GPU.

To further demonstrate the computational efficiency, we compared performance across different lattice sizes, as shown in Fig. 2. These results correspond to sampling a single $[T, \mu]$ point, where the reported values are averaged over 50 independent lattice configurations for MCMC and 100 configurations for the expTM model. The results show that the MCMC sampling time grows rapidly with lattice size L^2 , whereas both the training and sampling times of the expTM model remain comparatively insensitive to system size until the lattice size reaches $L^2 = 64^2$.

For small systems such as a $L^2 = 20^2$ size lattice at a single $[T, \mu]$ point, MCMC remains competitive. However, when sampling a large number of thermodynamic conditions, such as the 625 $[T, \mu]$ combinations considered here, our expTM model provides a substantial computational advantage. Once trained, the expTM model enables repeated sampling at new thermodynamic conditions with no additional training cost, yielding a $\sim 4\times$ reduction in wall-clock time at $L^2 = 20^2$, with an even greater advantage at larger lattice sizes. This has been added to the Supplementary Information (SI) Section III and the following paragraph is added in the main text:

In addition to accuracy, the expTM model provides a substantial computational advantage over MCMC simulations when sampling across multiple thermodynamic conditions. A detailed discussion

and wall-clock time benchmark comparison between MCMC simulations and the expTM model are presented in the SI Sec. III.

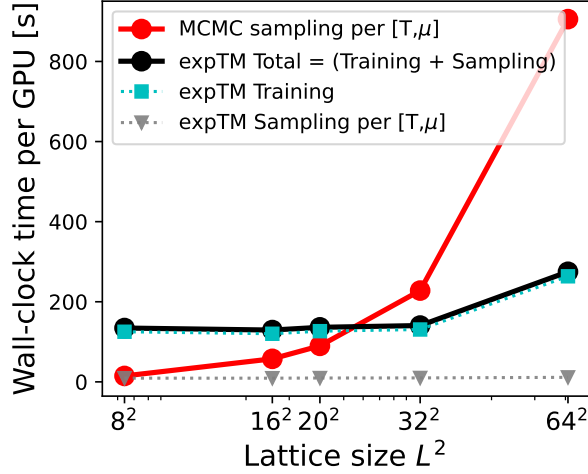


Figure 2: Wall-clock time per GPU required to generate samples using MCMC and diffusion-based (expTM) sampling as a function of lattice size L^2 . For the diffusion approach, training time, sampling time, and total cost (training + sampling) are shown separately.

Comment 1.5 — Overall, papers using deep networks are expected to perform ablation experiments to show that all elements of the proposed algorithm are really necessary. So the items above are examples of that.

Reply: We thank the reviewer for pointing this out. To demonstrate the effectiveness of our expTM model, we performed additional ablation experiments, in addition to the ones reported in response to previous comments, on the lattice gas system that removes thermodynamic conditioning. The following information and all figures below have been updated in the SI Sec. IV.

By construction, a key contribution of this work is the integration of an exponentially tilted prior by replacing the original unit normal Gaussian prior. To evaluate its role, we first revert the prior distribution to a unit normal Gaussian and examine the resulting generated samples in the absence of exponential tilting. In this setting, all training data are mapped to a single standard normal distribution, and the prior is no longer conditioned on temperature or chemical potential. Consequently, the model effectively learns an unconditional distribution over the feature space x (or lattice configurations in the lattice gas model), rather than the conditional distribution $P(x | T, \mu)$. To ensure that the observed behavior is not due to insufficient training, we trained models over a wide range of epochs, e.g., epochs = [5, 10, 25, 50, 100, 125, 150, 200, 250, 500].

As shown in Fig. 3 below, a model trained with a unit normal prior fails to reproduce accurate phase diagrams, even at extended training epochs. In contrast, as shown in Fig. 4, the exponentially tilted prior consistently yields accurate phase behaviors across all training epochs and random seeds.

