Interpretable and unsupervised phase classification

Julian Arnold[1,*] Frank Schäfer[1,†] Martin Žonda[2,3] and Axel U. J. Lode[2]
[1]*Department of Physics, University of Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland*
[2]*Institute of Physics, Albert-Ludwigs-Universität Freiburg, Hermann-Herder-Strasse 3, 79104 Freiburg im Breisgau, Germany*
[3]*Department of Condensed Matter Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 5, Praha 2 CZ-121 16, Czech Republic*

(Received 16 October 2020; accepted 15 June 2021; published 15 July 2021)

Fully automated classification methods that provide direct physical insights into phase diagrams are of current interest. Interpretable, i.e., fully explainable, methods are desired for which we understand why they yield a given phase classification. Ideally, phase classification methods should also be unsupervised. That is, they should not require prior labeling or knowledge of the phases of matter to be characterized. Here, we demonstrate an unsupervised machine-learning method for phase classification, which is rendered interpretable via an analytical derivation of the functional relationship between its optimal predictions and the input data. Based on these findings, we propose and apply an alternative, physically-motivated, data-driven scheme, which relies on the difference between mean input features. This mean-based method does not rely on any predictive model and is thus computationally cheap and directly explainable. As an example, we consider the physically rich ground-state phase diagram of the spinless Falicov-Kimball model.

DOI: 10.1103/PhysRevResearch.3.033052

I. INTRODUCTION

Phase diagrams and phase transitions are of paramount importance to physics [1–3]. While many-body systems have a large number of degrees of freedom, their phases are usually characterized by a small set of physical quantities like response functions or order parameters. However, the identification of phases and their order parameters is often a complex problem involving a large state space [4,5]. Machine-learning methods are apt for this task [3,6–15] as they can deal with large data sets and efficiently extract information from them. Ideally, such machine-learning methods should not require prior knowledge about the phases, e.g., in the form of samples that are labeled by their correct phase, or even the number of distinct phases. That is, the methods should be unsupervised [7,9,16–33].

Yet, they should also allow for a straightforward physical insight into the character of phases. Significant progress has been made recently [28–33], but some open issues remain regarding the interpretability [34,35] of phase classification methods, i.e., why a method yields a certain phase classification. Thus, unsupervised and interpretable phase classification stays a challenging, but highly rewarding task.

A good example of both progress in the field and relevant challenges regarding interpretability is the unsupervised method introduced in Ref. [21]. This approach is based on a predictive model trained to infer the parameters of a physical system from input data—obtained by experimental measurements or numerical simulations—that characterize the system's state. In the following, we refer to this approach as the *prediction-based method*. The predictions for the system parameters in the prediction-based method are changing most strongly near phase boundaries. Hence, the vector-field divergence of the deviations of the predicted system parameters with respect to their true values serves as an indicator (label I in Fig. 1) of phase transitions.

The prediction-based method was hitherto successfully applied to symmetry-breaking [21], driven-dissipative [21], quantum [22], and topological phase transitions [22,36] in various systems. The method requires a predictive model with sufficient expressive power [37,38] to resolve

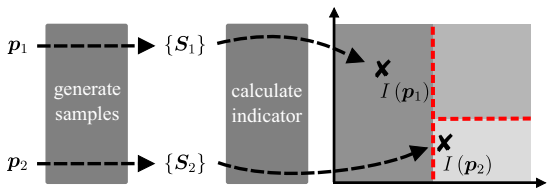

FIG. 1. Our workflow to predict a phase diagram with indicators I for phase transitions. Here, we illustrate the procedure for a two-dimensional parameter space. The parameter space is sampled on a grid, which yields a set of points $\{\boldsymbol{p}_i\}$ of fixed system parameters. At each such point $\boldsymbol{p}_i$ a set of samples $\{\boldsymbol{S}_i\}$ is generated. Based on these samples, a scalar indicator for phase transitions $I(\boldsymbol{p}_i)$ is calculated. This indicator highlights the boundaries (red) between phases (grey). Different unsupervised phase classification schemes are established via different indicators.

*julian.arnold@unibas.ch
†frank.schaefer@unibas.ch

different phases. Without prior system knowledge deep neural networks (DNNs) [38] constitute a good choice due to their capability of approximating arbitrary target functions efficiently [39]. However, the more expressive a machine-learning model such as a DNN, the more difficult it is to interpret the underlying functional dependence of the predictions on the input features [30,31]. Additionally, the training of DNNs is computationally demanding. Thus, the prediction-based method typically functions as a black-box model that solves a given phase classification task, but whose internal workings remain a mystery to the user.

Herein, we make the prediction-based method fully interpretable by deriving the form of its optimal predictions as a function of the input data. Therefore, we gain a full understanding of the resulting phase classification and the associated values of the indicator for phase transitions. These insights pave the way for the key result of this paper: A physically motivated, general, data-driven, unsupervised phase classification approach. It relies on the difference between mean input features as an indicator for phase transitions (Fig. 1); thus it is conceptually simple. In the following, we refer to this approach as the *mean-based method*. The mean-based method does not rely on a black-box predictive model and is thus computationally cheap and directly explainable.

This paper is organized as follows: In Sec. II, we introduce the Falicov-Kimball model (FKM) as our physical model of interest. In Sec. III, the FKM phase diagram is analysed with the prediction-based method. In particular, we derive the form of its optimal predictive model and discuss how this analytical expression makes the prediction-based method and its corresponding phase classification explainable. The mean-based method is introduced in Sec. IV. We demonstrate how the mean-based method reveals the FKM phase diagram given information about the prevalent correlations as input. Moreover, we discuss how the method can be used with other types of inputs and applied to different parameter spaces. A comparison of the prediction-based and mean-based method to another widespread unsupervised learning scheme for phase classification—namely principal component analysis (PCA) and k-means clustering—is provided in Sec. V. We conclude in Sec. VI and discuss potential future applications.

II. FALICOV-KIMBALL MODEL

As a physical system, we consider the two-dimensional spinless FKM [40–42]. This simple model of correlated electrons is used to address a broad range of contemporary physical problems including fractionalized metals [43], topological phenomena at finite temperature [44], nonthermal steady states [45], classical-quantum liquid transitions [46], or various quasiparticles [47,48]. It is also utilized as a standard test bed for the development of new methods in the context of strongly correlated systems [42,49–54] and recently machine learning [54,55].

The FKM ground-state phase diagram features a large number of different phases, e.g., charge stripes or various phase separations [56–58]. The properties of these phases play an important role in the investigation of numerous physical phenomena, e.g., metal-insulator and valence transitions [48,59–62], pattern formations in ultracold atoms in optical lattices [63–66], localization and correlations [67–73], or various nonequilibrium phenomena [45,49,74–79]. Hitherto, the classification of ground-state phases in the FKM was a manual and—due to the richness of the phase diagram [56–58]—lengthy and cumbersome task. The complexity of the FKM phase diagram makes it a challenging example for unsupervised and interpretable phase classification methods. To the best of our knowledge, neither supervised nor unsupervised phase classification methods have been applied to systems featuring a phase diagram with such a plethora of different orderings so far.

The Hamiltonian of the spinless FKM is

$$\mathcal{H} = -t \sum_{\langle ij \rangle} (d_i^\dagger d_j + d_j^\dagger d_i) + U \sum_i d_i^\dagger d_i f_i^\dagger f_i \,. \qquad (1)$$

Here, t is the hopping integral (energy unit throughout this paper), U is the on-site Coulomb interaction strength, $f_i^\dagger$ (f_i) and $d_i^\dagger$ (d_i) are the creation (annihilation) operators of heavy (f) and light (d) fermions at lattice site i. The number operator $n_{f,i} = f_i^\dagger f_i$ commutes with the Hamiltonian for all i, such that we can replace it by its eigenvalues $w_i \in \{0, 1\}$. The ground state is thus determined by the classical f-particle configuration $\boldsymbol{w}_0 = \{w_{0,i}\}$ that minimizes the system energy. We focus on the "neutral" case [56], characterized by an equal density of heavy and light particles $\rho = N_f/L^2 = N_d/L^2$. Here, N_f (N_d) is the total number of heavy (light) particles and $L = 20$—which we fix throughout this paper—is the linear size of the square two-dimensional lattice with periodic boundary conditions (plane symmetry group: $p4m$ [80]).

Figure 2(a) shows a sketch of the expected phase diagram in two-dimensional parameter space [56]. It highlights the regions of stability of three main types of orderings, namely, (1) segregated, (2) diagonal, and (3) axial orderings. A multitude of other phases with smaller stability regions are expected to be present in the full diagram [56,57].

Sample generation

We determine the ground-state configuration $\boldsymbol{w}_0$ approximately for a given $\boldsymbol{p} = (U, \rho)$ using an adaptive simulated annealing algorithm (see Appendix A) where ρ ranges from $1/L^2$ to half-filling ($\Delta\rho = 1/L^2$) and U ranges from 1 to 8 ($\Delta U = 0.2$). For each $\boldsymbol{p}$ we performed upwards of 64 independent simulations. For large systems, simulated annealing does not always converge to the ground state. This is akin to an experimental setting, where an experimentalist may not have access to all parameters of the underlying Hamiltonian and the system inevitably suffers from thermal (or other types of) noise. A crucial characteristic of any robust phase classification method is its ability to perform in such a "noisy" setting. Thus, we investigate two distinct cases: a "noise-free" case where the best estimate $\boldsymbol{w}_0$ is taken as the ground-state and a "noisy" case where we take into account 10 configurations with the smallest energies at each $\boldsymbol{p}$.

III. PREDICTION-BASED METHOD

In the first part of this paper, we analyze the FKM phase diagram using the prediction-based method with DNNs as predictive models $m : \boldsymbol{x} \to \hat{\boldsymbol{p}}(\boldsymbol{x}) = (\hat{U}(\boldsymbol{x}), \hat{\rho}(\boldsymbol{x}))$. For this, we

