Physics-based statistical learning approach to mesoscopic model selection

Søren Taverniers,[1] Terry S. Haut,[2] Kipton Barros,[3] Francis J. Alexander,[2] and Turab Lookman[3]
[1]*Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA*
[2]*Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
(Received 18 May 2015; published 9 November 2015)

In materials science and many other research areas, models are frequently inferred without considering their generalization to unseen data. We apply statistical learning using cross-validation to obtain an optimally predictive coarse-grained description of a two-dimensional kinetic nearest-neighbor Ising model with Glauber dynamics (GD) based on the stochastic Ginzburg-Landau equation (sGLE). The latter is learned from GD "training" data using a log-likelihood analysis, and its predictive ability for various complexities of the model is tested on GD "test" data independent of the data used to train the model on. Using two different error metrics, we perform a detailed analysis of the error between magnetization time trajectories simulated using the learned sGLE coarse-grained description and those obtained using the GD model. We show that both for equilibrium and out-of-equilibrium GD training trajectories, the standard phenomenological description using a quartic free energy does not always yield the most predictive coarse-grained model. Moreover, increasing the amount of training data can shift the optimal model complexity to higher values. Our results are promising in that they pave the way for the use of statistical learning as a general tool for materials modeling and discovery.

DOI: 10.1103/PhysRevE.92.053301 PACS number(s): 05.10.−a, 05.20.−y, 02.70.−c, 05.45.Tp

I. INTRODUCTION

Due to limitations in computational resources, the behavior of complex systems (e.g., climate, turbulent flow, materials under shock loading) often needs to be modeled using a coarse-grained description that captures the phenomenon of interest. Coarse-grained models cannot be perfect, of course, since many microscopic degrees of freedom are absent. The Mori-Zwanzig formalism [1,2] tells us that the relevant coarse-grained description should contain both noise [3,4] and memory kernels to represent the "integrated out" fine scale dynamics. Deriving an appropriate coarse graining analytically is therefore extremely difficult. Statistical learning provides a tractable way of finding a coarse-grained description that is able to predict the results of new experiments or simulations beyond those used in the model construction, which serves as a true objective test of the model. In fact, unlike traditional approaches, statistical learning can also serve as a coarse-graining strategy in cases where there is no clear separation of spatial and/or temporal scales. For example, it may be applied to problems that involve inhomogeneous flows (e.g., multicomponent fluids, complex fluids) and those in materials science where the coarse-grained description needs to account for inhomogeneities at a finer scale, such as microstructural defects. Techniques such as the heterogeneous multiscale method (HMM) [5] that demand a clean separation in time scales are not applicable.

In the statistical approach to coarse graining discussed here, the goal is to search over a certain class of coarse-grained models and find the complexity for which the description is "optimally predictive." This technique is called regularization, and to estimate the generalization error we use cross-validation. The latter involves randomly dividing data (either from experiments or simulations) into "training" and "test" samples. Ideally one would like both groups of data to be infinite, but in practice one only has a limited amount of data to work with. Experimentalists can only synthesize a small number of material samples, and in molecular dynamics simulations one is also limited to a finite number of samples. For the purposes of the current analysis, we will assume that the amount of training data is limited but that we can test our learned model on an infinite amount of data independent of the training samples. An extension to cases where both training *and* test data are finite will be the topic of future research.

Selecting the appropriate model regularization is of paramount importance because it can minimize both underfitting and overfitting. An underfitted model is too simplistic and therefore fails to capture much of the useful information available in the training data; hence, it will perform suboptimally on data independent of the training set. In contrast, overfitting refers to the case where an overly complex model describes the many irrelevant details that appear in the training data by chance. An overfitted model will therefore be also less successful in generalizing to new data from simulations or experiments that are outside the class of the training data. The model developed in this study avoids common issues associated with overfitting by using an effectively infinite amount of test data independent of the samples on which it was trained, and selecting the complexity that makes it most predictive of this test data.

As an error estimator, cross-validation has been used for a number of years. When the amount of data is very limited though, there can be a significant difference between the cross-validation error and the actual error [6,7]. Moreover, a detailed error analysis is often lacking in physics modeling applications. Our motivation is to learn mesoscale models from microstructural data incorporating prior domain knowledge and physical symmetries. We will focus on the time-dependent stochastic Ginzburg-Landau equation (sGLE) which provides a coarse-grained description of a kinetic nearest-neighbor Ising model with Glauber dynamics (GD) and of which the dynamics are expected to be particularly straightforward to

learn. While our approach is related to that in [8], we do not assume a prior distribution for the learned parameters and do not include a penalty for overfitting or complexity in the Bayesian information criterion. Moreover, instead of simply fitting a regular quartic free energy to a single or joint magnetization distribution function as in [9,10], we consider higher order terms and find the parameters that optimally predict GD data independent of the samples on which the model was trained. The inclusion of terms beyond fourth order in the free energy accounts for the fact that we are in a regime of finite coarse-graining block sizes, and hence not at a fixed point in the renormalization group theory [11,12]. Our current approach does not account for higher order spatial gradients, which can play an important role out of equilibrium; we plan to include these terms in future versions of the model.

Section II describes the microscopic GD model and its mesoscale description provided by the sGLE. Section III details our design loop used to select the optimal complexity of the sGLE for each amount of training data considered, after which Sec. IV discusses the results of the error analysis we performed in order to arrive at an optimally predictive model. Section V summarizes our conclusions and discusses possibilities for future work.

II. THE KINETIC ISING MODEL AND ITS COARSE-SCALE DESCRIPTION BY A STOCHASTIC GINZBURG-LANDAU EQUATION

A. Kinetic Ising model with Glauber dynamics

The Ising model with nearest-neighbor interactions [13] is a simple, yet very rich, model in statistical mechanics for describing ferromagnetic behavior. Consider a two-dimensional (2D) ferromagnet with atoms arranged on an $N_1 \times N_2$ square lattice. The spin $s_{i,j}$ (where $i = 0,\ldots,N_1-1$ and $j = 0,\ldots,N_2-1$) of each atom can be in one of two states, $s_{i,j} = \pm 1$, and can only interact with its four adjacent spins. We can add dynamics to this system by flipping spins with a certain transition rate w, and the result is a kinetic nearest-neighbor Ising model with spin-flip (Glauber) [14] dynamics (which we will refer to as GD). This allows us to express the time evolution of the spin system through a master equation given by

$$\frac{d}{dt}\mathbb{P}(\sigma;t) = \sum_{\sigma'}[w(\sigma' \to \sigma)\mathbb{P}(\sigma';t) - w(\sigma \to \sigma')\mathbb{P}(\sigma;t)], \tag{1}$$

where $\mathbb{P}(\sigma;t)$ is the joint probability of finding the system in spin configuration σ at time t, and the w's are the transition rates between two $N_1 \times N_2$-spin configurations differing only in the value of one spin, $s_{i,j}$. For w we choose the heat bath rate, given by

$$w_{\mathrm{HB}}(\sigma \to \sigma') = \kappa(1 + e^{-\beta[\mathcal{H}(\sigma)-\mathcal{H}(\sigma')]})^{-1}. \tag{2}$$

Here $\mathcal{H}(\sigma)$ represents the Hamiltonian of the spin system with configuration σ, $\beta = 1/(k_B T)$ with T the system temperature, and κ^{-1} sets the time scale of the spin-flip process and can depend both on T and the spins other than $s_{i,j}$. We simulate this kinetic Ising model via a Monte Carlo (MC) algorithm with one MC step per spin; i.e., to complete one step in our Ising run, we perform $N_1 \times N_2$ times the following procedure:

(1) Pick a random site (i,j) where $i = 0,\ldots,N_1-1$ and $j = 0,\ldots,N_2-1$.

(2) Draw a number r_1 from a uniform distribution on $[0,1]$.

(3) Flip the spin $s_{i,j}$ if $r_1 < w(\sigma \to \sigma')$, or leave it in its original state if $r_1 \geqslant w(\sigma \to \sigma')$.

In all of our work here, we initialize the lattice by selecting spins to be +1 or −1 randomly with equal probability.