To further validate our results, we quantified the density error between the ground-truth MCMC results and samples generated by expTM using either a unit normal prior or the exponentially tilted prior, measured by $\Delta\rho$. For each case, lattice configurations on a 25×25 grid (625 points) were

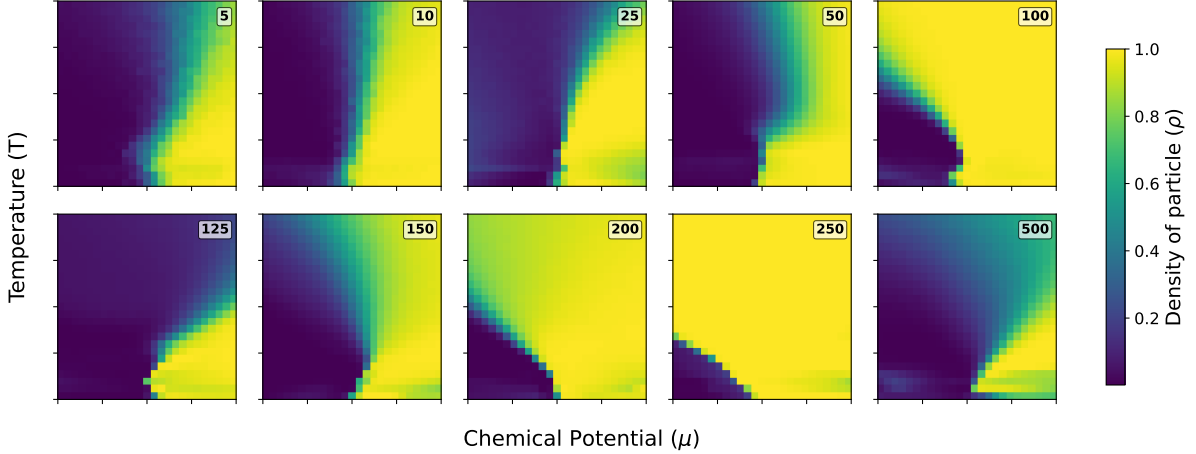


Figure 3: Lattice generation using a unit normal Gaussian prior distribution under different numbers of epochs.

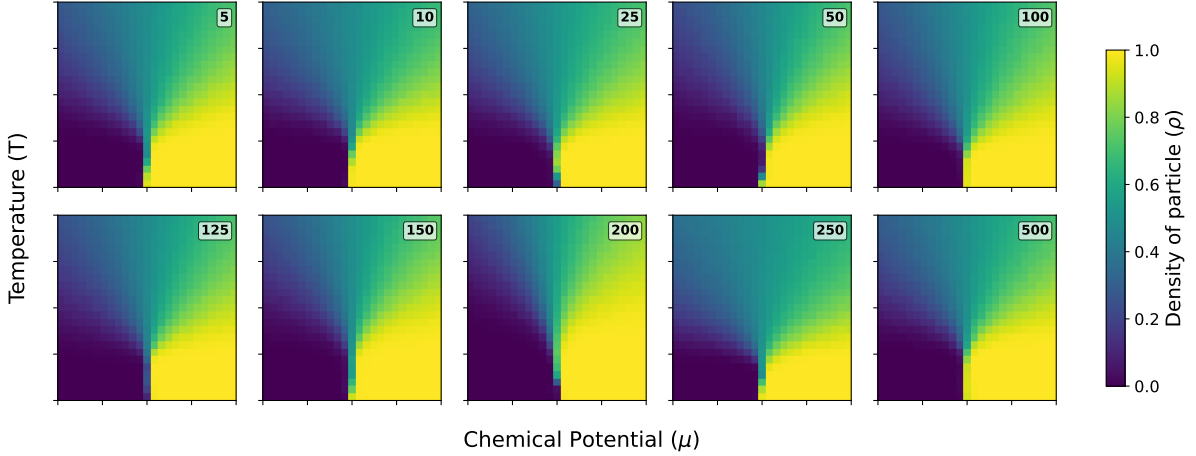


Figure 4: Lattice generation using exponentially tilted prior distribution under different numbers of epochs.

generated, and each model was retrained 10 times across all training epochs. We then computed the mean density error $\overline{\Delta\rho}$ and its variance. Overall, the results indicate that the exponentially tilted prior achieves the lowest error at approximately 25 training epochs. In contrast, models trained with a unit normal prior exhibit larger mean errors and substantially higher variance across random seeds, demonstrating a lack of robustness and reproducibility relative to the ground-truth MCMC results.

This ablation study demonstrates the necessity of exponential tilting of the prior in the diffusion model. Without the tilted prior, even when conditioning parameters are applied to a unit Gaussian distribution—as in the original diffusion—the model fails to reproduce the correct phase transition behavior. The exponentially tilted prior is therefore a key component that enables correct conditional generation, allowing the model to represent thermodynamic conditions.

For better comparison, we further computed the errors between the ground truth from the MCMC results, which is illustrated in Fig.5. Here, the error bars show the standard deviation across 10 training results.

