Ordered ground state configurations of the asymmetric Wigner bilayer system—Revisited with unsupervised learning

Cite as: J. Chem. Phys. 159, 204112 (2023); doi: 10.1063/5.0166822
Submitted: 7 July 2023 • Accepted: 1 October 2023 •
Published Online: 29 November 2023

Benedikt Hartl,[1,2,a)] Marek Mihalkovič,[3] Ladislav Šamaj,[3] Martial Mazars,[4] Emmanuel Trizac,[4,5] and Gerhard Kahl[1]

AFFILIATIONS

[1] Institute for Theoretical Physics and Center for Computational Materials Science (CMS), TU Wien, Vienna, Austria
[2] Allen Discovery Center, Tufts University, Medford, Massachusetts 02155, USA
[3] Institute of Physics, Slovak Academy of Sciences, Bratislava, Slovakia
[4] Université Paris-Saclay, Université Paris-Saclay, CNRS, LPTMS, Orsay, France
[5] ENS de Lyon, 46 allée d'Italie, 69364 Lyon, France

[a)] **Author to whom correspondence should be addressed:** benedikt.hartl@tuwien.ac.at

ABSTRACT

We have reanalyzed the rich plethora of ground state configurations of the asymmetric Wigner bilayer system that we had recently published in a related diagram of states [Antlanger *et al.*, Phys. Rev. Lett. **117**, 118002 (2016)], comprising roughly 60 000 state points in the phase space spanned by the distance between the plates and the charge asymmetry parameter of the system. In contrast to this preceding contribution where the classification of the emerging structures was carried out "by hand," we have used for the present contribution machine learning concepts, notably based on a principal component analysis and a k-means clustering approach: using a 30-dimensional feature vector for each emerging structure (containing relevant information, such as the composition of the configuration as well as the most relevant order parameters), we were able to reanalyze these ground state configurations in a considerably more systematic and comprehensive manner than we could possibly do in the previously published classification scheme. Indeed, we were now able to identify new structures in previously unclassified regions of the parameter space and could considerably refine the previous classification scheme, thereby identifying a rich wealth of new emerging ground state configurations. Thorough consistency checks confirm the validity of the newly defined diagram of states.

I. INTRODUCTION

Scientists nowadays are often confronted with huge datasets, may it be images or written text of any kind, scattering data from particle detectors, geometric data of lattice structures (i.e., particle arrangements) generated by experiments, theoretical frameworks, or via computer simulations.[1] Analyzing such datasets is usually a task far from being trivial, especially in high-dimensional spaces: let a dataset, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, consists of N data elements (or, equivalently, data points), termed $\mathbf{x}_i$ (with $i = 1, \ldots, N$); then each data element can be viewed as a vector, $\mathbf{x}_i = \{x_1, \ldots, x_{N_f}\}$ (also referred to as feature vector), which contains a large number of so-called features, x_j (with $j = 1, \ldots, N_f$); examples for such features are the values of different pixels of an image, the different channels of the measurement of a particle collision experiment or the coordinates, orientations and/or order parameters of a particle arrangement of a (lattice) structure. Under certain circumstances, a few clear signals in the data (i.e., a few characteristic features, x_j, in the feature vectors, $\mathbf{x}_i$, of a data set $\mathbf{X}$) may allow us to categorize the elements of a dataset, for instance, into different structural families with well-defined, characteristic order parameters. However, the shear size of typical datasets and the often immense complexity of the involved features usually render a classification scheme intractable to be manually carried out by a human being. In such cases and in an effort to obtain a more comprehensive picture of the properties of the underlying data, in general, it is highly advisable to use methods from unsupervised machine learning.[1–5] For example, an approach based on neural network potentials was used for local structure detection in polymorphic systems[6] and dimensionality reduction techniques were successfully utilized for crystal classification.[7] Very recently, an unsupervised topological learning approach was proposed for identifying atomic structures[8] and crystal nucleation[9] without *a priori* knowledge of the underlying physical system.

In this contribution, we reconsider and reanalyze the diagram of ground state configurations (occurring at vanishing temperature) of the so-called asymmetric Wigner bilayer system. In such a system, point charges form on each of the oppositely charged, confining planar walls ($\mathcal{L}_1$ and $\mathcal{L}_2$) ordered particle configurations; the respective surface charges of the plates, σ_1 and σ_2, which are separated by a reduced, dimensionless distance η are not necessarily identical—hence, the system is termed asymmetric; the ratio of the surface charge densities is defined as $A = \sigma_2/\sigma_1$ (with $A \in [0, 1]$). The entire system is charge neutral. Depending on the location in the parameter space (spanned by A and η) the system assumes ground state configurations that are characterized by the composition of the system $x = N_2/N$ (with N_2 being the number of particles occupying $\mathcal{L}_2$). Thus, for a given pair (A, η), the ground state configuration is specified by the value of x and by the lattices formed on the two layers (e.g., in terms of the respective lattice vectors). In preceding contributions,[10–13] these ground state configurations were determined via suitably adapted optimization tools (based notably on evolutionary algorithms),[14–18] limiting—for numerical reasons—the number of particles per unit cell to $N = 40$. In a subsequent step, the emerging ground state configurations were classified in terms of the emerging structures, based on suitably defined order parameters.[19,20] This classification procedure, which was realized by "hand," has led to a highly intricate diagram of states, where in total 14 different structures could be identified.

This "manual" scheme is of course prone to fail when aiming at an exhaustive classification of the emerging structures. To overcome this drawback, we have reanalyzed in this contribution the available set of data by utilizing dimensional reduction and clustering algorithms to automatically collect the corresponding data elements into distinct subgroups that share similar (structural) features. By applying these basic tools to a complex set of structural data, we want to demystify the involved concepts of unsupervised machine learning. Even more we want to encourage interested readers to actively use these approaches[21] as a versatile toolbox for extensive analysis of structural data.

We start the related analysis by defining a suitable feature vector (with length N_f) that captures for each specific lattice structure the relevant information: in our case, we choose $N_\mathrm{f} = 30$ features; the related vector contains as elements the composition of the system, a selected choice of order parameters, and some information about the radial distribution function.[22,64,65] These feature vectors are vectors in the N_f-dimensional feature space. The ultimate aim of this procedure is to identify in this space "spatially" separated clusters that collect similar data elements, i.e., a specific cluster contains similar (structural) data elements.[23] Starting off with the feature vectors, we reduce in a first step the complexity of the problem via a so-called principal component analysis (PCA),[24,25] which maps in this dimensional reduction step the huge amount of data into a lower dimensional latent space representation, thereby capturing and preserving the relevant aspects (or features) of each data element, while discarding the rather irrelevant information. Based on this reduced information, we then start to sort the different structures into clusters via the k-means algorithm.[26–29]

Thus, in this contribution, we use a rather simple (and therefore presumably the most applied) form of unsupervised machine learning,[1] namely *clustering algorithms*,[1,30,31] in order to organize datasets of lattice structures into families of structures. In the language of clustering algorithms, the procedure of categorizing data elements via a suitably defined similarity measure between data points in the feature space into different clusters is usually denoted as *clustering* or *labeling*: each of the N elements of a dataset is labeled by an identifier, which assigns each of its element to one of the categories (or *clusters*) identified by the clustering algorithm.[32]

Of course one might argue that other, possibly more refined machine learning based algorithms are available (notably autoencoders[33,34]) that might have been used for our purpose. We justify our decision in a rather length discussion, which we have shifted to the supplementary material, Sec. III.

In a first step, we have applied the above-outlined two-step procedure to the case of the *symmetric* Wigner bilayer system, where the charge densities on $\mathcal{L}_1$ and $\mathcal{L}_2$ are equal. We find that the principal component analysis indicates that only a five-dimensional latent space is required. We recover—not surprisingly, as anticipated by the exact results[35,36]—that the emerging structures can be classified into the five well-known clusters, each representing one of the well-known ordered ground state structures. With this confirmation of our procedure in mind, we proceed to the *asymmetric* Wigner bilayer system, where the aforementioned "by hand" classification[11–13] has led to 14 structural classes. Using the same 30-dimensional feature vector the principal component analysis provides evidence that the feature space can be mapped into a nine-dimensional latent space. Based on this reduced representation, we then perform the k-means clustering step of the structural data. Eventually and performing numerous consistency checks of this classification scheme, we end up with a reliable classification of the structural data into 32 clusters, which (i) identifies new, so far unclassified structures and (ii) does not leave any white regions in the diagram of states. We thereby demonstrate that even very basic tools from unsupervised machine learning can be utilized as a successful classification scheme of unlabeled structural data that can even be considerably more reliable (and systematic) as when done "by hand."

This manuscript is organized as follows: in the subsequent section, we briefly summarize the essential features of the asymmetric Wigner bilayer system and outline how the energy of the ordered configurations can be evaluated with high numerical accuracy. Furthermore, we define in this section the order parameters that we have used to characterize the emerging structures; we introduce our machine learning-based methods of how to identify structural similarities in our unlabeled set of data of ordered structures: the principal component analysis and the k-means algorithm. In Sec. III, the discussion of the results starts with a brief discussion of the originally derived diagram of states and the specific steps of how the unsupervised clustering algorithm is applied to the Wigner bilayer system. We then discuss the emerging results with the previously known exact results of the *symmetric* Wigner bilayer and then proceed to the *asymmetric* case, where particular focus is laid on the emerging new insights. The body of the manuscript is closed with a conclusion. Five appendices and three supplementary material sections close the manuscript: they are dedicated to conceptual details as well as to in-depth discussions of particular features of the machine learning-based approaches and provide more detailed background information.

II. MODEL AND METHODS

A. Model

In the Wigner bilayer system, classical, negative point charges $q = -e$ (e being the elementary charge) are confined between two parallel plates ($\mathcal{L}_1$ and $\mathcal{L}_2$) that are separated by a distance d. The plates carry uniform, positive surface charge densities, $e\sigma_1$ and $e\sigma_2$, which are not necessarily equal. The total system is electro-neutral.

Being interested in the ground state configurations of the system, the energetically most stable configurations that the particles form at vanishing temperature T, we can rely on Earnshaw's theorem,[37] which states that the particles have to be located on the plates. For a schematic view of the setup, we refer to Fig. 1.

The system is thus characterized by two parameters:

(i) the ratio of the surface charge densities, A,

$$A = \frac{\sigma_2}{\sigma_1},$$

which—without loss of generality—can be assumed to lie within the interval $[0, 1]$; a Wigner bilayer system with $A \equiv 1$ is termed symmetric, while otherwise it is called asymmetric;

(ii) the distance between the plates, d, which—for convenience—is reformulated via a dimensionless parameter η, defined as

$$\eta = d\sqrt{\frac{\sigma_1 + \sigma_2}{2}}.$$

Assuming that N particles populate the unit cell, we denote by N_1 and N_2 the number of particles that populate $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively; obviously, $N = N_1 + N_2$ and we define the composition of the system x via $x = N_2/N$. The surface particle number density is defined by $\rho = (\sigma_1 + \sigma_2)/2$. Since $1/\sqrt{\rho}$ only sets the length scale in the system, its particular value is irrelevant for our ground-state problem. For simplicity, and without loss of generality, we set $\rho = 1$ in numerical calculations, i.e., $1/\sqrt{\rho}$ can be considered as the unit length of our investigations.

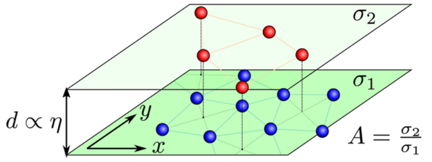

FIG. 1. Schematic view of the classical Wigner bilayer system: confined between two parallel plates, which are separated in the z-direction by the distance d, classical point charges (colored in blue and red) form ordered configurations on these plates (carrying homogeneous surface charge densities $e\sigma_1$ and $e\sigma_2$). The x- and y-directions of the Cartesian coordinate system are indicated.

B. Energy calculations

In the classical Wigner bilayer system, the point charges interact via the long-range Coulomb interaction with each other and with the uniformly charged plates. Furthermore, there is a distance-dependent, but otherwise, constant plate-to-plate interaction contributing to the total (internal) electrostatic energy of the unit cell of a bilayer structure, $E(\mathbf{r}^N;\ A, \eta)$, which is given in Gauss units by[11–13]

$$E(\mathbf{r}^N; A, \eta) = \sum_{i=1}^{N}\left[\sum_{j=1}^{N}\sum_{\mathbf{S_n}}{}^{*}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{S_n}|} - 2\pi e^2(\sigma_1 - \sigma_2)z_i\right] + \text{const.} \tag{1}$$

$\mathbf{r}^N = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is the set of position vectors $\mathbf{r}_i = (x_i, y_i, z_i)$ of the N point charges in the unit cell, $z_i = 0$ or $z_i = d$ ($\propto \eta$) specifies if particle i occupies $\mathcal{L}_1$ or $\mathcal{L}_2$; finally, $\mathbf{S_n}$ is a symbolic notation for periodic images of the unit cell in the x- and y-directions used to carry out the lattice summation.[38] In Eq. (1), the symbol $\sum^*$ indicates that for $\mathbf{S_n} = (0, 0, 0)$ the sum is carried out only for $j > i$ to avoid double counting within the unit cell (see Refs. 11–13 for details). For convenience, we chose the dielectric constant, ϵ, of the medium into which the particles are immersed, as well as the dielectric constant of the two plates, ϵ_1 and ϵ_2, to be equal; henceforward, we set this value to unity, i.e., $\epsilon = \epsilon_1 = \epsilon_2 = 1$. Following Ref. 11, we employ Ewald summation techniques[39]—specifically implemented for quasi-2D bilayer geometries[11,38]—to numerically evaluate the long-range electrostatic energy of the system in a highly reliable and computationally efficient manner.

Searching for a given pair of η and A for the global ground state configuration of the asymmetric Wigner bilayer system boils down to identifying simultaneously the correct number of particles per unit cell (N), to finding the optimal arrangement of the particles on the two plates, $\mathbf{r}^N$, and to identifying the correct unit cell geometry; from the positions, $\mathbf{r}^N$, one can extract N_1 and N_2 and thus the composition x. The components a_{11}, a_{21}, and a_{22} of the vectors $\mathbf{a}_1 = (a_{11}, 0, 0)$ and $\mathbf{a}_2 = (a_{21}, a_{22}, 0)$ that define the unit cell are subject to the structure optimization problem under the constraint of keeping the area of the unit cell, $S_0 = a_{11}a_{22}$, constant; the vector $\mathbf{a}_3 = (0, 0, d)$ is fixed by the plate separation distance d. With all this in mind, we minimize in our search for the ground state configuration the total energy per particle, $E(\mathbf{r}^N;\ A, \eta)/N$, as specified in Eq. (1).

Henceforward, we collect all these variational parameters via the following short-hand notation:

$$\mathcal{X} \equiv (\mathbf{r}^N, \mathbf{a}_1, \mathbf{a}_2). \tag{2}$$

In this spirit, we can also write $E(\mathbf{r}^N; A, \eta) \equiv E(\mathcal{X}; A, \eta)$ to parameterize the energy. If—at one occasion or the other—the particular values of A and η are not of relevance for the discussion, we then simply write $E(\mathcal{X})$ or even drop the argument of the energy completely, i.e., we simply use E.

The accuracy required for the evaluation of the energies of competing structures is tremendously high: anticipating that $E/(N\sqrt{\rho}e^2) \approx -1$, thus setting the energy of the system, we note that relative differences in the energies of competing structures down to the sevenths or eighths digit are quite frequent. In view of these accuracy requirements, the search for the ground state configurations thus becomes a very delicate optimization problem in a high dimensional search space; the first attempt to solve this challenge has been successfully carried out in Refs. 11–13 with the help of memetic evolutionary algorithms.

The computational cost for exploring the high-dimensional search space can be reduced since the energy, E/N, defined in Eq. (1), can be split into a (i) structure-dependent, but A-independent contribution and into an (ii) A-dependent, but structure-independent contribution. Following Refs. 11–13, we first define the reduced energy per particle as

$$\frac{E^*(\mathbf{r}^N; A, \eta)}{N} \equiv \frac{E(\mathbf{r}^N; A, \eta)}{N\sqrt{\rho}e^2}. \tag{3}$$

We then identify the structure-independent contribution to E^* as

$$\frac{E_A^*(A, \eta, x)}{N} = 2^{3/2}\pi\eta\frac{A}{(1+A)^2}[A - 2x(1+A)], \tag{4}$$

thus leading to

$$\frac{E^*(\mathbf{r}^N; A, \eta)}{N} = \frac{1}{N}\left[E^*(\mathbf{r}^N; A_0, \eta) - E_A^*(A_0, \eta, x) + E_A^*(A, \eta, x)\right] \tag{5}$$

with the reference asymmetry parameter A_0 (which, without loss of generality, we set to $A_0 = 0$) and the composition x introduced above. With this reformulation of E^*, the computational cost of the identification of ground state configurations can be substantially reduced but remains nevertheless quite high.

Using the above separation of the internal energy, the ground state configurations have been identified in Ref. 11 via independent evolutionary searches at a fixed value of $A(= 0)$ for different, numerically tractable values of the composition, x, on a fixed grid for the plate separation distance parameter η (≥ 0). The resulting set of structural ground state configurations obtained for different compositions, identified at $A = 0$ but for a particular value of η, provides all necessary information to identify subsequently the ground state configuration for any state point (η, A); in this manner, the approach becomes highly efficient. Limiting for computational reasons the number of particles per unit cell to $N = 40$, one can calculate[40] that at each value of η the total number of possible compositions is $N_{\rm tot} = 401$.[41] Further and following Ref. 11, we chose in our numerical analysis a uniform grid of $N_\eta = 141$ different values for $\eta \in [0, 1]$ (resulting in $N_{\rm tot} \times N_\eta = 56\,541$ different evolutionary optimized structures in total) and specify a uniform grid of $N_A = 201$ values for the asymmetry parameter $A \in [0, 1]$. The numerical values for the grid in η and A are thus $\Delta\eta = 10^{-2}/\sqrt{2}$ and $\Delta A = 5 \times 10^{-3}$.