B. Ginzburg-Landau equation

By invoking a phenomenological coarse-graining approach, it is possible to obtain a mesoscopic model of GD given by a time-dependent Ginzburg-Landau equation (GLE) [15]. The latter will describe the spatiotemporal evolution of an "order parameter" ϕ, a field variable which represents the instantaneous average of Ising spin values over some portion of a ferromagnetic material (also called "magnetization"). At finite temperatures, one needs to account for fluctuations, which can be added via a white noise term to obtain an overdamped stochastic relaxation equation,

$$\frac{\partial\phi(x,t)}{\partial t} = -M\frac{\delta F[\phi(x,t)]}{\delta\phi(x,t)} + \eta(x,t), \tag{3}$$

with M the mobility which sets the time scale of the dynamics. Here $\eta(x,t)$ is a zero-mean Gaussian white noise with a variance which, according to the fluctuation-dissipation theorem, scales linearly with M and the system temperature. Moreover, $F[\phi]$ is an effective free energy for the system, and is usually developed as a power series in ϕ and its derivatives,

$$F \sim \sum_k a_k\phi^k + b\nabla^2\phi + c\nabla^4\phi + d\,\phi\nabla^2\phi + \ldots. \tag{4}$$

In the context of late-stage domain growth, renormalization group arguments indicate that the Ising model with Glauber dynamics is in the universality class of Model A dynamics [12], and hence well represented at the coarse scale by having only a_2 and a_4 different from zero. However, since we are focusing on intermediate time and length scales, we relax this assumption and instead consider a model class for the free energy consisting of even-term polynomials with degree two or greater. The complexity that we eventually select is the one for which the model is optimally predictive of GD data independent of the samples from which it was learned.

III. NUMERICAL ALGORITHM

Our goal is to learn the parameters of a discrete version of the stochastic Ginzburg-Landau equation (sGLE) which evolves the magnetization ϕ from time t_n to time t_{n+1} according to

$$\begin{aligned}\phi_{n+1,i,j} = \phi_{n,i,j} &+ \alpha_0(\phi_{n,i+1,j} + \phi_{n,i-1,j} + \phi_{n,i,j+1} \\ &+ \phi_{n,i,j-1} - 4\phi_{n,i,j}) + \sum_{k=0}^{\frac{C-1}{2}}\alpha_{k+1}\phi_{n,i,j}^{2k+1} \\ &+ \alpha_{(C+3)/2}\,\xi_{n,i,j},\end{aligned} \tag{5}$$

where C is the model complexity [16] and the $\xi_{n,i,j}$ are independent, identically distributed standard normal random variables. Here n and $n+1$ refer to times t_n and t_{n+1}, while $i = 0, \ldots, \bar{N}_1 - 1$ and $j = 0, \ldots, \bar{N}_2 - 1$ indicate the spatial position of the spin blocks resulting from the coarse-graining procedure, with $\bar{N}_1$ and $\bar{N}_2$ the number of blocks in both spatial directions. Given block-averaged training data $S_{n,i,j}$ with $n = 0, \ldots, n_{\mathrm{eq}}$, where n_{eq} will be specified for each of our numerical experiments in Sec. IV, we would like to find the set of parameters $\boldsymbol{\alpha}_{\mathbf{opt}} \equiv \{\alpha_0, \alpha_1, \ldots, \alpha_{(C+3)/2}\}$ that maximizes the likelihood of observing this data using the sGLE model. For notational convenience, the $\bar{N}_1 \times \bar{N}_2$ block-averaged spin configuration after n steps will be denoted by S_n; the notation $S_{n,i,j}$ will refer to its (i,j)th matrix element. The same convention will be used for ϕ. It turns out $\boldsymbol{\alpha}_{\mathbf{opt}}$ is the solution to a linear system,

$$A\boldsymbol{\alpha}_{\mathbf{opt}} = \mathbf{b}, \tag{6}$$

where the components of A and the elements of $\mathbf{b}$ involve products of S_n, its powers, and its discrete Laplacian. The dimension of this system is given by the number of free parameters that make up the model (5), which is equal to $(C-1)/2 + 3$. For more details, we refer the reader to Appendix A where we have worked out the case of $C = 3$. We then test our learned sGLE model against independent GD data ("test" data) to ascertain how well learned models of different complexities perform on unseen data. By calculating the root-mean-square (RMS) error between GD test trajectories and those simulated using the learned sGLE, we then determine for which complexity C the latter is optimally predictive of the GD test data.

We will consider two error metrics in our analysis, which we will refer to as the "type 1" and "type 2" error. For the type 1 error, we calculate the sGLE grid at time t_{n+1} through (5) but replace ϕ_n with the block-averaged GD test data at time t_n. This error is of the same type as the error that we want to minimize when calculating $\boldsymbol{\alpha}_{\mathbf{opt}}$ from the GD training data, where we search for the set of parameters that maximizes the likelihood of observing the training data at t_{n+1} given the sGLE model and the training data at t_n, for every n (see Appendix A). For the type 2 error, we evolve the sGLE grid in time through (5) directly, i.e., we do not keep referring back to the GD test data at each time t_n.

A flowchart of the operational algorithm is shown in Fig. 1.

At a high level, our approach for computing one data point in the error probability density function (pdf) for a learned sGLE model of complexity C, given a number of training samples N_{train}, can be described as follows (see Appendix B for more details).

(1) We simulate N_{test} independent GD test sample trajectories. For each trajectory, we let the Ising system evolve over n_{mc} steps, after which we take another n_{eq} steps during which we record the block-averaged Ising configuration. Each of these steps represents one MC step per spin as detailed in Sec. II A.

(2) We simulate N_{train} independent GD training sample trajectories. We let the spins evolve over n_{mc} steps, and then record their block-averaged configuration over the next n_{eq} steps.

(3) Using the data gathered during the last n_{eq} steps of each training trajectory, we compute the coefficients of the learned sGLE polynomial using a log-likelihood analysis (see Appendix A).

(4) With the parameters calculated in step 3, we now simulate N_{test} sGLE trajectories. Each trajectory consists of n_{eq} steps, with each step involving the advancement of the discrete sGLE (5) from one discrete point in time to the next.

(5) For the kth sGLE trajectory, we calculate the RMS error ε_k between this trajectory and the kth block-averaged GD test trajectory.

(6) Finally, we compute the test-averaged error,

$$\varepsilon = \frac{1}{N_{\mathrm{test}}} \sum_{k=1}^{N_{\mathrm{test}}} \varepsilon_k, \tag{7}$$

which we will call the "type 1" or "type 2" test error depending on how the sGLE trajectory has been calculated (see our above definition of these errors).

The quantity ε represents one point in the error pdf for the considered complexity C and number of training samples N_{train}. The entire pdf is then obtained by repeating the above procedure except for step 1 (we use the same GD test trajectories for each point in the pdf) $n_{\mathrm{real}} \gg 1$ number of times. We will denote the sample (i.e., realization) mean and variance of this pdf by $\bar{\varepsilon}$ and s_ε^2, respectively.

IV. ERROR ANALYSIS AND MAIN RESULTS

We now present the pdfs of the type 1 and type 2 errors, defined in Sec. III, for different complexities $C = 3,5,7$ or 9, given a finite number of GD training samples N_{train}. We will use $N_{\mathrm{test}} = 1000$ GD test samples, which provides an accurate generalization error [17]. Moreover, we will build up the error histograms using $n_{\mathrm{real}} = 5000$ independent realizations, and consider both training data in equilibrium and out of equilibrium. For the equilibrium case, we measure the energy of the spin system and choose n_{mc} as the number of steps after which thermal equilibrium has been reached. Out of equilibrium, we choose n_{mc} such that after n_{mc} steps domains have started to form. Learning the sGLE model parameters from GD data for which the block size is smaller than the size of a typical domain allows the Laplacian in the sGLE to better capture gradients in the GD data, and should hence yield a more accurate coarse-grained description. For both cases, we determine an appropriate value for n_{eq} through trial and error, and choose values that provide a sufficient amount of input data to our log-likelihood solver. We will use $n_{\mathrm{mc}} + 1 = 2000$ and $n_{\mathrm{eq}} + 1 = 100$ for the equilibrium case, and $n_{\mathrm{mc}} + 1 = n_{\mathrm{eq}} + 1 = 500$ out of equilibrium (these values include the number of steps plus the starting condition at $t = 0$ or after n_{mc} steps). In all cases, we simulate the Glauber dynamics on a 256×256 spin lattice with periodic boundary conditions, and coarse grain using 16×16 blocks. We set the parameter κ in (2) equal to 1. Finally, next to each error distribution, we show a plot of the free energy constructed using the learned model parameters. These plots are meant to serve as a check of the physical soundness of our approach in the sense that, consistent with common knowledge, a single-well potential should be observed above the phase transition and a double-well potential below the phase transition. However, they do not convey any information