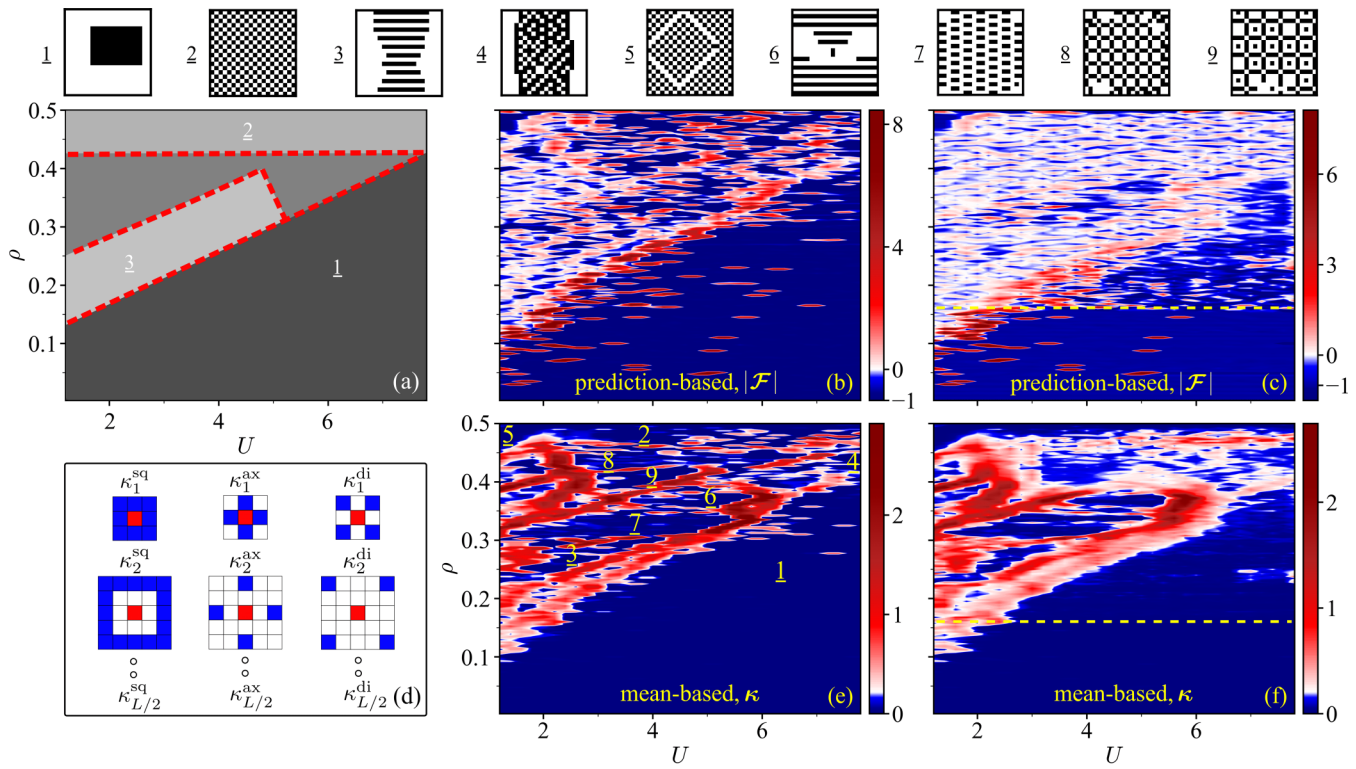

FIG. 2. (a) Sketch of the ground-state phase diagram of the spinless FKM. Red-dashed lines highlight the boundaries of the phases with (1)–(3) segregated, (2) diagonal, and (3) axial orderings. For each ordering (1)–(3), an example of a typical ground-state configuration $\boldsymbol{w}_0$ ($L = 20$) is shown on top. Here, the absence ($w_{0,i} = 0$) and presence ($w_{0,i} = 1$) of an f particle at lattice site i is denoted by a white or black square, respectively. [(b), (c)] $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ [Eq. (4)] based on the predictions of a DNN trained using $|\mathcal{F}|$ as input in the (b) noise-free and (c) noisy case. The color scale in (b) and (c) was cut off at –1 and –2, respectively, for better visualization. There were very few distinct points in parameter space with a divergence signal $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ below these cut-off values. (d) Illustration of the correlation functions that measure square (κ_n^{sq}), axial (κ_n^{ax}), and diagonal (κ_n^{di}) correlation at a distance n from the origin [cf. Eq. (14)]. Blue squares denote the lattice sites marked by the corresponding stencil, where red denotes the origin. [(e), (f)] Correlation indicator $\Delta\bar{\kappa}$ [Eq. (15)] in the (e) noise-free and (f) noisy case. Both $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ and $\Delta\bar{\kappa}$ serve as indicators for phase transitions (Fig. 1). The dashed line in (c) and (f) marks the cut along $\rho = 63/400 \approx 0.16$ analyzed in Fig. 3. Representative configurations (1)–(9) for some of the largest predicted regions of stability (connected regions marked in blue by the indicator), i.e., phases, are shown on the top. These regions connect configurations of the same character. We checked this manually and selected a representative example configuration for some of the regions as a guidance for the reader. This post-process labeling is necessary as the system's phases are, in principle, not known beforehand. A finite-size scaling of the FKM phase diagram with square lattices of linear size $L = 10$ and $L = 16$ is provided in Appendix E.

train a DNN to predict the underlying set of system parameters $\{\boldsymbol{p} = (U, \rho)\}$ from input data $\{\boldsymbol{x}\}$ for each sampled point $\boldsymbol{p}$ in parameter space (see Sec. II). Here the input data may be the raw ground-state configuration samples $\{\boldsymbol{w}_0\}$ or alternative representations of the system's state derived thereof. We use DNNs with convolutional layers to process image-like inputs, and DNNs with fully-connected layers otherwise. During training, the weights and biases of the DNNs are optimized through minimization of a mean-square error (MSE) loss function defined as

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N_p N_x} \sum_{\boldsymbol{p}} \sum_{\boldsymbol{x}} \|\hat{\boldsymbol{p}}(\boldsymbol{x}) - \boldsymbol{p}\|^2, \tag{2}$$

where the sum runs over all N_p sampled points $\{\boldsymbol{p}\}$ in parameter space and all N_x inputs $\{\boldsymbol{x}\}$ at each point $\boldsymbol{p}$. To incorporate configurations related through transformations of $p4m$, we use *online* data augmentation, i.e., each time a configuration is revisited during training a random transformation of $p4m$ is performed. Note that if we use inputs that are invariant under transformations of $p4m$, we do not need to use data augmentation. In case of inputs, which are only invariant under transformations of a particular subgroup of $p4m$, we instead perform data augmentation using random transformation of the corresponding subgroup. For further details on the DNN architectures and training procedure, see Appendix B.

Given a trained DNN, the predictions $\hat{\boldsymbol{p}}$ as a function of the system parameters $\boldsymbol{p}$ are obtained by averaging over the predictions for all N_x inputs $\{\boldsymbol{x}\}$ at $\boldsymbol{p}$:

$$\hat{\boldsymbol{p}}(\boldsymbol{p}) = \frac{1}{N_x} \sum_{\boldsymbol{x}} \hat{\boldsymbol{p}}(\boldsymbol{x}). \tag{3}$$

The phase diagram can be revealed by analyzing the vector field $\boldsymbol{\delta p}(\boldsymbol{p}) = \hat{\boldsymbol{p}}(\boldsymbol{p}) - \boldsymbol{p}$ whose individual components are given by $\delta U(\boldsymbol{p}) = \hat{U}(\boldsymbol{p}) - U$ and $\delta\rho(\boldsymbol{p}) = \hat{\rho}(\boldsymbol{p}) - \rho$. In particular, because the predictions $\hat{\boldsymbol{p}}(\boldsymbol{p})$ are most susceptible near the phase boundaries, maxima in the vector-field divergence

$$\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p} = \left.\frac{\partial \delta U}{\partial U}\right|_{\boldsymbol{p}} + \left.\frac{\partial \delta \rho}{\partial \rho}\right|_{\boldsymbol{p}} \tag{4}$$

serve as an indicator $I(\boldsymbol{p})$ (Fig. 1) of phase transitions. We approximate the corresponding derivatives using the symmetric difference quotient as

$$\left.\frac{\partial \delta U}{\partial U}\right|_{\boldsymbol{p}} \approx \frac{\delta U(U+\Delta U,\rho)-\delta U(U-\Delta U,\rho)}{2\Delta U},$$
$$\left.\frac{\partial \delta \rho}{\partial \rho}\right|_{\boldsymbol{p}} \approx \frac{\delta \rho(U,\rho+\Delta\rho)-\delta \rho(U,\rho-\Delta\rho)}{2\Delta\rho}. \quad (5)$$

A. Phase diagrams

We start the analysis of the FKM phase diagram with the prediction-based method using a DNN to predict $\boldsymbol{p}=(U,\rho)$ that takes the magnitude of the two-dimensional discrete Fourier transform $|\mathcal{F}|$ of each ground-state configuration $\boldsymbol{w}_0$ as input. Here, the elements of $\mathcal{F}$ are given as

$$\mathcal{F}_{u,v}=\sum_{x,y=0}^{L-1} w_{x,y}\cdot \mathrm{e}^{-2\pi i(ux+vy)/N}, \quad (6)$$

where $w_{x,y}\in\{0,1\}$ denotes the absence (0) or presence (1) of an f particle at site (x,y) on the 2D square lattice and $u,v\in\{0,1,\ldots,L-1\}$ [81]. The lattice sites in x and y direction are labeled from 0 to $L-1$ and L is the number of lattice sites in x and y direction. In practice, we calculate $\mathcal{F}$ in Eq. (6) using the fast Fourier transform algorithm [82]. Figures 2(b) and 2(c) show the resulting divergence signal in the noise-free and the noisy case, respectively. The usage of $|\mathcal{F}|$ instead of $\boldsymbol{w}_0$ results in a shorter training time because data augmentation by lattice translations is not necessary. Moreover, it yields an improved signal-to-noise ratio for $\nabla_{\boldsymbol{p}}\cdot\boldsymbol{\delta p}$, because $|\mathcal{F}_{0,0}|$ corresponds to N_f, i.e., ρ is directly fed into the DNN. The vector field $\boldsymbol{\delta p}$ exhibits a horizontal structure (see Fig. 6 in Appendix B), meaning that ρ is predicted with near-perfect accuracy in both cases. The horizontal structure of $\boldsymbol{\delta p}$ implies that maxima in $\nabla_{\boldsymbol{p}}\cdot\boldsymbol{\delta p}$ indicate phase transitions along U at fixed ρ. The (U,ρ) parameter space can thus be analyzed with cuts along U—we will do this later in this section.

The largest connected region with a negative divergence signal covering the bottom right half of the sampled parameter space displayed in Fig. 2(b) coincides with the main region of segregated orderings labeled 1 in the sketched phase diagram displayed in Fig. 2(a). Its character can be confirmed by a simple order parameter analysis (see Appendix C). However, the remaining phase boundaries are not reproduced with a large contrast. It is difficult to identify stability regions besides the main region of segregated orderings labeled 1, in particular the main regions of the diagonal and axial orderings [2, 3 in Fig. 2(a)]. Specifically, the prediction-based method indicates changes of the phase at several points within stable phase regions [Fig. 2(b)]. These artefacts intensify in the noisy case [Fig. 2(c)], where, in addition, the training of the DNN becomes computationally more demanding. Moreover, a large(er) amount of input data is needed to obtain the phase diagram with sufficient accuracy. To resolve these problems, it is first necessary to understand the DNN predictions.

B. Optimal predictive model

For this purpose, we analytically derive the form of the optimal model predictions when using the prediction-based method for phase classification. Here, an optimal predictive model m_{opt} is any model m, which minimizes the MSE loss in Eq. (2). In the general noisy case, the prediction of m_{opt} trained for the prediction-based method given the input $\boldsymbol{x}$ is

$$\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{x})=\frac{\sum_i \mathrm{P}_i(\boldsymbol{x})\boldsymbol{p}_i}{\sum_i \mathrm{P}_i(\boldsymbol{x})}. \quad (7)$$

Here, the sum runs over all sampled points $\{\boldsymbol{p}_i\}$ in parameter space. The probability of drawing the input $\boldsymbol{x}$ at $\boldsymbol{p}_i$ is governed by the distribution $\mathrm{P}_i(\boldsymbol{x})$.

Proof—In general, the system to be analyzed is characterized by a set of d tunable parameters $\boldsymbol{p}=(p^{(1)},p^{(2)},\ldots,p^{(d)})$, which we sample on an equidistant grid with grid spacings $\Delta\boldsymbol{p}=(\Delta p^{(1)},\Delta p^{(2)},\ldots,\Delta p^{(d)})$. At each grid point $\boldsymbol{p}_i$ in parameter space, we have N_x inputs $\{\boldsymbol{x}\}$, which constitute our training data and we train a predictive model $m:\boldsymbol{x}\to\hat{\boldsymbol{p}}(\boldsymbol{x})$ to minimize the MSE loss function $\mathcal{L}_{\mathrm{MSE}}$ specified in Eq. (2). Now, consider a particular input $\boldsymbol{x}_j$: We can determine the optimal model prediction $\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{x}_j)$ for this input by minimizing the loss function in Eq. (2) with respect to $\hat{\boldsymbol{p}}(\boldsymbol{x}_j)$, i.e., by solving

$$\frac{\partial \mathcal{L}_{\mathrm{MSE}}}{\partial \hat{\boldsymbol{p}}(\boldsymbol{x}_j)}=\frac{2}{N_p N_x}\sum_{\boldsymbol{p}} N_x^j(\boldsymbol{p})(\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{x}_j)-\boldsymbol{p})\ =0. \quad (8)$$

Here, $N_x^j(\boldsymbol{p}_i)$ denotes the number of times the particular input $\boldsymbol{x}_j$ is drawn at point $\boldsymbol{p}_i$, and $\mathrm{P}_i(\boldsymbol{x}_j)\approx N_x^j(\boldsymbol{p}_i)/N_x$ is the associated probability. Solving Eq. (8) yields

$$\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{x}_j)=\frac{\sum_i \mathrm{P}_i(\boldsymbol{x}_j)\boldsymbol{p}_i}{\sum_i \mathrm{P}_i(\boldsymbol{x}_j)}. \quad (9)$$

Repeating this step for all available training data $\{\boldsymbol{x}\}$, we recover Eq. (7). ■

Equation (7) implies that an optimal model m_{opt} predicts the center of mass for a particular input $\boldsymbol{x}_j$, where each grid point i is weighted according to the probability to draw the input $\boldsymbol{x}_j$ given by $\mathrm{P}_i(\boldsymbol{x}_j)/\sum_i \mathrm{P}_i(\boldsymbol{x}_j)$. Consequently, the prediction of an optimal model m_{opt} at a sampled point $\boldsymbol{p}$ is given by $\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{p})=1/N_x\sum_{\boldsymbol{x}}\hat{\boldsymbol{p}}_{\mathrm{opt}}(\boldsymbol{x})$ [see Eq. (3)], where the sum runs over all N_x inputs $\{\boldsymbol{x}\}$ at $\boldsymbol{p}$. The optimal divergence signal at a point $\boldsymbol{p}$ is thus given by $\nabla_{\boldsymbol{p}}\cdot\boldsymbol{\delta p}_{\mathrm{opt}}$, where $\boldsymbol{\delta p}_{\mathrm{opt}}=\hat{\boldsymbol{p}}_{\mathrm{opt}}-\boldsymbol{p}$ and the derivatives are approximated using the symmetric difference quotient [see Eq. (5)].

We have thus derived a simple analytical expression for the divergence signal of an optimal model m_{opt} in the general noisy case. Because this describes the optimal relationship between the input data and the indicator of the prediction-based method, we have made its phase classification—and thus the method itself—fully explainable. This is because the expression removes the need for further interpretation of the DNN, as it merely serves to approximate m_{opt}. The key to this analytical analysis of the prediction-based method is the fact that the method only requires the underlying predictive model to be evaluated on the training data; both during training and when calculating the indicator value. One may first choose to evaluate the prediction-based method on a test set to assess the

quality of the training data. However, for best performance the model should eventually be retrained using the entire available data. As such, the predictions for inputs that lie outside the training data, i.e., where the model needs to generalize, do *not* need to be analysed. This would constitute a much harder task. Nevertheless, note that the predictive models utilized in the prediction-based method, such as DNNs, *can* typically generalize to inputs that lie outside the training data after training and may thus still be of interest for other tasks.