For given values of η and A, the configuration that minimizes $E^*(\mathcal{X})/N$ is considered as the related ground state configuration and we denote the ground state energy as $E_{\rm GS}^*(A, \eta)/N$. Henceforward, we usually drop the arguments of the energy—unless we want to emphasize its dependency on certain arguments—and we synonymously use E^*/N for the expression given in Eq. (5) and $E_{\rm GS}^*/N$ for the ground state energy, respectively.

C. Order parameters

The identification of the ordered ground state configurations was based in Refs. 11–13. In this contribution, we specifically make use of the composition of the system, x, and on bond-orientational order parameters (BOOPs) $\Psi_\nu = \Psi_\nu(\mathcal{X})$. In their most elementary version, these parameters are defined as

$$\Psi_\nu(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{\mathcal{N}_i}\sum_{j=1}^{\mathcal{N}_i}\exp[\imath\nu\phi_{ij}]\right|. \tag{6}$$

For a tagged particle with index i, the angles ϕ_{ij} are enclosed by the bond of particle i to one of its $\mathcal{N}_i$ neighboring particles j and some reference axis $\hat{\mathbf{e}}_{\rm ref}$; thus, the angles ϕ_{ij} are given by $\cos\phi_{ij} = \hat{\mathbf{r}}_{ij}\cdot\hat{\mathbf{e}}_{\rm ref}$, with $\hat{\mathbf{r}}_{ij} = (\mathbf{r}_j - \mathbf{r}_i)/|\mathbf{r}_j - \mathbf{r}_i|$. The neighbors are identified via a standard Voronoi construction[42,43] using an open-source software package.[44]

The orientational symmetry of the neighborhood of particle i is characterized by the (integer) variable ν: the ν-fold BOOP $\Psi_\nu(\mathcal{X})$ assumes the value one if the angles between neighbors are multiples of $2\pi/\nu$ and attain values close to zero for a disordered particle arrangement or if there is no ν-fold symmetry.

Due to small, inherent numerical inaccuracies, the lattices that we deal with are never perfect; consequently, the exact number of nearest neighbors can be strongly influenced by minute changes in the particle positions, making the actual evaluation of the BOOPs numerically unstable. In an effort to guarantee better numerical stability in the evaluations of BOOPs, we modify via a simple remedy the BOOPs specified in Eq. (6) by including a weight factor that is related to the polygon side length, l_{ij}, that neighboring particles (with indices i and j) share[45]

$$\Psi_\nu(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{L_i}\sum_{j=1}^{\mathcal{N}_i} l_{ij}\exp[\imath\nu\phi_{ij}]\right|, \tag{7}$$

where $L_i = \sum_{j=1}^{\mathcal{N}_i} l_{ij}$; the l_{ij} are again extracted from the Voronoi construction.

Henceforward, we usually drop the argument, $\mathcal{X}$, and often simply write $\Psi_\nu \equiv \Psi_\nu(\mathcal{X})$, unless the explicit indication of a particular realization of a structure, $\mathcal{X}$, is required for the discussion.

The above definition of the BOOPs has been extended by considering separately particles in different layers (or combinations thereof). Following Refs. 11–13, we use the following four variations of BOOPs:

(1) $\Psi_\nu^{(1)}$, quantifying ν-fold bond-orientational order of particles in $\mathcal{L}_1$;

(2) $\Psi_\nu^{(2)}$, quantifying ν-fold bond-orientational order of particles in $\mathcal{L}_2$;

(3) $\Psi_\nu^{(3)}$, quantifying ν-fold bond-orientational order of particles of both layers projected onto the same plane;

(4) $\Psi_\nu^{(4)}$, quantifying ν-fold bond-orientational order of particles of layer two, considering only layer one particles (projected onto the same layer) as neighbors.

For convenience, we introduce here a short-hand notation for addressing a set of $u_i = u_1, \ldots, u_n$ different BOOPs, $\Psi_{\nu_j}^{(u_i)}$, of different bond-orientational order, $\nu_j = \nu_1, \ldots, \nu_m$, via $\Psi_{[\nu_1,\ldots,\nu_m]}^{(u_1,\ldots,u_n)} = \{\Psi_{\nu_1}^{(u_1)}, \ldots, \Psi_{\nu_m}^{(u_1)}, \Psi_{\nu_1}^{(u_2)}, \ldots, \Psi_{\nu_m}^{(u_2)}, \ldots, \Psi_{\nu_1}^{(u_n)}, \ldots, \Psi_{\nu_m}^{(u_n)}\}$; if only one lower index is used, e.g., $[\nu]$, the lower brackets may also be omitted and we may write $\Psi_\nu^{(u_1,\ldots,u_n)} = \{\Psi_\nu^{(u_1)}, \ldots, \Psi_\nu^{(u_n)}\}$.

D. Identifying similarities in unlabeled data: Clustering algorithms

1. General remarks

In this contribution, we are ultimately interested to analyze systematically the ground state configurations of the asymmetric Wigner bilayer system in terms of families (or clusters) of similar (or related) structures, with the configurations being characterized notably via their composition and their order parameters. We emphasize that there is no prior information in our data set that relates the different data elements (i.e., the different structures) to certain crystalline phases; thus, our task is to classify unlabeled data. In an effort to achieve the above goal, we have used clustering algorithm techniques from unsupervised learning that follow the objective of (automatically) grouping the elements of huge data sets into distinct subgroups, i.e., clusters, that share similar features, and have thus received a rapidly growing share of interest in recent years.

We introduce a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ that consists of $i = 1, \ldots, N$ data elements $\mathbf{x}_i$. Each such feature vector $\mathbf{x}_i$ may contain a large number (N_f) of features x_i, i.e., $\mathbf{x}_i = \{x_1, \ldots, x_{N_\mathrm{f}}\}$; the $\mathbf{x}_i$ are elements in the N_f-dimensional feature space. Thus, $\mathbf{X}$ can also be viewed as an $(N \times N_\mathrm{f})$-matrix. The huge size and the complexity of the features of typical data sets stored in $\mathbf{X}$ usually render a conventional classification scheme of these data in terms of families or clusters of data intractable to be manually carried out by a human being. To obtain a more comprehensive picture of the properties of the underlying data, it has turned out to be advantageous to involve concepts from unsupervised machine learning.[1–5]

Here, we use the possibly simplest form of unsupervised machine learning, namely (unsupervised) clustering algorithms;[1,30,31] we will show that such an approach turns out to be very helpful to organize data sets of lattice structures of the classical Wigner bilayer system into families of structures in an unsupervised manner, i.e., without prior knowledge about structure- or feature-dependent relations in our unlabeled data set.

We will introduce the two basic concepts that help us to implement our approach: (i) the principal component analysis (PCA) and (ii) the k-means clustering. More specifically, we utilize PCA to identify the characteristic features in our data set in an effort to effectively reduce the dimensionality of the features for the subsequent k-means clustering with the aim to improve the performance and stability of the latter. As our data set builds upon order parameters, the particular choice of dimensional reduction and clustering tools[46] allows us to directly provide physical insight via the characteristic features and the identified phases (cf., supplementary material, Secs. I and II).

2. Principal component analysis (PCA)

In an effort to reduce the dimensionality of the N_f-dimensional feature space to a considerably smaller, N_ℓ-dimensional latent space, we use the PCA. Thus, we transform a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{N \times N_\mathrm{f}}$ into a low-dimensional latent space representation, $\mathbf{L} = \mathbb{P}_{XL}(\mathbf{X}) = \{\mathbf{l}_1, \ldots, \mathbf{l}_N\} \in \mathbb{R}^{N \times N_\ell}$; the $\mathbf{l}_i \in \mathbb{R}^{N_\ell}$ are the latent space representations of the $\mathbf{x}_i$. While the mapping $\mathbb{P}_{XL}(\mathbf{X})$ is not bijective for $N_\ell < N_\mathrm{f}$, it is crucial that the low-dimensional representation of the data in $\mathbf{L}$ is able to address the essential correlations of the features of the original $\mathbf{X}$. In short, local structures in the feature space representation should be conserved as good as possible in the latent space representation, a requirement that is fulfilled by the popular PCA.[1]

In describing the PCA, we first assume—without loss of generality—zero empirical mean and unit variance of the $\mathbf{x}_i$ along the columns of $\mathbf{X}$. We consider the data set $\mathbf{X}$ as an $(N \times N_\mathrm{f})$ "design" matrix, whose rows are the N data points and whose columns are the N_f features and then construct the $(N_\mathrm{f} \times N_\mathrm{f})$, symmetric, positive-semidefinite covariance matrix, $\boldsymbol{\Sigma}(\mathbf{X})$, defined via

$$\boldsymbol{\Sigma}(\mathbf{X}) = \frac{1}{(N-1)} \mathbf{X}^\mathrm{T} \mathbf{X}, \tag{8}$$

where superscript "T" denotes the transpose of the matrix. The diagonal elements of $\boldsymbol{\Sigma}(\mathbf{X})$ measure the variance of features and the off-diagonal elements measure the covariance between features i and j. $\boldsymbol{\Sigma}(\mathbf{X})$ can be diagonalized as follows:

$$\boldsymbol{\Sigma}(\mathbf{X}) = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T. \tag{9}$$

It is assumed that the (real-valued, positive) eigenvalues λ_i, $i = 1, \ldots, N_\mathrm{f}$ of the diagonal matrix $\boldsymbol{\Lambda}$ are arranged in descending order. The related eigenvectors are denoted by $\mathbf{v}_i$.

$\boldsymbol{\Lambda}$ can now be used for dimensional reduction: large values of λ_i label along the associated eigenvectors $\mathbf{v}_i$ directions of high variance in the feature space, i.e., those directions that contain the relevant information of the data. In contrast, directions associated with small values of λ_i are usually related to noise and can potentially be ignored. The eigenvector $\mathbf{v}_i$ with the largest (second largest, ...) eigenvalue λ_i is referred to as the first (second, ...) principal component,[1] denoted in the following as PC (hence the term "principal component analysis").

Often, only very few of the λ_i have a significant value. Selecting the N_ℓ largest eigenvalues and the associated eigenvectors provides us with an effective way to project the original data points into a low-dimensional (but representative) latent space $\mathbf{L} = \mathbb{P}_{XL}(\mathbf{X})$, where the transformation $\mathbb{P}_{XL}$ is simply a linear projection from $\mathbb{R}^{N_f}$ onto $\mathbb{R}^{N_\ell}$. To quantify the amount of information encoded in each PC direction $\mathbf{v}_i$, we rely on the percentage of the explained variance (PEV), defined as

$$\lambda_i^{(\mathrm{e})} = \left(\sum_{j=1}^{N_\mathrm{f}} \lambda_j \right)^{-1} \lambda_i. \tag{10}$$

For a more comprehensive discussion of the PCA, see, e.g., Refs. 3, 4, and 47.

3. k-means clustering

Probably, the simplest form of unsupervised learning is clustering algorithms, whose objective is to identify groups in unlabeled data according to similarity or distance measures of one kind or another.[1,30,31] In the following, we introduce the k-means algorithm,[26–29] which we have used for our problem.

Starting again from N data points, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, in an N_f-dimensional feature space, $\mathbf{x}_i \in \mathbb{R}^{N_\mathrm{f}}$, the objective is to distribute a certain number of K cluster centers, called the cluster means $\mathbf{K} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$ with $\boldsymbol{\mu}_k \in \mathbb{R}^{N_\mathrm{f}}$, in the feature space, such that

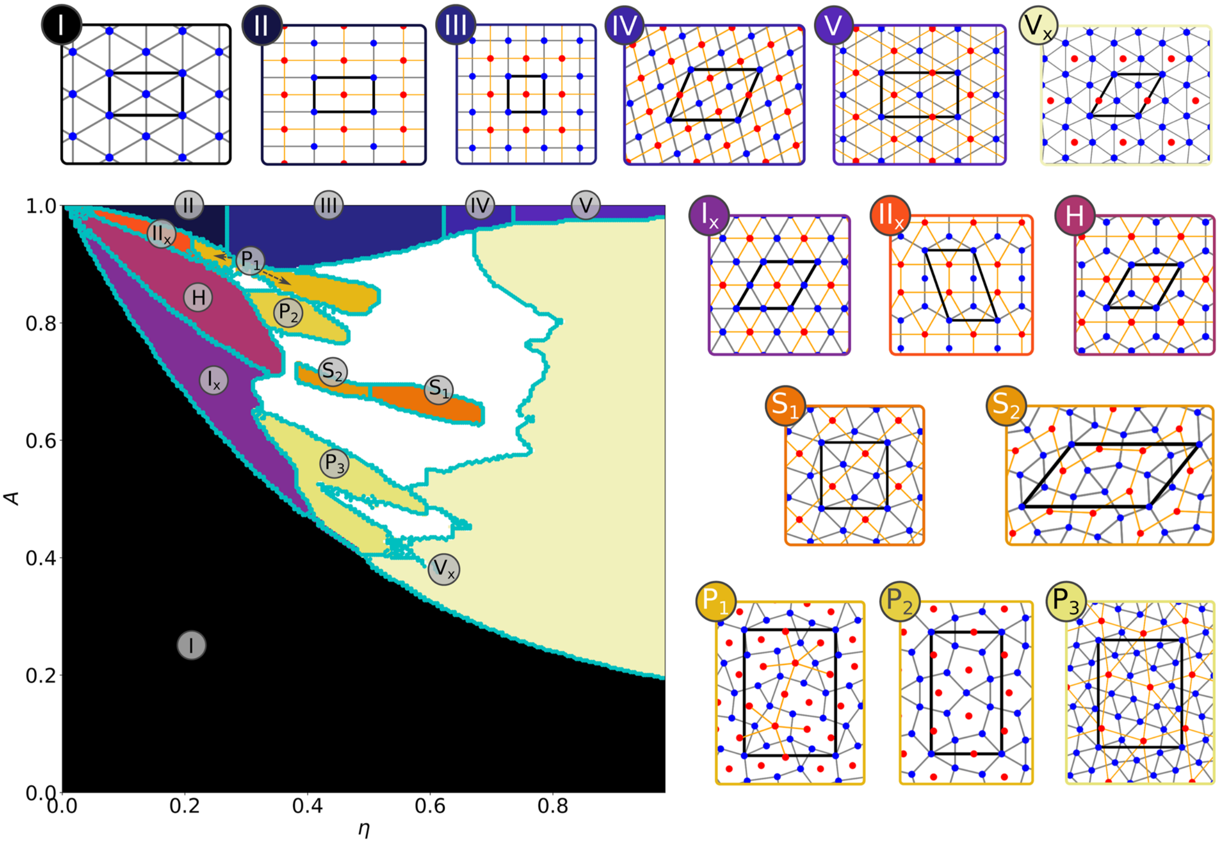

FIG. 2. Diagram of states of the ground state configurations of the asymmetric Wigner bilayer system in the (A, η)-plane, redrawn from Refs. 11–13. Colored and labeled regions denote the 14 phases that have been identified "by hand" according to the specifications summarized in Table I. The corresponding structures are shown in separate panels, specified by the same labeling and the same color code. In these panels, layer one $(\mathcal{L}_1)$ particles are colored in blue and layer two $(\mathcal{L}_2)$ particles are colored red; the thick black frames indicate the unit cells of the bilayer structures. The cyan lines in the main panel highlight the phase boundaries. Structures within the white region have not been classified, yet.

data points assigned to the different clusters minimize the following cost function:

$$\mathcal{C}(\mathbf{X}, \mathbf{K}) = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^2. \tag{11}$$

In this relation, the assignment of data point i to cluster k is realized via the binary variable $r_{ik} = 1$ (and $r_{ik'} = 0$ for all $k' \neq k$). $\sum_{i=1}^{N} r_{ik} = N_k$ defines the size of cluster k, i.e., the number of data points associated with it. The set of assignments $\mathbf{k} = \{r_{ik}\}$ is also called labeling or clustering of the data points.

Minimizing Eq. (11) can be interpreted as finding $\mathbf{K}$ and assigning via the r_{ik} the N data points to different clusters, k, such that the (scaled) variance of each cluster, $\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^2$, is minimized. In practice, this task is performed in a two-step procedure:[1]

1. Equation (11) is minimized with respect to $\boldsymbol{\mu}_k$ given a set of assignments $\{r_{ik}\}$, i.e., $(\partial \mathcal{C} / \partial \boldsymbol{\mu}_k)|_{\{r_{ik}\}} = 0$, yielding the update rule for $\boldsymbol{\mu}_k = N_k^{-1} \sum_{i=1}^{N} r_{ik} \mathbf{x}_i$; thus, $\boldsymbol{\mu}_k$ is the geometric center of the members $r_{ik} \mathbf{x}_i$ of cluster k;
2. given the cluster means $\mathbf{K}$, we want to find the assignments $\mathbf{k} = \{r_{ik}\}$ that minimize Eq. (11) by assigning each data point to its nearest cluster-mean: $r_{ik} = 1$ if $k = \arg\left[\min_{k'} (\mathbf{x}_i - \boldsymbol{\mu}_{k'})^2\right]$ and $r_{ik} = 0$ otherwise.

These two steps are performed in an alternating manner until some convergence criterion is met: this can, for instance, be the case if the change of the object function, given by Eq. (11), between two iteration steps, is smaller than a predefined threshold value.

The k-means algorithm scales linearly with the size of the data set and can therefore be used for a large amount of data. However, Eq. (11) is in general a non-convex function and the result for the minimization may largely depend on the initial (random) choice of the means $\mathbf{K}$ and the assignments $\mathbf{k} = \{r_{ik}\}$. In practice, the k-means algorithm is therefore applied several times with different (random) initial conditions that may result in different assignments (see discussion in Subsection III D and Appendix D). Eventually, the particular assignment with the minimal value of $\mathcal{C}(\mathbf{X}, \mathbf{K})$—as compared to all other assignments—is chosen to be the "best" solution to the clustering problem.

TABLE I. List of the 14 ordered ground state configurations as identified in Refs. 11–13 "by hand" (see also Fig. 2): short-hand notation of the respective configurations (left column), short description of the characteristic features (central column), and specification in terms of the composition of the system, x, and order parameters $\Psi_{\nu}^{(u)}$ (right column). Note that all structures—except for structure I which refers to a monolayer—are bilayer structures.