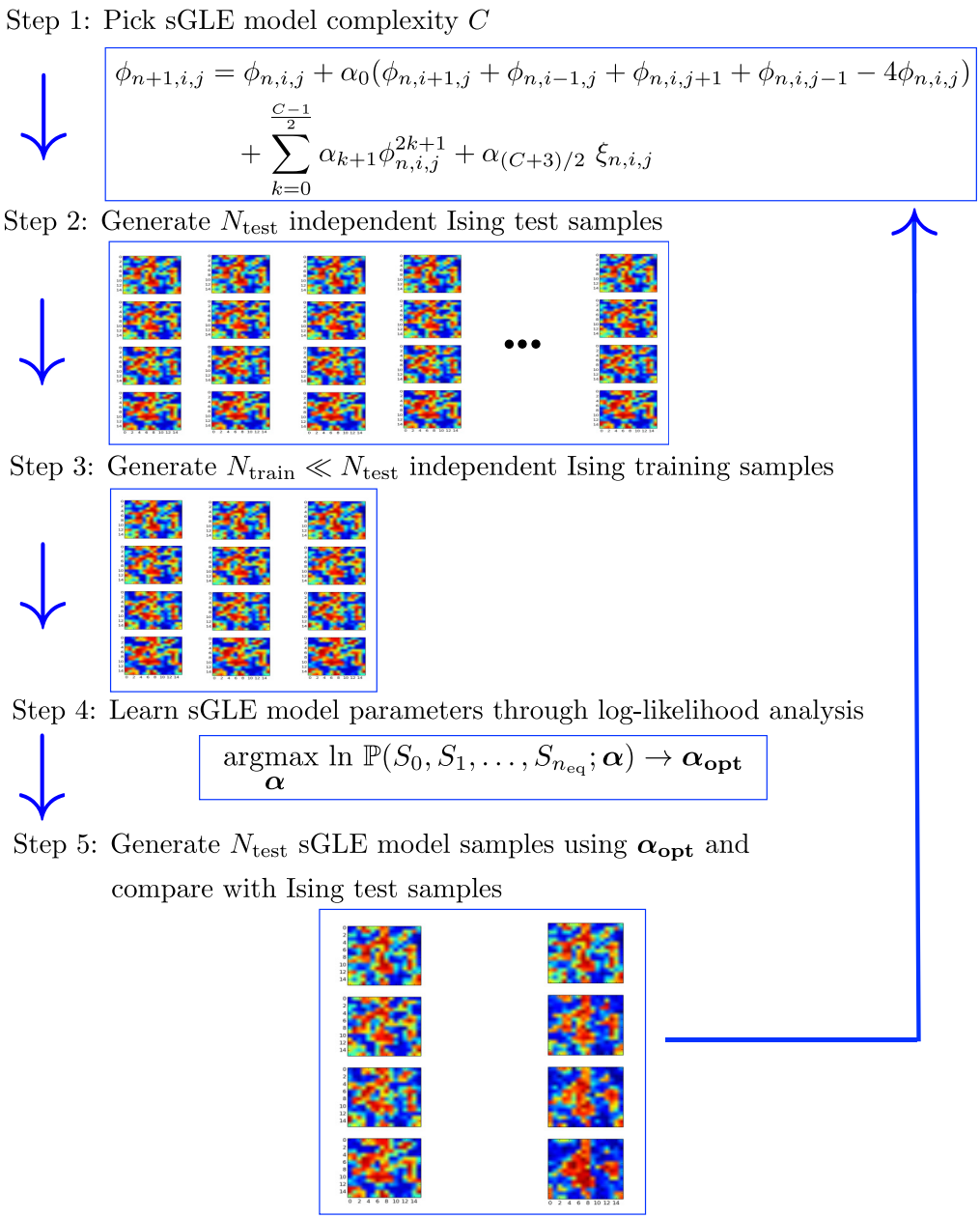

FIG. 1. (Color online) Flowchart for the model complexity selection algorithm.

regarding the predictiveness of our model, which follows entirely from the error distributions.

A. Type 1 test error pdfs for training data in equilibrium

We first consider the case where we train on $N_{\text{train}} = 8$ GD trajectories in equilibrium, obtained by quenching the spin lattice from infinite temperature to either $T = 1.6\,J/k_B$ (below $T_c = 2.269\,J/k_B$ [18]) or $T = 2.8\,J/k_B$ (above T_c). As Fig. 2 shows, at $T = 1.6\,J/k_B$ the error pdf's sample mean $\bar{\varepsilon}$ clearly decreases with complexity C. Hence, the most predictive sGLE model is that with the highest complexity considered, $C = 9$.

At $T = 2.8\ J/k_B$, however, the error pdfs for all the complexities overlap almost completely (see Fig. 3), indicating that the regular third-order sGLE polynomial is adequate to predict the coarse-grained Glauber dynamics.

B. Type 1 test error pdfs for training data out of equilibrium

Next, we look at the case where we train on various amounts of GD trajectories out of equilibrium, obtained by quenching the spin lattice from infinite temperature to $T = 2.2\,J/k_B$ (just below T_c).

As Figs. 4–7 show, regardless of the amount of training data the regular ϕ^4 form of the free energy ($C = 3$) is not optimally predictive of the GD test data. The complexity for which the sGLE model best predicts the GD test trajectories varies with the amount of training data. For small amounts of training data (i.e., $N_{\text{train}} = 1$ or 2), the histograms for the different complexities largely overlap. As the number of training samples is increased to 16, the pdfs for $C = 5$ and higher can be more clearly distinguished from that for $C = 3$. When one further increases the amount of training samples to 128, the

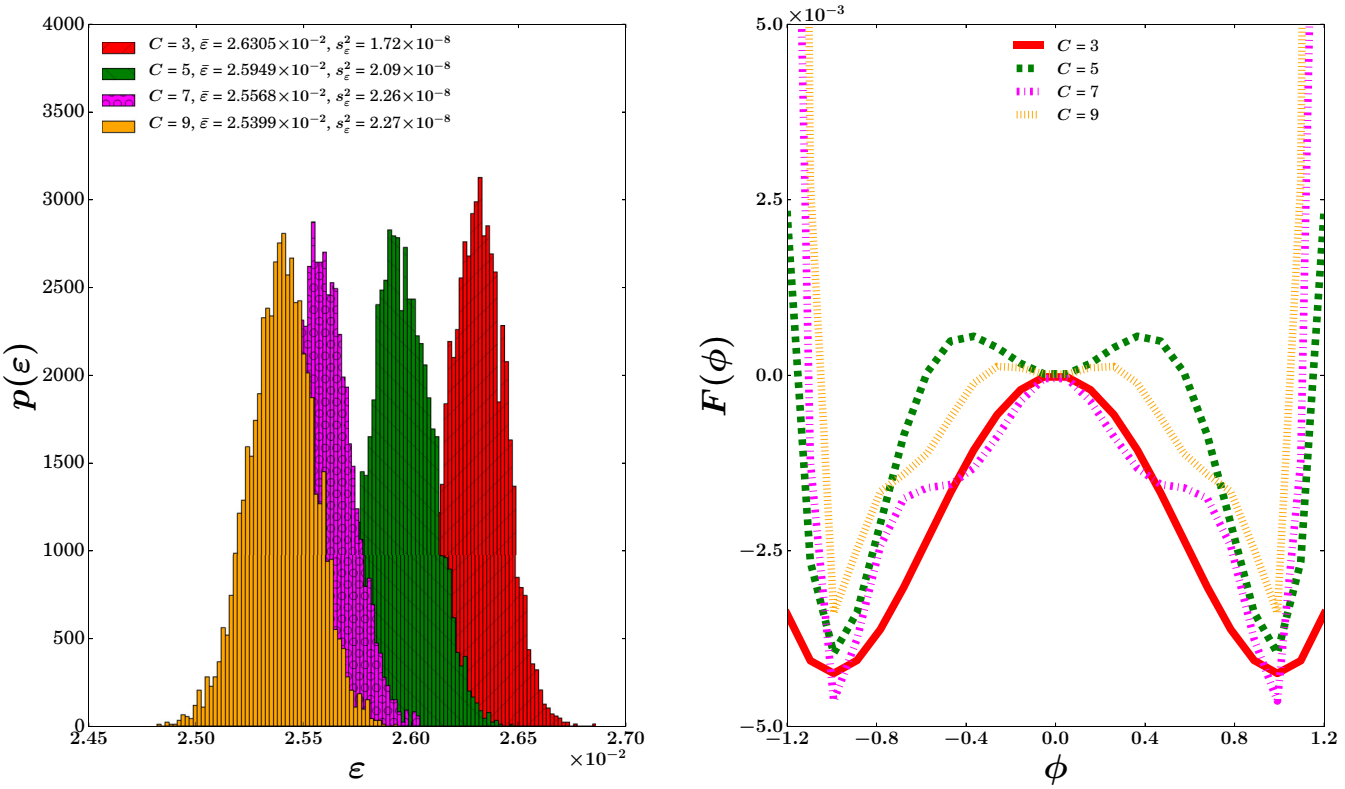

FIG. 2. (Color online) Type 1 error pdfs (left) and learned Ginzburg-Landau free energy (right) for different complexities of an sGLE model learned from eight equilibrium GD training samples at $T = 1.6\ J/k_B$ (below the phase transition). The sGLE with $C = 9$ predicts the GD test data best.