C. Interpretation

The gain of interpretability granted by the analytical expressions can be best illustrated for the noise-free case, where the parameter space is divided into regions along U (at $\rho = \text{const.}$) with distinct input data: The prevalent ground-state configuration and all configurations related to it through transformations of $p4m$. That is, the probability distributions $\{\mathrm{P}_i(\boldsymbol{x})\}$ governing the input data in a given region are identical and the probability to draw the same input data outside the region vanishes. Thus, the optimal model predictions are identical for all grid points in such a region and are placed at its center of mass according to Eq. (7). In particular, ρ is predicted with perfect accuracy at all sampled points $\boldsymbol{p}$. This results in a vanishing optimal derivative along that direction in parameter space

$$\left.\frac{\partial \delta\rho_{\text{opt}}}{\partial \rho}\right|_{\boldsymbol{p}} \approx 0, \tag{10}$$

and the optimal divergence signal corresponds to the derivative along U

$$\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\text{opt}} \approx \left.\frac{\partial \delta U_{\text{opt}}}{\partial U}\right|_{\boldsymbol{p}}. \tag{11}$$

In what follows, we consider regions that contain at least two points in parameter space. This becomes the prevalent case when the parameter space is sampled sufficiently dense. For all other cases, see the analogous analysis in Appendix B. At all points in the *interior* of a region, the optimal divergence signal is then given by

$$\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\text{opt}} \approx -1. \tag{12}$$

At the two points in parameter space that make up the *boundary* of two neighboring regions along U, labeled I and II, the divergence signal is

$$\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\text{opt}} \approx \frac{\langle U\rangle_{\text{II}} - \langle U\rangle_{\text{I}}}{2\Delta U} - 1 \geqslant 0. \tag{13}$$

Here, $\langle U\rangle_{\text{I/II}} = 1/N_p^{\text{I/II}} \sum_{U\in \text{I/II}} U$ denotes the center of mass in U of the two regions, which each contain $N_p^{\text{I/II}} \geqslant 2$ grid points. As such, the value of $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\text{opt}}$ spikes at the points that constitute a region's boundary. The prediction-based method thus classifies these regions in parameter space as distinct phases. In particular, the method predicts a phase boundary whenever neighboring configurations in U (at $\rho = \text{const.}$) are not related by transformations of $p4m$. The divergence signal at the phase boundaries measures the mean extent of the two neighboring phases (along U). Equivalently, it assesses the stability of the two phases, i.e., their robustness against variations in the system parameters.

One can confirm that the divergence signal $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ shown in Fig. 2(b) obtained using a DNN to approximate the optimal predictive model m_{opt} matches this description, i.e., it coincides with the indicator obtained based on m_{opt} [see Fig. 7 in Appendix B]. Consequently, one can identify the large extent of the segregated phase in parameter space as the cause for the large, isolated maxima in $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ within it, see Fig. 2(b). Such isolated points are physically meaningless.

The noisy case can be understood from a single line scan. At $\rho = 63/400$ [dashed line in Fig. 2(c)] a broad transition from a nonsegregated to a segregated ordering occurs. Figures 3(a) and 3(d) show that the predictions $\hat{U}$ and the corresponding divergence $\partial\delta U/\partial U$ obtained with a DNN trained on the line scan indicate the corresponding phase boundary. Finite-sample statistics cause significant fluctuations in the local probability distributions $\mathrm{P}_i(|\boldsymbol{\mathcal{F}}|)$. This can lead to varying predictions $\hat{U}$ and divergence signals close to zero that, again, correspond to misleading predictions of phase boundaries within the segregated phase.

IV. MEAN-BASED METHOD

The analytical expression for the optimal model predictions has significantly strengthened our understanding of the prediction-based method. While we have so far relied on DNNs to approximate m_{opt}, it opens up the possibility to compute the indicator of the prediction-based method $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\text{opt}}$ directly from the input data, i.e., the underlying probability distributions $\{\mathrm{P}_i(\boldsymbol{x})\}$ without the need of universal function approximators. More generally, it paves the way for a class of alternative, computationally cheap algorithms for unsupervised phase classification without any predictive model.

To reproduce the phase classification in the noise-free case obtained using the prediction-based method [Fig. 2(b)], for example, a method simply needs to detect changes in neighboring configurations (up to transformations of $p4m$) in U (at $\rho = \text{const.}$). If a change is detected, a new phase is declared. Once all the points in parameter space are analysed, the phase diagram can be fully recovered by setting the indicator values according to Eqs. (12) and (13) given the phase classification at hand. To detect changes in neighboring configurations one can, for example, search for an appropriate symmetry transformation that relates the two or compare their energy. For further details on these two approaches, see Appendix D. While the first approach is computationally inefficient, the second requires perfect knowledge of the system Hamiltonian, which is typically absent in experimental settings. Instead, one can also detect changes in the configurations (up to transformations of $p4m$) by a direct comparison of observables derived from the configurations, which are invariant under transformations of $p4m$. In this section, we will discuss this approach in detail.

A. Correlation indicator

In Figs. 2(b) and 2(c) we have seen that the indicator of the prediction-based method can result in misleading predictions of phase boundaries within the large stability regions, such as the segregated phase. This is because the value of

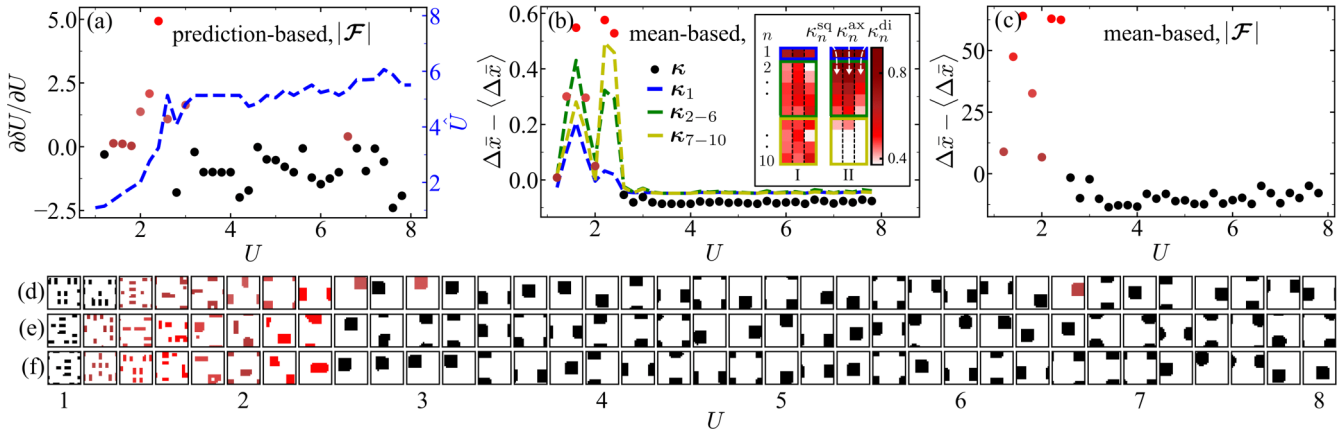

FIG. 3. Analysis of the transition from nonsegregated to segregated orderings occurring at $U \approx 2$ along the line scan [dashed line in Figs. 2(c) and 2(f)] from $U_{\min} = 1$ to $U_{\max} = 8$ at fixed $\rho = 63/400 \approx 0.16$. [(a), (d)] Predictions $\hat{U}$ and corresponding divergence $\partial\delta U/\partial U$ of a DNN trained on the line scan with $|\mathcal{F}|$ as input, as well as the indicator $\Delta\bar{x}$ [Eq. (16)] based on (b), (e) the entire set of correlation functions $\boldsymbol{\kappa}$ (points) or a subset thereof denoted as $\boldsymbol{\kappa}_{1,2-6,7-10}$ (dashed lines), and (c), (f) $|\mathcal{F}|$. Here, $\boldsymbol{\kappa}_{1,2-6,7-10}$ denotes the set of correlation functions measuring square (sq), axial (ax), and diagonal (di) correlations over 1, 2–6, and 7–10 lattice sites, respectively. $\langle\Delta\bar{x}\rangle$ denotes the average difference signal over the entire line scan and is subtracted to account for noise arising due to finite sample statistics. The inset in panel (b) shows the average input $\boldsymbol{\kappa}$ for the nonsegregated phase (I) and segregated phase (II) obtained from averaging over all inputs for U ranging from $U = 1.0$ to $U = 1.2$ and $U = 2.6$ to $U = 8.0$, respectively. The results shown in panels (c) and (f) are obtained using $n_{\text{trafo}} = 20$ for offline data augmentation at which we find the indicator to converge. The degree of red in (a)–(c) denotes an increasingly positive value of the respective indicator for phase transitions (Fig. 1); (d)–(f) configurations visualized using the same color scale as for the points in (a)–(c), respectively (see Fig. 5 in Appendix A for all configurations).

the indicator at phase boundaries reflects the mean extent of the neighboring phases, irrespective of how small or large the actual change in the ground-state configurations $\boldsymbol{w}_0$ is when moving from one stability region to another. To resolve these problems we propose an alternative indicator, which is instead given by the magnitude of the change in the configurations as measured in a representation that is invariant under transformations of $p4m$. Ideally, the resulting observable should not be very sensitive to small changes in $\boldsymbol{w}_0$ (up to transformations of $p4m$) within a stability region. This establishes an alternative, physically motivated, data-driven, unsupervised phase classification approach, which we call the *mean-based method*.

The probing of correlations is a standard procedure for studying phase transitions. As such, correlations also represent a suitable physically-motivated choice for a representation in any data-driven phase classification method, irrespective of the system at hand. Given that the FKM is defined on a square lattice with periodic boundary conditions, we consider the following set of correlation functions that measure the order of a given classical configuration:

$$\kappa_n^{\text{sq}} = \frac{1}{8nL^2}\sum_{i,j=0}^{L-1}\sum_{\alpha,\beta=-n}^{n}\mathcal{S}^{\text{sq}}_{n,\alpha,\beta}\tilde{w}_{i,j}\tilde{w}_{i+\alpha,j+\beta},$$

$$\kappa_n^{\text{ax}} = \frac{1}{4L^2}\sum_{i,j=0}^{L-1}\sum_{\alpha,\beta\in\{-n,n\}}\mathcal{S}^{\text{ax}}_{n,\alpha,\beta}\tilde{w}_{i,j}\tilde{w}_{i+\alpha,j+\beta}, \qquad (14)$$

$$\kappa_n^{\text{di}} = \frac{1}{4L^2}\sum_{i,j=0}^{L-1}\sum_{\alpha,\beta\in\{-n,n\}}\mathcal{S}^{\text{di}}_{n,\alpha,\beta}\tilde{w}_{i,j}\tilde{w}_{i+\alpha,j+\beta},$$

where $\tilde{w}_{i,j} = 2w_{i,j} - 1$ and $\mathcal{S}^{\text{sq}}_{n,\alpha,\beta} = (\delta_{|\alpha|,n} + \delta_{|\beta|,n} - \delta_{|\alpha|+|\beta|,2n})$, $\mathcal{S}^{\text{ax}}_{n,\alpha,\beta} = \delta_{|\alpha|+|\beta|,n}$, $\mathcal{S}^{\text{di}}_{n,\alpha,\beta} = \delta_{|\alpha|+|\beta|,2n}$ are the corresponding stencils with lattice points matching three types of orders that measure square (sq), axial (ax), and diagonal (di) correlations over n lattice sites. Given periodic boundary conditions, the largest unique n is given by $n_{\max} = L/2$. The computation of the correlation functions from Eq. (14) is illustrated in Fig. 2(d). Note that we choose these correlation functions as one possible representation out of many, which render configuration samples related through transformations of $p4m$ identical.

With the correlation functions in Eq. (14), we define the following *correlation indicator* for the mean-based method:

$$\Delta\bar{\kappa}(U,\rho) = \|\bar{\boldsymbol{\kappa}}(U+\Delta U,\rho) - \bar{\boldsymbol{\kappa}}(U-\Delta U,\rho)\|, \qquad (15)$$

at each point $\boldsymbol{p} = (U,\rho)$, where $\bar{\boldsymbol{\kappa}} = (\bar{\kappa}_1^{\text{sq}},\ldots,\bar{\kappa}_{L/2}^{\text{sq}},\bar{\kappa}_1^{\text{ax}},\ldots,\bar{\kappa}_{L/2}^{\text{ax}},\bar{\kappa}_1^{\text{di}},\ldots,\bar{\kappa}_{L/2}^{\text{di}})$. Here, the $\bar{\cdot}$ notation indicates the average over all inputs at a given point $\boldsymbol{p}$, if multiple inputs are considered. The indicator for phase transitions $\Delta\bar{\kappa}$ (Fig. 1) measures the magnitude of the mean change of order quantified by $\bar{\boldsymbol{\kappa}}$. To account for variations in the grid spacing one may rescale the correlation indicator [Eq. (15)] by an additional factor of $1/2\Delta U$ (that we omit here). Note that the correlation indicator can be computed from the input data via a simple analytical expression without relying on a black-box predictive model. Thus, the mean-based method is computationally cheap and fully interpretable.