	Characteristic features	Composition and order parameters
I	Hexagonal monolayer	$x = 0$
II	Rectangular bilayer	$x = \frac{1}{2}, \Psi_4^{(1,2)} = 1, 0 < \Psi_6^{(1,2)} < 1$
III	Square bilayer	$x = \frac{1}{2}, \Psi_4^{(1,2)} = 1, \Psi_6^{(1,2)} = 0$
IV	Rhombic bilayer	$x = \frac{1}{2}, 0 < \Psi_4^{(1,2)} < 1, 0 < \Psi_6^{(1,2)} < 1$
V	Hexagonal bilayer	$x = \frac{1}{2}, \Psi_4^{(1,2)} = 0, \Psi_6^{(1,2)} = 1$
I_x	Trihexagonal (layer one)	$0 < x < \frac{1}{3}, \Psi_6^{(3)} > 0.9$
H	Honeycomb (layer one)	$x = \frac{1}{3}, \Psi_6^{(3)} > 0.9$
II_x	Modified rectangular bilayer	$\frac{1}{3} < x < \frac{1}{2}, \Psi_6^{(3)} > 0.9$
V_x	Hexagonal bilayer	$0 < x < \frac{A}{1+A}, (1-x)\Psi_6^{(1)} + x\Psi_6^{(2)} > 0.9$
S_1	Snub square (layer one)	$x = \frac{2}{6}, \Psi_5^{(1)} > 0.7, \Psi_4^{(2)} > 0.9$
S_2	Snub square like (layer two)	$x = \frac{2}{6}, \Psi_5^{(2)} > 0.45$
P_1	Pentagonal type two	$\frac{1}{3} < x < \frac{1}{2}, \Psi_5^{(2)} > 0.45$
P_2	Pentagonal holes	$\frac{1}{3} < x < \frac{1}{2}, \Psi_5^{(4)} > 0.9$
P_3	Pentagonal holes	$0 < x < \frac{1}{3}, \Psi_5^{(4)} > 0.9$

III. RESULTS

A. The original diagram of states

The original diagram of states of the asymmetric Wigner bilayer system, presented in Refs. 11–13, has been redrawn in Fig. 2. The indicated 14 ordered configurations have been identified "by hand" according to the criteria summarized in Table I, based on the composition x and the BOOPs $\Psi_{[4,5,6]}^{(1,2,3,4)}$. From Fig. 2, it is obvious that quite extended regions in the (A, η)-plane remain unidentified with this classification scheme.

B. Unsupervised clustering algorithms applied to the Wigner bilayer

The huge amount of data accumulated in Refs. 11–13 in the search for equilibrium structures and the fact that—despite considerable efforts—white regions still remain in the diagram of states has motivated us to revisit and to reanalyze this data set with the help of a tool that is able to perform such a classification in a more systematic and more efficient manner as one could possibly do "by hand." To this end, we have introduced a more systematic classification scheme based on unsupervised clustering (as introduced in Sec. II D).

In order to apply this unsupervised clustering algorithm scheme, we first have to define the N_{f}-dimensional feature vector $\mathbf{x}(\mathcal{X})$, which is built up by the following components:

- the set of BOOPs, originally used in Refs. 11–13, $\Psi_{[4,5,6]}^{(1,2,3,4)}$, has been extended by the set $\Psi_{[3,8,10,12]}^{(1,2,3,4)}$ (using the short-hand notation introduced above); this new set of in total 28 BOOPs offers now the possibility to identify structures with eightfold, tenfold, or twelvefold symmetries;
- the composition $x = x(\mathcal{X})$;
- in an effort to quantify the ratio of the average nearest neighbor distance in $\mathcal{L}_1$, $r_{\mathrm{nn}}^{(1)}(\mathcal{X})$, and the average nearest neighbor distances in $\mathcal{L}_2$, $r_{\mathrm{nn}}^{(2)}(\mathcal{X})$, for a certain bilayer configuration, $\mathcal{X}$, we define the "intralayer nearest neighbor ratio" order parameter, $r_g(\mathcal{X})$, via

$$r_g(\mathcal{X}) = \frac{r_{\mathrm{nn}}^{(1)}(\mathcal{X})}{r_{\mathrm{nn}}^{(2)}(\mathcal{X})}. \tag{12}$$

The values of $r_g(\mathcal{X})$ are not bound to a maximum value; however, we find empirically an upper limit of $\simeq 1.07$ for all considered bilayer ground state configurations.

With the 28 BOOPs, the values for x and $r_g(\mathcal{X})$ we end up with a feature vector that has $N_{\mathrm{f}} = 30$ components $f_i(\mathcal{X})$,

$$\mathbf{x}(\mathcal{X}) = \{f_1(\mathcal{X}), \ldots, f_{N_{\mathrm{f}}=30}(\mathcal{X})\}. \tag{13}$$

In the following, we will reanalyze the classification scheme of phases used in Refs. 11–13 with the help of unsupervised machine learning techniques in order to automatically identify different families of structures directly from the feature vector, $\mathbf{x}$, given in Eq. (13). To be more specific: (i) we first perform a principal component analysis[24,25] (PCA) on the feature vectors of all structures from Refs. 11–13, which defines our basic data set. This allows us to identify directions of large variance in the data set, which capture the most relevant information among the different features. To this end, we transform the data set of feature vectors to unit-variance and zero-mean coordinates; this technique is termed "whitening" in literature and is used to decouple the PCA from the relative scales of different features.[1] (ii) We then apply the k-means[26–29] clustering algorithm to the latent space representation of the data set, which is spanned by the leading PCs. This will help us to identify new, previously unclassified phases that are potentially hidden in the huge set of the original structural data.

As a benchmark for this approach, we start our analysis with the simplest problem within the topic of Wigner bilayers, namely, with the *symmetric* Wigner bilayer system (as considered in Refs. 35 and 36), where $\sigma_1 = \sigma_2$ or, equivalently, $A \equiv 1$.

C. The symmetric Wigner bilayer system—A benchmark

For the symmetric case, the identification of the ground state configurations has been solved completely and analytically in Refs. 35 and 36 with five emerging structures, labeled I through V; these phases are depicted in the top row of Fig. 2. Furthermore, the exact η-values where the transitions between these phases occur as well as the nature of the related transitions could be identified with high accuracy in the above contributions: the hexagonal monolayer (I) is stable only at $\eta = 0$ and transforms for an infinitesimally small value of η into a rectangular bilayer, termed II. This structure is stable within the range $0 < \eta \lesssim 0.263$ and then transforms via a second-order transition into a square bilayer (III), which is stable within the range $0.263 \lesssim \eta \lesssim 0.621$. This structure then turns—again via a second order transition—into a rhombic bilayer phase (IV), stable within $0.621 < \eta \le 0.728$. Eventually, a hexagonal bilayer (V) emerges at $\eta \simeq 0.728$ via a first-order transition [see also the line $(A = 1)$ in Fig. 2].

As a first step, we have tested our clustering approach for this particular case knowing that we have the exact solution already at hand. As detailed in Appendix A, we first perform a principal component analysis (PCA) (cf. Subsection II D 2) on the set of the (unit-variance and zero-mean) feature vectors, $\mathbf{X}^{(\mathrm{sym})}$, of the $N_{\mathrm{sym}} = 141$ ground state configurations that were identified via the memetic evolutionary algorithm in Refs. 11–13 for different values of $\eta \in [0, 1]$ and for $A \equiv 1$.

The percentage of the so-called explained variance (PEV), defined in the Eq. (10), quantifies the amount of information encoded in each PC direction. This measure provides a threshold in the expressive quality of the PCs so that we can safely restrict our further analysis to the five leading PCs; in terms of the PEV, the PCs of order six (and higher) contribute by orders of magnitude less information as compared to the leading five ones (cf., Appendix, Fig. 8).

Next, we apply the k-means clustering algorithm (cf. Subsection II D 3) to the now $(N_\ell = 5)$-dimensional latent space representation $\mathbf{L}^{(\mathrm{sym})}$ of the data $\mathbf{X}^{(\mathrm{sym})}$ and assign to all $i = 1, \ldots, N_{\mathrm{sym}}$ data points a cluster label $c_i \in \{1, \ldots, K\}$, defining thereby the labeling (or clustering) $\mathbf{k}^{(\mathrm{sym})} = \{c_1, \ldots, c_{N_{\mathrm{sym}}}\}$ of the data set. In the particular case of the symmetric Wigner bilayer system we already know the number of phases and therefore set $K = 5$. The emerging k-means clustering $\mathbf{k}^{(\mathrm{sym})}$ of the data is in excellent agreement with the phase-assignment known from literature,[35,36] as can be seen in Fig. 9 in the Appendix.

D. The asymmetric Wigner bilayer system

Now that we know from the *symmetric* case of the Wigner bilayer system we can firmly rely on the analysis approaches detailed in Subsection II D, we proceed to the *asymmetric* case; this structural database was generated for the preceding contributions[11–13] by independent evolutionary structure optimization of configurations with up to 40 particles per unit cell and considering all related possible values of the composition, x, on an A- and η-grid as specified in Subsection II B. We perform the same analysis—i.e., first a PCA and then a subsequent k-means clustering—on the set of feature vectors, $\mathbf{X}^{(\mathrm{asym})} = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_{\mathrm{asym}}})$ with the $\mathbf{x}_i \in \mathbb{R}^{N_f = 30}$ being taken initially from the entire set of $N_{\mathrm{asym}} \sim 56\,541$ configurations considered in the asymmetric Wigner bilayer system.

1. Principal component analysis (PCA)

Via the PCA, we first transform the data set $\mathbf{X}^{(\mathrm{asym})}$ to a latent space representation $\mathbf{L}^{(\mathrm{asym})} = (\mathbf{l}_1, \ldots, \mathbf{l}_{N_{\mathrm{asym}}})$ of the data (for which we again assume unit-variance and zero-mean coordinates): $\mathbf{l}_i (\in \mathbb{R}^{N_\ell})$ is the projection of the data point $\mathbf{x}_i \in \mathbb{R}^{N_f}$ into the latent space of dimension $N_\ell (\le N_f)$; note that the actual value of N_ℓ has not been fixed, yet. In Fig. 3, we present the PEVs, $\lambda_j^{(\mathrm{e})}$, defined in Eq. (10), for each of the $N_{\mathrm{f}} = 30$ PCs of the feature vectors, $\mathbf{X}^{(\mathrm{asym})}$, of the structures considered in Refs. 11–13. We conclude from the PEVs that—similar to the symmetric case (cf. Fig. 8)—only very few principal components are expected to carry relevant information, i.e., will have significant $\lambda_i^{(\mathrm{e})}$-values. Thus, we restrict our analysis to the leading nine principal components, located left of the (second) "elbow"[48] occurring for i-values larger than $i = 9$, whose $\lambda_i^{(\mathrm{e})}$-values are larger than 0.02; hence, we set $N_\ell = 9$ in what follows.

In Fig. 4, we present the leading nine PCs, $\mathbf{v}_1, \ldots, \mathbf{v}_9$, which are vectors in the feature space spanned by $\mathbf{x} = (f_1, \ldots, f_{N_f})$. The 30 elements of each PC indicate the direction of the PC in feature space; we refer to the values of these elements of PCs as *feature weights*. Large positive or large negative values of certain feature weights of a particular PC indicate important features that quantify information in the data set with directions of high variance. These characteristic features of PCs can be used to identify important order parameters—or combinations of order parameters if several feature weights are dominant in a particular PC. On the other hand, feature weights close to zero indicate less relevant directions (see the supplementary material, Sec. I, for more details).

From the data presented in Fig. 4, we see that the first PC, $\mathbf{v}_1$, exhibits large positive feature weights of the six- and twelvefold order parameters $\Psi_{[6,12]}^{(1,3,4)}$ and medium to large negative weights of $\Psi_{[4,5,8,10]}^{(1,3,4)}$, r_g, and x. The PCs $\mathbf{v}_2$ and $\mathbf{v}_3$ exhibit medium to large feature weights distributed over a range of order parameters that makes them more difficult to interpret than the feature weights of $\mathbf{v}_1$. From the fourth PC $(\mathbf{v}_4)$ onward, single (or very few) directions in feature-space are relevant: this applies, for instance, for $\Psi_{[10,12]}^{(2)}$ in the case of $\mathbf{v}_4$ and for $\Psi_{[10,12]}^{(1)}$ in the case of $\mathbf{v}_5$. Furthermore, we learn from Fig. 4 that for all PCs the feature weights for the intralayer nearest neighbor order parameter, r_g, and for the composition, x, are strongly correlated. Thus, one could interpret our results such that the nearest

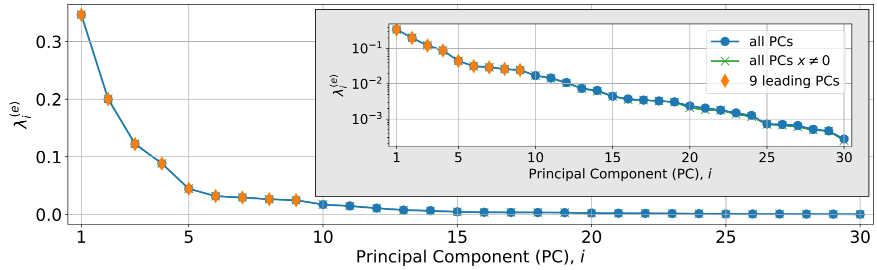

FIG. 3. Percentage of the explained variance (PEV), $\lambda_i^{(e)}$ [as defined in Eq. (10)], of the principal components (PCs) of the data, set $\mathbf{X}^{(\text{asym})}$ of feature vectors of all $N_{\text{asym}} = 56\,541$ configurations of the asymmetric Wigner bilayer system, considered in Refs. 11–13 (blue dots); the leading nine PCs (with the PEV $\lambda_j^{(e)} > 0.02$) are shown in orange. We also present (via green crosses) the PEV of the PCs of the data set $\mathbf{X}^{(*)}$, i.e., the data set that contains all data points of $\mathbf{X}^{(\text{asym})}$ except for those which correspond to hexagonal monolayer configurations (i.e., where $x = N_2/N = 0$). It should be noted that structures with $x = 0$ can uniquely be characterized as trigonal monolayers, i.e., phase I, in the investigated data set. Thus, the PEV results emphasize that the PCA is only marginally affected by the large proportion of phase I structures in the data set. The gray inset shows the related data in a semi-logarithmic presentation.

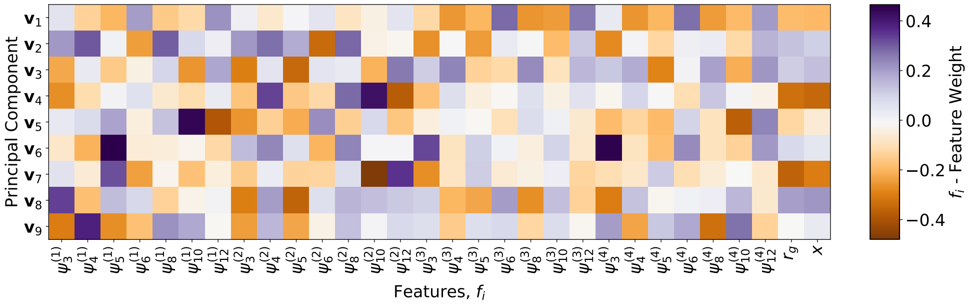

FIG. 4. Leading nine PCs, $\mathbf{v}_1, \ldots, \mathbf{v}_9 \in \mathbb{R}^{N_f=30}$, of the (whitened) data set $\mathbf{X}^{(\text{asym})}$ represented in the feature space spanning $\mathbf{x} = (f_1, \ldots, f_{N_f})$ [cf. Eq. (13)]. The features $f_1, \ldots, f_{30}$ are listed along the horizontal axis and represent the order parameters used to evaluate $\mathbf{X}^{(\text{asym})}$ from the structural database of the asymmetric Wigner bilayer system from literature;[11–13] the respective 30 elements of the PCs, $\mathbf{v}_j$, related to the directions of f_j are presented via the color-coding specified by the color bar on the right-hand side: large positive (negative) values of feature weights are emphasized by dark purple (orange) coloring, while values close to zero are colored in white. It should be noted that prior to the PCA, all features are separately scaled to zero mean and unit variance (i.e., the data are "whitened"). Thus, even positive valued features, such as r_g and x, can have negative feature weights in the corresponding principal components.

neighbor distances, r_g, of particles in both layers are largely governed by the composition x for low energy configurations of the system. This, in turn, might suggest that the particles tend to be distributed in ground state configurations of the system as uniformly as possible [constraint by the (A,η)-specific lattice formation] in both layers.

To provide a first impression of the descriptive power of the PCA, we present in Fig. 5 the revised diagram of states of the ground state configurations of the asymmetric Wigner bilayer system in the (A,η)-plane in a new [R, G, B]-scheme, which is now based on the leading three PCs: to this end, we consider the latent space representations, $\mathbf{l}_g$, of the feature vectors, $\mathbf{x}_g$, which correspond to the suggested ground state configurations, $\mathcal{X}_g$, of the asymmetric Wigner bilayer system for different values of the system parameters, A and η. For each of these data points, $\mathbf{l}_g = (v_{g1}, \ldots, v_{gN_\ell})$, we use the first three coordinates, $[v_{g1}, v_{g2}, v_{g3}]$, i.e., the coordinates of $\mathbf{l}_g$ associated with the first three PCs $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, to define the relative contribution of the colors red, green, and blue to the color of each pixel in the (A,η)-plane. Moreover, we have transformed the values of the related coordinates to the interval [0, 1] via $\hat{v}_{gi} = \frac{1}{2}[\tanh(v_{gi}) + 1]$. This first, admittedly simplistic view[49] demonstrates that the phase boundaries as suggested in Refs. 11–13 nicely correlate with the values of the PCs; however, we can also spot out regions in the (A,η)-plane, which call for a closer inspection: for instance, the region of phase I_x is likely to have a more sophisticated internal structure than previously assumed, as indicated by the different greenish (i.e., dominant $\mathbf{v}_2$) and bluish (i.e., dominant $\mathbf{v}_3$) regions, a feature that we will further investigate in the following.

