A machine learning approach to graph-theoretical cluster expansions of the energy of adsorbate layers

Emanuele Vignola, Stephan N. Steinmann, Bart D. Vandegehuchte, Daniel Curulla, Michail Stamatakis, and Philippe Sautet

Citation: The Journal of Chemical Physics **147**, 054106 (2017);

View online: https://doi.org/10.1063/1.4985890

View Table of Contents: http://aip.scitation.org/toc/jcp/147/5

Published by the American Institute of Physics

A machine learning approach to graph-theoretical cluster expansions of the energy of adsorbate layers

Emanuele Vignola,[1,2] Stephan N. Steinmann,[1] Bart D. Vandegehuchte,[3] Daniel Curulla,[3] Michail Stamatakis,[4,a)] and Philippe Sautet[1,5,a)]
[1]*Université Lyon, ENS de Lyon, CNRS, Université Lyon 1, Laboratoire de Chimie UMR 5182, F-69342 Lyon, France*
[2]*Total Research and Technology Gonfreville, BP 27, F-76700 Harfleur, France*
[3]*Total Research and Technology Feluy, Zone Industrielle Feluy C, Seneffe, Belgium*
[4]*Department of Chemical Engineering, University College of London, Torrington Place, London WC1E7JE, United Kingdom*
[5]*Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, California 90095, USA*

(Received 30 May 2017; accepted 17 July 2017; published online 7 August 2017)

The accurate description of the energy of adsorbate layers is crucial for the understanding of chemistry at interfaces. For heterogeneous catalysis, not only the interaction of the adsorbate with the surface but also the adsorbate-adsorbate lateral interactions significantly affect the activation energies of reactions. Modeling the interactions of the adsorbates with the catalyst surface and with each other can be efficiently achieved in the cluster expansion Hamiltonian formalism, which has recently been implemented in a graph-theoretical kinetic Monte Carlo (kMC) scheme to describe multi-dentate species. Automating the development of the cluster expansion Hamiltonians for catalytic systems is challenging and requires the mapping of adsorbate configurations for extended adsorbates onto a graphical lattice. The current work adopts machine learning methods to reach this goal. Clusters are automatically detected based on formalized, but intuitive chemical concepts. The corresponding energy coefficients for the cluster expansion are calculated by an inversion scheme. The potential of this method is demonstrated for the example of ethylene adsorption on Pd(111), for which we propose several expansions, depending on the graphical lattice. It turns out that for this system, the best description is obtained as a combination of single molecule patterns and a few coupling terms accounting for lateral interactions. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4985890]

I. INTRODUCTION

Many chemical production processes for fuels, plastics, fine chemicals, fertilizers, and pharmaceuticals are catalytic in nature.[1] Catalyst development remains crucial to reduce the energy intensive nature of the processes and improve the chemical selectivity of the reaction. Three performance criteria can be identified: activity, selectivity, and stability. While catalyst development is mostly driven by experiments, computational studies provide valuable insights into the nature of the active sites and the reaction mechanism. They have become reliable over the last decade with respect to predictive power, computational speed, and can now actively participate in the catalyst design.[2–7]

Single crystal surfaces approximated as a periodic repetition of a unit cell (typically in the order of 10-20 surface atoms) are commonly applied to modeling heterogeneous catalysts and their energy is often described by Density Functional Theory (DFT).[8,9] This approach provides a fair trade-off between accuracy and computational speed. Surface structures, adsorption energies, and reaction mechanisms can be investigated rather easily at low coverage, e.g., one adsorbate per unit cell. At higher coverage, however, lateral interactions between adsorbates develop and cause adsorbate rearrangements on the surface, requiring larger and therefore more complex supercells.[10] The ordering of adsorbate structures depends on the adopted supercell. For instance, consider the adsorption of 3 molecules on 4 indistinguishable sites.[11] Only one (symmetry distinct) arrangement is possible in the smallest unit cell. However, a supercell with 16 sites and 12 adsorbates suggests a much higher number of possible arrangements, making the computations much more complex especially when chemical kinetics are involved. As this task becomes insurmountable for explicit first principles models, approximate methods are recommended for this endeavor.