Our results for both noisy and noise-free cases demonstrate that the mean-based method with the indicator $\Delta\bar{\kappa}$ reveals the phase diagram more clearly than $\nabla_{\boldsymbol{p}}\cdot\boldsymbol{\delta p}$, compare Figs. 2(e), 2(f) and 2(b), 2(c). The indicator $\Delta\bar{\kappa}$ reproduces the main characteristics of the FKM phase diagram [Fig. 2(a)]: $\Delta\bar{\kappa}$

almost vanishes within the stability region of segregated orderings, in the presence or absence of noise [Figs. 2(e) and 2(f)], and marks all phase boundaries of Fig. 2(a) (order parameter analysis in Appendix C). Moreover, we obtain a detailed, physical subdivision of the phase diagram—see the identified orderings in Fig. 2 (top) and labels (1)–(9) in Fig. 2(e). In particular, the method is able to distinguish between dimer structures (7), different stable tile patterns (8), (9), or orderings exhibiting a complicated phase separation (4). Thus, the mean-based method with the correlation indicator $\Delta\bar{\kappa}$ in Eq. (15) is an excellent tool to detect phase transitions.

Let us take a closer look at the line scan $\rho = 63/400$ where a broad transition from a nonsegregated to a segregated ordering occurs at $U \approx 2$ [Figs. 3(b) and 3(e)]. More precisely, we see that the system undergoes complete segregation starting from a most homogeneous ordering of dimers (which we call nonsegregated for simplicity) with increasing U. In the process of segregation, we first observe the formation of independent clusters. The correlation indicator of the mean-based method shows two distinct peaks (or "one broad peak"), which correctly highlights that the transition from nonsegregated to segregated orderings proceeds via this intermediate ordering of independent clusters. Looking at Figs. 2(e) and 2(f), we see that the investigated line scan at $\rho = 63/400 \approx 0.16$ (dashed line) indeed passes through the edge of a small stability region at $U \approx 2$ comprised of this intermediate ordering of independent clusters. These results highlight the performance of the mean-based method, as it is capable of resolving even such small stability regions and capture the competition of distinct orders in the vicinity of phase boundaries. Note that such a competition of different orders at transitions is of contemporary interest [83]. A more detailed study focused on this region of parameter space together with a proper finite-size scaling could be done to confirm that this stability region does not correspond to a finite-size effect but persists in the thermodynamic limit.

Using the mean-based method, we can obtain information about the relevant change in order governing a phase transition straightforwardly by calculating an indicator, Eq. (15), based on individual features of the overall input (here $\boldsymbol{\kappa}$). Calculating individual indicators based on the correlation functions measuring short-, medium-, and long-range correlations, for example, one can identify which type changes the most at a given phase boundary [Fig. 3(b)]. The prevalent order in a predicted phase can be characterized by calculating the mean input for the corresponding region in parameter space [Fig. 3(b), inset]. For the transition along $\rho = 63/400$ (Fig. 3), this approach reveals the decrease in long-range correlations in the transition from nonsegregated to segregated orderings and quantifies its importance compared to changes in short-range correlations. This demonstrates that the application of the mean-based method allows for direct physical insight into the predicted phase diagram.

B. Generic indicators

We now ask the question whether the mean-based method can be applied to the phase classification problem without a specific physically-motivated input. To this end, we extend the mean-based method to arbitrary inputs by defining an *input-generic indicator*:

$$\Delta\bar{x}(\boldsymbol{p}) = \|\bar{\boldsymbol{x}}(U + \Delta U, \rho) - \bar{\boldsymbol{x}}(U - \Delta U, \rho)\|. \tag{16}$$

Here, $\bar{\boldsymbol{x}}(\boldsymbol{p}_i) = \sum_j \mathrm{P}_i(\boldsymbol{x}_j)\boldsymbol{x}_j \approx \sum_j N_x^j(\boldsymbol{p}_i)/N_x \boldsymbol{x}_j$ denotes the average input at a point $\boldsymbol{p}_i$. Because the inputs $\{\boldsymbol{x}\}$ do not generally need to be invariant under transformations of $p4m$, we perform *offline* data augmentation by applying n_{trafo} random symmetry transformations to the configurations analogous to online data augmentation (see Sec. III). Based on the augmented set of configurations, we then compute the input data $\{\boldsymbol{x}\}$ (such as $\{\boldsymbol{\kappa}\}$ or $\{|\mathcal{F}|\}$) and the corresponding averages. Similar to the prediction-based method, the computational cost of data augmentation can be significantly reduced by choosing a representation in which the inputs are invariant under transformations of $p4m$ (or a subgroup thereof).

In a data-driven approach such as the mean-based method, the choice of representation for the input data also crucially affects the performance of the phase classification; Figs. 2(e) and 2(f) show that the set of correlation functions $\boldsymbol{\kappa}$ are an appropriate choice in the case of the FKM. In general, based on the knowledge that the system is defined on a square lattice with periodic boundary conditions one can identify the Fourier transformed configuration $|\mathcal{F}|$ as a suitable representation for the input because it removes the translational symmetry. Even with such a generic choice of input, the difference signal is a good indicator for phase transitions and reveals the main characteristics of the phase diagram both in the noise-free and noisy case, see line scan along U in Figs. 3(c), 3(f), and 4. This underpins the generality and robustness of the mean-based method and shows its possible applicability to other models beyond the FKM, where different data representations may be chosen.

The increased level of noise when using $|\mathcal{F}|$ compared to $\boldsymbol{\kappa}$ as input [compare Fig. 4 to Figs. 2(e) and 2(f)] can be attributed to the fact that $|\mathcal{F}|$ is not fully invariant under transformations of $p4m$. As such, differences in the input that can be resolved through application of transformations of $p4m$ nevertheless contribute to the indicator. This effect is even more pronounced when using the raw configuration samples $\boldsymbol{w}_0$ as input. In this case, we observe that the phase boundaries get washed out. This highlights the importance of a representation in which samples that are related through the system's symmetries are identical for the success of the mean-based method. In contrast, the prediction-based method does not fundamentally rely on such a representation. Knowledge of the system's symmetries is, however, required to perform data augmentation that saves computation time during sample generation and training of the predictive model.

In case of the FKM, the parameter space can be effectively analyzed with cuts along a single parameter (here U). We made use of this fact when defining the indicators of the mean-based method [Eq. (15) and (16)]. Note, however, that these indicators can easily be extended to include changes in ρ. This can, for example, be accomplished by a *parameter-generic indicator* of the form

$$\Delta\bar{x}_{\mathrm{tot}}(\boldsymbol{p}) = \Delta\bar{x}(\boldsymbol{p}) + \Delta\bar{x}_\rho(\boldsymbol{p}), \tag{17}$$

where $\Delta\bar{x}(\boldsymbol{p})$ [Eq. (16)] measures changes along U and $\Delta\bar{x}_\rho(\boldsymbol{p}) = \|\bar{\boldsymbol{x}}(U, \rho + \Delta\rho) - \bar{\boldsymbol{x}}(U, \rho - \Delta\rho)\|$ measures

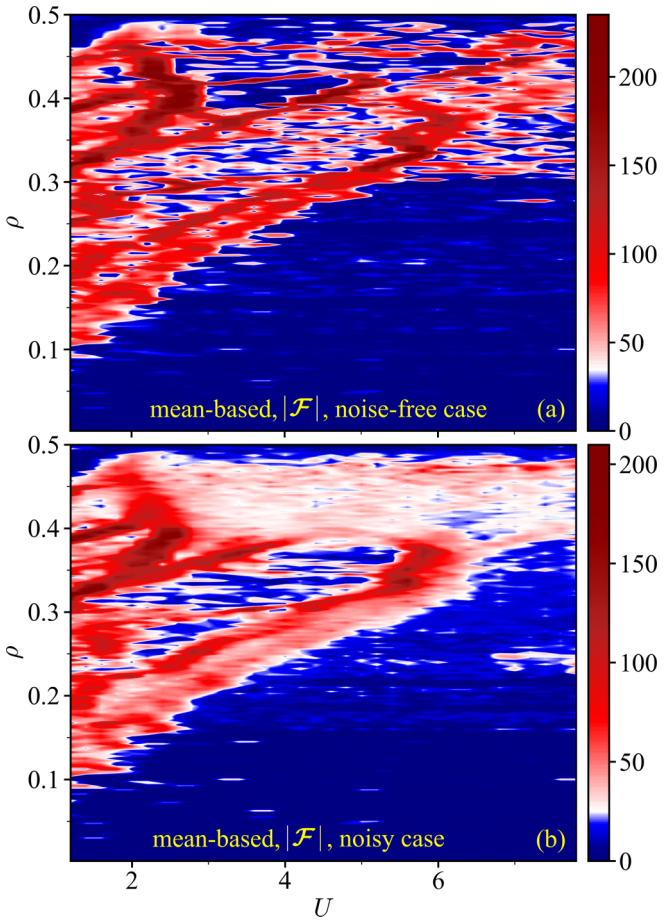

FIG. 4. Indicator $\Delta\bar{x}$ [Eq. (16)] of the mean-based method based on $|\mathcal{F}|$ as input on the two-dimensional parameter space of the FKM in the (a) noise-free and (b) noisy case. For offline data augmentation we use $n_{\text{trafo}} = 20$ at which we find the indicator to converge.

changes along ρ. As expected, the FKM phase diagrams obtained with the parameter-generic indicator in Eq. (17) of the mean-based method do not differ significantly from the phase diagrams obtained using Eq. (16), see Fig. 11 in Appendix F. Similarly, one can extend the indicator to parameter spaces of arbitrary dimension. This establishes the mean-based method as a general, unsupervised phase classification method that can be applied to arbitrary phase diagrams.

V. COMPARISON WITH PRINCIPAL COMPONENT ANALYSIS AND k-MEANS CLUSTERING

As a final step, let us compare the prediction-based and mean-based method to PCA and k-means clustering—a simple but widespread unsupervised learning scheme for phase classification [9,84]. We find that PCA and k-means clustering assigns configuration samples related through transformations of $p4m$ to different clusters (phases) when using the raw configuration samples as input in the noise-free case. Here, we used the scikit-learn implementation of PCA and k-means clustering with default settings [85]. Thereby, without adopting a different data representation, this method fails at classifying the phases of the FKM.

These issues can be explained by the fact that configuration samples related through transformations of $p4m$ are not necessarily close in configuration space. Because the k-means clustering algorithm relies on Euclidean distance as a measure of similarity, such configurations do not necessarily get clustered together as the number of distinct clusters increases. When performing dimensionality reduction using PCA, the data is projected into the subspace that contains most of the variance present in the data. If two configuration samples related through transformation of $p4m$ are separated far in the original configuration space, they are likely to also going to be separated far after performing PCA. This is because the corresponding direction is of large variance and therefore of high priority when reducing the dimensionality. Note that this problem will remain in more elaborate nonlinear dimensionality reduction methods, such as t-distributed stochastic neighbor embedding [86] (t-SNE) or uniform manifold approximation and projection [87] (UMAP), as they all try to preserve the distance (similarity) of the data within the original space [84].

A way to resolve these symmetry-related issues is to adapt an alternative representation in which configurations related through transformation of $p4m$ are identical such as the set of correlation functions $\boldsymbol{\kappa}$. In particular, using inputs that are only invariant under transformations of a subgroup of $p4m$ such as $|\mathcal{F}|$ is not sufficient. However, we have seen that the phase classification problem in the noise-free case is rendered trivial given such a representation and can equally be solved by other simple algorithms—in particular the mean-based method. Moreover, the indicator signals of the prediction-based and mean-based method entail additional information about the nature of the phases and corresponding phase transitions. Recall that it is precisely the alternative indicator signal of the mean-based method that allowed us to address the remaining issues of the prediction-based method. The application of PCA and k-means clustering does not yield such insights. Note that a possible approach to extend PCA and k-means clustering towards different inputs and the general noisy case is to apply it on averaged inputs, similar to the mean-based method. However, it is unclear what advantages such an approach would have over using the mean-based method directly.