We can see from Fig. 5 that phase I can uniquely be identified via the black color. Furthermore, the leading three principal components of structures II, II_x, H, and I_x are clearly different from those of the pentagonal structures P_1, P_2, P_3, and S_2: (i) the former structures (i.e., II, II_x, H, and I_x) are characterized by large [G, B]-values (associated with the second and third PCs), thus indicating large values of the latent space coordinates into directions $\mathbf{v}_2$ and $\mathbf{v}_3$; the corresponding bilayer structures have the property that—when projecting the particle positions onto a single plate—a hexagonal monolayer is

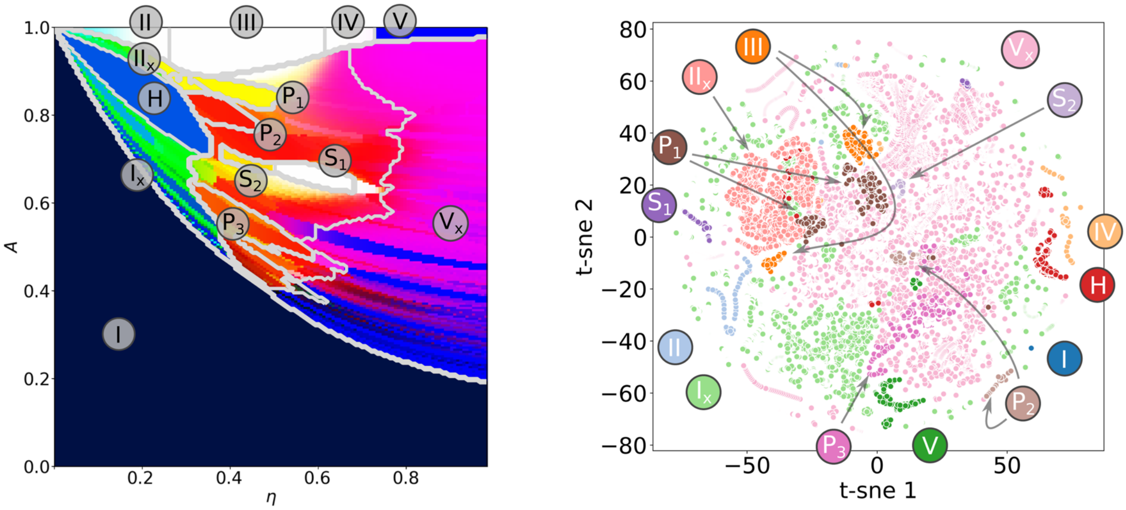

FIG. 5. Left: Diagram of states of the ground state configuration of the asymmetric Wigner bilayer system in the (A, η)-plane, colored via an [R, G, B]-scheme that is based on the first three PC vectors $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$ of the data set $\mathbf{X}^{(\mathrm{asym})}$: for every (A, η)-pair, we define the relative amount of red, green, and blue color [R, G, B] of the corresponding pixel in the (A, η)-plane by the coordinates, $[\hat{v}_{g1}, \hat{v}_{g2}, \hat{v}_{g3}]$, of the latent space data point, $\mathbf{l}_g = (v_{g1}, \ldots, v_{gN_\ell})$, of the associated ground state configuration of the asymmetric Wigner bilayer system, using the database available in literature;[11–13] the values of the coordinates v_{gi} are transformed to the interval $[0, 1]$ via $\hat{v}_{gi} = \frac{1}{2}[\tanh(v_{gi}) + 1]$, to establish an [R, G, B]-scheme for the entire range of v_{gi}-values. The light gray lines indicate phase boundaries taken from Refs. 11–13. Phases from the above references are labeled according to the classification of the 14 phases summarized in Table I. Right: t-SNE[50] analysis (see also Refs. 1 and 40), mapping the $(N_{\ell\ell} = 9)$-dimensional latent space representation of all bilayer configurations from the literature database[11–13] onto a two-dimensional t-SNE manifold spanned by "t-sne 1" and "t-sne 2." Each point represents a structure from the database embedded into the two-dimensional t-SNE plot and is colored according to the labeling given in Table I (unknown phases are omitted).

formed; (ii) in contrast, the pentagonal structures (P$_1$, P$_2$, and P$_3$) are throughout more complicated to interpret. Still, we can see that their symmetry (cf. Fig. 4) is either dominated by the first principal component $\mathbf{v}_1$ (red) or by combinations of $\mathbf{v}_1$ and $\mathbf{v}_2$ (yellow, which is generated by adding red and green in the [R, G, B]-notation). Furthermore, structures in the V$_x$ region show strong signals either from the third principal component $\mathbf{v}_3$ (blue) or from combinations of $\mathbf{v}_1$ and $\mathbf{v}_3$ (where red and blue become magenta); in that way, they can be distinguished from the (red and yellow) pentagonal region in the diagram of states depicted in Fig. 5. Summarizing we note that the region in the diagram of states depicted in Fig. 5, where the latent space representation of the related ground state structures are dominated either by $\mathbf{v}_1$ or by combinations of $\mathbf{v}_1$ and $\mathbf{v}_2$, largely corresponds to the unclassified white region occurring in Fig. 2. In an effort to obtain an even more profound insight into the so far hidden structures, we apply in a subsequent step a k-means clustering analysis of the data.

Before proceeding in this direction, we note that our set of data has also been inspected with the "t-stochastic neighbor embedding" (t-SNE)[50] method (see the right panel of Fig. 5), which represents another quite useful non-linear technique for visually representing a set of data from a high-dimensional feature space in a low-dimensional latent space, while preserving the local structure of the data in a few embedding coordinates; for details, we refer the interested reader to Refs. 1 and 40. The coloring of the depicted t-SNE embedding in the right panel of Fig. 5 corresponds to the phase classification from literature[11–13] and indeed some spatially separated clusters of the t-SNE can be associated with these structural phases. However, in our t-SNE analysis of the entire database of potential ground state structures from the asymmetric Wigner bilayer system, no clear (i.e., visual) separation of structural families into spatially separated clusters (which could be further analyzed with the DBSCAN method,[51] for instance) could be identified in the two t-SNE dimensions.

In a next step, we have analyzed our structural data with the k-means algorithm, thereby identifying the optimum value for the number of clusters, K. We have shifted the details of this rather technical procedure to Appendix B. As specified there in detail, we eventually come up with $K = 32$ as the optimal number of structural clusters.

E. New insights to the Wigner bilayer system from unsupervised learning

The preceding PCA provides already clear evidence that so far unexplored and unidentified ground state phases are hidden in the incredibly rich plethora of ordered bilayer structures in the asymmetric Wigner bilayer system. A step toward a more systematic analysis of the ground state configurations can be realized by applying a subsequent k-means clustering analysis (see Subsection II D 3) of the representation of the data set in terms of the nine leading principal components (for more details on the precise way how we applied k-means clustering to the structural database of the asymmetric Wigner bilayer system, we refer to Appendix B).

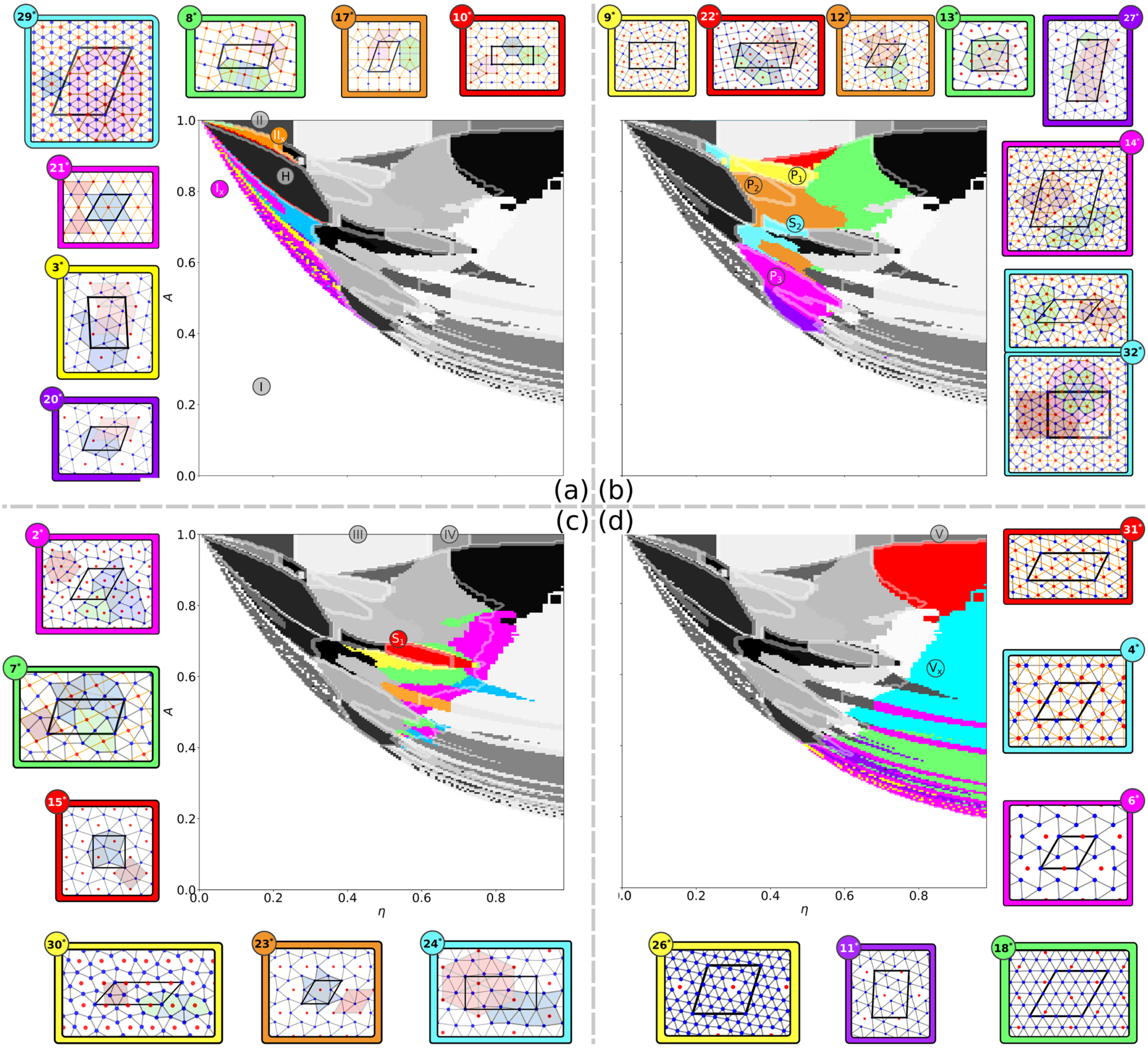

FIG. 6. Ground state phase diagram of the asymmetric Wigner bilayer system (see Refs. 11–13) in the (A, η)-plane as identified by the $K^* = 32$-means clustering algorithm; the respective 32 families of structures, $\mathbf{k}_{32}^{*c}$, are color-coded in different gray scales, ranging from white to black. The presentation of the ground state families of this phase diagram was subdivided into four (A, η)-subpanels (a)–(d) by symmetry arguments of the respective occurring ground states (see discussion of the subpanels in the text). In each subpanel, we highlight the respective parameter regions in bright colors (cyan, yellow, green, orange, red, magenta, or purple—in no particular order) where certain $\mathbf{k}_{32}^{*c}$ families form the ground state; archetypical structures of the respective $\mathbf{k}_{32}^{*c}$ families are shown as insets. In each panel, these structures are labeled by the corresponding value of $c = (1, \ldots, 32)$ in the upper left corner to address their association to a certain family $\mathbf{k}_{32}^{*c}$. For convenience, the frames of the insets are color-coded in the same way as the ground state regions of the respective $\mathbf{k}_{32}^{*c}$ family in the phase diagram. Particles in $\mathcal{L}_1$ $(\mathcal{L}_2)$ are always colored blue (red) and connections between nearest neighbors in each layer are drawn. Special tiles and features of the different structures are highlighted by colored shapes and the respective unit cell of each structure is emphasized by a thick black frame. The phase-boundaries, as documented in literature[11–13] (cf. also Fig. 2) are indicated by opaque white lines in each panel; furthermore, the corresponding phases (known from literature) are labeled by their acronyms in circles, which are associated with the considered parameter region of the different panels. Colored, disk-shaped labels in subpanels (a) to (d) indicate that archetypical structures of the corresponding $\mathbf{k}_{32}^{*c}$ families are displayed as insets in the respective panels, while gray, disk-shaped labels (i.e., I through V and H) indicate that the corresponding structures are not shown (although the phases have been identified by the $\mathbf{k}_{32}^{*}$ – clustering algorithm).

We first replace the original diagram of states (cf. Fig. 2) with a diagram of states based on the $\mathbf{k}_{32}^{*}$ − clustering results, summarized in Fig. 6. Indeed, the 32 families of structures provide information about new, so far unidentified ground state configurations: we observe that some regions (such as the ones originally occupied by the I_x or V_x phases) are obviously subdivided into sub-regions, indicating the emergence of so far unclassified phases. For each of these families (and except for the "trivial" phases I through V and H), we present in Fig. 6 typical examples of ground state configurations of the system for different values of A and η, based on the $\mathbf{k}_{32}^{*c}$ clustering. In an effort to obtain a better overview of the different structures, we have split in Fig. 6 the presentation of the diagram of states into four qualitatively different subpanels [labeled (a) to (d)]:

- in **panel (a)**, we focus on the region that hosts the phases I_x, II, II_x, and H, occurring at small to medium values of η (i.e., $0 < \eta \lesssim 0.4$) and medium to large values of A (i.e., $0.4 \lesssim A \leq 1$); the ground state structures in this region have in common that they form a hexagonal monolayer if all particles were projected onto the same layer;
- in **panel (b)**, we collect structural families that feature pentagonal tiles in $\mathcal{L}_1$, i.e., configurations that belong to the broader family of pentagonal structures: suggested ground states candidates that belong to this category are P_1, P_2, P_3, and S_2, the associated range of the system parameters can be roughly given by $0.3 \lesssim \eta \lesssim 0.7$ and $0.4 \lesssim A \lesssim 0.9$;
- in **panel (c)**, we address $\mathbf{k}_{32}^{*c}$ families that have tilings in $\mathcal{L}_1$ that are similar to the snub-square structure, S_1, which, in turn, can potentially give rise to ground state configurations with a global twelvefold symmetry (see Refs. 52–54);
- eventually, in **panel (d)**, we present $\mathbf{k}_{32}^{*}$ − clustering results that can be related to the V_x region in the parameter space of the asymmetric Wigner bilayer system, i.e., at large plate separation distances ($\eta \gtrsim 0.7$) and covering a large range of A.

A comparison of the identified phases of the asymmetric Wigner bilayer system via methods from literature[11–13] (see Fig. 2) and via labeling by unsupervised $\mathbf{k}_{32}^{*}$ − clustering techniques (see Fig. 10 in the Appendix) shows—on one hand—an excellent agreement for several structural families: not only the symmetric cases (i.e., phases I through V), but also more complex configurations such as II_x, H, S_1, S_2, P_1, P_2, and P_3 are faithfully reproduced. On the other hand, the clustering technique is able to reveal that for several regions in the parameter space (such as the regions of the phases I_x and V_x) a reanalysis of the data is in order; this will be done in the following.

(a) **phase $\mathbf{I}_x$**: based on the clustering (see Fig. 6), we can indeed identify a rich variety of phases within this region, which is correlated with the dodecagonal $\Psi_{12}^{(4)}$ "hole" BOOP; note in this context, in particular the blue region within the I_x phase near the boundary to the H phase and the relatively large region populated by the $\mathbf{k}_{32}^{*29}$ family [see panel (a) of Fig. 6, marked in cyan];

(b) **pentagonal phases**: regions that were declared in the original diagram of states to be populated exclusively by the P_3 structure (see Fig. 2) now show—as a consequence of the more refined analysis of the clustering approach—a considerably more complex internal structure: we can see from panel (b) of Fig. 6 that not only the $\mathbf{k}_{32}^{*14}$ and the $\mathbf{k}_{32}^{*27}$ families but also the S_2 configuration (represented by the $\mathbf{k}_{32}^{*32}$ family) form similar patterns in $\mathcal{L}_2$ (notably distorted rectangles and triangles organized in a distorted snub-square vertex); however, the decoration of the $\mathcal{L}_2$ tiles with particles in $\mathcal{L}_1$ becomes increasingly complex: more particles of $\mathcal{L}_1$ are involved per tile in $\mathcal{L}_2$ the more we approach the boundary to phase I in the parameter space. The structure in $\mathcal{L}_1$ thus approximates a trigonal lattice when approaching the phase boundary of the I phase and exhibits only pentagonal defects around the x, y-locations of $\mathcal{L}_2$ particles. Interestingly, the structural families $\mathbf{k}_{32}^{*27}$ and $\mathbf{k}_{32}^{*14}$ appear to belong to a family of increasingly complex super-structures of the top panel of family $\mathbf{k}_{32}^{*32}$, i.e., the original S_2 phase. This might hint at the existence of a low-energy quasi-crystalline structure of this family in the asymmetric Wigner bilayer system.

(c) **phases $\mathbf{S}_1$ and $\mathbf{S}_2$**: with the help of the clustering analysis, we are now able to classify also the previously unclassified structures in the (A, η)-regions in the vicinity of phases S_1 and S_2 (cf. related white regions in Fig. 2) by several different structural families, as illustrated in panel (c) of Fig. 6: these structures have in common that their basic tiles (such as equilateral triangles and squares arranged in a snub-square vertex) form in $\mathcal{L}_1$ a structure that might indicate the existence of a quasicrystalline state with a global dodecagonal symmetry.[52,53,55]

(d) **phase $\mathbf{V}_x$**: eventually, in the V_x region of the diagram of states, some interesting new structural families are identified as a consequence of the $\mathbf{k}_{32}^{*}$ − clustering procedure, as can be seen in panel (d) of Fig. 6. Characteristic values and boundaries of the order parameters and of the corresponding principal component representation of all newly identified structural families $\mathbf{k}_{32}^{*c}$ depicted in Fig. 6 are collected in Subsection 3.1.7 of Ref. 40 (which also provides more detailed information about the symmetries of these structural families).

For more details on characteristic values of the order parameters and principal components for the families of structures described by the $\mathbf{k}_{32}^{*}$ − clustering, see the supplementary material, Sec. II.