Recent years have witnessed a surge of lattice based (kinetic) Monte Carlo (kMC) methods that rely on model Hamiltonians parameterized by DFT.[12–14] The advantage of model Hamiltonians is that they yield results much faster than the underlying DFT computations, while they are able to properly assess the configurational average and evolution of a surface under reactive conditions. In principle, the model Hamiltonian needs to account for three energy contributions: (i) The energy of the surface, which includes deformation energies, (ii) the interaction of adsorbates with the surface, and (iii) the interaction between adsorbates (also called lateral

a)Electronic addresses: m.stamatakis@ucl.ac.uk and sautet@ucla.edu

interactions). In this work, we focus on (ii) and (iii), with the surface distortion energy being implicitly incorporated in the interaction energies. Such a model is appropriate for surfaces that do not significantly restructure under reactive conditions. The *interaction model* for adsorbate-surface systems aimed at in this work is rather general and could be applied to surface reconstruction as well.

A two-dimensional Ising-type model is typically selected as the basic model Hamiltonian.[15,16] An Ising-type model is defined as a set of lattice points in a given spatial arrangement, which are mapped onto a set of spin-like values. The energy is derived from an effective Hamiltonian, which incorporates the effect of an external field and the coupling between different spins, e.g., given a positive coupling constant, if a site has positive spin, the neighboring site "prefers" a negative spin. The energy thus depends on the configuration of spins within the lattice and on the intensity of an external field. Such Ising-type models (and the closely related lattice-gas Hamiltonians) are very effective in the description of phase transitions of multicomponent materials.[17,18] They have also been employed to represent statistical models of adsorbed layers.[19] In general, the partition function of Ising-type models is not known analytically except in special cases.[20] Yet, Monte Carlo algorithms have been found to be very effective in converging to the ground state or in evaluating ensemble averages. The parameters of the model Hamiltonian have to be known beforehand and are either determined from truly *ab initio* energy evaluations or, in some systems, fitted to reproduce experiments.

Model Hamiltonians with good accuracy can be established for many problems, especially bulk materials. They are also applicable to systems larger than the scope of direct DFT computations. The well-known ATAT suite of programs by Ceder and co-workers automatically computes phase diagrams of alloys, based on model Hamiltonians that are fitted to DFT data.[21–23] The ATAT tool pertains to equilibrium phenomena not to the kinetics of a system. However, for catalysis, both the surface structure, i.e., coverage under realistic conditions, and the kinetics of the corresponding processes are of great significance. kMC has proven its potential in this respect having been successfully applied to a large variety of catalytic systems.[24–37] A recent kMC framework is the graph-theoretical *Zacros* code of Stamatakis *et al.*,[38,39] which is particularly suited for applications in catalysis, since it is able to deal with multidentate species, complex elementary steps (e.g., involving more than two sites in specific geometric arrangements), and reaction barriers changing in the presence of spectator species that exert lateral interactions. As is common in lattice-based kMC, the evolution of the system in time is represented as a set of transformations of a graph. The structural parameters of a catalytic surface are mapped to the graph's vertices, with occupation variables representing the presence of adsorbates.

Establishing the model Hamiltonian of such a complex system is more difficult than for a bulk alloy because the number of possible configurations is enormous (even though finite). Molecular species bond specifically to a surface layer and can adopt many geometries. Furthermore, the link between the adsorbate geometry and its pattern on the graphical lattice (synonymously called "cluster" herein) is not trivial. In the past, users had to construct a model Hamiltonian by carefully analyzing the computed configurations, tracking the system's relevant features, and mapping them to a graphical lattice. Herein, we instead rely on a *machine learning*[40] approach, which is based on pattern recognition. Once the recognition selection process has been defined, the state of the lattice is inferred automatically without significant knowledge of the system. Our contribution first sketches the theoretical background and then proposes a chemically meaningful, operative framework to automatically determine model Hamiltonians in terms of a cluster expansion (CE), given a set of configurations, their adsorption energies, and a graphical lattice. We apply our approach to the adsorption of ethylene on Pd(111) as a particularly important model system in heterogeneous catalysis.