VI. CONCLUSION AND OUTLOOK

In conclusion, we have made the prediction-based method fully interpretable with a derivation of its optimal model predictions. This opens up the possibility to compute the indicator of the prediction-based method directly from the input data $\{\boldsymbol{x}\}$, i.e., the corresponding probability distributions $\{\mathrm{P}_i(\boldsymbol{x})\}$, without the need of predictive models, such as DNNs. Moreover, the analytical expressions of the optimal model predictions have guided us to propose the mean-based method that works outstandingly well as an unsupervised phase classification approach for various types of inputs and in the presence of noise. We infer that applications of our mean-based method to arbitrary phase diagrams featuring, e.g., quantum or topological phase transitions are feasible. Specifically, applications to quantum-classical systems such as the FKM and its numerous generalizations [44,49,58,88,89] are now straightforward. Finally, we note that the indicators of

phase transitions in the prediction-based [Eq. (4)] and mean-based method [Eq. (16)] differ fundamentally. They constitute two distinct approaches to characterize changes in the probability distributions $\{P_i(\boldsymbol{x})\}$ that govern the input data and provide complementary insights into the phase diagram. The success of the mean-based method suggests extensions to unsupervised phase classification methods whose indicator is, e.g., based on the magnitude of the change in the higher-order moments of the underlying probability distributions or measures of similarity such as the Hellinger distance [90].

The code for the prediction-based and mean-based method that was utilized in this paper is open source [91].

ACKNOWLEDGMENTS

We would like to thank Niels Lörch, Eliska Greplova, Michael Thoss, and Christoph Bruder for inspiring discussions. J.A. and F.S. acknowledge financial support from the Swiss National Science Foundation (SNSF) and the NCCR Quantum Science and Technology. M.Z. acknowledges financial support through Grant No. INTER-COST LTC19045. A.U.J.L. acknowledges financial support by the Austrian Science Foundation (FWF) under Grant No. P-32033-N32. Computation time on the Hawk cluster at the HLRS Stuttgart and at sciCORE (scicore.unibas.ch) scientific computing core facility at University of Basel, as well as support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through Grants No. INST 40/467-1 FUGG (JUSTUS cluster), No. INST 39/963-1 FUGG (bwForCluster NEMO), and No. INST 37/935-1 FUGG (bwForCluster BinAC) is gratefully acknowledged.

APPENDIX A: DETAILS ON THE SAMPLE GENERATION

In this Appendix, we provide further details on the sample generation procedure (see Sec. II of the main text). For a fixed f-particle configuration $\boldsymbol{w}$, the Hamiltonian of the FKM in Eq. (1) of the main text can be transformed into

$$\mathcal{H}^{\boldsymbol{w}} = \sum_{j,j'} h_{jj'} d_j^\dagger d_{j'} = \sum_{\alpha} \lambda_\alpha^{\boldsymbol{w}} b_\alpha^\dagger b_\alpha, \tag{A1}$$

where we introduce the matrix elements $h_{jj'} = U w_j \delta_{jj'} - t\delta_{\langle jj'\rangle}$. Its eigenvalues $\lambda_\alpha^{\boldsymbol{w}}$ are obtained by numerical diagonalization. Finding the ground state of the FKM then means to find the configuration $\boldsymbol{w}$, which leads to the lowest energy

$$E_{\mathrm{gs}}(\boldsymbol{w}) = \sum_{\alpha=1}^{N_d} \lambda_\alpha^{\boldsymbol{w}}. \tag{A2}$$

However, even after accounting for the lattice symmetries, the ground-state configurations of systems with linear size $L = 20$, in general, cannot be determined exactly by comparing the energies of all possible configurations $\boldsymbol{w}$. An approximate method is required. Instead of using a reduced set of chosen orderings, as was done in previous studies [56,57] of the model, we determine the corresponding f-particle ground-state configuration $\boldsymbol{w}_0$ using simulated annealing.

We use an algorithm based on a semiclassical Metropolis Monte Carlo [92] with $E_{\mathrm{gs}}(\boldsymbol{w})$ in the statistical weight's energy instead of the free energy. This means, that the candidate configuration $\boldsymbol{w}_c$, generated by a random displacement of a single f particle from the current configuration $\boldsymbol{w}$, is accepted as a new $\boldsymbol{w}$ if $E_{\mathrm{gs}}(\boldsymbol{w}_c) \leqslant E_{\mathrm{gs}}(\boldsymbol{w})$ or $\min\left(1, \exp[-\beta(E_{\mathrm{gs}}(\boldsymbol{w}_c) - E_{\mathrm{gs}}(\boldsymbol{w}))]\right) > r$, where r is a random number drawn from a uniform distribution $r \in [0, 1]$ and $\beta = 1/T$ is the inverse temperature. We first used a classical protocol, where we started at relatively high temperature $T \sim 0.1t$ and cooled the sample in $20 - 40$ discrete temperature steps to zero. A thermalization process consisting of $10^2 - 10^3 \times L^2$ updates was done at every time step.

However, we found that an alternative adaptive protocol is much more efficient in lowering the energy. Namely, we started the annealing with a long thermalization at a low temperature (typically $T = 0.003t$). In the next steps, depending whether the algorithm has found a configuration with lower energy at the current temperature or not, the temperature was either lowered by dividing it by a factor between one and two (typically 1.25) or increased by multiplying it by the same factor. The modified protocol is better in escaping local minima and has less troubles with the fact, that the FKM can go through more than one ordered phase with decreasing temperature [93,94].

We have typically used a number of independent runs with random initial conditions. For small lattices ($L \leqslant 10$) all simulations converged to configurations identical up to transformations of $p4m$. For $L = 20$ we used 64 runs with random initial conditions, plus several runs with initial configurations reflecting typical ground-state orderings identified for smaller lattices ($L \leqslant 16$). In this case, we obtained several distinct configuration samples at each investigated point in parameter space $\boldsymbol{p} = (U, \rho)$ from independent simulated annealing runs that converged to nearby local energy minima, as opposed to the global minimum. Thus, we distinguished between the "noise-free" and "noisy" case.

In the noisy case, we considered the 10 configurations with the lowest energies at each sampled point in parameter space. Figure 5 shows all such 10 configurations for each sampled point in parameter space along the line scan at fixed $\rho = 63/400$, which was analysed in Fig. 3 in the main text. In the noise-free case, we performed one additional step: Namely, at each investigated $\boldsymbol{p} = (U, \rho)$ we took the configuration with the lowest energy and compared it with the energy calculated using the configuration obtained as the ground state for the same ρ, but different (neighboring) U. The configuration with the lowest energy was then taken as the final ground-state approximation.

APPENDIX B: DETAILS ON THE PREDICTION-BASED METHOD

Here, we provide further details on the prediction-based method (see Sec. III of the main text). In particular, we discuss the architecture of the DNNs employed in this paper and how they were trained. Moreover, we discuss the vector field whose divergence corresponds to the indicator of phase transitions in the prediction-based method and we extend our analysis of the optimal model predictions in the noise-free case.

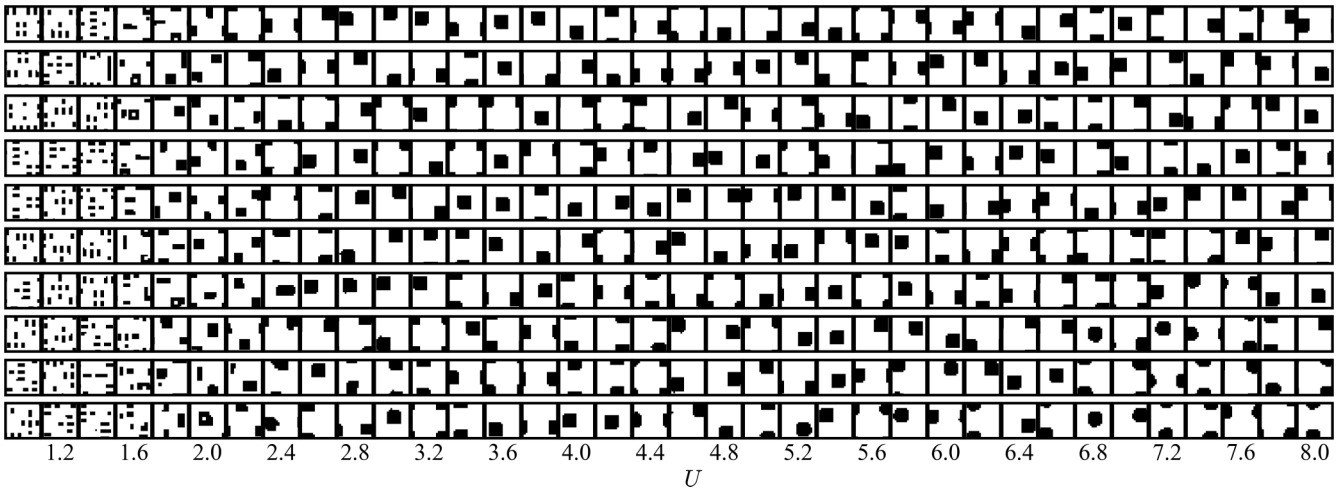

FIG. 5. Ground-state configurations ($L = 20$) along the line scan from $U_{\min} = 1$ to $U_{\max} = 8$ at fixed $\rho = 63/400 \approx 0.16$ [cf. dashed line in Figs. 2(c) and 2(f) in the main text]. At each value of U, we show the 10 configurations arising from independent simulated annealing runs that are considered in the noisy case.

1. Neural network architecture

The DNNs utilized in this paper are built as follows: if the inputs are image-like, such as ground-state configurations $\{\boldsymbol{w}_0\}$ or the magnitude of their two-dimensional discrete Fourier transform $\{|\mathcal{F}|\}$, we first apply K different square filters with the same linear size L as the input image. Subsequently, we apply a rectified linear unit, $\mathrm{ReLU}(z) = \max(0, z)$, as an activation function. This results in an output feature map of size $1 \times 1 \times K$, which is then flattened to a feature vector with K elements. In case of vector-like inputs, we skip this step. In both cases, we feed the corresponding vectors into a series of fully-connected layers (FCLs), where ReLUs are used as activation functions [38]. While this architecture remains to be optimized systematically to achieve a similar or improved accuracy at lower computational cost, such DNNs satisfy a universal approximation theorem [95].

2. Training procedure

For training the DNNs, each input $\boldsymbol{x} = \{x_i\}$ is standardized by the map $\mathfrak{S} : \boldsymbol{x} \to \boldsymbol{x}'$ whose element-wise action is given by the following affine transformation

$$x_i' = \frac{x_i - \langle x_i \rangle}{\sigma_{x_i}}. \tag{B1}$$

Each output $\boldsymbol{p} = \{p_i\}$ is normalized by the map $\mathfrak{N} : \boldsymbol{p} \to \boldsymbol{p}'$, where each element is transformed as

$$p_i' = \frac{p_i}{\sigma_{p_i}}. \tag{B2}$$

Here, $\{\langle x_i \rangle\}$ ($\{\langle p_i \rangle\}$) and $\{\sigma_{x_i}\}$ ($\{\sigma_{p_i}\}$) denote the mean values and standard deviations of the distributions of the inputs (outputs) over the entire training data, respectively. Standardization ensures that the distribution of each transformed input x_i' over the entire training data is characterized by ($\langle x_i' \rangle = 0, \sigma_{x_i'} = 1$). Whereas normalization results in the distribution of each transformed output p_i' over the entire training data being characterized by ($\langle p_i' \rangle = \langle p_i \rangle / \sigma_{p_i}, \sigma_{p_i'} = 1$). Scaling of the inputs, here by means of standardization, is common practice in the data pre-processing step of machine-learning tasks relying on gradient descent for optimization, because it generally leads to a faster rate of convergence [96]. The additional normalization of the outputs generally improves the model accuracy when training with a mean-square error (MSE) loss function, as it ensures that the outputs do not differ in size or spread and consequently enter the problem with equal weight during the optimization. Here, the MSE loss function is defined as

$$\mathcal{L}'_{\mathrm{MSE}} = \frac{1}{N_p N_x} \sum_{\boldsymbol{p}} \sum_{\boldsymbol{x}} \| \hat{\boldsymbol{p}}'(\mathfrak{S}(\boldsymbol{x})) - \mathfrak{N}(\boldsymbol{p}) \|^2, \tag{B3}$$

where the sum runs over all N_p sampled points $\boldsymbol{p}$ in parameter space and all N_x inputs $\boldsymbol{x}$ at each point $\boldsymbol{p}$. Here, $\hat{\boldsymbol{p}}'(\boldsymbol{x}') = (\hat{U}'(\boldsymbol{x}'), \hat{\rho}'(\boldsymbol{x}'))$ denotes the prediction of the $\mathrm{DNN} : \boldsymbol{x}' \to \hat{\boldsymbol{p}}'(\boldsymbol{x}')$ given a transformed input $\boldsymbol{x}' = \mathfrak{S}(\boldsymbol{x})$. The function composition $m = \mathfrak{N}^{-1} \circ \mathrm{DNN} \circ \mathfrak{S} : \boldsymbol{x} \to \hat{\boldsymbol{p}}(\boldsymbol{x}) = (\hat{U}(\boldsymbol{x}), \hat{\rho}(\boldsymbol{x}))$ then yields the desired predictive model, which maps an untransformed input $\boldsymbol{x}$ to a prediction $\hat{\boldsymbol{p}} = (\hat{U}, \hat{\rho})$ that approximates the underlying system parameters $\boldsymbol{p}$. In particular, given a DNN that minimizes the MSE loss function in Eq. (B3) ($\mathrm{DNN}_{\mathrm{opt}}$) the resulting predictive model m minimizes the MSE loss function in Eq. (2) of the main text: $m_{\mathrm{opt}} = \mathfrak{N}^{-1} \circ \mathrm{DNN}_{\mathrm{opt}} \circ \mathfrak{S}$.