IV. CONCLUSIONS

In this contribution, we have reanalyzed the ordered ground state configurations of the asymmetric Wigner bilayer system where identical point charges are immersed into the space confined between two parallel plates of opposite charge. The ratio of the (not necessarily equal) surface charges of the two plates (σ_1 and σ_2), i.e., $A = \sigma_2/\sigma_1$ and the reduced, dimensionless distance η between the plates uniquely define each state point of the system. A previous classification scheme of the emerging configurations[11–13] into structural families, was done by "hand": such an approach is not only a tedious, possibly hopeless task, but it is also—and even more relevant—prone to faulty analysis, preventing thereby a faithful identification and sorting of the emerging structures.

In an effort to overcome these drawbacks, we have reanalyzed this huge set of data (comprising ~60 000 data points) and have used instead machine learning based tools, notably on the principal component analysis and a subsequent k-means clustering algorithm. In a first step, we have assigned to each emerging structure a feature vector with 30 entries: (i) predominantly suitably defined order parameters that characterize the ground state configuration, (ii) the composition of the system, and (iii) further structural information based on the radial distribution function. With the help of a principal component analysis, we have extracted from this representation the most relevant information by projecting the underlying 30-dimensional feature space on a reduced representation in the so-called latent space; in our case, we found out that a dimension of nine of the latent space is sufficient to capture the relevant features. Eventually, we have classified this reduced information via a k-means algorithm and have collected ground state configurations into families of structures that occupy "neighboring" regions in this latent space. Applying different types of thorough internal consistency checks we eventually found that the sorting of the emerging structures into 32 families provides the most reliable and most consistent classification scheme of these structures, as compared to 14 structural families that were identified in the preceding, "by hand" classification approach.[11–13]

In view of the achieved results presented in this contribution, we can righteously conclude that with our machine learning based tool at hand we are now able to provide a systematic, reliable, and thorough classification scheme for the emerging structures. The new insights into the diagram of states comprise now particle configurations that were previously hidden (or even not accessible) in the zoo of structures in the database of Refs. 11–13: (i) within the region that was originally assigned to the I_x structure we could identify a rich variety of phases that are characterized by a dodecagonal $\Psi_{12}^{(4)}$ bond order parameter; (ii) the region that was originally thought to be populated exclusively by the P_3 structure has a very rich internal structure and other configurations could be identified that are formed by distorted rectangles and triangles, organized in a distorted snub-square vertex; (iii) it could be shown that the previously unclassified ("white") regions in the vicinity of phases S_1 and S_2 are populated by several different structural families; all these structures have in common that their basic tiles arrange in $\mathcal{L}_1$ into a structure that might possibly indicate the existence of a quasicrystalline state with global dodecagonal symmetry; (iv) eventually, the huge, and previously unexplored region populated by the V_x structure reveals the existence of quite a few interesting subtle new structural families. With all these new findings, we conclude that the reanalyzed diagram of states is bare of any white (i.e., unexplored) regions.

Apart from a more systematic and thorough classification of the structures, our approach offers several other attractive features that turned out to be very useful.

A particular challenge when identifying the ground state configurations for our system is the issue of degeneracy: this applies to structures that were obtained in Refs. 11–13 via evolutionary algorithms, involving different numbers of particles per unit cell but characterized by the same composition; they eventually end up as identical structures with the same (i.e., degenerate) energies but parameterized by different (but equivalent) unit cells. In our previous "by hand" classification scheme, it was very difficult to prove the structural equivalence of two such particle configurations, while with the clustering-based labeling of the available structural database as it has been used in this contribution, such degenerate configurations are automatically grouped into the same structural family via the information emerging from the feature vectors.

Another advantage of our approach is that we easily obtain within each structural family an energy-based ranking of the structures: thus, for a given state point [defined by a pair (A, η)], we have clear information about an energy-based ranking of the structures, starting at the lowest level with the ground state configuration. In this manner, we can identify those structures that energy-wise are sometimes very close to the ground state configuration, but which possibly are structure-wise distinctively different from the latter one; such a ranking was essentially inaccessible in our previous approach, while in our present approach, they are automatically labeled. With this information at hand, we can then focus in a subsequent step on different structural families, where we have direct access to the properties of the energetically competing families of structures for any (A, η)-state (see a more in-depth discussion in Appendix C).

From a more formal point of view, it should be mentioned that, in general, clustering of structural data using PCA and k-means clustering (or any other, suited clustering algorithm or classification algorithm) provides us with an additional attractive feature: PCA is a linear transformation from the feature space to the latent space and k-means is a mapping of a data point in the latent space to a cluster label. Once the clustering algorithm is trained (i.e., once it has converged), it can be used as a classification model[56,67] and we can ask the following questions for an arbitrary structure: "what family would it belong to?," "where would it appear in the phase-diagram," and "what would be its characteristic features?"; see Appendix A.1.2 of Ref. 40 for related numerical details on the characteristic features of the here employed $\mathbf{k}_{32}^{*}$ – clustering classification scheme of the structural data from Refs. 11–13.

When re-exploring the data set of configurations of the asymmetric Wigner bilayer system, we also encountered situations where particular numerical care had to be taken: this applies in particular when exploring regions where first-order transitions between competing structures have to be identified, characterized by distinct discontinuous changes in order parameters. In an effort to locate the transition point accurately, particular numerical care has to be taken. Even more challenging are second-order phase transitions (such as those between phases II → III and III → IV) where some of the features (order parameters, etc.) change continuously.

Summarizing, we can righteously state that the clustering tools discussed in this contribution represent an indispensable help in classifying complex emerging structures and undoubtedly offer a deeper insight into the complexity of the phase diagram of the asymmetric Wigner bilayer system.

SUPPLEMENTARY MATERIAL

In the supplementary material, we provide further physical and conceptual insights into the unsupervised learning approach put forward in this manuscript. First, we specifically visualize in supplementary material Sec. I the relative contributions and the expressed physical symmetries of all characteristic features (i.e., order parameters) to the most relevant principal components (PCs) of the structural data set of the asymmetric Wigner bilayer system. Second, we present in supplementary material Sec. II the characteristic values of the order parameters and of the principle components related to each family of ground state structures of the asymmetric Wigner bilayer system that is either known from literature or has been newly discovered with the methods presented in this contribution. Third, we provide a discussion about alternative dimensional reduction and clustering tools and a justification for opting for PCA and *k*-means clustering in this manuscript in supplementary material Sec. III.

ACKNOWLEDGMENTS

We gratefully thank Moritz Antlanger for helpful discussions. B.H. acknowledges funding via a DOC Fellowship and an APARTMINT stipend of the Austrian Academy of Sciences. M. Mihalkovič and L. Šamaj are supported by the Slovak Grant Agency with Projects VEGA Grant Nos. 2/0144/21 and 2/0092/21, respectively. The computational results presented have been achieved (in part) using the Vienna Scientific Cluster (VSC).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Benedikt Hartl: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (equal); Investigation (lead); Methodology (lead); Resources (equal); Software (lead); Validation (equal); Visualization (lead); Writing – original draft (equal); Writing – review & editing (equal). **Marek Mihalkovič**: Conceptualization (equal); Data curation (supporting); Formal analysis (supporting); Investigation (equal); Methodology (supporting); Supervision (equal); Validation (equal); Writing – original draft (supporting); Writing – review & editing (supporting). **Ladislav Šamaj**: Formal analysis (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Martial Mazars**: Formal analysis (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Emmanuel Trizac**: Conceptualization (supporting); Formal analysis (supporting); Investigation (supporting); Project administration (equal); Supervision (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Gerhard Kahl**: Conceptualization (supporting); Formal analysis (supporting); Funding acquisition (equal); Investigation (supporting); Project administration (equal); Resources (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

Computational protocols and numerical data that support the findings of this study are shown in this article, in the appendix and in the supplementary material.

APPENDIX A: THE SYMMETRIC WIGNER BILAYER SYSTEM—A BENCHMARK

For the symmetric case, the identification of the ground state configurations has been solved analytically,[35,36] with five emerging structures, labeled I through V; these phases are depicted in the top row of Fig. 2 (of the main text) and Fig. 7, respectively. Furthermore, the exact η-values where the transitions between these phases occur as well as the nature of these transitions could be identified with high accuracy in the above contributions: the hexagonal monolayer (I) is stable only at $\eta = 0$ and transforms for an infinitesimally small value of η into a rectangular bilayer, termed II. This structure is stable within the range $0 < \eta \lesssim 0.263$ and then transforms via a second-order transition into a square bilayer (III), which is stable within the range $0.263 \lesssim \eta \lesssim 0.621$. This structure then turns—again via a second order transition—into a rhombic bilayer phase (IV), stable within $0.621 < \eta \le 0.728$. Eventually, a hexagonal bilayer (V) emerges at $\eta \simeq 0.728$ via a first-order transition (see also the line $(A = 1)$ in Fig. 2.

We now test the clustering approach for this particular case where we have the solution already at hand. These calculations are based on the $N_{\text{sym}} = 141$ ground state configurations that were identified via the memetic evolutionary algorithm in Refs. 11–13 for different values of $\eta \in [0, 1]$ and for $A \equiv 1$.

We first perform for all structures a principal component analysis (PCA)[25] on the set of the (unit-variance and zero-mean) feature vectors, $\mathbf{X}^{(\text{sym})}$, defined in main-text Eq. (13) with $N_{\text{f}} = 30$. The data set is then transformed into an N_ℓ-dimensional latent space representation, $\mathbf{L}^{(\text{sym})} = \{\mathbf{l}_1, \ldots, \mathbf{l}_{N_{\text{sym}}}\}$ with $\mathbf{l}_i = (v_{i1}, v_{i2}, \ldots, v_{iN_\ell}) \in \mathbb{R}^{N_\ell}$ and $N_\ell \le N_{\text{f}}$. The actual value of N_ℓ defines how many leading PCs are considered in the latent space representation of the data.

When investigating the N_{sym} data points as a function of the first three PCs (corresponding to the data points $\mathbf{x}_i$ projected onto the first three latent space directions $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$) one can already distinguish the different phases by eye: structures belonging to a specific phase form clusters in such a representation, which are spatially separated from each other (data not shown here, cf. top panel in Fig. 3.5 of Ref. 40).

In Fig. 8, we present the percentage of the explained variance (PEV), $\lambda_i^{(\text{e})}$ defined in Eq. (10), contained in each PC with index i. The PEV quantifies the amount of information encoded in each PC direction $\mathbf{v}_i$. We see that the values of $\lambda_i^{(\text{e})}$ quickly drop from $\lambda_1^{(\text{e})} \sim 1/3$ to $\lambda_6^{(\text{e})} < 0.05$, and further on by several orders of magnitudes such that the higher PC (i.e., for $i \gtrsim 6$) are insignificant as compared to the leading ones. Thus, we can safely

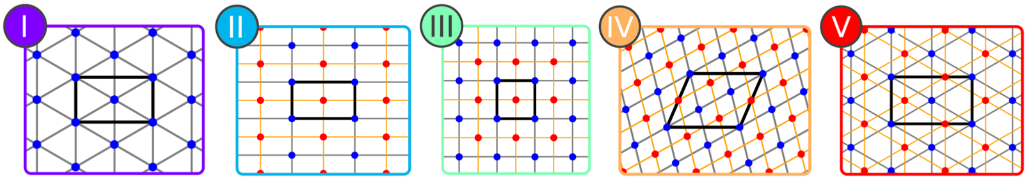

FIG. 7. Phases I through V representing the ground state configurations of the symmetric Wigner bilayer system.[35,36] Blue and red symbols represent particles from layers $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. The respective unit cells are indicated by black frames.

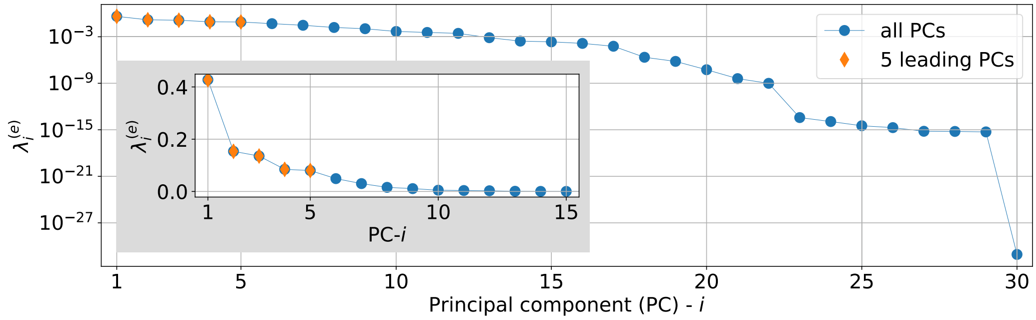

FIG. 8. Percentage of the explained variance, $\lambda_i^{(e)}$, for each PC i as defined in main text [Eq. (10)] for all 30 PCs (blue) of the symmetric Wigner bilayer system. The leading five principal components are shown in orange. Inset: same data as in the main plot but with linear scale for $\lambda_i^{(e)}$ for the first 15 PCs.

restrict ourselves to the five leading PCs and set in the following $N_\ell = 5$.

We now apply the k-means clustering algorithm (cf. Subsection II D 1) to the $(N_\ell = 5)$-dimensional latent space representation $\mathbf{L}^{(\mathrm{sym})}$ of the data $\mathbf{X}^{(\mathrm{sym})}$ and assign to all $i = 1, \ldots, N_{\mathrm{sym}}$ data points a cluster label $c_i \in \{1, \ldots, K\}$, defining thereby the labeling (or clustering) $\mathbf{k}^{(\mathrm{sym})} = \{c_1, \ldots, c_{N_{\mathrm{sym}}}\}$ of the data set. In the particular case of the symmetric Wigner bilayer system, we already know the numbers of phases and therefore set $K = 5$.

Results are shown in Fig. 9. It can be seen that the emerging k-means clustering $\mathbf{k}^{(\mathrm{sym})}$ of the data is in excellent agreement with the phase-assignment known from literature,[35,36] $\mathbf{w}^{(\mathrm{sym})} = \{C_1, \ldots, C_{N_{\mathrm{sym}}}\}$; here, the C_i (=1 through 5) label the corresponding phases (I through V), respectively, for every data point i (as specified in Table I). It should be noted that the particular numerical values that associate the data points with a certain cluster are usually arbitrarily chosen by the k-means algorithm; however, they are unique: in Fig. 9, we see that the clusters related to phases I through V are, respectively, labeled by c_i. Furthermore, the assignment of the data points into the different clusters is almost perfect and the labels c_i can be redefined to match the numerical values of C_i by mapping $C_i(= Ł1, 2, 3, 4, 5) \leftrightarrow c_i(= Ł5, 1, 3, 4, 2)$, respectively.

In an effort to test the reliability of the partitioning of a data set, we use the so-called mutual information score, introduced and discussed in Appendix E. For this particular case, we report an (adjusted) mutual information score of $I_K(\mathbf{k}^{(\mathrm{sym})}, \mathbf{w}^{(\mathrm{sym})}) = 0.95$, as defined in Eq. (E2) in Appendix E, between the clustering $\mathbf{k}^{(\mathrm{sym})}$ and the phase-assignment from literature $\mathbf{w}^{(\mathrm{sym})}$.[35,36] The small discrepancies arising from two data points (denoted in Fig. 9 as *outliers*) are discussed below.

Usually, the results from the k-means clustering depend on the initial conditions of the algorithm such as (i) the initial (and usually arbitrary) placement of the K different cluster centers in the latent space of the data set $\mathbf{X}^{(\mathrm{sym})}$ and (ii) the initial data point assignments to the clusters. To justify the results shown in Fig. 9, we thus perform 100 independent runs of k-means clustering (labeled with an index l) on the leading five principal components of the data set $\mathbf{X}^{(\mathrm{sym})}$. We find that all corresponding labels, $\mathbf{k}_l^{(\mathrm{sym})}$, share the same adjusted mutual information score, $I_K(\mathbf{k}_l^{(\mathrm{sym})}, \mathbf{w}^{(\mathrm{sym})}) = 0.95$ with the results $\mathbf{w}^{(\mathrm{sym})}$ being known from literature.

Furthermore, we also increased the number of leading PCs from five to 30 without observing significant changes in the results; however, when using less than five principal components, the results become unreliable. These observations confirm that the clustering shown in Fig. 9, indeed, represents the optimal k-means clustering to group the data points of $\mathbf{X}^{(\mathrm{sym})}$ into the phases I through V.

In Fig. 9, two outlier structures are highlighted by arrows that are located at η-values close to transition boundaries from phase II to phase III as well as from phase III to phase IV, respectively. The reason that these structures are, erroneously, attributed by the k-means algorithm to phase III lies in the fact that for the respective η-values both a rectangular bilayer structure (phase II) or a rhombic bilayer structure (phase IV) can righteously be con-

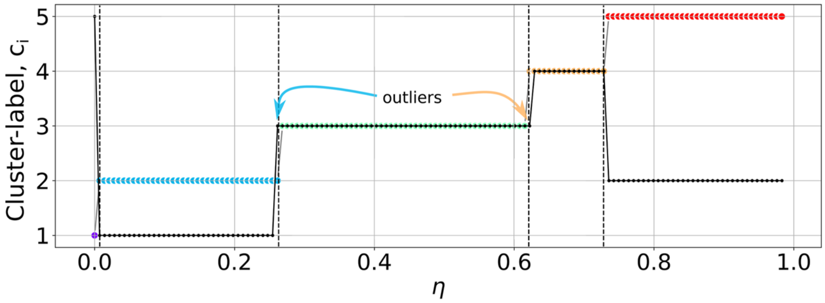

FIG. 9. Labeling $\mathbf{w}^{(\mathrm{sym})}$[35,36] of the ground state configurations of the symmetric Wigner bilayer system, i.e., I: 1, II: 2, III: 3, IV: 4, and V: 5 (with symbols color-coded according to Fig. 7), and labeling $\mathbf{k}^{(\mathrm{sym})}$ by the k-means clustering (black symbols) for each of the N_{sym} data points of the data set $\mathbf{X}^{(\mathrm{sym})}$ identified in Ref. 11 for different values of $\eta \in [0, 1]$ at $A = 1$. Note that the numerical value of a particular cluster label c_i (=1 through 5) assigned by the k-means algorithm to all data points belonging to one particular cluster is arbitrary but unique. An adjusted mutual information score, $I_K(\mathbf{k}^{(\mathrm{sym})}, \mathbf{w}^{(\mathrm{sym})}) = 0.95$ [as defined in Eq. (E2) in Appendix E], is realized. The two outliers (highlighted by the arrows) are discussed in the text. The vertical dashed lines mark the η-values where phase transitions from phase I through V occur: from left to right, $\eta = 1/141$ represents the emergence of phase II (given the discrete steps in η), while $\eta = 0.263$, $\eta = 0.621$, and $\eta = 0.728$ mark the transitions from phase II to III, III to IV, and IV to V, respectively.

sidered (within numerical accuracy) as "nearly" square structures (phase III).