II. THE CORRESPONDENCE BETWEEN REAL AND GRAPHICAL INTERACTIONS

Consider a system of M atoms. Graph theory implies that the total energy of the system can be expanded in a series of atomic interaction energy terms between n atoms, with $n \le M$,[41] taking the form of the following equation:

$$E_{tot}^{atm}(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_M}) = \sum_i W_i(\boldsymbol{r_i}) + \sum_{i>j} W_{ij}\left(\boldsymbol{r_i}, \boldsymbol{r_j}\right) + \sum_{i>j>k} W_{ijk}\left(\boldsymbol{r_i}, \boldsymbol{r_j}, \boldsymbol{r_k}\right) + \sum_{i>j>k>l} W_{ijkl}\left(\boldsymbol{r_i}, \boldsymbol{r_j}, \boldsymbol{r_k}, \boldsymbol{r_l}\right) + \cdots, \quad (1)$$

where the indexes $i, j, k, l, \ldots$ go from 1 to M, $W_i(\boldsymbol{r_i})$ is the (electronic or potential) energy of a given atom at location $\boldsymbol{r_i}$, and $W_{ijkl\ldots}$ is the energy contribution of the $i, j, k, l, \ldots$ group of atoms, which is a function of their location in (continuous) space, $\boldsymbol{r_i}, \boldsymbol{r_j}, \boldsymbol{r_k}, \boldsymbol{r_l}, \ldots$. These groups of atoms are also called "clusters."

Though this expansion is exact, it is not well suited to describe molecular systems. It can, however, be transformed from an atomic to a molecular expression. As an example, consider the interaction of two molecules (Fig. 1 depicts the case of two diatomics). The energy U_α of a molecule or "body" α is obtained as the sum of all terms involving atoms from that body

$$U_\alpha(\{\boldsymbol{r}\}_\alpha) = \sum_{i\in\alpha} W_i + \sum_{i\in\alpha, j\in\alpha} W_{ij} + \cdots, \quad (2)$$

where we use curly brackets to indicate the set of atomic coordinate ($\mathbf{r}$) belonging to molecule α and we dropped the argument ($\mathrm{r_i}$) for the sake of clarity. The interaction energy between two molecules α and β is similarly given by

$$U_{\alpha\beta}\left(\{\boldsymbol{r}\}_\alpha, \{\boldsymbol{r}\}_\beta\right) = \sum_{i\in\alpha, j\in\beta} W_{ij} + \sum_{i\in\alpha, j\in\alpha, k\in\beta} W_{ijk} + \sum_{i\in\alpha, j\in\beta, k\in\beta} W_{ijk} + \cdots. \quad (3)$$

This allows us to express the total energy of a system in terms of molecular cluster energies, i.e., in this case, a "cluster" refers

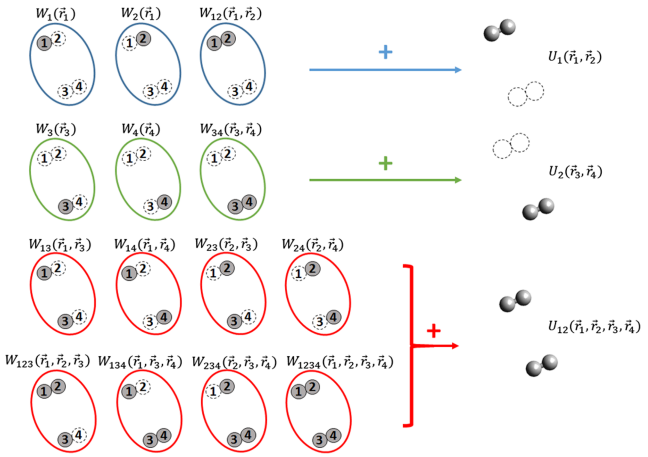

FIG. 1. From atomic to molecular cluster expansions. Two top rows: 1-body clusters, bottom two rows: 2-body clusters, where "body" refers to a molecule.