The DNNs are implemented in PyTorch [97], where the weights and biases are optimized using the stochastic gradient-based optimizer Adam [98] to minimize the loss function [Eq. (B3)] over a series of training epochs. The learning rate is reduced by a fixed factor f_{r} if the loss $\mathcal{L}'_{\mathrm{MSE}}$ does not drop below a certain relative threshold value within a given number of epochs, referred to as "patience". Gradients are calculated using backpropagation. During training, weights and biases are updated batch-wise, i.e., during each epoch the entire training data is randomly split into batches of equal size. For each batch, the predictions and the resulting loss are calculated and the NN parameters are then updated accordingly. To incorporate configurations related through transformations of $p4m$ we use *online* data augmentation (see Sec. III of the

TABLE I. DNN hyperparameters employed in this paper. Here, the number of inputs n_{in} and outputs n_{out} of each fully-connected layer (FCL) is denoted as $(n_{\text{in}}, n_{\text{out}})$. The total number of NN parameters (weights and biases) is denoted as N_{tot}. Default settings are used except where explicitly stated.

Figure	2(b)	2(c)	3(a)
K	2048	2048	512
FCL 1	(2048,1024)	(2048,1024)	(512,256)
FCL 2	(1024,512)	(1024,512)	(256,64)
FCL 3	(512,512)	(512,256)	(64,1)
FCL 4	(512,256)	(256,2)	
FCL 5	(256,2)		
N_{tot}	3838722	3576066	353153
learning rate	0.001	0.0001	0.0001
batch size	700	7000	355
f_{r}	0.5	0.5	0.5
patience	50	50	50
epochs	1576	1485	770

main text). The DNN hyperparameters employed in this paper are collected in Table I.

3. Vector field

Figure 6 shows the vector field $\boldsymbol{\delta p} = \hat{\boldsymbol{p}} - \boldsymbol{p}$ whose divergence signal is shown in Figs. 2(b) and 2(c) in the main text and serves as an indicator for phase transitions. All vectors in the vector field are horizontal; this demonstrates that ρ is predicted with near-perfect accuracy. The largest stability regions identified based on the prediction-based method, i.e., the connected regions with large $\|\boldsymbol{\delta p}\|$, roughly coincide with the three main stability regions of the FKM where segregated (1), diagonal (2), and axial ordering (3) are prevalent, see sketched phase diagram displayed in Fig. 2(a) in the main text. These regions are particularly well highlighted in the noisy case [Fig. 6(b)], whereas it is more difficult to identify these regions based on the vector-field divergence signal [Fig. 2(c) in the main text]. These discrepancies may be resolved by using an alternative indicator derived from the vector-field, as has already been suggested in Ref. [21].

4. Optimal predictive model in noise-free case

In the main text (see Sec. III B), we have derived the optimal divergence signal in the noise-free case where the parameter space is divided into regions along U (at $\rho = \text{const.}$) with distinct input data. We restricted our analysis to regions that contain at least two grid points. In the following, we call such regions large (L) to distinguish them from small (S) regions that only contain a single grid point. In general, the parameter space in the noise-free case consists of regions along U of both types. Note that Eqs. (12) and (13) derived in the main text still hold for all points in large regions, even in the presence of small regions. Thus, we only need to derive analogous expressions for the single points within small regions. All possible types of region boundaries can be characterized by a three letter code $X_{\text{I}}X_{\text{II}}X_{\text{III}}$, where $X \in \{\text{L}, \text{S}\}$. Here, $X_{\text{I}}X_{\text{II}}X_{\text{III}}$ denotes the case where a region labeled I of type X_{I} is followed by a region labeled II of type X_{II} and a

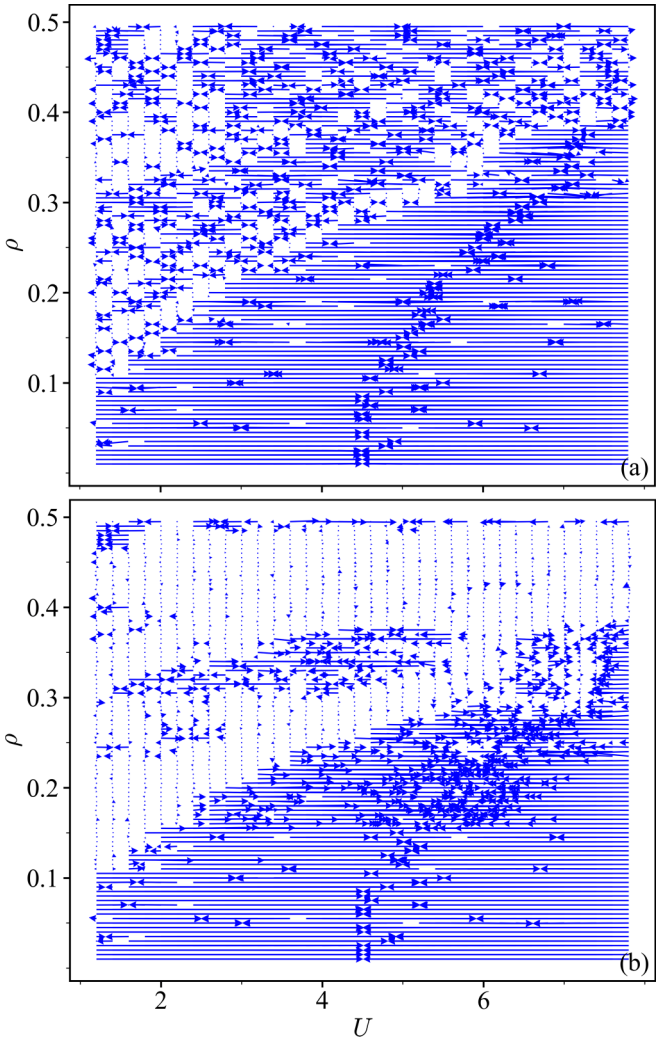

FIG. 6. Vector field $\boldsymbol{\delta p} = \hat{\boldsymbol{p}} - \boldsymbol{p}$ obtained using a DNN trained with $|\mathcal{F}|$ as input for the FKM phase diagram in (a) the noise-free and (b) noisy case [see Figs. 2(b) and 2(c) in the main text for the vector-field divergence signal, respectively].

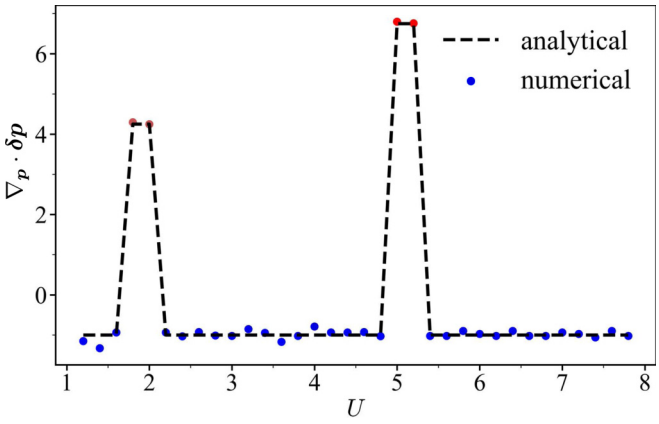

FIG. 7. Vector-field divergence $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ as a function of U at at fixed $\rho = 63/400 \approx 0.16$ obtained analytically based on Eqs. (12) and (13), as well as numerically using a DNN trained with $|\mathcal{F}|$ as input for the two-dimensional ground-state phase diagram in the noise-free case [see Fig. 2(b) in the main text for full phase diagram].

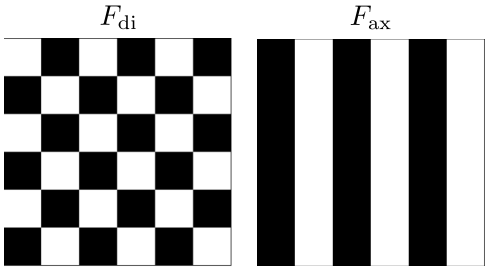

FIG. 8. Representative filters for computing the order parameter for diagonal (F_{di}) or axial (F_{ax}) orderings. Here, black denotes 1 and white denotes –1.

region labeled III of type X_{III} along U. The cases that remain to be analysed correspond to letter codes of the form $X_{\mathrm{I}}\mathrm{S}_{\mathrm{II}}X_{\mathrm{III}}$ with $X \in \{\mathrm{L}, \mathrm{S}\}$, i.e., {SSS, LSL, SSL, LSS}. The optimal divergence signal at the point that constitutes region II (S_{II}) is then given as

$$\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}_{\mathrm{opt}} \approx \frac{\langle U\rangle_{\mathrm{III}} - \langle U\rangle_{\mathrm{I}}}{2\Delta U} - 1 \geqslant 0. \tag{B4}$$

Figure 7 shows the vector-field divergence $\nabla_{\boldsymbol{p}} \cdot \boldsymbol{\delta p}$ as a function of U at $\rho = 35/400$ for the DNN trained using $|\mathcal{F}|$ as input in the noise-free case. The corresponding predicted two-dimensional ground-state phase diagram of the FKM is shown in Fig. 2(b) in the main text. The values of the divergence match the results in Eqs. (12) and (13) in the main text with near perfect accuracy. This confirms that our trained predictive model is indeed optimal, i.e., minimizes $\mathcal{L}_{\mathrm{MSE}}$ [Eq. (2) in the main text].

APPENDIX C: ORDER PARAMETER ANALYSIS

Here, we discuss order parameters for segregated, diagonal and axial orderings [cf. labels (1)–(3) in Fig. 2 in the main text]. To define order parameters for the diagonal (di), as well as axial (ax) orderings, we introduce appropriate filters F_ξ, $\xi \in \{\mathrm{di}, \mathrm{ax}\}$. The values of the order parameters are obtained by taking the Frobenius scalar product of the raw configurations $\boldsymbol{w}$ with the corresponding filters. To account for configurations that are related through transformations of $p4m$, we also subject the filters to the corresponding transformations. Ultimately, we take the maximum value over all symmetry-related filters $\{F_\xi\}$ as the value of the order parameter O_ξ for a particular configuration sample $\boldsymbol{w}$:

$$O_\xi(\boldsymbol{w}) = \max_{F_\xi} \frac{1}{L^2} \sum_{i,j=0}^{L-1} (F_\xi \odot (2\boldsymbol{w} - \mathbb{1}))_{ij}, \tag{C1}$$

where $\odot$ denotes the element-wise product and $\mathbb{1}$ is the identity matrix. Figure 8 displays representative filters for the order parameters of diagonal and axial orderings, where all other filters can be obtained from these examples through transformations of $p4m$. Note that the filters have the same size as the configurations they are applied to, here $L = 20$. The filters for lattices of different size can be defined analogously by retaining the same patterns as in Fig. 8. If necessary, order parameters for other orderings [cf. labels (4)–(9) in Fig. 2 in the main text] can be defined in a similar manner.