We point out that small numerical variations in the data, which are often related to artifacts (such as noise), can trigger undesired effects in clustering approaches and may lead to an artificial partitioning of data in a clustering or classification task. Therefore, a proper preparation of the data with, for instance, PCA can help to reduce the effects of noise on the outcome of a clustering approach of a particular data set. On the other hand, sometimes small variations in the data do have a physical meaning such as, for instance, when continuous phase transitions occur; in such a case, particular caution has to be taken for correctly distinguishing between different clusters of data points.

APPENDIX B: *k*-MEANS CLUSTERING OF STRUCTURAL DATA

As detailed in Subsection III D, the PCA provides clear evidence that so far unexplored and unidentified ground state phases are hidden in the incredibly rich plethora of ordered bilayer structures in the asymmetric Wigner bilayer system. A step toward a more systematic analysis of the ground state configurations can be realized by applying a subsequent k-means clustering analysis (see Subsection II D 3) of the representation of the data set in terms of the nine leading principal components.

Before proceeding, three comments are in order:

(i) first, we have to fix the actual parameter of the k-means clustering, namely, the number of clusters, K, which is not known *a priori*;

(ii) when applying the k-means algorithm the choice of the initial location of the K different clusters is usually arbitrary; however, the final result may depend on the particular choice of the initial cluster coordinates and on the initial assignment of the different data points to these clusters. It is therefore good practice to apply the k-means clustering several times with independent initial conditions. The results of these independent clustering (termed $\mathbf{k}_1, \mathbf{k}_2, \ldots$) can then be analyzed, for instance, in terms of adjusted mutual information, $I_K(\mathbf{k}_i, \mathbf{k}_j)$ (as defined and discussed in Appendix E); in our investigations, we have used 40 independent clustering for a given K-value; the detailed results of these investigations are presented in Appendix D;

(iii) to simplify the analysis, we have reduced the data set $\mathbf{X}^{(\mathrm{asym})}$ by eliminating those trivial data points that are hexagonal monolayers and which are unambiguously characterized by $x = 0$. Ruling out these data points [which cover a large portion of the (A, η)-plane] leads to the data set $\mathbf{X}^*$, which thus covers all data points of the original data set except for those feature vectors of hexagonal monolayers. This elimination of data reduces the size of data but does not have any influence on the PCA part of our approach, as shown by the results obtained for the explained variance PEV, $\lambda_i^{(e)}$ (see Fig. 3). Henceforward, all quantities based on the reduced data set, $\mathbf{X}^*$, are specified by an asterisk and we refer to the k-means clustering of the data set $\mathbf{X}^*$ as k^*-means clustering.

With the PCA results for the available structures at hand, we can now proceed to the obvious subsequent steps: we have to find an appropriate value for K for the k-means classification of the structural data for which we have to analyze the results of several independent clustering in order to identify an accurate labeling of the data.

In practice, we proceed as follows: first, we define a reasonable range of K values, ranging in our case from 14 to 42 in integer steps. For each K-value, we then perform 40 independent clustering from which we choose the most suitable one; the rather lengthy procedure of how to compare the different clustering and how to select the "best" labeling for a given K-value has been deferred to Appendix D. This results in a set of "best" clustering, one for each K-value. In an effort to identify the most appropriate number of clusters (or, in our case, the number of structural families), K, we compare the clustering of this set of K-dependent "best"

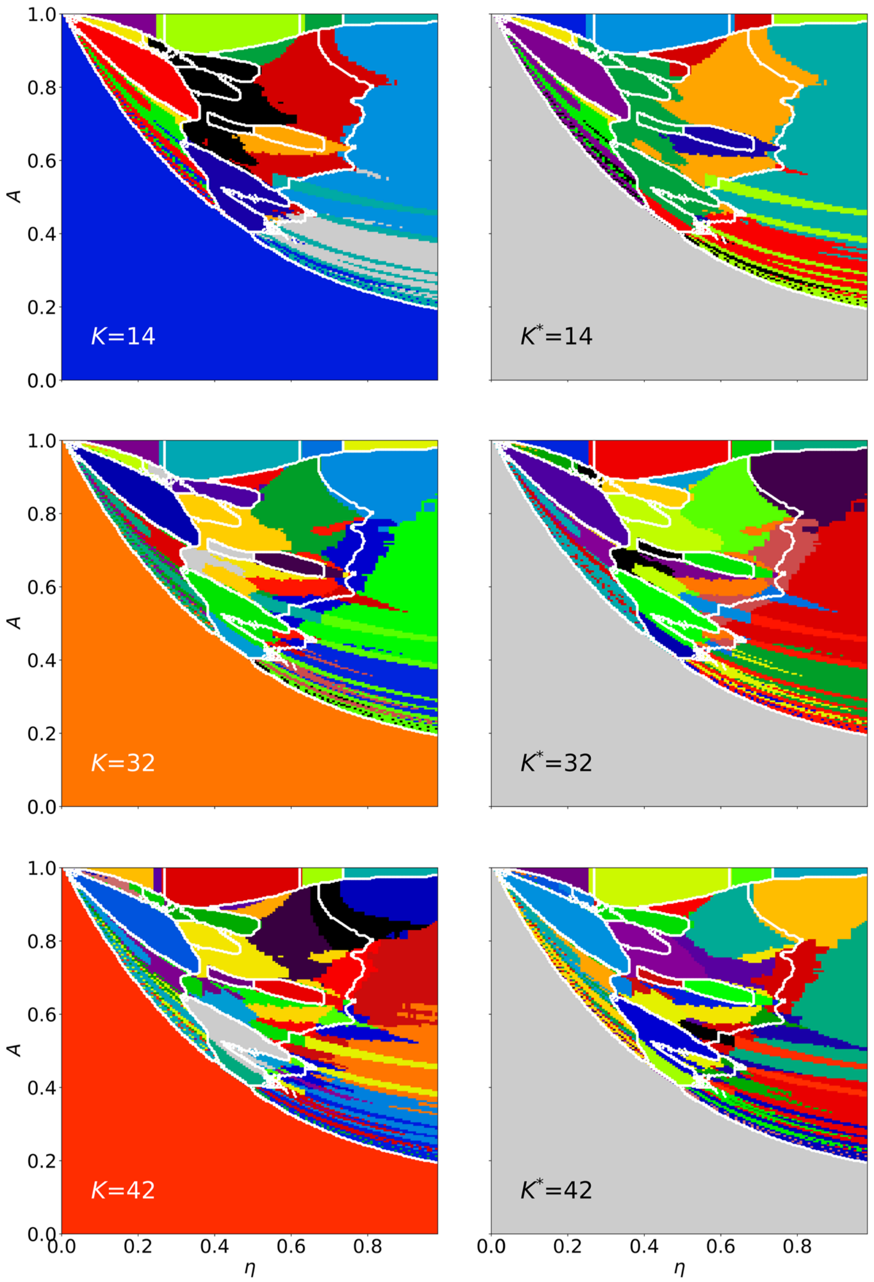

FIG. 10. Labeling the ground state configurations of Refs. 11–13 into K = 14 (top row), K = 32 (middle row), and K = 42 (bottom row) families by k-means (left) and k^*-means (right) clusters. The color scheme is arbitrary, and white lines indicate phase boundaries as specified in Refs. 11–13. While there are clearly differences in the clustering between the left and right columns for a given value of K and K^*, respectively, we emphasize that these are not relevant for the discussion here.

clustering in an analogous procedure as described in Appendix D: instead of comparing the results of independent clustering for a given value of K, we now compare the "best" clustering for different K-values (for more details, see also Ref. 40). The "best" clustering of the latter step is then labeled as the "best" k-means (or k^*-means) clustering of the data set $\mathbf{X}$ (or $\mathbf{X}^*$) and represents our revised mapping of the structural data set of Refs. 11–13 into families of structures.

In an effort to visualize the impact of the value of K on the k- (or k^*-)means results, we briefly summarize in the following the results obtained for three selected values of K, namely, K = 14, 32, and 42 (for a more detailed and graphical representation, we refer to

Ref. 40). To this end, we have redrawn in Fig. 10 the diagram of states of our system for these three values of K, showing the respective phase labeling as suggested by the best clustering results of several k-means (left panels) and k^*-means (right panels) clustering procedures; note in this context that the color-coding of the different families is arbitrary. From the panels, it is obvious that the actual value of K has a major impact on the final k-means (or k^*-means) results.[57]

The value $\mathbf{K} = \mathbf{14}$ corresponds to the number of phases that have been identified in Refs. 11–13 and that are specified in Table I; the corresponding clustering are shown in the top row of Fig. 10. Already for $K = 14$, most of the phases specified so far in literature[11–13] are correctly identified. Phases I through V are clearly visible with (almost) correct boundaries; also the honeycomb phase H, the phase II$_x$, as well as phase P$_3$ are identified essentially in a correct manner. Furthermore, the phase boundary of the V$_x$ phase is resolved with good accuracy. However, a total number of $K = 14$ (or $K^* = 14$) clusters is definitely too small to resolve appropriately further details of the phases S$_1$, S$_2$, P$_1$, and P$_2$; it also becomes obvious that the vast white regions of so far unclassified structures (see, e.g., Fig. 2) and/or the rich variety of yet unidentified substructures in phases I$_x$ and V$_x$ call for a closer and more refined analysis; thus, a value $K = 14$ is definitely not appropriate to reach these goals.

Proceeding to $\mathbf{K} = \mathbf{32}$ we find that—and, admittedly, for the time being at the qualitative level—our above requirements are met at a more satisfactory level. By a careful inspection of the related panels of Fig. 10, it seems that a set of 32 structural families is able to capture the variety of emerging structures: on one hand, this clustering does indeed accurately resolve the different phases identified in Refs. 11–13 (such as phases I through V as well as the phases II$_x$, H, P$_1$, P$_3$, and S$_2$); on the other hand, a value of $K = 32$ provides clear evidence of a rich variety of substructures in phases I$_x$ and V$_x$ that have not been classified so far in Refs. 11–13 and that are not captured by a $(K = 14)$-clustering, either.

Eventually, a value of $\mathbf{K} = \mathbf{42}$ was—from the conceptual point of view—considered as the upper limit: increasing further the value of K has led to the emergence of "new" phases that were—after all—only an artificial subdivision of well-defined phases. However, a closer comparison of the information contained for $(K = 42)$- and for $(K = 32)$-structure families—via an analogous procedure as described in Appendix D—reveals that the former one does not provide more substantial information on the emerging structure families than the latter one (we again refer to Ref. 40 for details).

Thus, we eventually select the "best" k^*-means clustering (cf., Appendix D) for a total number of $(K = 32)$ clusters as the "best" labeling of the structural data set from literature[11–13] of the asymmetric Wigner bilayer system into structure families. We henceforward refer to this result as $\mathbf{k}^*_{32}$ – clustering results and to the $c = (1, \ldots, 32)$ different clusters (i.e., to the different categories of structural families) as $\mathbf{k}^{*c}_{32}$ families, respectively.

APPENDIX C: ANALYZING PHASE-BOUNDARIES BY COMPARING ENERGETICALLY DEGENERATE BUT GEOMETRICALLY DIFFERENT FAMILIES OF STRUCTURES ACROSS THE PHASE DIAGRAM

In this Appendix, we discuss in more detail and on a more quantitative level how our combined approach (of a PCA and a subsequent k-means clustering) can cope with the issue of possibly degenerate structures emerging in the database of structures of our Wigner bilayer system.

As noted in the body of the text, the main challenge is to identify for a given pair of (A, η)-values the energetically most favorable configuration: if such an identification is made "by hand" the following implications have to be expected: (i) extremely small energy differences between competing structures might occur and (ii) the fact that the genetic algorithm produces energetically degenerate structures, which are characterized by different unit cells that describe an equivalent lattice. These implications are discussed in the following.

The clustering algorithm helps us to classify at each (A, η)-point *all* the structures provided by the evolutionary algorithm within a certain number of structure families (in our case—and as argued in the body of the text—we have chosen 32 families). In Fig. 11, we display in a color-coded manner the energy difference between the energetically most favorable structure (i.e., the energy of the ground state, termed E^*_{GS}) and the structure that pertains to the structure family with the energetically second best structure.[58] Thus, dark-colored areas (notably in black and purple) in Fig. 11 highlight regions in the diagram of states where the energetically best and second best structures exhibit very small differences in their energies (going down to values as small as 10^{-8} in relative units); note that this feature is particularly pronounced at phase boundaries. In contrast, orange to yellow areas in the (A, η)-plane indicate a large energetic gap between the ground state and energetically subsequent, competing structure.

An alternative view on the dataset provided in literature[11–13] via the structural families (as obtained via the clustering algorithm) is shown in Fig. 12: here, we display—again via a color code—the number of $\mathbf{k}^{*c}_{32}$ families whose energetically most favorable structure lies within an energy interval $\Delta E^*/N$ above the energy of the respective ground state, i.e., E^*_{GS}/N. Assuming different values of $\Delta E^*/N$, ranging in relative units from 10^{-5} down to 10^{-7} (as labeled) to the ground state energy, some two or three structures of competing families have been identified with the genetic algorithm. These findings indicate, in turn, the high numerical accuracy that is required to distinguish between energetically competing structures (see also discussion in Refs. 11–13).

APPENDIX D: ON THE RELIABILITY OF THE CLUSTERING ALGORITHM

In Fig. 13, we present the adjusted mutual information, $I_K(\mathbf{k}^*_i, \mathbf{k}^*_j)$, of $N_c = 40$ independent clustering results (with $i, j = 0, \ldots, N_c - 1$) of the k^*-means clustering algorithm for $K = 14$ and $K = 32$ clusters, respectively.

For a smaller number of clusters (i.e., $K = 14$), the algorithm is more stable: many samples exhibit a perfect score of the adjusted mutual information, i.e., $I_K(\mathbf{k}^*_i, \mathbf{k}^*_j) \simeq 1$ (cf. yellow pixels in the left panel of Fig. 13), indicating that the algorithm has identified the same results several times. For a larger number of clusters (i.e., $K = 32$), the situation is more complicated since the number of possible clustering results grows rapidly with the number of clusters.

For both values of K, there is evidence of qualitatively different clustering results to the clustering problem as depicted in Fig. 13. In order to elucidate this issue, we present in Fig. 14 histograms of

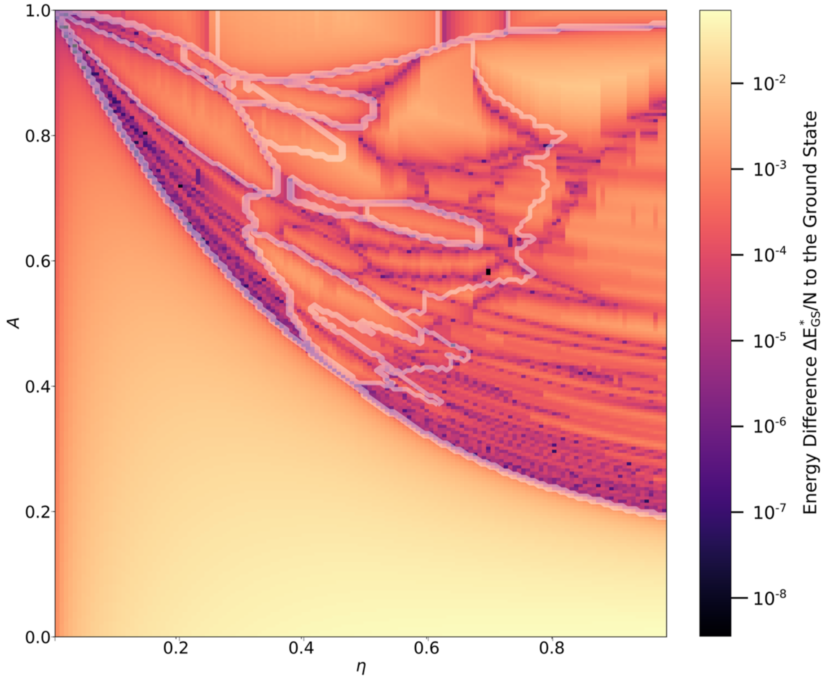

FIG. 11. Difference, $\Delta E^*_{\mathrm{GS}}/N = (E^*/N - E^*_{\mathrm{GS}}/N)$, between the energy (per particle), E^*/N, of the energetically most favorable structure amongst all non-ground state structural families with respect to the respective ground state energy (per particle), E^*_{GS}/N, of the asymmetric Wigner bilayer system for the data set taken from Refs. 11–13 for every state point in the (A, η)-plane. Opaque white lines indicate phase boundaries as presented in literature (see also Fig. 2). In the proximity of the boundaries of the H and I$_x$ phases and within the I$_x$ region, we observe very small values of $\Delta E^*_{\mathrm{GS}}/N$ (ranging typically from $\approx 10^{-8}$ to 10^{-6}), corresponding to nearly degenerate competing structures of different structural families; the related structures exhibit large values of the twelvefold symmetric order parameter $\Psi^{(4)}_{12}$.[12,40] This region corresponds to the newly identified ground state candidate family $\mathbf{k}^{*29}_{32}$ illustrated by the top left inset structure and in the cyan-emphasized area in the (η, A)-plane of Fig. 6(a); the details of this family of structures are summarized in Subsection III E and will be discussed in more detail in a forthcoming contribution.

the adjusted mutual information score, $I_K(\mathbf{k}^*_i, \mathbf{k}^*_j)$ (as depicted in Fig. 13 for selected clustering samples), some of which share little information with other clustering results (dark regions in Fig. 13 and orange distributions in Fig. 14) and others with a more consistent clustering result (bright regions in Fig. 13 and green distributions in Fig. 14).