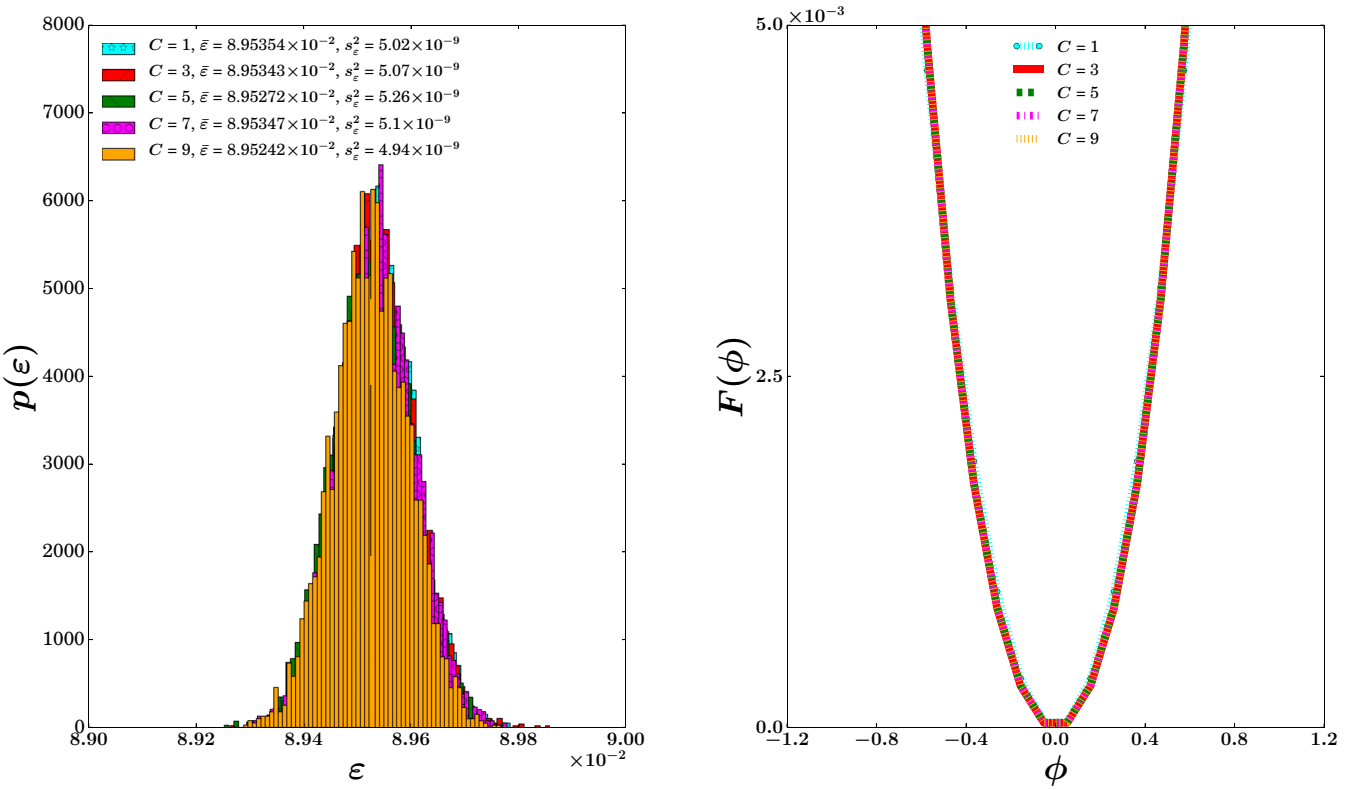

FIG. 3. (Color online) Type 1 error pdfs for different complexities of an sGLE model learned from eight equilibrium GD training samples at $T = 2.8\ J/k_B$ (above the phase transition). All considered model complexities are equally predictive of the GD test data.

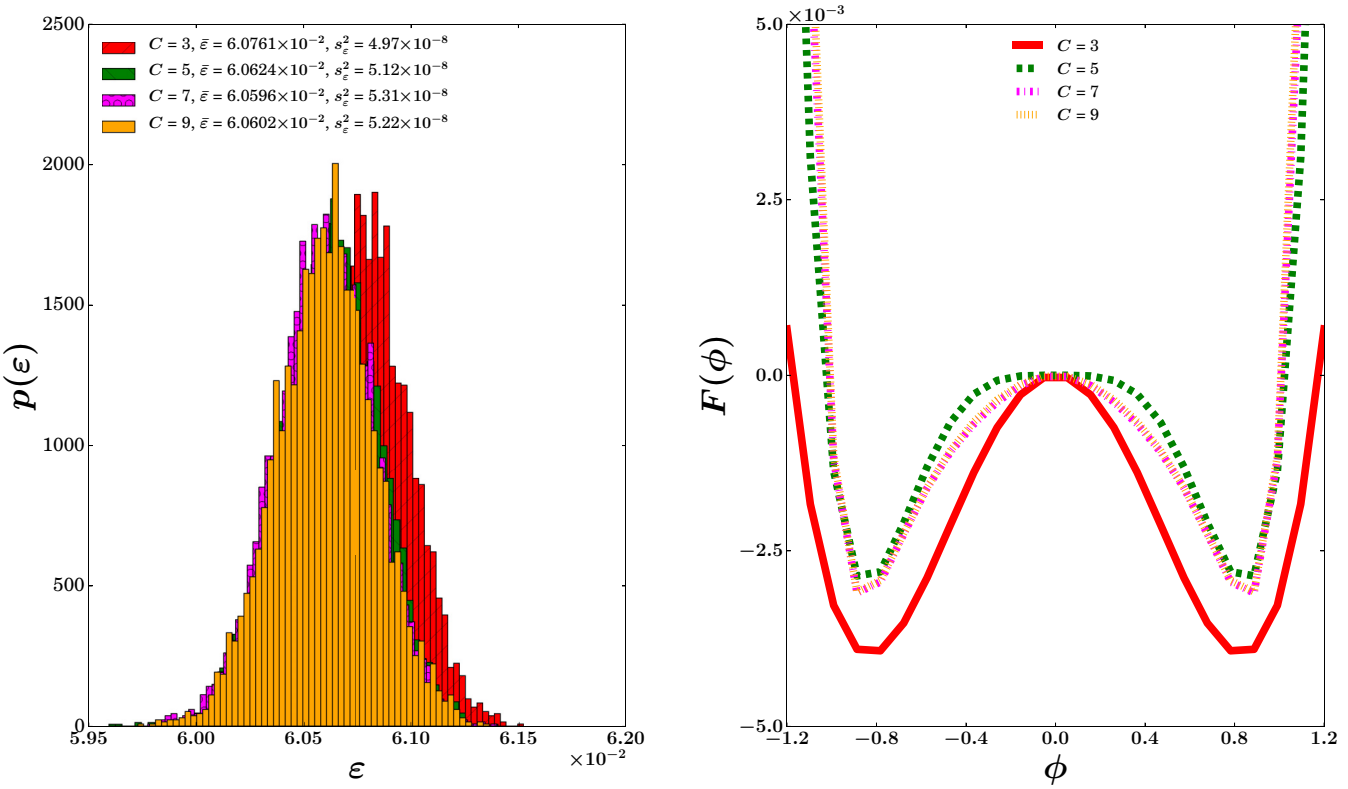

FIG. 4. (Color online) Type 1 error pdfs for different complexities of an sGLE model learned from one out-of-equilibrium GD training sample at $T = 2.2\ J/k_B$ (just below the phase transition). Complexities 5 and higher are optimally predictive of the GD test data, but the corresponding error pdfs still largely overlap with that for $C = 3$.