to a group of molecules

$$E_{tot}^{mol}(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_M}) = \sum_{\alpha} U_{\alpha}(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}) + \sum_{\beta>\alpha} U_{\alpha\beta}\left(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}, \{\boldsymbol{r}\}_{\boldsymbol{\beta}}\right) + \cdots, \quad (4)$$

where the first term accounts for the energy of each molecule α in the particular geometry $\{\mathbf{r}\}_{\boldsymbol{\alpha}}$, the second for the interaction energy between all pairs of molecules, the following term for triplets, and so on. This formulation was also deduced in dedicated books.[42]

The present approach is now specifically applied to adsorption energies. We divide atoms into two sets: the ones of adsorbates (ads, N atoms) and those of the catalyst surface (cat, M-N atoms). The adsorption energy is then given by

$$\Delta E_{ads} = E_{tot}(ads@cat) - E_{tot}(ads) - E_{tot}(cat). \quad (5)$$

For adsorption on relatively rigid surfaces, we assume the coordinates of the catalyst as fixed during the adsorption event. This assumption is standard in the theory of lattice statistics.[43] It is not compulsory and surface reconstructions can be described even with lattice based model Hamiltonians.[44] Furthermore, we account for the deformation energy of both the catalyst and the molecules in their respective interaction energy, as these deformations are rather local. Doing so, we can focus on the adsorption energy only,

$$\Delta E_{ads}(\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_N}) \approx E(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}, \ldots, \{\boldsymbol{r}\}_{\boldsymbol{\omega}}), \quad (6)$$

where $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r_N}$ are the coordinates of the atoms within adsorbed molecules labeled by $\alpha \ldots \omega$. In this case, the interaction energy is given by formally two or more body terms

$$E(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}, \ldots, \{\boldsymbol{r}\}_{\boldsymbol{\omega}}) = \sum_{\alpha} \left[U_{\alpha\tau}^{int}(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}})\right] + \sum_{\beta>\alpha} \left[U_{\alpha\beta\tau}^{int}\left(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}, \{\boldsymbol{r}\}_{\boldsymbol{\beta}}\right)\right] + \cdots, \quad (7)$$

where τ stands for the catalyst. In this equation, the first term accounts for the adsorption energy of "isolated" molecules and the second term for lateral interactions.

In the following, we describe how Eq. (7) is transformed into a cluster expansion on a fixed lattice and how its energy coefficients are determined. The first essential step is mapping Eq. (7) from real space onto a lattice in order to build Ising-type (or lattice gas) Hamiltonians. This coarse-graining is performed by using a two-dimensional lattice of points in real space. In the framework of the Zacros code, strictly positive occupation variables are used instead of pseudo-spins to identify what kind of species occupy which lattice point so that molecular configurations have a lattice representation. Let S be the set of occupation variables σ assigned to each molecular adsorption mode, i.e., di-σ and π-bound ethylene have a distinct occupation variable. Then, a mapping $\mathcal{M}$ from the lattice coordinates to the set of occupations will generate the lattice configuration. Let s_v be the occupation variable relative to the lattice point at positions $\boldsymbol{p}_v$, where v runs over all lattice points; then, we can write the energy of each configuration as

$$\hat{E}(s_1, s_2, \ldots, s_V) \approx \sum_{\alpha} \left[\hat{U}_{\alpha\tau}^{int}(\mathcal{M}(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}))\right] + \sum_{\beta>\alpha} \left[\hat{U}_{\alpha\beta\tau}^{int}\left(\mathcal{M}(\{\boldsymbol{r}\}_{\boldsymbol{\alpha}}, \{\boldsymbol{r}\}_{\boldsymbol{\beta}})\right)\right] + \cdots, \quad (8)$$

where we have used the hat to indicate the transition from continuous ($\boldsymbol{r}$) to discrete ($\boldsymbol{p}$) space. The approximate equality comes from the "coarse-graining," i.e., several configurations in real space can be projected into the same $\boldsymbol{s}$ vector, which results in an average energy for $E(\boldsymbol{s})$,