Defining an order parameter for the segregated (sg) ordering is conceptually simple. It amounts to determining whether the configuration sample contains a single, connected cluster of f particles. This is implemented by a backtracking algorithm [99]. We define a binary order parameter O_{sg} taking on the value 1(0) if the configuration sample does(not)

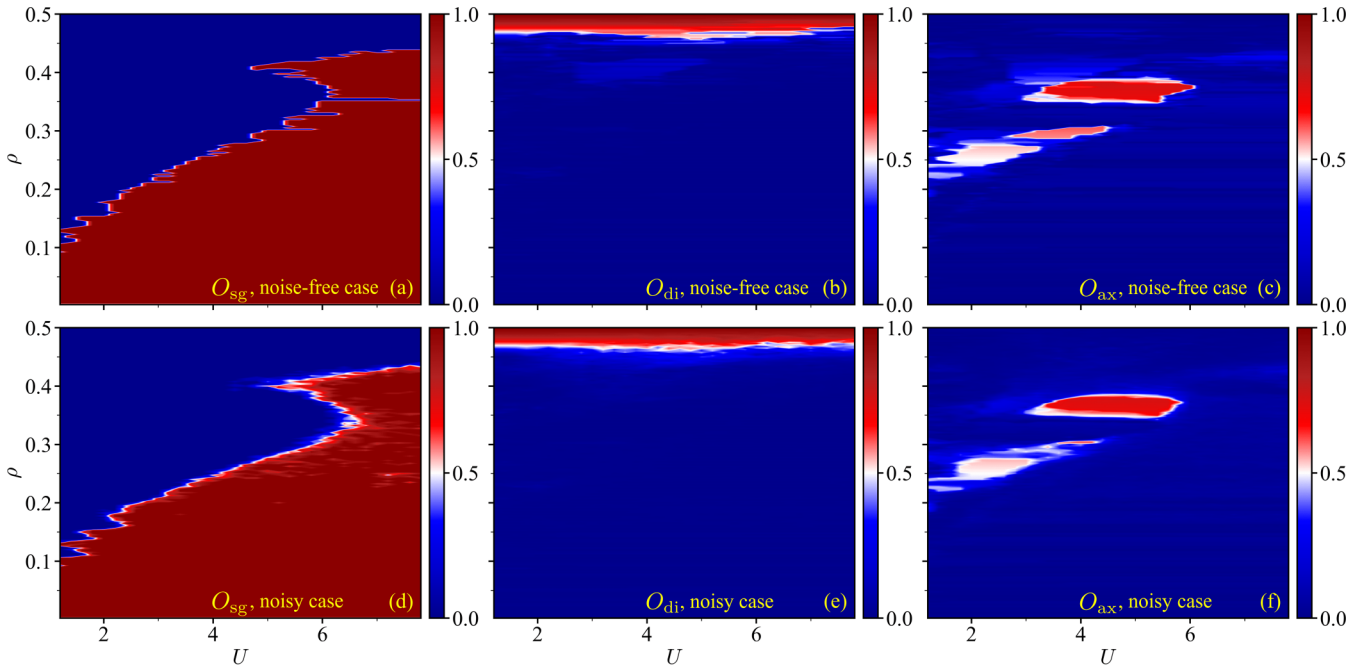

FIG. 9. Values of order parameters [(a), (d)] O_{sg}, [(b), (e)] O_{ax}, and [(c), (f)] O_{di} on the two-dimensional parameter space of the FKM in the (a)–(c) noise-free and (d)–(f) noisy case.

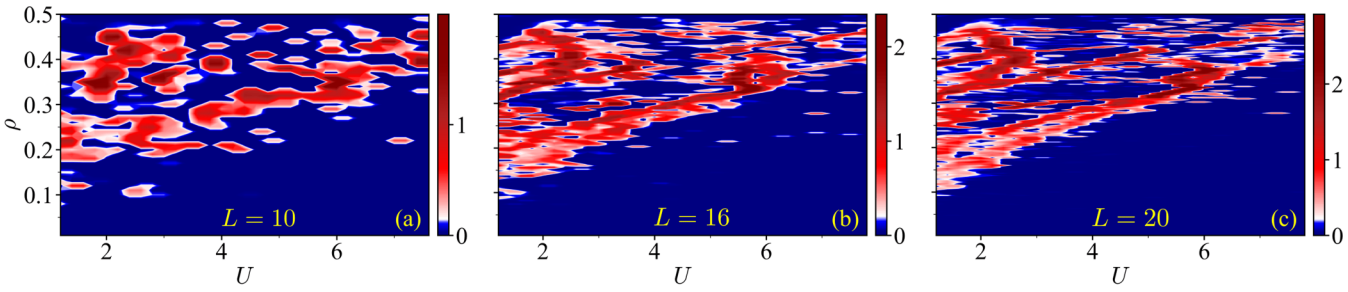

FIG. 10. Indicator $\Delta\bar{x}$ [Eq. (16) in the main text] of the mean-based method based on $\boldsymbol{\kappa}$ as input on the two-dimensional parameter space of the FKM in the noise-free case for two-dimensional square lattices of linear size (a) $L = 10$, (b) $L = 16$, and (c) $L = 20$. The density of heavy particles ρ ranges from $1/L^2$ to half-filling ($\Delta\rho = 1/L^2$) and U ranges from 1 to 8 ($\Delta U = 0.2$). This allows for an assessment of finite-size effects on the FKM phase diagram. In particular, we observe that the phase diagram is converging with increasing lattice size but still displays some finite-size effects at $L = 20$.

contains a single, connected cluster of f particles (as determined by the algorithm).

Clearly, all three order parameters O_ξ, $\xi \in \{\mathrm{di}, \mathrm{ax}, \mathrm{sg}\}$ share a common set of desired properties [100]. In particular, the order parameters remain unchanged when the input configuration samples are subjected to transformations of $p4m$. Furthermore, the maximum value of the order parameters is $\max_{\boldsymbol{w}} O_\xi(\boldsymbol{w}) = 1$, which is only achieved for samples showing perfect ordering of type ξ. Additionally, the three order parameters defined by means of filters can take on values ranging from 0 to 1, indicating the partial presence of the corresponding pattern.

Figure 9 shows the values of all three order parameters O_ξ, $\xi \in \{\mathrm{di}, \mathrm{ax}, \mathrm{sg}\}$ for each sampled point $\boldsymbol{p} = (U, \rho)$ in parameter space for the FKM in the noise-free and noisy case. In the noisy case, we average the value of a given order parameter over all available configurations at each point $\boldsymbol{p}$ to recover a scalar quantity. The order parameters reveal the presence of segregated, diagonal, and axial orderings marked as (1), (2), and (3) in Fig. 2 in the main text, respectively.

APPENDIX D: DETAILS ON ALTERNATIVE PHASE CLASSIFICATION METHODS

In Sec. IV of the main text we discussed alternative phase classification methods, which reproduce the results of the prediction-based method in the noise-free case. Such methods simply need to detect changes in neighboring configurations (up to transformations of $p4m$) in U (at $\rho = \text{const.}$). Here, we provide further details on the two approaches that were discussed to detect such changes.

In a naïve first approach, one searches for an appropriate symmetry transformations that relates neighboring configurations. That is, one compares the ground-state configuration samples of two neighboring points in parameter space along U. If a symmetry transformation is found that relates the configuration samples, the corresponding points belong to the same phase. Otherwise, a new phase is declared. Clearly, the computational complexity of such an approach is reduced significantly by using $|\mathcal{F}|$ instead of $\boldsymbol{w}_0$, because the Fourier representation removes the need to consider lattice translations.

In the second approach that is motivated by the simulated annealing procedure, we propose to use the system Hamiltonian. For a given point in parameter space (point I), we take the corresponding ground-state configuration sample and calculate its energy using the system Hamiltonian at a neighboring point along U (point II). Additionally, we evaluate the energy of the ground-state configuration sample at point II

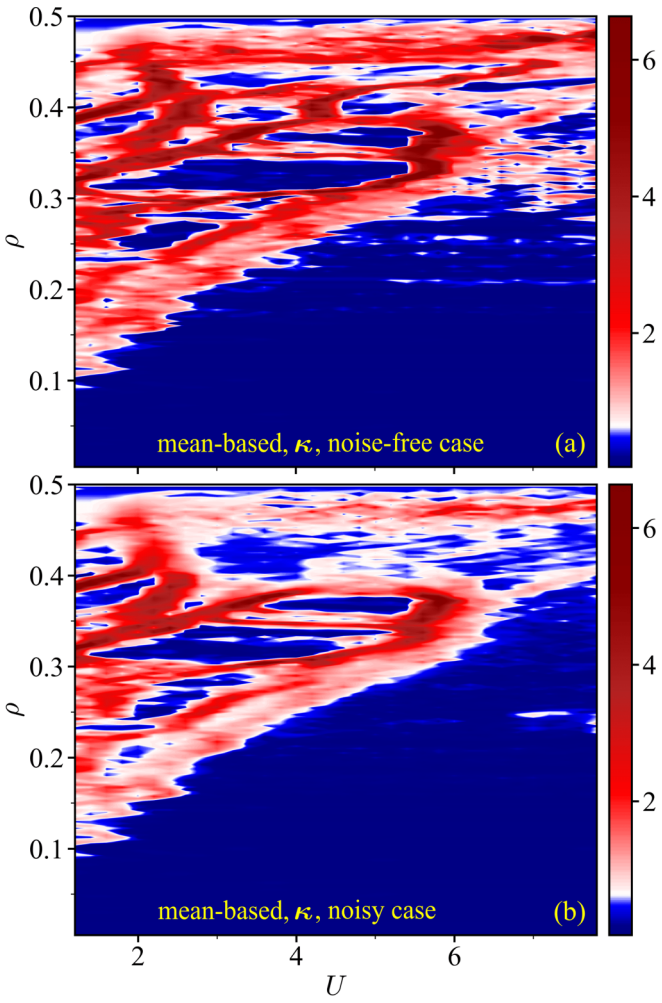

FIG. 11. Parameter-generic indicator $\Delta\bar{x}_{\mathrm{tot}}$ [Eq. (17) in the main text] of the mean-based method based on $\boldsymbol{\kappa}$ as input on the two-dimensional parameter space of the FKM in the (a) noise-free and (b) noisy case.

using the system Hamiltonian at point II. If the difference in energy is smaller than an appropriate threshold value, we can consider the two samples to be degenerate and assign them to the same phase. This is valid, since both samples are equally likely to be generated using the simulated annealing procedure. Otherwise, a new phase is declared. Note that an extension of these two approaches to the general noisy case may not be straightforward.

APPENDIX E: FINITE-SIZE SCALING

In this Appendix, we discuss the effect of a finite lattice size on the FKM phase diagram. Figure 10 displays the indicator [Eq. (16) in the main text] based on κ as input on the two-dimensional parameter space of the FKM in the noise-free case for square lattices of linear size $L = 10, 16, 20$. These results show that the phase boundaries become sharper with increasing lattice size. Hence, the corresponding stability regions become more defined. In particular, most of the largest stability regions discussed in the main text are stable both at $L = 16$ and $L = 20$. This points to the fast convergence of the phase diagram. Note, however, that the phase diagram at $L = 20$ still displays finite-size effects. An extraction of a sharper and more detailed phase diagram that truthfully reflects the expected complexity of the thermodynamic limit result would require an investigation of even larger lattices. This, however, goes beyond the scope of our paper.

APPENDIX F: DETAILS ON THE PARAMETER-GENERIC MEAN-BASED METHOD

In Eq. (17) of the main text, we have extended the indicator of the mean-based method to measure both changes along U and ρ. Figure 11 shows the FKM phase diagram obtained using this indicator with κ as input both in the noise-free and noisy case. The resulting phase diagrams closely match the results shown in Figs. 2(e) and 2(f) obtained with the correlation indicator in Eq. (15) of the main text that is only sensitive to changes in the input along U.

[1] S. Sachdev, *Quantum Phase Transitions* (Cambridge University Press, Cambridge, 2011).

[2] N. Goldenfeld, *Lectures On Phase Transitions and The Renormalization Group* (CRC Press, Boca Raton, FL, 2018).

[3] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[4] J. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, Oxford, 2006).

[5] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 1995).

[6] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[7] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).

[8] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Machine Learning Phases of Strongly Correlated Fermions, Phys. Rev. X **7**, 031038 (2017).

[9] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).

[10] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, Identifying quantum phase transitions using artificial neural networks on experimental data, Nat. Phys. **15**, 917 (2019).

[11] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap, Classifying snapshots of the doped Hubbard model with machine learning, Nat. Phys. **15**, 921 (2019).

[12] V. Dunjko and H. J. Briegel, Machine learning & artificial intelligence in the quantum domain: A review of recent progress, Rep. Prog. Phys. **81**, 074001 (2018).

[13] T. Ohtsuki and T. Ohtsuki, Deep learning the quantum phase transitions in random electron systems: Applications to three dimensions, J. Phys. Soc. Jpn. **86**, 044708 (2017).

[14] J. Carrasquilla, Machine learning for quantum matter, Adv. Phys. X **5**, 1797528 (2020).

[15] A. Bohrdt, S. Kim, A. Lukin, M. Rispoli, R. Schittko, M. Knap, M. Greiner, and J. Léonard, Analyzing non-equilibrium quantum states through snapshots with artificial neural networks, arXiv:2012.11586.

[16] S. J. Wetzel, Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders, Phys. Rev. E **96**, 022140 (2017).

[17] Y.-H. Liu and E. P. L. van Nieuwenburg, Discriminative Cooperative Networks for Detecting Phase Transitions, Phys. Rev. Lett. **120**, 176401 (2018).

[18] P. Huembeli, A. Dauphin, and P. Wittek, Identifying quantum phase transitions with adversarial neural networks, Phys. Rev. B **97**, 134109 (2018).

[19] J. F. Rodriguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, Nat. Phys. **15**, 790 (2019).

[20] K. Liu, J. Greitemann, and L. Pollet, Learning multiple order parameters with interpretable machines, Phys. Rev. B **99**, 104410 (2019).

[21] F. Schäfer and N. Lörch, Vector field divergence of predictive model output as indication of phase transitions, Phys. Rev. E **99**, 062107 (2019).

[22] E. Greplova, A. Valenti, G. Boschung, F. Schäfer, N. Lörch, and S. D. Huber, Unsupervised identification of topological phase transitions using predictive models, New J. Phys. **22**, 045003 (2020).

[23] Y. Che, C. Gneiting, T. Liu, and F. Nori, Topological quantum phase transitions retrieved through unsupervised machine learning, Phys. Rev. B **102**, 134213 (2020).