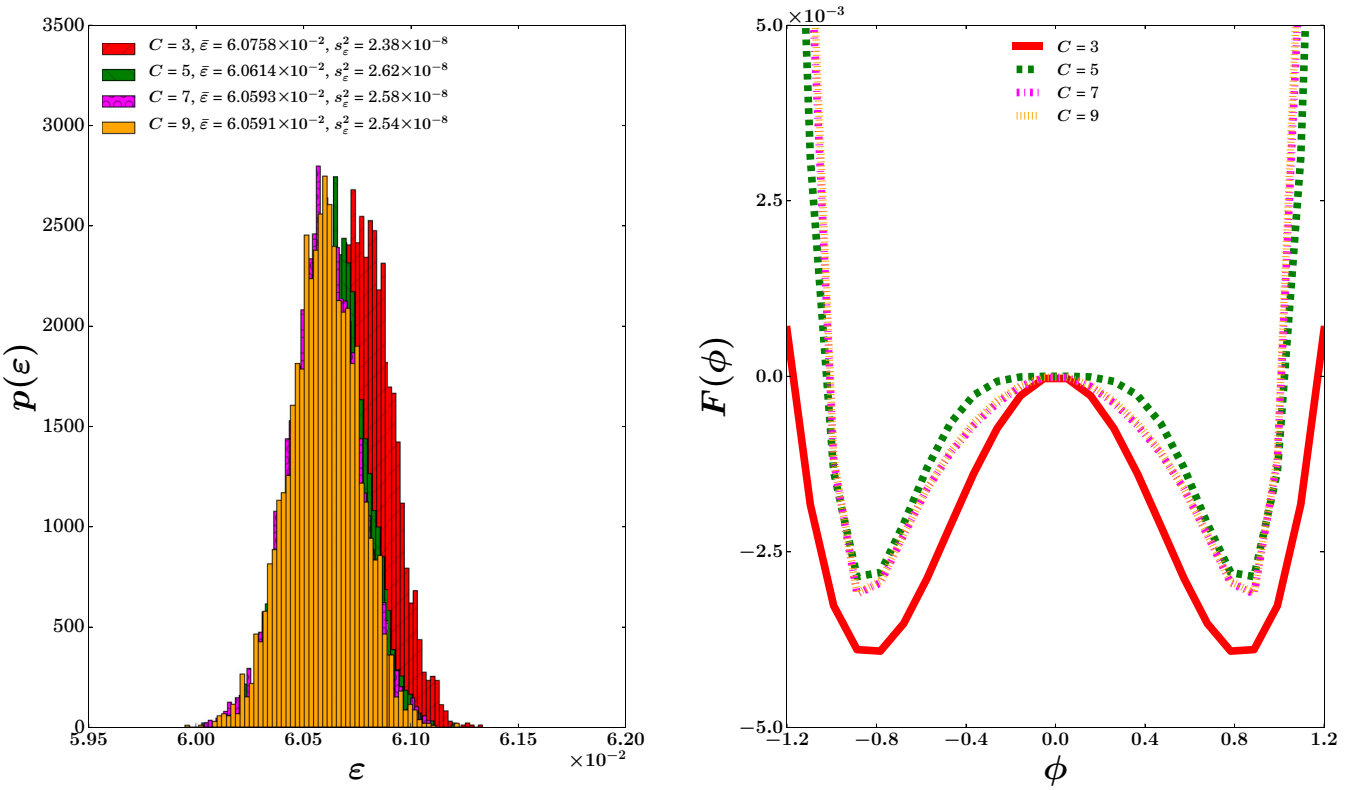

FIG. 5. (Color online) Type 1 error pdfs for different complexities of an sGLE model learned from two out-of-equilibrium GD training samples at $T = 2.2\ J/k_B$.

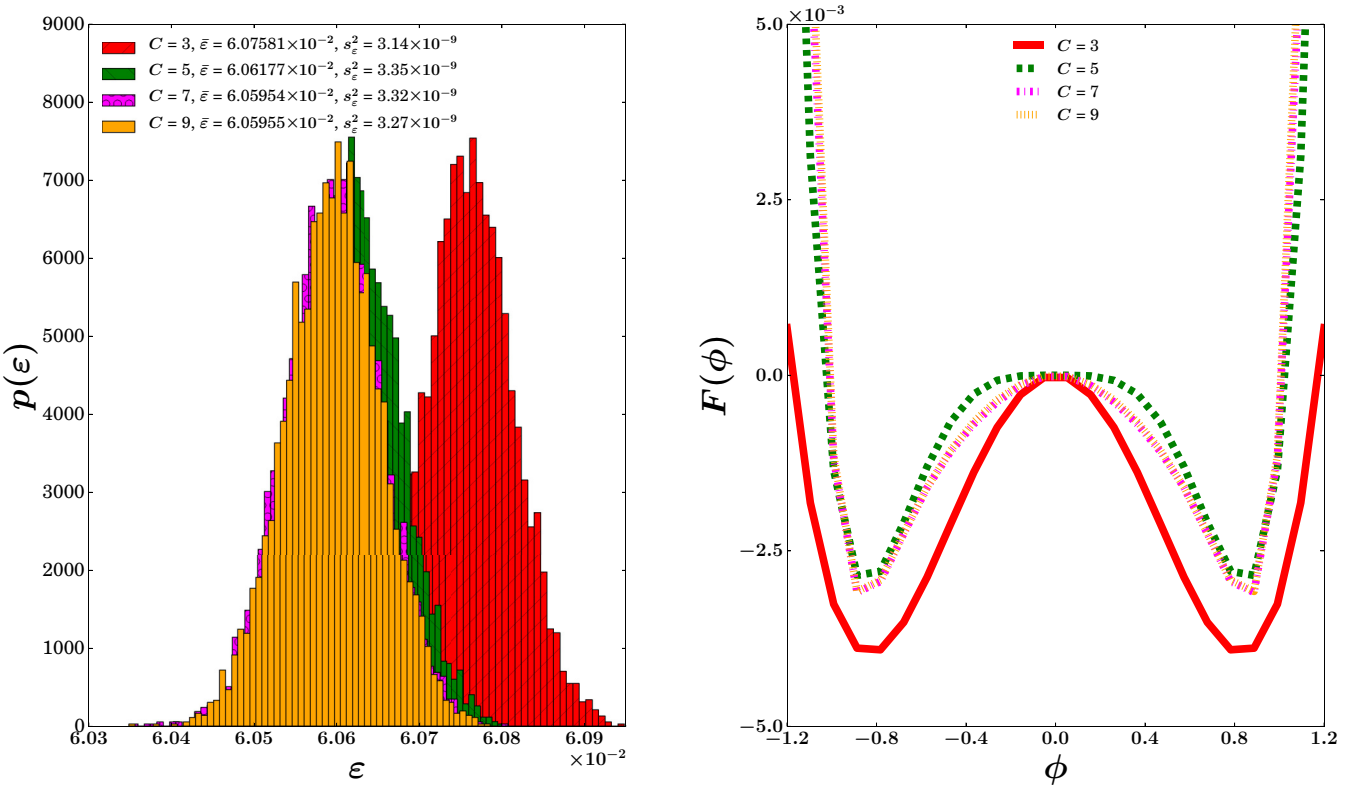

FIG. 6. (Color online) Type 1 error pdfs for different complexities of an sGLE model learned from 16 out-of-equilibrium GD training samples at $T = 2.2\ J/k_B$. The error pdfs for $C = 5$ and higher are now clearly distinct from that for $C = 3$.

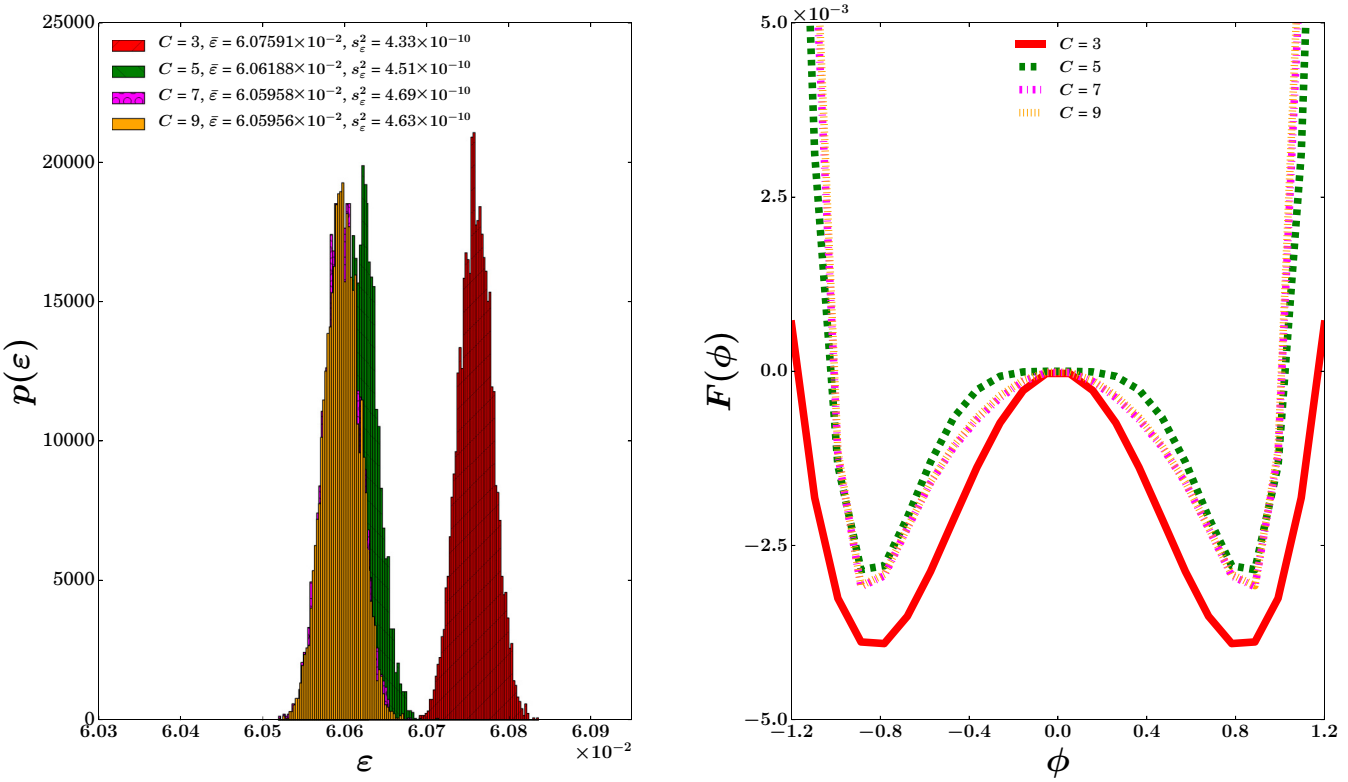

FIG. 7. (Color online) Type 1 error pdfs for different complexities of an sGLE model learned from 128 out-of-equilibrium GD training samples at $T = 2.2\ J/k_B$. The error pdf for $C = 5$ and those for $C = 7$ and higher are becoming more distinct, shifting the optimally predictive model complexity to $C = 7$.