$$\hat{E}(\boldsymbol{s}) = \langle E(\mathcal{M}(\boldsymbol{r}))\rangle_{\{\mathcal{M}(\boldsymbol{r})=\boldsymbol{s}\}}, \quad (9)$$

where the brackets indicate the average over all acceptable realizations of a given configuration mapped into the same coarse-grained occupation vector representation, see curly brackets. As a result, the total energy may be written as

$$\hat{E}(s_1, s_2, \ldots, s_V) = \sum_{\mu} \hat{U}_{\mu}^{int}(\boldsymbol{s}) = \sum_{\mu} \varepsilon_{\mu}, \quad (10)$$

where the index μ stands for a collection of molecular indices $(\alpha, \beta, \ldots, \alpha\beta, \ldots)$, i.e., for the clusters that are encoded in $\boldsymbol{s}$, and ε_{μ} is the cluster energy coefficient relative to μ.

Notice that redundancies may appear in Eq. (10), i.e., some of the clusters nested in the sums may be equal. Combining identical terms together, Eq. (10) is converted into Eq. (11),

$$\hat{E} = \sum_{\kappa} \xi_{\kappa}\varepsilon_{\kappa}, \quad (11)$$

where ξ is the cluster multiplicity and κ runs over non-identical molecular clusters (also called patterns herein); in this scheme, ξ plays the role of a cluster basis function. Thus, κ ranges from 1 to |P|, the number of unique clusters detected in the set of configurations analyzed. Section III discusses the identification of the different clusters, with Sec. IV devoted to the determination of the cluster energies (ε) through inversion.

III. A RECOGNITION MEASURE TO MAP MOLECULAR CONFIGURATIONS INTO PATTERNS ON THE GRAPHICAL LATTICE

The projection of molecular coordinates onto the graphical lattice and the associated coarse-graining of the energy

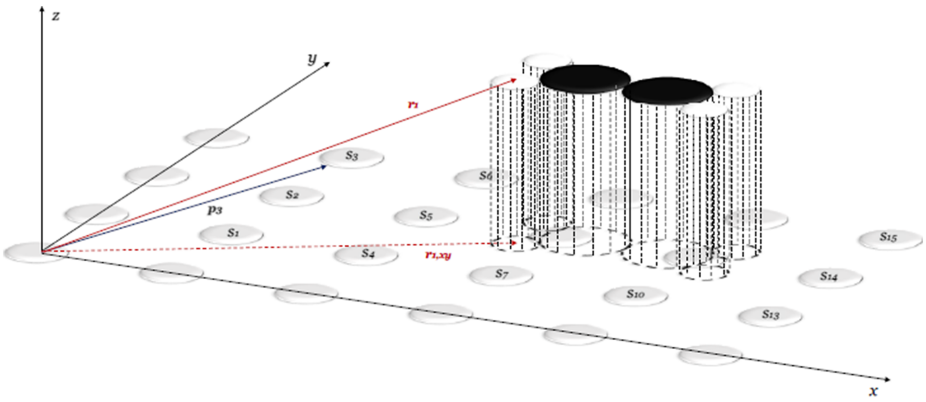

FIG. 2. Illustration of $\mathcal{M}$, the projection of adsorbate coordinates $\{\boldsymbol{r}\}$ on the lattice points at $\{\boldsymbol{p}\}$, applied to the case of ethylene adsorption on a Pd surface.