Furthermore, and in order to compare the quality of different clustering results, we present in the bottom panels of Fig. 14 the column-wise average value, i.e., $\langle I_K(\mathbf{k}^*_i, \mathbf{k}^*_j)\rangle_j = \sum_{j=0}^{N_c} I_K(\mathbf{k}^*_i, \mathbf{k}^*_j)/N_c$, and the median (analogously defined) of the adjusted mutual information score of each clustering sample, $\mathbf{k}^*_i$, with all other clustering, $\mathbf{k}^*_{j\neq i}$, and mark both the maximum of the mean and the median. For $K = 14$, the clustering sample $i = 31$ seems to be a good choice for the final clustering result (cf. left panels of Figs. 13 and 14). However, for a larger number of clusters, e.g., $K = 32$, it is harder to decide what the optimal clustering might be: both samples (i.e., for $i = 0$ and $i = 38$) appear to have qualitatively similar traits as can be seen in the right panels of Figs. 13 and 14.

However, there is another ingredient that we can include in our analysis to bias the adjusted mutual information score into a physically motivated direction, namely, the ground state solutions of the *symmetric* Wigner bilayer system, which also shows up in the phase-diagram of the asymmetric Wigner bilayer system at $A = 1$. We can identify the fraction of the data points in the data sets $\mathbf{X}^{(\mathrm{asym})}$, which correspond to the ground state solutions of the symmetric Wigner bilayer system and collect them in a separate data set $\mathbf{X}^{(\mathrm{sym})}$. We assign all data points in $\mathbf{X}^{(\mathrm{sym})}$ to the phases I through V following the Table I[35,36] and collect the corresponding phase labels in the set $\mathbf{w}^{(\mathrm{sym})}$. Analogously, we collect in the set $\mathbf{k}^{(\mathrm{sym})}_i$ the particular clustering labels from the clustering result $\mathbf{k}_i$ (performed on the full data set $\mathbf{X}^{(\mathrm{asym})}$ after PCA), which correspond to the data points in $\mathbf{X}^{(\mathrm{sym})}$. Hence, the adjusted mutual information score $I_K(\mathbf{w}^{(\mathrm{sym})}, \mathbf{k}^{(\mathrm{sym})}_i)$ quantifies the overlap between the clustering result, $\mathbf{k}^{(\mathrm{sym})}_i$, and the analytically known labeling, $\mathbf{w}^{(\mathrm{sym})}$ (i.e., the amount of commonly labeled data points), of the data set, $\mathbf{X}^{(\mathrm{sym})}$, of the feature vectors of the ground states of the symmetric case. We now define the *biased adjusted mutual information score*, $S(\mathbf{k}_i, \mathbf{k}_j | \mathbf{w}^{(\mathrm{sym})})$, via

$$S(\mathbf{k}_i, \mathbf{k}_j | \mathbf{w}^{(\mathrm{sym})}) = I_K(\mathbf{k}_i, \mathbf{k}_j) \times \sqrt{I_K(\mathbf{w}^{(\mathrm{sym})}, \mathbf{k}^{(\mathrm{sym})}_i)} \times \sqrt{I_K(\mathbf{w}^{(\mathrm{sym})}, \mathbf{k}^{(\mathrm{sym})}_j)}, \tag{D1}$$

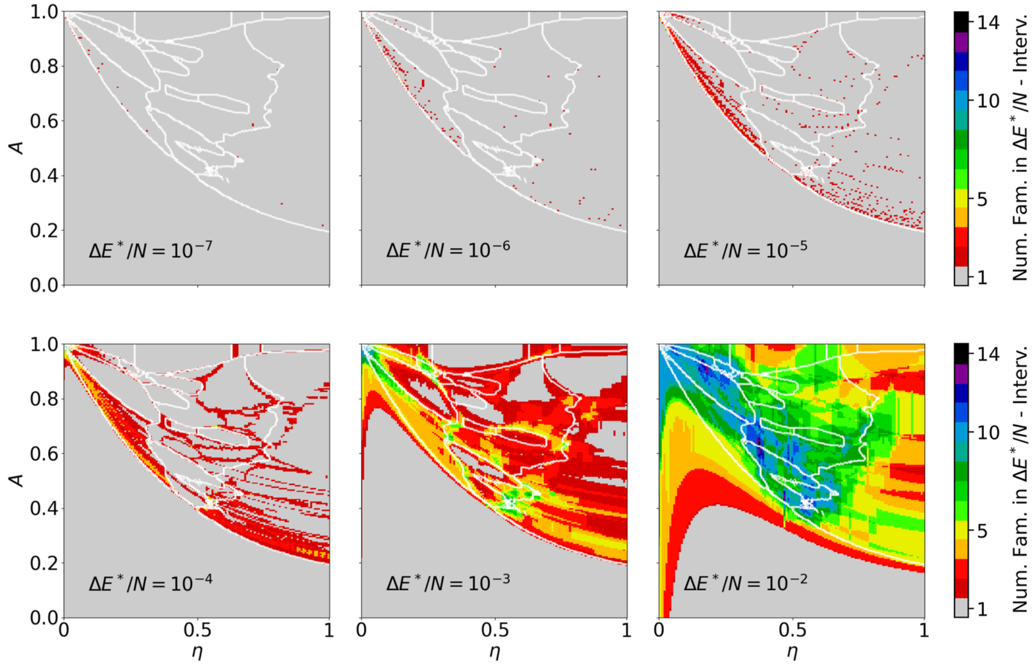

FIG. 12. Color-coded plot (cf. color bars on the right-hand side) of the numbers of families $(K = 32)$ that show an energy difference of at most $\Delta E^*/N$ with respect to the respective ground state candidates of the asymmetric Wigner bilayer system (suggested by Refs. 11–13) in the (A, η)-plane. The values of $\Delta E^*/N$, range from 10^{-7} to 10^{-2}, and are chosen separately for each panel (as labeled). Light-gray regions visualize regions where only one structure is identified within the respective $\Delta E^*/N$-interval; the colors according to the color bar emphasize the level of "$\Delta E^*/N$-degeneracy" at a given pair of the system parameters, i.e., the number of $\mathbf{k}_{32}^{*c}$ families that exhibit an energy difference to the ground state—at a given (η, A)-pair—of at most $\Delta E^*/N$. Phase boundaries from Refs. 11–13 (cf. Fig. 2) are emphasized by white lines.

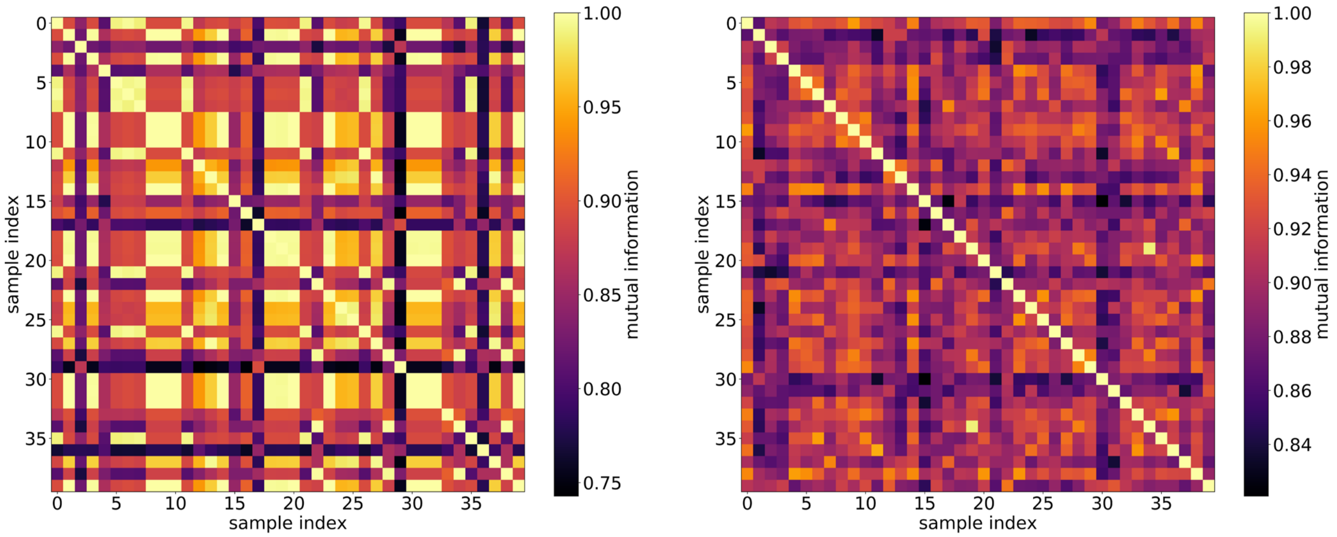

FIG. 13. Adjusted mutual information score, $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$, as defined by in Eq. (E2) of 40 different and randomly initialized k^*-means clustering results with $K = 14$ (left) and $K = 32$ (right) clusters, respectively. Values of $I(\mathbf{k}_i^*, \mathbf{k}_j^*)$ close to unity (bright, yellow regions) identify large overlap between the different clustering, $\mathbf{k}_i^*$ and $\mathbf{k}_j^*$, while smaller values, i.e., $I(\mathbf{k}_i^*, \mathbf{k}_j^*) \approx 0.8$ (black and purple) indicate less consistent results. Note that $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*) = I_K(\mathbf{k}_j^*, \mathbf{k}_i^*)$.

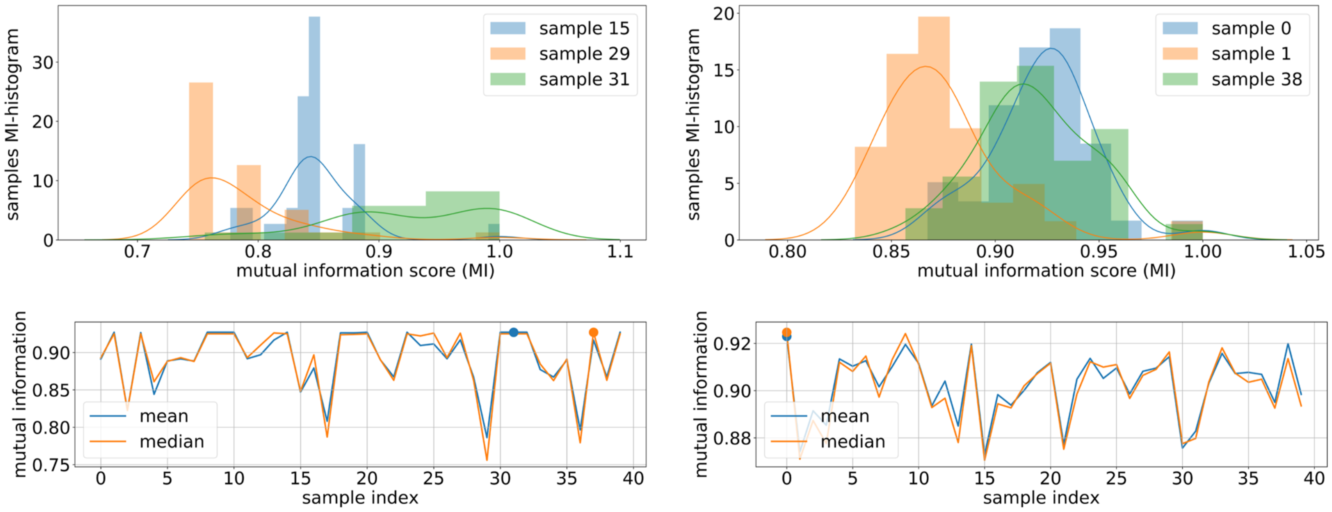

FIG. 14. Top-left panel: Column-wise histogram of $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$, as defined by Eq. (E2), of clustering samples $\mathbf{k}_{i=15}^*$ (blue), $\mathbf{k}_{i=29}^*$ (orange), and $\mathbf{k}_{i=31}^*$ (green) for the $K = 14$ clustering results shown in the left panel of Fig. 13. Top-right panel: Column-wise histogram of $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$ of clustering samples $\mathbf{k}_{i=0}^*$ (blue), $\mathbf{k}_{i=1}^*$ (orange) and $\mathbf{k}_{i=38}^*$ (green) for the $K = 32$ clustering results shown in the right panel of Fig. 13. Bottom panels: mean (blue) and median (orange) of the adjusted mutual information score, $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$, for each of the 40 clustering samples, $\mathbf{k}_i^*$, with respect to all other 39 clustering samples, $\mathbf{k}_{j\neq i}^*$ shown in Fig. 13 for $K = 14$ (left) and $K = 32$ (right) clusters (i.e., column-wise average and mean of the data shown in Fig. 13). Maxima of the mean and median of the adjusted mutual information score, $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$, as a function of the 40 sample indices are indicated by filled circles that emphasize clustering results that potentially share the most information with other results on average (or represent the maximum median thereof).

which weighs the adjusted mutual information, $I_K(\mathbf{k}_i, \mathbf{k}_j)$, of different k-means (or analogously k^*-means[59]) clustering results, $\mathbf{k}_i$ and $\mathbf{k}_j$, with the square root of the respective adjusted mutual information scores of $\mathbf{k}_i^{(\mathrm{sym})}$ and $\mathbf{k}_j^{(\mathrm{sym})}$ with $\mathbf{w}^{(\mathrm{sym})}$.

In our case, the biased adjusted mutual information score, $S(\mathbf{k}_i, \mathbf{k}_j|\mathbf{w}^{(\mathrm{sym})})$, is an important measure for the quality of the clustering results $\mathbf{k}_i$ and $\mathbf{k}_j$ since we demand of a corresponding labeling to be as accurate as possible, especially for the fraction of the data, $\mathbf{X}^{(\mathrm{sym})}$, that can be labeled analytically via $\mathbf{w}^{(\mathrm{sym})}$. In Fig. 15, we present the *biased* adjusted mutual information score, $S(\mathbf{k}_i^*, \mathbf{k}_j^*|\mathbf{w}^{(\mathrm{sym})})$, of the same selected samples as used in Fig. 14 and we also present the corresponding mean and median values of all biased sample scores [i.e., $S(\mathbf{k}_i, \mathbf{k}_j|\mathbf{w}^{(\mathrm{sym})})$] as we have already shown for the unbiased case [i.e., $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$] in the bottom panels of Fig. 14.

By comparing the biased, $S(\mathbf{k}_i^*, \mathbf{k}_j^*|\mathbf{w}^{(\mathrm{sym})})$, and the unbiased scores, $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$, we see that in general scaling the adjusted mutual information according to Eq. (D1) leads to smaller values of $S(\mathbf{k}_i^*, \mathbf{k}_j^*|\mathbf{w}^{(\mathrm{sym})})$ score as compared to $I_K(\mathbf{k}_i^*, \mathbf{k}_j^*)$. Especially the diagonal terms, $S(\mathbf{k}_i^*, \mathbf{k}_i^*|\mathbf{w}^{(\mathrm{sym})})$, whose adjusted mutual information scores are $I_K(\mathbf{k}_i^*, \mathbf{k}_i^*) = 1$ by definition (cf. Fig. 13), are now weighed by $I_K(\mathbf{w}^{(\mathrm{sym})}, \mathbf{k}_i^{(\mathrm{sym})}) \leq 1$, accounting for the quality of the clustering result with respect to the labels of the ground states of the symmetric case. Consequently, the scaling of the adjusted mutual information score, $I(\mathbf{k}_i^*, \mathbf{k}_j^*)$, via Eq. (D1) also causes an additional bias to larger values of the $S(\mathbf{k}_i^*, \mathbf{k}_j^*|\mathbf{w}^{(\mathrm{sym})})$ score for clustering results with large respective overlaps between $\mathbf{k}_i^{(sym)}, \mathbf{k}_j^{(\mathrm{sym})}$ and $\mathbf{w}^{(\mathrm{sym})}$ (i.e., commonly labeled ground states of the symmetric case); results with corresponding smaller overlaps of $\mathbf{k}_i^{(\mathrm{sym})}, \mathbf{k}_j^{(\mathrm{sym})}$, and $\mathbf{w}^{(\mathrm{sym})}$ are biased toward smaller values of $S(\mathbf{k}_i^*, \mathbf{k}_j^*|\mathbf{w}^{(\mathrm{sym})})$ (cf. rightmost bins of sample 0 and sample 38 in the top right panel of Figs. 14 and 15).

We now assume that "good" clustering results, which are biased toward large values of the $S(\mathbf{k}_i, \mathbf{k}_i|\mathbf{w}^{(\mathrm{sym})})$ score by labeling the symmetric part in the data set as well as possible, occur frequently and perform similarly in terms of the overall quality of the clustering. For such good clustering too, the mean (and the median) of the $S(\mathbf{k}_i, \mathbf{k}_i|\mathbf{w}^{(\mathrm{sym})})$ scores are biased toward larger values while being biased toward smaller values for qualitatively poor clustering results. We define the mean value, $\bar{k}_i$, of the biased adjusted mutual information score, $S(\mathbf{k}_i, \mathbf{k}_i|\mathbf{w}^{(\mathrm{sym})})$, of the $i, j = 0, \ldots, (N_c - 1)$ different clustering samples (cf. Fig. 13), by

$$\bar{k}_i = \frac{1}{N_c} \sum_{j=0}^{N_c-1} S(\mathbf{k}_i, \mathbf{k}_j|\mathbf{w}^{(\mathrm{sym})}). \tag{D2}$$

With $\bar{k}_i$, we have a reasonably good measure for comparing different clustering results for one given number of clusters, K: we here rely on $\bar{k}_i$ to quantify the quality of a clustering result, $\mathbf{k}_i^*$, of assigning the total number of K clusters correctly, given N_c independent clustering results (cf. Fig. 14). We evaluate $\bar{k}_i$ separately for all independent k-means and k^*-means clustering for several different values of K = 14 to K = 43: for a given value of K, the one sample from the respective $i = 0, \ldots, (N_c - 1)$ clustering with the maximum

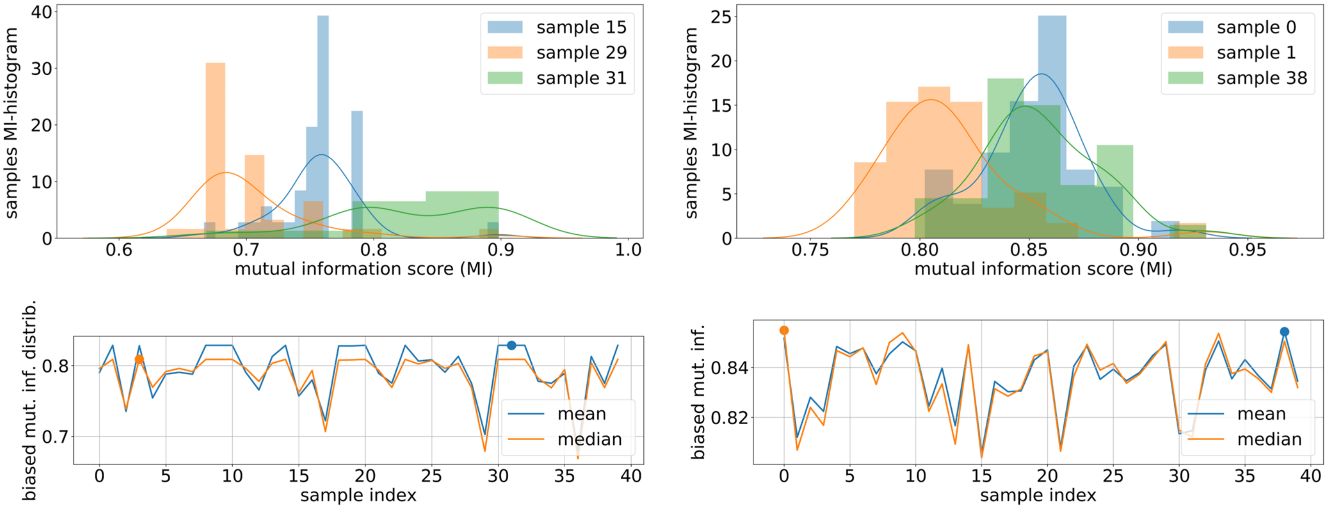

FIG. 15. Top and bottom rows: same as Fig. 14 but now for the *biased* adjusted mutual information score, $S(\mathbf{k}_i^*, \mathbf{k}_j^* | \mathbf{w}^{(\text{sym})})$, defined in Eq. (D1).

value of $\bar{k}_i$, given by Eq. (D2), is considered as the best clustering results.