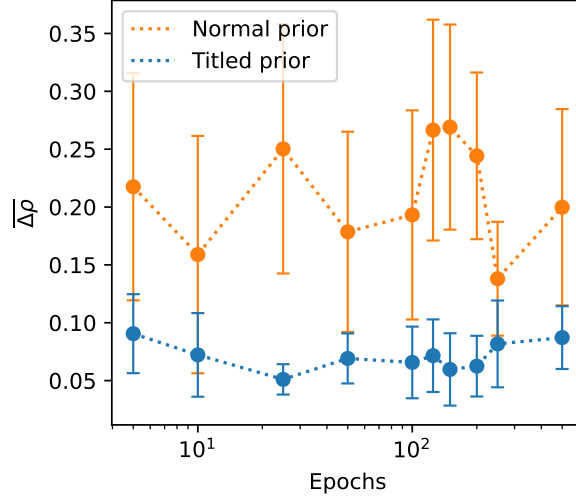


Figure 5: Mean density error $\overline{\Delta\rho}$ between generated samples and MCMC ground truth results as a function of training epochs, comparing models trained with a unit normal prior (orange) and an exponentially tilted prior (blue). Error bars indicate the standard deviation over 10 independent training runs.

runs at each epoch.

Comment 1.6 — In the supplemental information they provide an pseudo code algorithm but it actually confused me. In the paper there is a clear distinction between the temperature and pressure of the training sample and the gaussian population of temperatures and pressures. This somehow disappears in the pseudo code. The derivative of the density should be in terms of all the variables $\mathbf{x}, \beta, \alpha$. This need to be clarified.

Reply: We thank the reviewer for pointing out the ambiguity in our previous algorithm. The reviewer is correct; the confusion arose from (i) the distinction on the joint space over $(\mathbf{x}, \beta, \alpha)$ and (ii) the conditional notation used when introducing the tilted prior was written in a potentially misleading manner. We have revised the algorithm to clarify this distinction.

To clarify, the expTM model is trained on a dataset $\mathcal{D} = \{(\mathbf{x}_i, \beta_i, \alpha_i)\}_{i=1}^N$, which defines a joint distribution $p_{\text{data}}(\mathbf{x}, \beta, \alpha)$. During the forward process, Gaussian noise is applied to each component of $(\mathbf{x}, \beta, \alpha)$. During the reverse process, the neural network learns the score on a joint space defined as $s_{\theta}(\mathbf{x}, \beta, \alpha, \tau) \approx \nabla_{\mathbf{x}, \beta, \alpha} \log p_{\tau}(\mathbf{x}, \beta, \alpha)$. In other words, the model approximates the gradient of the log joint density rather than a conditional density in $\mathbf{x}$ alone.

When generating samples at the thermodynamic condition of our interest (β^*, α^*) , we evaluate the x -component of the learned joint score at fixed (β^*, α^*) , i.e., $\nabla_x \log p_t(x, \beta^*, \alpha^*)$, which is equivalent to sampling from $p(x | \beta^*, \alpha^*)$. To address these issues, we have revised Algorithm 1 in Section II of the SI.

Reviewer 2

The paper extends the thermodynamic maps framework to capture systems with an exponentially tilted prior. This leads to a second fluctuation variable related to the mean of the prior distribution. The fluctuation variables undergo diffusion in the forward process, and contribute to the score function. The justification for the diffusion on β is that it provides a temperature dependent correction to the score function. However, given that both α and β diffuse from their initial value to a normal distribution it is unclear what diffusing on α is adding to the model.

Reply: We thank the reviewer for pointing this out and acknowledge that the role of the β and α deserves a clear explanation and distinction. First, in the original TM model [2], β was introduced as a parameter that controls the width, i.e., variance, of the Gaussian prior distributions, which have a physical meaning as thermal fluctuations. The presence of the α , on the other hand, controls the mean of the Gaussian prior that entails the physical meaning of the shift of chemical potential or pressure due to the equilibrium changes. The role of the α is essential; otherwise, it cannot condition on different chemical potentials or pressures, so it loses the ability to generate samples at new thermodynamic conditions.

Also, to emphasize that our expTM model differs from the traditional conditional diffusion model, which is label-based on the training data and lacks physical interpretations. The significance of our model lies in the presence of (α, β) physical parameters that jointly diffuse with the feature space x , and learns the mapping between the prior and the input space (x, α, β) . Furthermore, note that the feature space x does not diffuse to a normal distribution; at the final time t in the diffusive time, it converges to the newly titled prior distribution $N(\alpha, \beta)$, not to the unit normal distribution. If we only diffused x and kept α and β in fixed states, the expTM model would not be able to determine which features of x correspond to the initial states of α_0 or β_0 and further would not be able to learn the joint score along the backward process. We have also discussed this in the context of ablation tests performed in response to reviewer 1, and added as extra explanations in Section IV in the SI.