require the function $\mathcal{M}$. In other words, the projection $\mathcal{M}$ allows us to assign the binding mode of molecules without arbitrary choices by the user, simply by applying a set of rules to identify the clusters, to which we also refer as patterns, on the lattice that correspond to a given adsorption mode (see Fig. 2). In our case, we treat flat surfaces; hence, each vertex of the lattice is identified in Euclidean space by a 2-dimensional vector $\boldsymbol{p}_v = (\mathrm{x}_v, \mathrm{y}_v)$. Then, the coordinates of the atoms are projected on this plane. However, in general, the atom will not be projected exactly on top of one of the vertexes, but only in its vicinity. Additionally, depending on the spacing of the lattice, a single atom might be large enough to occupy more than one site. Hence, we need to assign a size to the atoms and the vertexes to reliably identify the occupied vertexes by computing the overlap between the area of the atom and the vertex. Many ways for estimating the size of an atom exist.[45] Two extreme, but simple, choices are available: van der Waals radii[46] ($\mathrm{R_{vdw}}$) provide an upper limit on the size of an atom and therefore impose significant separations between neighboring adsorbates. Covalent radii[47] ($\mathrm{R_{cov}}$), on the other hand, provide a lower limit of the size of atoms, which ensures that high coverages are accessible without leading to overlapping patterns. The area of the vertex could be chosen as the Wigner-Seitz cell, which is obtained by connecting the mid-points between vertexes by straight lines. However, the overlap between the atoms and the Wigner Seitz cell and the atoms would be cumbersome to compute. Hence, we decided to use the largest, non-overlapping circles to approximate the Wigner Seitz cells as illustrated in Fig. 3. Together with the circular atoms, the overlap $A_{i,v}$ of atom i with a given site v can be evaluated analytically

$$A_{i,v} = A(\boldsymbol{r_i}, \boldsymbol{p}_v) = \rho_i^2 \cos^{-1}\left(\frac{|\boldsymbol{r_i} - \boldsymbol{p}_v|^2 + \rho_i^2 - \rho_v^2}{2d\rho_i}\right) + \rho_v^2 \cos^{-1}\left(\frac{|\boldsymbol{r_i} - \boldsymbol{p}_v|^2 - \rho_i^2 + \rho_v^2}{2d\rho_v}\right) - \frac{1}{2}\sqrt{(\rho_i - |\boldsymbol{r_i} - \boldsymbol{p}_v| + \rho_v)(|\boldsymbol{r_i} - \boldsymbol{p}_v| - \rho_i + \rho_v)(\rho_i + |\boldsymbol{r_i} - \boldsymbol{p}_v| - \rho_v)(\rho_i + |\boldsymbol{r_i} - \boldsymbol{p}_v| + \rho_v)}, \tag{12}$$

where ρ_i and ρ_v are the radii of the atom and the vertex, respectively. To obtain a transferable measure, the overlap of each atom with vertexes is expressed as the fraction of the vertex that is covered by the atom

$$\tilde{A}_{i,v} = \frac{A_{i,v}}{A_v}, \tag{13}$$

where A_v is the area of the circle associated with vertex v, and $A_{i,v}$ is the overlap with atom i. Note that this framework only addresses the case of direct interaction with the catalyst surface and is not adapted for multi-layer adsorptions (e.g., adsorption on a H pre-covered surface), except if the definition of the catalyst lattice includes the information of, for example, an add-layer of hydrogen.

With a measure at hand, a recognition criterion can be finally cast on a quantitative basis. The [0, 1] range of fractional overlap must be divided into classes: occupied and unoccupied.

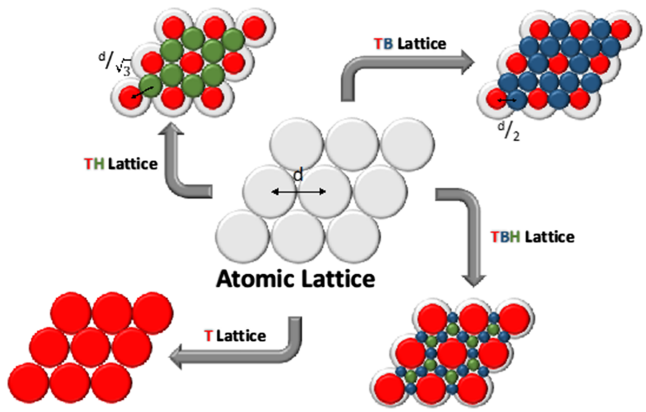

FIG. 3. Illustration of the construction of the different lattices: the size of the sites is chosen at the largest, non-overlapping circles, i.e., they touch at the midpoints. Therefore, the size of the top sites (in red) is the largest in the T lattice and the smallest in the TB lattice.