[24] M. S. Scheurer and R.-J. Slager, Unsupervised Machine Learning and Band Topology, Phys. Rev. Lett. **124**, 226401 (2020).

[25] O. Balabanov and M. Granath, Unsupervised learning using topological data augmentation, Phys. Rev. Research **2**, 013354 (2020).

[26] Y. Long, J. Ren, and H. Chen, Unsupervised Manifold Clustering of Topological Phononics, Phys. Rev. Lett. **124**, 185501 (2020).

[27] N. Käming, A. Dawid, K. Kottmann, M. Lewenstein, K. Sengstock, A. Dauphin, and C. Weitenberg, Unsupervised machine learning of topological phase transitions from experimental data, Mach. Learn.: Sci. Technol. (2021), doi: 10.1088/2632-2153/abffe7.

[28] C. Casert, T. Vieijra, J. Nys, and J. Ryckebusch, Interpretable machine learning for inferring the phase boundaries in a nonequilibrium system, Phys. Rev. E **99**, 023304 (2019).

[29] S. Blücher, L. Kades, J. M. Pawlowski, N. Strodthoff, and J. M. Urban, Towards novel insights in lattice field theory with explainable machine learning, Phys. Rev. D **101**, 094507 (2020).

[30] Y. Zhang, P. Ginsparg, and E.-A. Kim, Interpreting machine learning of topological quantum phase transitions, Phys. Rev. Research **2**, 023283 (2020).

[31] A. Dawid, P. Huembeli, M. Tomza, M. Lewenstein, and A. Dauphin, Phase detection with neural networks: Interpreting the black box, New J. Phys. **22**, 115001 (2020).

[32] A. Cole, G. J. Loges, and G. Shiu, Quantitative and interpretable order parameters for phase transitions from persistent homology, arXiv:2009.14231.

[33] N. Rao, K. Liu, M. Machaczek, and L. Pollet, Machine-learned phase diagrams of generalized Kitaev honeycomb magnets, arXiv:2102.01103.

[34] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. **51**, 1 (2018).

[35] C. Molnar, Interpretable Machine Learning (2019) https://christophm.github.io/interpretable-ml-book/.

[36] J. Singh, M. S. Scheurer, and V. Arora, Conditional generative models for sampling and phase transition indication in spin systems (unpublished), https://scipost.org/submissions/scipost_202103_00010v1/.

[37] Y. Bengio and O. Delalleau, On the Expressive Power of Deep Architectures, in *Algorithmic Learning Theory*, edited by J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann (Springer, Berlin, 2011), pp. 18–36.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Boston, 2016).

[39] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, The expressive power of neural networks: A view from the width, in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), pp. 6232–6240.

[40] L. M. Falicov and J. C. Kimball, Simple Model for Semiconductor-Metal Transitions: SmB_6 and Transition-Metal Oxides, Phys. Rev. Lett. **22**, 997 (1969).

[41] J. Hubbard, Electron correlations in narrow energy bands, Proc. R. Soc. London, Sec. A **276**, 238 (1963).

[42] J. K. Freericks and V. Zlatic, Exact dynamical mean-field theory of the Falicov-Kimball model, Rev. Mod. Phys. **75**, 1333 (2003).

[43] M. Hohenadler and F. F. Assaad, Fractionalized Metal in a Falicov-Kimball Model, Phys. Rev. Lett. **121**, 086601 (2018).

[44] M. Gonçalves, P. Ribeiro, R. Mondaini, and E. V. Castro, Temperature-Driven Gapless Topological Insulator, Phys. Rev. Lett. **122**, 126601 (2019).

[45] M. Eckstein and M. Kollar, Nonthermal Steady States after an Interaction Quench in the Falicov-Kimball Model, Phys. Rev. Lett. **100**, 120404 (2008).

[46] M. M. Oliveira, P. Ribeiro, and S. Kirchner, Classical and Quantum Liquids Induced by Quantum Fluctuations, Phys. Rev. Lett. **122**, 197601 (2019).

[47] C. Prosko, S.-P. Lee, and J. Maciejko, Simple $\mathbb{Z}_2$ lattice gauge theories at finite fermion density, Phys. Rev. B **96**, 205104 (2017).

[48] A. Kauch, P. Pudleiner, K. Astleithner, P. Thunström, T. Ribic, and K. Held, Generic Optical Excitations of Correlated Systems: π-tons, Phys. Rev. Lett. **124**, 047401 (2020).

[49] J. K. Freericks, V. M. Turkowski, and V. Zlatić, Nonequilibrium Dynamical Mean-Field Theory, Phys. Rev. Lett. **97**, 266408 (2006).

[50] H. Aoki, N. Tsuji, M. Eckstein, M. Kollar, T. Oka, and P. Werner, Nonequilibrium dynamical mean-field theory and its applications, Rev. Mod. Phys. **86**, 779 (2014).

[51] T. Maier, M. Jarrell, T. Pruschke, and M. Hettler, Quantum cluster theories, Rev. Mod. Phys. **77**, 1027 (2005).

[52] V. Turkowski and J. K. Freericks, Nonequilibrium perturbation theory of the spinless Falicov-Kimball model: Second-order truncated expansion in U, Phys. Rev. B **75**, 125110 (2007).

[53] J. Kaye and D. Golež, Low rank compression in the numerical solution of the nonequilibrium Dyson equation, SciPost Phys. **10**, 91 (2021).

[54] L. Huang and L. Wang, Accelerated Monte Carlo simulations with restricted Boltzmann machines, Phys. Rev. B **95**, 035105 (2017).

[55] S. Zhang, P. Zhang, and G.-W. Chern, Anomalous phase separation and hidden coarsening of super-clusters in the Falicov-Kimball model, arXiv:2105.13304.

[56] R. Lemański, J. K. Freericks, and G. Banach, Stripe Phases in the Two-Dimensional Falicov-Kimball Model, Phys. Rev. Lett. **89**, 196403 (2002).

[57] R. Lemański, J. K. Freericks, and G. Banach, Charge stripes due to electron correlations in the two-dimensional spinless Falicov-Kimball model, J. Stat. Phys. **116**, 699 (2004).

[58] H. Cencarikova and P. Farkasovský, Formation of charge and spin ordering in strongly correlated electron systems, Condens. Matter Phys. **14**, 42701 (2011).

[59] M. Plischke, Coherent-Potential-Approximation Calculation on the Falicov-Kimball Model of the Metal-Insulator Transition, Phys. Rev. Lett. **28**, 361 (1972).

[60] P. Farkašovský, Falicov-Kimball model and the problem of valence and metal-insulator transitions, Phys. Rev. B **51**, 1507 (1995).

[61] P. Farkašovský, Pressure-induced insulator-metal transitions in the spinless Falicov-Kimball model, Phys. Rev. B **52**, R5463(R) (1995).

[62] P. Haldar, M. S. Laad, and S. R. Hassan, Universal dielectric response across a continuous metal-insulator transition, Phys. Rev. B **99**, 125147 (2019).

[63] M. M. Maśka, R. Lemański, J. K. Freericks, and C. J. Williams, Pattern Formation in Mixtures of Ultracold Atoms in Optical Lattices, Phys. Rev. Lett. **101**, 060404 (2008).

[64] M. M. Maśka, R. Lemański, C. J. Williams, and J. K. Freericks, Momentum distribution and ordering in mixtures of ultracold light- and heavy-fermion atoms, Phys. Rev. A **83**, 063631 (2011).

[65] A. Hu, M. M. Maśka, C. W. Clark, and J. K. Freericks, Robust finite-temperature disordered Mott-insulating phases in inhomogeneous Fermi-Fermi mixtures with density and mass imbalance, Phys. Rev. A **91**, 063624 (2015).

[66] T. Qin, A. Schnell, K. Sengstock, C. Weitenberg, A. Eckardt, and W. Hofstetter, Charge density wave and charge pump of interacting fermions in circularly shaken hexagonal optical lattices, Phys. Rev. A **98**, 033601 (2018).

[67] D. O. Maionchi, A. M. C. Souza, H. J. Herrmann, and R. N. da Costa Filho, Anderson localization on Falicov-Kimball model with next-nearest-neighbor hopping and long-range correlated disorder, Phys. Rev. B **77**, 245126 (2008).

[68] A. E. Antipov, Y. Javanmard, P. Ribeiro, and S. Kirchner, Interaction-Tuned Anderson versus Mott Localization, Phys. Rev. Lett. **117**, 146601 (2016).

[69] P. Haldar, M. S. Laad, and S. R. Hassan, Real-space cluster dynamical mean-field approach to the Falicov-Kimball model: An alloy-analogy approach, Phys. Rev. B **95**, 125116 (2017).

[70] A. Smith, J. Knolle, D. L. Kovrizhin, and R. Moessner, Disorder-Free Localization, Phys. Rev. Lett. **118**, 266601 (2017).

[71] M. Žonda, J. Okamoto, and M. Thoss, Gapless regime in the charge density wave phase of the finite dimensional Falicov-Kimball model, Phys. Rev. B **100**, 075124 (2019).

[72] T. Ribic, G. Rohringer, and K. Held, Nonlocal correlations and spectral properties of the Falicov-Kimball model, Phys. Rev. B **93**, 195105 (2016).

[73] T. Ribic, G. Rohringer, and K. Held, Local correlation functions of arbitrary order for the Falicov-Kimball model, Phys. Rev. B **95**, 155130 (2017).

[74] M. Eckstein, M. Kollar, and P. Werner, Thermalization after an Interaction Quench in the Hubbard Model, Phys. Rev. Lett. **103**, 056403 (2009).

[75] A. J. Herrmann, N. Tsuji, M. Eckstein, and P. Werner, Nonequilibrium dynamical cluster approximation study of the Falicov-Kimball model, Phys. Rev. B **94**, 245114 (2016).

[76] A. J. Herrmann, A. E. Antipov, and P. Werner, Spreading of correlations in the Falicov-Kimball model, Phys. Rev. B **97**, 165107 (2018).

[77] J. Freericks, *Transport in Multilayered Nanostructures: The Dynamical Mean-Field Theory Approach* (Imperial College Press, London, 2006), pp. 1–328.

[78] M. Žonda and M. Thoss, Nonequilibrium charge transport through Falicov-Kimball structures connected to metallic leads, Phys. Rev. B **99**, 155157 (2019).

[79] R. Smorka, M. Žonda, and M. Thoss, Electronic transport through correlated electron systems with nonhomogeneous charge orderings, Phys. Rev. B **101**, 155116 (2020).

[80] D. Schattschneider, The plane symmetry groups: Their recognition and notation, Am. Math. Mon. **85**, 439 (1978).

[81] W. L. Briggs and V. E. Henson, *The DFT: An Owner's Manual for the Discrete Fourier Transform* (SIAM, Philadelphia, 1995).

[82] J. W. Cooley and J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comput. **19**, 297 (1965).

[83] E. Maniv, R. A. Murphy, S. C. Haley, S. Doyle, C. John, A. Maniv, S. K. Ramakrishna, Y.-L. Tang, P. Ercius, R. Ramesh *et al.*, Exchange bias due to coupling between coexisting antiferromagnetic and spin-glass orders, Nat. Phys. **17**, 525 (2021).

[84] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, Phys. Rep. **810**, 1 (2019).

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[86] L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9**, 2579 (2008).

[87] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv:1802.03426.

[88] M. D. Petrović, B. S. Popescu, U. Bajpai, P. Plecháč, and B. K. Nikolić, Spin and charge pumping by a steady or pulse-current-driven magnetic domain wall: A self-consistent multiscale time-dependent quantum-classical hybrid approach, Phys. Rev. Applied **10**, 054038 (2018).

[89] X.-H. Li, Z. Chen, and T. K. Ng, Generalized Falicov-Kimball models, Phys. Rev. B **100**, 094519 (2019).

[90] A. W. v. d. Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 1998).

[91] J. Arnold, F. Schäfer, M. Žonda, and A. U. J. Lode, Interpretable and unsupervised phase classification (2020), https://github.com/arnoldjulian/Interpretable-and-unsupervised-phase-classification.

[92] M. M. Maśka and K. Czajka, Thermodynamics of the two-dimensional Falicov-Kimball model: A classical Monte Carlo study, Phys. Rev. B **74**, 035109 (2006).

[93] M.-T. Tran, Inhomogeneous phases in the Falicov-Kimball model: Dynamical mean-field approximation, Phys. Rev. B **73**, 205110 (2006).

[94] M. Žonda, Phase transitions in the Falicov-Kimball model away from half-filling, Phase Transit. **85**, 96 (2012).

[95] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Netw. **6**, 861 (1993).

[96] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, Efficient backprop, in *Neural Networks: Tricks of the Trade* (Springer, Berlin, 2012), pp. 9–48.

[97] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2019), pp. 8026–8037.

[98] D. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[99] D. L. Kreher and D. R. Stinson, Combinatorial algorithms: Generation, enumeration, and search, SIGACT News **30**, 33 (1999).

[100] P.-L. Chau and A. Hardwick, A new order parameter for tetrahedral configurations, Mol. Phys. **93**, 511 (1998).