To clarify these explanations, we add to the explanations Section II.D.3 as follows:

It is not necessary to go through essentially identical derivations for both of the considered ensembles (grand canonical and isothermal-isobaric). Similarly the sections on diffusion models, exponentially tilted diffusion models, TMs, and exponentially tilted TMs do not all need a full section of explanation and could be more consolidated. Also, there is a small typo in the text between Eqs 17 and 18 – “exponential titling factor”

Reply: We thank the reviewer for the careful reading and catching the typo. The typo has now been corrected. Regarding structural similarities in the derivations, we respectfully prefer to maintain the current structure with explicit derivations for both ensembles. We acknowledge that the derivations for the NPT and μ VT ensembles are mathematically very similar, with only a difference in the plus and minus signs. However, for readers without a background in thermodynamic ensembles, which is often the case for interdisciplinary audiences such as at Nature Communications, we believe and humbly argue that it is important to present both cases explicitly. While the mathematical forms are nearly identical, the physical interpretations — particularly how the linear approximations arise for different thermodynamic terms — warrant separate discussion. While we would like to emphasize these differences, respecting

the reviewer’s comments we nevertheless reduced the overall length of the discussion to improve clarity.

The thermodynamic justification for using this approach for the CO_2 phase transition is not clearly established. It is not clear why the collective variables extracted from the MD simulations should have an exponentially tilted prior, given the analysis steps that separate them from the isothermal-isobaric system. Further explanation could provide more insight into the range of systems to which this approach could be applied.

Reply: We thank the reviewer for the question regarding the choice of collective variables (CVs) we have used for expTM, under the isothermal-isobaric ensemble, which has not been clearly addressed in the main text. In our theoretical framework, x represents the feature space characterizing the system’s states and need not correspond to raw atomic coordinates. In the case of CO_2 , we use the order parameters λ_1 and λ_3 (as defined in the manuscript) as our features. These CVs were adopted by Gimondi et al. [1] in MD simulations to describe the local structural order of each molecule. Using the same selected CVs, we can extend the features extracted from the MD simulations under the thermodynamic conditions of interest and recover configurations. Furthermore, since CVs are correlated with the phases of CO_2 , and the phase is correlated with the free energy ($G=U+pV-TS$), which is why we introduce a tilting term to mimic the pressure effects. As a result, we believe that, though λ_1 and λ_3 are not directly calculated from pressure, the CVs contain information about pressure.

Using selected CVs rather than raw atomic coordinates provides several advantages. First, it enhances physical interpretability: λ_1 and λ_3 can be directly mapped to the target CO_2 phases. Furthermore, it improves computational efficiency by reducing the problem to a two-dimensional CV space, rather than operating in the full, high-dimensional atomic-coordinate space, which would be more complex.

More broadly, we believe that the isothermal-isobaric ensemble is particularly well suited to a CV-based expTM framework, especially for pressure-driven systems. As demonstrated here for CO_2 , crystal polymorphism provides a clear example with well-defined phase boundaries under different thermodynamic conditions.

To improve the clarity of the work, we have added the following sentence in the main text:

Instead of using the full atomic coordinates, we employ the CVs λ_1 and λ_3 to characterize the phase states, effectively reducing the description to a two-dimensional space.

Not only limited to the given applications, more broadly, we believe that the isothermal-isobaric ensemble is particularly well-suited to a CV-based expTM framework, especially for pressure-driven systems. The successful application of CO_2 crystal polymorphism demonstrates that the expTM model applies to systems with well-defined phase boundaries across different thermodynamic conditions.

While this work does extend the existing framework to a larger range of systems, the assumptions still require equilibrium systems with harmonic Hamiltonians (or those that can be collectively described as such). The ability of the framework to generate samples from unseen points remains impressive, but a comparison to a simpler diffusion model would make the impact of the thermodynamic approach clear. The idea of adding an offset to the prior mean is novel, but it’s significantly less of a step than the original thermodynamic maps framework, making the improvements described in the work largely incremental. This paper is not currently suited for this journal, but a rigorous analysis of what this approach learns above a simpler diffusion model and a stronger thermodynamic analysis of how the fluctuating variables guide the diffusion could make it suitable. I expect that doing this expanded analysis rigorously will take substantial effort but I am happy

to revisit my decision and re-evaluate the work if the authors do.