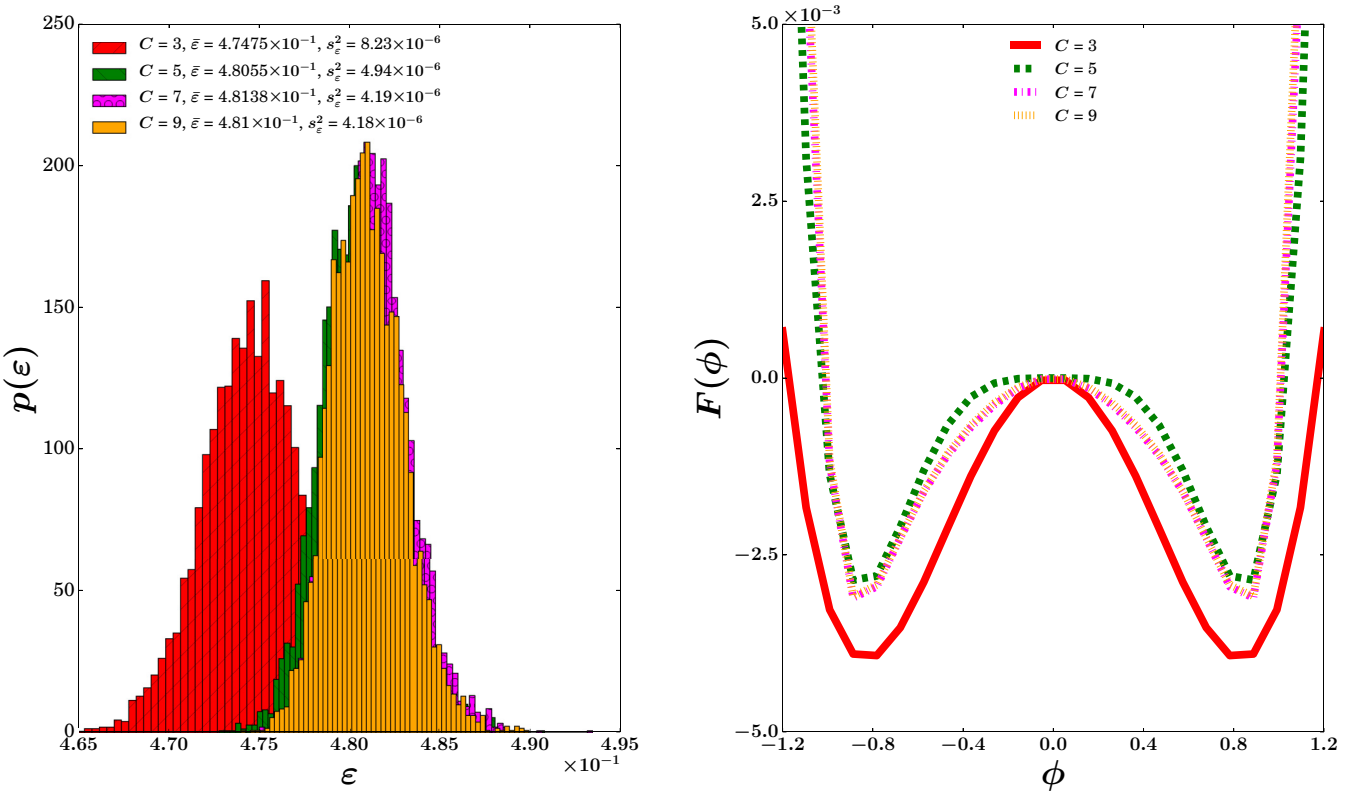

FIG. 8. (Color online) Type 2 error pdfs for different complexities of an sGLE model learned from one out-of-equilibrium GD training sample at $T = 2.2\ J/k_B$. The sGLE with $C = 3$ is most predictive of the GD test data.

pdf for $C = 3$ becomes fully distinct from those for higher complexities, and the pdf for $C = 5$ is becoming more distinct from those for $C = 7$ and $C = 9$. In sum, increasing the amount of GD training data causes the complexity at which the sGLE is optimally predictive to shift toward higher values.

C. Type 2 test error pdfs for training data out of equilibrium

Finally, we repeat the simulations in Sec. IV B for the type 2 error. We find that for this error type the regular ϕ^4 free energy *does* optimally predict the GD test data, regardless of the amount of training data. Figure 8 shows this for the case of one GD training sample. We note here that it is to be expected that $\bar{\varepsilon}$ is bigger for the type 2 error than for the type 1 error, since the latter is calculated in the same way as the optimization error for obtaining the $\boldsymbol{\alpha}_{\mathbf{opt}}$ is calculated from the GD training data, while the former is not.

V. SUMMARY AND CONCLUSIONS

By performing a detailed error analysis in the context of a statistical learning approach using cross-validation, we derive an optimally predictive coarse-grained description of a two-dimensional kinetic nearest-neighbor Ising model with Glauber dynamics (GD) based on the stochastic Ginzburg-Landau equation (sGLE). The latter is learned from microscopic GD "training" data through a log-likelihood analysis, and its capacity to predict GD "test" data independent of the training data is analyzed for various model complexities using two error metrics and varying amounts of training data.

Our analysis yields the following major conclusions:

(1) For the type 1 error, a complexity of 3 in the sGLE force equation does *not* yield an optimally predictive model for any amount of training data that we investigated. Moreover, the model complexity yielding the most predictive coarse-grained description increases with the amount of GD training data.

(2) For the type 2 error, the regular Ginzburg-Landau description using a ϕ^4 mean-field free energy *does* yield the most predictive model irrespective of the amount of GD training data.

The principled methodology developed here for simple model A dynamics can be applied to more complicated problems such as model H dynamics [12]. A particular application which might benefit from this work is the use of data generated from experiments, e.g., ultrafast x-ray diffraction patterns of structural phase transitions in semiconductor crystals to generate models describing crystal disordering [19]. In general, our approach can be utilized in any application using a Ginzburg-Landau functional, e.g., in phase field simulations of materials.

Directions for future work include studying the effects of the coarse-graining block size (in both space and time), performing a rigorous analysis of the stability and discretization error of our numerical scheme, and expanding the model class by including operator terms that account for higher order spatial gradients. Moreover, it is desirable to complement the current computational analysis with a rigorous theoretical derivation of an expression for the error probability density function in terms of model complexity, number of training samples, and coarse-graining block size.

ACKNOWLEDGMENTS

This work, LA-UR-15-23538, was carried out under the auspices of the U.S. Department of Energy by Los Alamos National Laboratory (LANL) under Contract No. DE-AC52-06NA25396, and was supported by Laboratory Directed Research and Development (LDRD) Project No. 20140013DR. Numerical simulations were performed using a hybrid C++/Python code run on LANL computing facilities. The authors would also like to thank Avadh Saxena for helpful discussions.