APPENDIX E: ADJUSTED MUTUAL INFORMATION

Here, we provide a tool that is able to test the reliability of the partitioning of a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, of N data elements, into subsets, $\mathbf{U}^R = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_R\}$, with the following requirements: $\cup_{i=1}^R \mathbf{U}_i = \mathbf{X}$ and $\mathbf{U}_i \cap \mathbf{U}_j = \emptyset$ for all $i \neq j$.[60]

Commonly used clustering algorithms, such as k-means clustering or DBSCAN,[51] are, on the one hand, applicable in a variety of problems, but, on the other hand, not unique in their predictions: the final result of such algorithms usually depends on the clustering algorithm (such as the initial—usually random—choice of assigning data points to clusters, etc.), on the choice of the parameters of the algorithm, or on noise in the data.[61]

It is thus of particular relevance to comparing the results of different clustering of a given data set $\mathbf{X}$. Thus, we want either (i) to compare the results emerging from different clustering algorithms or (ii) to compare the results of the same algorithm but with different initial conditions. Assuming two different partitioning of a data set $\mathbf{X}$, i.e., $\mathbf{U} \equiv \mathbf{U}^R = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_R\}$ and $\mathbf{V} \equiv \mathbf{V}^C = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_C\}$ (satisfying both the above requirements), we want to quantify their overlap or, in other words, quantify the shared information of the two different clustering.

A fundamental class of techniques for comparing clustering of labeled data sets is based on information theoretic measures.[60] In our contribution, we use the concept of adjusted mutual information.[60,62]

In a first step, we define the $(R \times C)$-dimensional contingency table $M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}$ (see Table II), whose elements, $n_{ij} = |\mathbf{U}_i \cap \mathbf{V}_j|$, quantify the number of common objects in $\mathbf{U}_i$ and $\mathbf{V}_j$. The mutual information, $I_M(\mathbf{U}, \mathbf{V})$, of two different clustering, $\mathbf{U}$ and $\mathbf{V}$, is defined as[60,62]

TABLE II. Contingency table between two different clustering, $\mathbf{U}^R = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_R\}$ and $\mathbf{V}^C = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_C\}$, with $n_{ij} = |\mathbf{U}_i \cap \mathbf{V}_j|$ being the number of common objects in clustering $\mathbf{U}_i$ and $\mathbf{V}_j$; further $a_i = \sum_{j=1}^C n_{ij}$ and $b_j = \sum_{i=1}^R n_{ij}$.

$\mathbf{U}^R/\mathbf{V}^C$	$\mathbf{V}_1$	$\mathbf{V}_2$	$\ldots$	$\mathbf{V}_C$	Sums
$\mathbf{U}_1$	n_{11}	n_{12}	$\ldots$	$n_{1\,C}$	a_1
$\mathbf{U}_2$	n_{21}	n_{22}	$\ldots$	$n_{2\,C}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{U}_R$	n_{R1}	n_{R2}	$\ldots$	n_{RC}	a_R
Sums	b_1	b_2	$\ldots$	b_C	$\sum_{ij} n_{ij} = N$

$$I_{\mathrm{M}}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{R} \sum_{j=1}^{C} P_{\mathrm{UV}}(i,j) \log \frac{P_{\mathrm{UV}}(i,j)}{P_{\mathrm{U}}(i) P_{\mathrm{V}}(j)}, \tag{E1}$$

where $P_{\mathrm{UV}}(i,j) = |\mathbf{U}_i \cap \mathbf{V}_j|/N$ is the probability that a (random) data point belongs to both clusters $\mathbf{U}_i$ (in $\mathbf{U}$) and $\mathbf{V}_j$ (in $\mathbf{V}$); $P_{\mathrm{U}}(i) = |\mathbf{U}_i|/N$ and $P_{\mathrm{V}}(j) = |\mathbf{V}_j|/N$ denote the probabilities that randomly chosen data points fall into the cluster $\mathbf{U}_i$ and $\mathbf{V}_j$, respectively. In that way, $I_{\mathrm{M}}(\mathbf{U}, \mathbf{V})$ quantifies the information that is shared between two clustering and thus can be interpreted as a similarity measure for clustering; notably, the upper bounds of $I_{\mathrm{M}}(\mathbf{U}, \mathbf{V})$ are the quantities $H(\mathbf{U}) = -\sum_{i=1}^R P_{\mathrm{U}}(i) \log P_{\mathrm{U}}(i)$ and $H(\mathbf{V}) = -\sum_{j=1}^C P_{\mathrm{V}}(j) \log P_{\mathrm{V}}(j)$.[60]

The adjusted mutual information, $I_K(\mathbf{U}, \mathbf{V})$, corrects the information–theoretic measures of mutual information agreement of clustering for chance (see Refs. 60, 62, and 63 for details) and can be given by

$$I_K(\mathbf{U},\mathbf{V}) = \frac{I_{\mathrm{M}}(\mathbf{U},\mathbf{V}) - E_{\mathrm{MI}}(\mathbf{U},\mathbf{V})}{\max\left[H(\mathbf{U}),H(\mathbf{V})\right] - E_{\mathrm{MI}}(\mathbf{U},\mathbf{V})}, \tag{E2}$$

where the expected mutual information, $E_{\mathrm{MI}}(\mathbf{U},\mathbf{V})$, between two (random) clustering is defined by

$$E_{\mathrm{MI}}(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}=\max(1,a_i+b_j-N)}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{N\,n_{ij}}{a_i b_j}\right) \times \frac{a_i!\,b_j!\,(N-a_i)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_j-n_{ij})!\,(N-a_i-b_j+n_{ij})!}, \tag{E3}$$

with $a_i = \sum_{j=1}^{C} n_{ij}$ and $b_j = \sum_{i=1}^{R} n_{ij}$ being the partial sums over the contingency table $M[n_{ij}]_{j=1\ldots C}^{i=1\ldots R}$ defined in Table II.

A value of $I_K(\mathbf{U},\mathbf{V}) = 1$ indicates perfect overlap between two different clustering (i.e., the two clustering label the data equivalently but potentially use different numerical values to label the different clusters), a value smaller than one indicates differences in the clustering.

REFERENCES

[1] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, Phys. Rep. **810**, 1 (2019).

[2] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook* (Springer, 2010).

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, *Springer Series in Statistics* (Springer, New York, 2001).

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer, New York, 2013).

[5] M. Ceriotti, J. Chem. Phys. **150**, 150901 (2019).

[6] P. Geiger and C. Dellago, J. Chem. Phys. **139**, 164105 (2013).

[7] R. van Damme, G. M. Coli, R. van Roij, and M. Dijkstra, ACS Nano **14**, 15144 (2020).

[8] S. Becker, E. Devijver, R. Molinier, and N. Jakse, Phys. Rev. E **105**, 045304 (2022).

[9] S. Becker, E. Devijver, R. Molinier, and N. Jakse, Sci. Rep. **12**, 3195 (2022).

[10] M. Antlanger, M. Mazars, L. Šamaj, G. Kahl, and E. Trizac, Mol. Phys. **112**, 1336 (2014).

[11] M. Antlanger, "Ordered equilibrium structures in systems with long-range interactions," Ph.D. thesis, TU Wien, 2015.

[12] M. Antlanger, G. Kahl, M. Mazars, L. Šamaj, and E. Trizac, Phys. Rev. Lett. **117**, 118002 (2016).

[13] M. Antlanger, G. Kahl, M. Mazars, L. Šamaj, and E. Trizac, J. Chem. Phys. **149**, 244904 (2018).

[14] D. Gottwald, G. Kahl, and C. N. Likos, J. Chem. Phys. **122**, 204503 (2005).

[15] J. Fornleitner, F. Lo Verso, G. Kahl, and C. N. Likos, Soft Matter **4**, 480 (2008).

[16] G. J. Pauschenwein and G. Kahl, Soft Matter **4**, 1396 (2008).

[17] G. J. Pauschenwein and G. Kahl, J. Chem. Phys. **129**, 174107 (2008).

[18] G. Doppelbauer, E. Bianchi, and G. Kahl, J. Phys.: Condens. Matter **22**, 104105 (2010).

[19] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).

[20] W. Lechner and C. Dellago, J. Chem. Phys. **129**, 114707 (2008).

[21] We refer to Ref. 1 to a review on various machine learning tools applied to physical systems.

[22] At this point, it should be mentioned that other related techniques are also available, which may even directly operate on the structural data (i.e., on a data set of coordinates of the particles)—see, for instance, Refs. 64 and 65.

[23] See Ref. 2 for an in-depth discussion on different, problem specific similarity measures in data science problems. Methods from unsupervised machine learning[1] can be used to analyze a data set of feature vectors (or of order parameters in our case) for certain similarity measures in the features that may permit us to algorithmically organize the elements of the data set into an initially unknown set of categories.[2]

[24] K. Pearson, London, Edinburgh Dublin Philos. Mag. J. Sci. **2**, 559 (1901).

[25] I. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002).

[26] H. Steinhaus, B. Acad. Pol. Sci. **IV**, 801 (1956).

[27] E. Forgy, Biometrics **21**, 768 (1965).

[28] J. MacQueen, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (University of California Press, 1967), pp. 281–297.

[29] S. Lloyd, IEEE Trans. Inf. Theory **28**, 129 (1982).

[30] P. Z. Moghadam, S. M. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck, and D. Fairen-Jimenez, Matter **1**, 219 (2019).

[31] E. Boattini, M. Dijkstra, and L. Filion, J. Chem. Phys. **151**, 154901 (2019).

[32] Depending on the particularly applied clustering algorithm the number of clusters may be a preset parameter to the algorithm or may even be identified by the algorithm during execution.

[33] M. A. Kramer, AIChE J. **37**, 233 (1991).

[34] M. A. Kramer, Comput. Chem. Eng. **16**, 313 (1992).

[35] L. Šamaj and E. Trizac, Phys. Rev. B **85**, 205131 (2012).

[36] L. Šamaj and E. Trizac, Europhys. Lett. **98**, 36004 (2012).

[37] S. Earnshaw, Trans. Cambridge Philos. Soc. **7**, 97 (1842).

[38] M. Mazars, Phys. Rep. **500**, 43 (2011).

[39] P. P. Ewald, Ann. Phys. **369**, 253 (1921).

[40] B. Hartl, "Confinement-driven self-assembly of charged particles," Ph.D. thesis, TU Wien, 2020.

[41] The total number of possible compositions at each value of η is $N_{\mathrm{tot}} = 1 + \frac{1}{2}\sum_{n=2}^{N}(n - n \mod 2)$ given that $0 < N_2 \leq \frac{N_1}{2}$ for $N_1 > 1$, and only counting the monolayer structure with $N_2 = 0$ and $N_1 = 1$ once. For $N = 40$, we thus have $N_{\mathrm{tot}} = 401$.

[42] G. Lejeune Dirichlet, J. Reine Angew. Math. **40**, 209 (1850).

[43] G. Voronoi, J. Reine Angew. Math. **133**, 97 (1908).

[44] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1. 0 Contributors, Nature Meth. **17**, 261–272 (2020).

[45] W. Mickel, S. Kapfer, G. Schröder-Turk, and K. Mecke, J. Chem. Phys. **138**, 044501 (2013).

[46] For a discussion about alternative dimensional reduction and clustering tools and a justification for opting for PCA and k-means clustering in this manuscript, we refer to the supplementary material, Sec. III.

[47] J. Shlens, arXiv:1404.1100 (2014).

[48] The "elbow" in a linear scree plot,[66] such as shown in the main panel of Fig. 3, is defined as the point where the contribution of the significant eigenvalues (here realized via the PEVs) seems to level off.[4] Notably, this is a subjective measure as multiple elbows can occur in a scree plot. In this contribution, we opted for the second elbow (i.e., for i-values larger than $i = 9$) in Fig. 3—as opposed to the first elbow (at $i = 4$)—in an effort to include more information and to be on the safe side in our further analysis (cf., the supplementary material, Figs. 7 and 8, for the significance of $\mathrm{PC}_{i>4}$ in the state diagram of the asymmetric Wigner bilayer system).

[49] It should be noted that we do not claim that the first three PCs are sufficient to fully describe the structural diversity of the asymmetric Wigner bilayer system. We have decided to use PC_1 through PC3 in Fig. 5 simply for illustrative purposes in an effort to motivate the need for a more refined investigation of the diagram of states. Using other permutations of PC_1 through PC_9 will lead to similar plots that differ, however, in significant details, which, in turn, allows for a more systematic classification of structural phases in the system.

[50] L. V. D. Maaten and G. Hinton, J. Mach. Learn. Res **9**, 2579 (2008).

[51] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96* (AAAI Press, 1996), pp. 226–231.

[52] P. Stampfli, Helv. Phys. Acta **59**, 1260 (1986).

[53] M. Oxborrow and C. L. Henley, Phys. Rev. B **48**, 6966 (1993).

[54] B. Grünbaum and G. C. Shephard, *Tilings and Patterns* (Courier Dover Publications, 1987).

[55] T. Janssen, G. Chapuis, and M. de Boissieu, *Aperiodic Crystals–From Modulated Phases to Quasicrystals* (Oxford University Press, Oxford, 2007).

[56] Nowadays, it is easily possible to train a neural network in a supervised way with the objective of performing classification tasks.[1] For our purposes, such a task would be to classify structural data into a number of K different categories (identified, for instance, by unsupervised clustering), which would allow us to directly classify a structure from its geometric, structural data, e.g., via coordinates and lattice vectors.[67] The output of the classifier would then be the probability of a structure falling into any of the K clusters or families (when using "softmax" activation in the output layer of the neural network and "categorical cross-entropy loss" during training[1]), which may give additional insight when comparing competing structures.

[57] We report at this occasion that for all K-values both k- and k^*-means clustering results nicely correlate when evaluating the adjusted mutual information scores between the k- and k^*-means clustering, i.e., $I_K(\mathbf{k}_i, \mathbf{k}_j^*)$ as defined in the Appendix E; for more details and graphical representations, we refer to Ref. 40.

[58] Although most of the suggested ground state candidates identified by the evolutionary algorithm in Refs. 11–13 are very likely to represent the true ground state configurations of the asymmetric Wigner bilayer system at a given state point, there is no rigorous proof that they are, indeed, the ground states. Thus, whenever we use the term "ground state solutions," we refer to "ground state candidate solutions" of the asymmetric Wigner bilayer system.

[59] Also, for the k^*-means clustering, we rely on the set of analytically labeled data, $\mathbf{w}^{(\mathrm{sym})}$, of the entire data set, $\mathbf{X}^{(\mathrm{asym})}$, which correspond to the ground state solutions of the symmetric case, $A = 1$, in the evaluation of $S(\mathbf{k}_i^*, \mathbf{k}_j^* | \mathbf{w}^{(\mathrm{sym})})$, given by Eq. (D1): we, respectively, compare in $S(\mathbf{k}_i^*, \mathbf{k}_j^* | \mathbf{w}^{(\mathrm{sym})})$ the labels $\mathbf{w}^{(\mathrm{sym})}$ with $\mathbf{k}_i^{(\mathrm{sym})}$ and $\mathbf{k}_j^{(\mathrm{sym})}$, i.e., the fraction of the samples $\mathbf{k}_i^*$ and $\mathbf{k}_j^*$, which, respectively, corresponds to the known ground state structures of the symmetric Wigner bilayer system.

[60] N. X. Vinh, J. Epps, and J. Bailey, J. Mach. Learn. Res. **11**, 2837 (2010).

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[62] N. X. Vinh, J. Epps, and J. Bailey, in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09* (Association for Computing Machinery, 2009), pp. 1073–1080.

[63] W. M. Rand, J. Am. Stat. Assoc. **66**, 846 (1971).

[64] R. B. Jadrich, B. A. Lindquist, and T. M. Truskett, J. Chem. Phys. **149**, 194109 (2018).

[65] R. B. Jadrich, B. A. Lindquist, W. D. Piñeros, D. Banerjee, and T. M. Truskett, J. Chem. Phys. **149**, 194110 (2018).

[66] R. B. Cattell, Multivar. Behav. Res. **1**, 245 (1966).

[67] N. Guttenberg, N. Virgo, O. Witkowski, H. Aoki, and R. Kanai, "Permutation-equivalent neural networks applied to dynamics prediction," arXiv:1612.04530 [cs.CV] (2016).