Reply: We thank the reviewer for recognizing the novelty of this work and for providing constructive feedback. We agree that a more rigorous comparison with simpler baseline models is necessary to demonstrate the impact of the additional framework, which is now added extensively. In response to Reviewer 1’s comment, we performed several ablation studies to rigorously evaluate the roles of α and β . (1) We removed the conditioning on α and β , training the model to approximate the marginal distribution of the feature states without thermodynamic information. (2) We removed the exponentially tilted prior and instead mapped the data to a unit Gaussian distribution, thereby eliminating the thermodynamic parameterization in the prior distribution.

In the first ablation test, as discussed also for Reviewer 1, we trained a diffusion model only on lattice configurations at two fixed thermodynamic conditions, without embedding thermodynamic variables and without modeling the joint space. This model failed to generate samples even at the same temperature as in the training data. We believe this is due to confusion arising from mixed-temperature training data. When we further introduced temperature labels, the model could only recover the original training temperatures, failed to extrapolate across conditions, and could not generate meaningful samples beyond the training points. This demonstrates that learning the joint score and the exponentially tilted prior mapping is essential.

In the second study, we removed the exponential tilting component mapped to the prior and instead mapped it to a unit Gaussian distribution, such that the prior is no longer parameterized by the thermodynamic conditions. In this setting, the model is analogous to the original diffusion model, in which input spaces are mapped to a unit Gaussian distribution without conditioning on specific thermodynamic parameters. The model can extrapolate at the thermodynamic conditions of interest during sampling, but the mapping no longer correctly encodes thermodynamic information. As a result, it loses the ability to preserve thermodynamic conditioning and fails to recover the phase boundaries, regardless of the number of training epochs. The results are shown in Fig. 3, which demonstrates that the recovered phase boundary is poorly defined and does not match the ground truth.

To further demonstrate the effectiveness of our expTM model, we performed additional wall-clock time comparison on the lattice gas model, which is now added in Section III of SI. The results have shown how the current approach was at least four times faster than running explicit Monte Carlo simulations, especially as the lattice size increases. We believe these analyses significantly complement the previous version of the manuscript bringing out its advances even better. To complement the following work, which has not been addressed in the previous version of the manuscript, in addition to the response to the detailed descriptions, we further added the following explanations in the main manuscript:

To assess the necessity of the exponentially tilted prior, we performed an ablation study in which the exponential tilting term was removed from the prior distribution and replaced with a unit Gaussian independent of the thermodynamic parameters. In this setup, the model reduces to a standard diffusion model without thermodynamic conditioning embedded in the prior. During sampling, one can still attempt to extrapolate to thermodynamic conditions of interest; however, the model fails to recover the phase boundaries. This demonstrates that exponential tilting of the prior distribution is essential for preserving thermodynamic consistency and enabling accurate phase transition prediction. The following results are demonstrated in SI Fig.S3 and Fig.S4.

References

- [1] Ilaria Gimondi and Matteo Salvalaglio. Co2 packing polymorphism under pressure: Mechanism and thermodynamics of the i-iii polymorphic transition. *The Journal of Chemical Physics*, 147(11):114502, 09 2017.
- [2] Lukas Herron, Kinjal Mondal, John S. Schneekloth, and Pratyush Tiwary. Inferring phase transitions and critical exponents from limited observations with thermodynamic maps. *Proceedings of the National Academy of Sciences*, 121(52):e2321971121, 2024.

Author Response:

We are pleased that the reviewers are satisfied with the revised manuscript and recommend it for publication. We sincerely appreciate their positive assessment.

Reviewer #3 (Remarks to the Author):

Thank you for clarifying the structural basis of the relationship between the collective variables and the thermodynamically relevant quantities. The connection to the order parameters adopted from Gimondi and Salvalaglio, and the point that the CVs track the CO₂ phases and thus the free energy, resolves my question about why the tilting term is appropriate here.

The added ablation studies are what most strengthen the manuscript. They establish that the tilted prior and joint score are necessary rather than incremental additions to the original framework. Removing all thermodynamic conditioning leaves the model unable to recover even the training conditions, showing that the conditioning is necessary rather than incidental. Replacing the tilted prior with a unit Gaussian then isolates the role of the thermodynamic quantities in tilting the prior, supported by the quantitative error analysis against MCMC. Together these establish what the expTM framework contributes. The clarified score function and corrected algorithm resolve the earlier ambiguity about how the feature space couples to the thermodynamic variables, and the wall-clock comparison against MCMC shows a concrete improvement in efficiency.

On balance, the revisions clarify the method and show that it is a significant enough expansion of the existing method to make it suitable for this journal. I recommend acceptance.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.