APPENDIX A: COMPUTATION OF THE PARAMETERS IN THE LEARNED sGLE MODEL

In order to calculate the coefficients of the learned discrete sGLE, we employ the widely used statistical technique of maximizing the log likelihood that the model will predict the GD training data. The discrete sGLE we are trying to learn has the form,

$$\begin{aligned}\phi_{n+1,i,j} = \phi_{n,i,j} + \alpha_0(\phi_{n,i+1,j} + \phi_{n,i-1,j} + \phi_{n,i,j+1} \\ + \phi_{n,i,j-1} - 4\phi_{n,i,j}) + \sum_{k=0}^{\frac{C-1}{2}} \alpha_{k+1}\phi_{n,i,j}^{2k+1} \\ + \alpha_{(C+3)/2}\,\xi_{n,i,j},\end{aligned} \tag{A1}$$

where C is the model complexity (we only consider odd complexities), the $\xi_{n,i,j}$ are independent, identically distributed standard normal random variables. Subscripts n and $n+1$ refer to times t_n and t_{n+1}, while $i = 0, \ldots, \bar{N}_1 - 1$ and $j = 0, \ldots, \bar{N}_2 - 1$ are the spatial coordinates of the blocks obtained by coarse-graining the spin lattice, with $\bar{N}_1$ and $\bar{N}_2$ the number of blocks in both spatial directions. Given the block-averaged training data $S_{n,i,j}$ with $n = 0, \ldots, n_{\rm eq}$, we would like to find the set of parameters $\boldsymbol{\alpha}_{\mathbf{opt}} \equiv \{\alpha_0, \alpha_1, \ldots, \alpha_{(C+3)/2}\}$ that maximizes the probability of observing $S_0, S_1, \ldots, S_{n_{\rm eq}}$ from model (A1). Here S_n refers to the $\bar{N}_1 \times \bar{N}_2$ block-averaged spin configuration after n steps following the convention introduced in Sec. III. This probability is given by

$$\begin{aligned}&\mathbb{P}(S_0, S_1, \ldots, S_{n_{\rm eq}}; \boldsymbol{\alpha}) \\ &\quad = \mathbb{P}(S_0)\mathbb{P}(S_1 \mid S_0; \boldsymbol{\alpha}) \ldots \mathbb{P}(S_{n_{\rm eq}} \mid S_1, \ldots, S_{n_{\rm eq}-1}; \boldsymbol{\alpha}) \\ &\quad = \mathbb{P}(S_0)\mathbb{P}(S_1 \mid S_0; \boldsymbol{\alpha}) \ldots \mathbb{P}(S_{n_{\rm eq}} \mid S_{n_{\rm eq}-1}; \boldsymbol{\alpha}).\end{aligned} \tag{A2}$$

Given S_n, we can see from (A1) that for each i and j, $S_{n+1,i,j}$ is normally distributed with mean,

$$\begin{aligned}Y_{n,i,j} = S_{n,i,j} + \alpha_0(S_{n,i+1,j} + S_{n,i-1,j} + S_{n,i,j+1} \\ + S_{n,i,j-1} - 4S_{n,i,j}) + \sum_{k=0}^{\frac{C-1}{2}} \alpha_{k+1} S_{n,i,j}^{2k+1},\end{aligned} \tag{A3}$$

and variance $\alpha^2_{(C+3)/2}$. Therefore, we have

$$\begin{aligned}&\mathbb{P}(S_{n+1,i,j} = s \mid S_n; \boldsymbol{\alpha}) \\ &\quad = \frac{1}{\alpha_{(C+3)/2}\sqrt{2\pi}} \exp\left[-\frac{(s - Y_{n,i,j})^2}{2\alpha^2_{(C+3)/2}}\right],\end{aligned} \tag{A4}$$

and since the $\xi_{n,i,j}$ are independent,

$$\begin{aligned}&\mathbb{P}(S_{n+1} \mid S_n; \boldsymbol{\alpha}) \\ &\quad = \frac{1}{\left(\alpha^2_{(C+3)/2} 2\pi\right)^{(\bar{N}_1\bar{N}_2)/2}} \exp\left[-\frac{\|S_{n+1} - Y_n\|^2}{2\alpha^2_{(C+3)/2}}\right],\end{aligned} \tag{A5}$$

where Y_n is defined according to the notational convention in Sec. III. Hence,

$$\begin{aligned}&\ln \mathbb{P}(S_0, S_1, \ldots, S_{n_{\rm eq}}; \boldsymbol{\alpha}) \\ &\quad = -\sum_{n=0}^{n_{\rm eq}-1} \frac{\|S_{n+1} - Y_n\|^2}{2\alpha^2_{(C+3)/2}} \\ &\qquad - n_{\rm eq}\bar{N}_1\bar{N}_2 \ln(\alpha_{(C+3)/2}) + \text{constant}.\end{aligned} \tag{A6}$$

For notational simplicity, let us now focus on the case of $C = 3$; the generalization to higher order nonlinearity is straightforward. From (A6), it follows that we need to minimize

$$\begin{aligned}\mathcal{L}(S_1, \ldots, S_{n_{\rm eq}}; \boldsymbol{\alpha}) = \tfrac{1}{2}\alpha_3^{-2}\, f(\alpha_0, \alpha_1, \alpha_2) \\ + n_{\rm eq}\bar{N}_1\bar{N}_2 \ln(\alpha_3),\end{aligned} \tag{A7}$$

where

$$\begin{aligned}f(\alpha_0, \alpha_1, \alpha_2) = \sum_{n=0}^{n_{\rm eq}-1} \sum_{i=0}^{\bar{N}_1-1} \sum_{j=0}^{\bar{N}_2-1} \big[(S_{n+1,i,j} - S_{n,i,j}) \\ - \alpha_0(S_{n,i+1,j} + S_{n,i-1,j} + S_{n,i,j+1} \\ + S_{n,i,j-1} - 4S_{n,i,j}) - \alpha_1 S_{n,i,j} - \alpha_2 S_{n,i,j}^3\big]^2.\end{aligned} \tag{A8}$$

From

$$\frac{\partial \mathcal{L}}{\partial \alpha_3} = -\alpha_3^{-3} f(\alpha_0, \alpha_1, \alpha_2) + n_{\rm eq}\bar{N}_1\bar{N}_2\alpha_3^{-1} = 0, \tag{A9}$$

we can solve for α_3,

$$\alpha_3 = \sqrt{\frac{f(\alpha_0, \alpha_1, \alpha_2)}{n_{\rm eq}\bar{N}_1\bar{N}_2}}. \tag{A10}$$

Also, for $k \neq 3$, we have

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = \frac{1}{2}\alpha_3^{-2}\frac{\partial f}{\partial \alpha_k}. \tag{A11}$$

If we now define

$$\begin{aligned}&D_{n,i,j} = S_{n+1,i,j} - S_{n,i,j}, \\ &A_{n,i,j} = S_{n,i+1,j} + S_{n,i-1,j} + S_{n,i,j+1} + S_{n,i,j-1} - 4S_{n,i,j}, \\ &B_{n,i,j} = S_{n,i,j},\ C_{n,i,j} = S_{n,i,j}^3,\end{aligned} \tag{A12}$$

then

$$\begin{aligned}\frac{\partial f}{\partial \alpha_0} = \partial_{\alpha_0} \sum_{n=0}^{n_{\rm eq}-1} \sum_{i=0}^{\bar{N}_1-1} \sum_{j=0}^{\bar{N}_2-1} (D_{n,i,j} - \alpha_0 A_{n,i,j} \\ - \alpha_1 B_{n,i,j} - \alpha_2 C_{n,i,j})^2\end{aligned}$$

$$
\begin{aligned}
&= -2\left(\sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} D_{n,i,j}A_{n,i,j}\right)\\
&\quad + 2\left(\sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} A^2_{n,i,j}\right)\alpha_0\\
&\quad + 2\left(\sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} B_{n,i,j}A_{n,i,j}\right)\alpha_1\\
&\quad + 2\left(\sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} C_{n,i,j}A_{n,i,j}\right)\alpha_2\\
&= -2(a_0 - a_{00}\alpha_0 - a_{01}\alpha_1 - a_{02}\alpha_2). \qquad \text{(A13)}
\end{aligned}
$$

Similarly,

$$
\frac{\partial f}{\partial \alpha_1} = -2(a_1 - a_{10}\alpha_0 - a_{11}\alpha_1 - a_{12}\alpha_2), \qquad \text{(A14)}
$$

where

$$
\begin{aligned}
a_1 &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} D_{n,i,j}B_{n,i,j},\\
a_{10} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} A_{n,i,j}B_{n,i,j},\\
a_{11} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} B^2_{n,i,j},\\
a_{12} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} C_{n,i,j}B_{n,i,j}.
\end{aligned} \qquad \text{(A15)}
$$

Finally,

$$
\frac{\partial f}{\partial \alpha_2} = -2(a_2 - a_{20}\alpha_0 - a_{21}\alpha_1 - a_{22}\alpha_2), \qquad \text{(A16)}
$$

where

$$
\begin{aligned}
a_2 &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} D_{n,i,j}C_{n,i,j},\\
a_{20} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} A_{n,i,j}C_{n,i,j},\\
a_{21} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} B_{n,i,j}C_{n,i,j},\\
a_{22} &= \sum_{n=0}^{n_{\rm eq}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1} C^2_{n,i,j}.
\end{aligned} \qquad \text{(A17)}
$$

Hence, we need to solve the linear system,

$$
\begin{aligned}
a_{00}\alpha_0 + a_{01}\alpha_1 + a_{02}\alpha_2 &= a_0,\\
a_{10}\alpha_0 + a_{11}\alpha_1 + a_{12}\alpha_2 &= a_1,\\
a_{20}\alpha_0 + a_{21}\alpha_1 + a_{22}\alpha_2 &= a_2.
\end{aligned} \qquad \text{(A18)}
$$

We can see that despite the nonlinearity of the sGLE, optimizing the log-likelihood function can be reduced to solving a linear system.

APPENDIX B: DETAILS OF THE OPERATIONAL PROCEDURE FOR CALCULATING THE ERROR PDFS

To compute one data point in the error pdf for a learned sGLE model of complexity C given a number of training samples $N_{\rm train}$, we do the following.

(1) We simulate $N_{\rm test}$ independent GD test sample trajectories. The kth trajectory is obtained as follows:

(a) Starting from a random initial Ising configuration, we march over $n_{\rm mc}$ steps.

(b) Starting from the resulting Ising configuration, we march over $n_{\rm eq}$ steps and store the block-averaged time history over these steps in a three-dimensional (3D) matrix $s^{\rm av,test}$ with dimensions $(n_{\rm eq}+1)\times\bar{N}_1\times\bar{N}_2$. Here $\bar{N}_1$ and $\bar{N}_2$ represent the number of spin blocks in each spatial direction.

(c) This matrix $s^{\rm av,test}$ will be the kth element of a four-dimensional matrix $s^{\rm av,test,all}$ with dimensions $N_{\rm test}\times(n_{\rm eq}+1)\times\bar{N}_1\times\bar{N}_2$.

The initial spins $s_{i,j}$ ($i=0,\ldots,N_1-1$ and $j=0,\ldots,N_2-1$) are given by $s_{i,j}=1-2\,r_{i,j}$, where the $r_{i,j}$ are drawn from a discrete uniform distribution on the half-open interval [0,2). Furthermore, with "block-averaged time history," we refer to the time evolution of the block-averaged spin configuration of the Ising lattice. At each discrete point in time, we group the individual spins into blocks of a certain size, and then calculate the average spin values over the different blocks. The resulting coarsened grid is then recorded.

(2) We simulate $N_{\rm train}$ independent GD training sample trajectories. A trajectory is calculated as follows.

(a) Starting from a random initial Ising configuration, we march over $n_{\rm mc}$ steps.

(b) Starting from the final Ising configuration, we march over $n_{\rm eq}$ steps and store the block-averaged time history over these steps in a 3D matrix $s^{\rm av,train}$ with dimensions $(n_{\rm eq}+1)\times\bar{N}_1\times\bar{N}_2$.

(c) We concatenate $s^{\rm av,train}$ of the current training sample with the corresponding matrices of the previous training samples along the first (time) dimension, and hence obtain a bigger matrix $s^{\rm av,train,all}$ with dimensions $N_{\rm train}\times(n_{\rm eq}+1)\times\bar{N}_1\times\bar{N}_2$.

(3) Using the training data stored in $s^{\rm av,train,all}$, we compute the coefficients of the learned sGLE polynomial using a log-likelihood analysis (see Appendix A).

(4) With the parameters calculated in step 3, we now simulate $N_{\rm test}$ sGLE trajectories. The kth trajectory is obtained as follows:

(a) We define a 3D matrix ϕ with dimensions $(n_{\rm eq}+1)\times\bar{N}_1\times\bar{N}_2$.

(b) We take the block-averaged Ising test configuration stored in $s^{\rm av,test,all}_{k,0,:,:}$ as the initial condition and define $s^{\rm av,test}\equiv s^{\rm av,test,all}_{k,:,:,:}$. We then define $\phi_{0,:,:}\equiv s^{\rm av,test}_{0,:,:}$.

(c) We march over $n_{\rm eq}$ steps and store the time history over these steps in ϕ.

(d) The configuration $\phi_{n+1,:,:}$ is calculated according to

$$\phi_{n+1,i,j} = x_{i,j} + \mathcal{G}(x_{:,:};\boldsymbol{\alpha}), \tag{B1}$$

where $x_{:,:}$ is either $s^{\text{av,test}}_{n,:,:}$ or $\phi_{n,:,:}$, $i = 0,\ldots,\bar{N}_1 - 1$, and $j = 0,\ldots,\bar{N}_2 - 1$.

In (d), we have defined $\mathcal{G}(x_{:,:};\boldsymbol{\alpha})$ as

$$\mathcal{G}(x_{:,:};\boldsymbol{\alpha}) = \alpha_0(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1} - 4x_{i,j}) + \sum_{k=0}^{\frac{C-1}{2}} \alpha_{k+1}(x_{i,j})^{2k+1} + \alpha_{(C+3)/2}\,\xi_{n,i,j}. \tag{B2}$$

(5) For the kth sGLE trajectory, we calculate the RMS error ε_k

$$\varepsilon_k = \sqrt{\frac{1}{a}\sum_{n=0}^{n_{\text{eq}}-1}\sum_{i=0}^{\bar{N}_1-1}\sum_{j=0}^{\bar{N}_2-1}\left(\phi_{n+1,i,j} - s^{\text{av,test}}_{n+1,i,j}\right)^2}, \tag{B3}$$

where $a = n_{\text{eq}}\bar{N}_1\bar{N}_2$.

(6) Finally, we compute the test-averaged error,

$$\varepsilon = \frac{1}{N_{\text{test}}}\sum_{k=1}^{N_{\text{test}}}\varepsilon_k, \tag{B4}$$

which we will call the "type 1" test error if the $\phi_{n+1,i,j}$ are calculated using $x_{:,:} = s^{\text{av,test}}_{n,:,:}$ in (B1), or the "type 2" test error if the $\phi_{n+1,i,j}$ are calculated using $x_{:,:} = \phi_{n,:,:}$ in (B1).

[1] H. Mori, Transport, collective motion and Brownian motion, Prog. Theor. Phys. **33**, 423 (1965).

[2] R. Zwanzig, Nonlinear generalized Langevin equations, J. Stat. Phys. **9**, 215 (1973).

[3] J. B. Bell, J. Foo, and A. L. Garcia, Algorithm refinement for the stochastic Burgers' equation, J. Comput. Phys. **223**, 451 (2007).

[4] S. Taverniers, F. J. Alexander, and D. M. Tartakovsky, Noise propagation in hybrid models of nonlinear systems: The Ginzburg-Landau equation, J. Comput. Phys. **262**, 313 (2014).

[5] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, Heterogeneous multiscale methods: A review, Commun. Comput. Phys. **2**, 367 (2007).

[6] L. A. Dalton and E. R. Dougherty, Optimal mean-square-error calibration of classifier estimators under Bayesian models, Pattern Recogn. **45**, 2308 (2012).

[7] L. A. Dalton and E. R. Dougherty, Optimal classifiers with minimum expected error within a Bayesian framework—Part I: Discrete and Gaussian models, Pattern Recogn. **46**, 1301 (2013).

[8] B. C. Daniels and I. Nemenman, Automated adaptive inference of phenomenological dynamical models, Nat. Commun. **6**, 8133 (2015).

[9] K. Kaski, K. Binder, and J. D. Gunton, A study of a coarse-grained free energy functional for the three-dimensional Ising model, J. Phys. A: Math. Gen. **16**, 623 (1983).

[10] K. Kaski, K. Binder, and J. D. Gunton, Study of cell distribution functions of the three-dimensional Ising model, Phys. Rev. B **29**, 3996 (1984).

[11] A. J. Bray, Theory of phase-ordering kinetics, Adv. Phys. **43**, 357 (1994).

[12] P. C. Hohenberg and B. I. Halperin, Theory of dynamic critical phenomena, Rev. Mod. Phys. **49**, 435 (1977).

[13] E. Ising, Beitrag zur Theorie des Ferromagnetismus, Z. Phys. **31**, 253 (1925).

[14] R. J. Glauber, Time-dependent statistics of the Ising model, J. Math. Phys. **4**, 294 (1963).

[15] A Ginzburg-Landau model for the free energy can also be used for other systems; see, e.g., [20].

[16] We only consider odd complexities in our analysis.

[17] The notation S_n indicates the 2D block-averaged spin configuration after n time steps.

[18] In future work, we will carry out an error analysis when only a small number of test samples is available [6,7].

[19] T_c refers to the critical temperature for a second-order phase transition following Onsager's [21] solution.

[20] K. J. Gaffney *et al.*, Observation of structural anisotropy and the onset of liquidlike motion during the nonthermal melting of InSb, Phys. Rev. Lett. **95**, 125701 (2005).

[21] M. E. Gracheva, J. M. Rickman, and J. D. Gunton, Coarse-grained Ginzburg-Landau free energy for Lennard-Jones systems, J. Chem. Phys. **113**, 3525 (2000).

[22] L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition, Phys. Rev. **65**, 117 (1944).