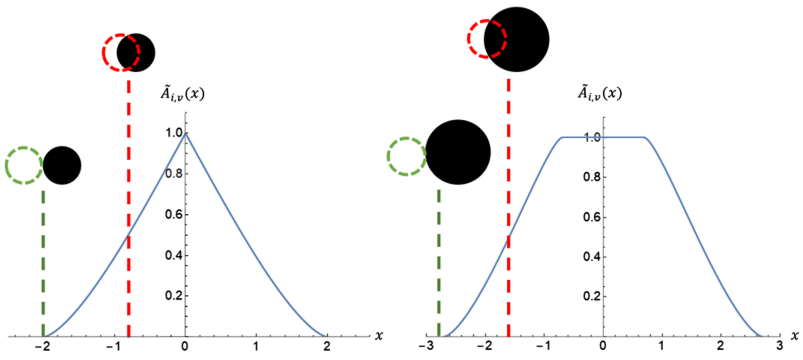

FIG. 4. Normalized overlap as a function of relative position, starting at circles that touch each other, indicated by the green line. The red line indicates the 0.5 threshold. The scheme on the left refers to two circles with a radius of unity, the one on the right to one circle with a unity radius and one of 1.7 Å (vdW radius of carbon).

The criterion is stated as follows: a molecule α recognizes and therefore occupies a site v if at least one of its atoms generates an overlap percentage greater than 50%, i.e., a vertex is assigned to a given molecule if at least one atom of that molecule produces $\tilde{A}_{i,v}$>0.5. Figure 4 shows the evolution of the overlap for the recognition of a site (circle in broken line) of radius 1.0 Å by an atom of identical size (full line circle) and by a carbon atom of van der Waals radius of 1.7 Å (left and right part of the scheme, respectively). The x-axis describes the distance between the centers of the site and the atom, with negative and positive values indicating that the site is on the "left" and "right" of the atom, respectively. The role of the radius of recognition is clearly captured, as the carbon atom shows an earlier recognition threshold (center separated by 1.6 Å; indicated by the red line) than the smaller site (centers separated by 0.8 Å). Of course, the size of the atoms and the vertexes need to be compatible, i.e., if $\rho_i < \sqrt{(2)}/2\ \rho_v$, then the vertex cannot be recognized.

The recognition process can be summarized as a sum over the atomic coordinates of Heaviside functions of the overlap-threshold difference, $\Theta(\tilde{A}_{i,v} - 0.5)$. Note that the geometrical threshold of 0.5 is a "logical" choice but implies that the size of the atoms needs to be compatible with the size of the sites. Since the size of the atoms is not uniquely defined anyway, we will test several linear combinations between R_{vdw} and R_{cov} in Sec. V. In practice, one probably would like that a hydrogen atom can be recognized by the smallest site in the lattice, but does not occupy more than one site, which gives quite simple geometric rules for the minimum and maximum combination of [x* R_{vdw} + (1 − x)* R_{cov}], given that the covalent and vdW radii of hydrogen are 0.32 and 1.2 Å, respectively. Thus, the occupation variable assigned to lattice point $v(s_v)$ is built as

$$s_v = \sum_{\alpha} \sigma_\alpha \max_{i \in \alpha} \left\{ \Theta \left[\tilde{A}_{i,v} - 0.5 \right] \right\}, \tag{14}$$

where r represents the collective vector of positions of all adsorbed atoms, $\boldsymbol{p}_v$ is the position of the vth vertex, α indexes the different adsorbates that are represented by the occupation variable s_α, and σ_α is the occupation value associated with adsorbate α. While Eq. (14) formally contains a sum, in the lattice gas description no vertex can be shared by two different adsorbates. In cases where such a sharing is chemically meaningful, the corresponding (pair of) species needs to be treated as a distinct molecule. In other words, the sum simply stands for the loop over molecules, which is necessary to perform the assignment of occupation variables to vertexes.

IV. THE OCCURRENCE MATRIX AND ITS INVERSION

Going one step back, we can recognize that Eq. (11) is a dot product between a vector ξ that counts the number of instances of each cluster in a given lattice configuration and a vector of cluster interactions ε. For instance, the given configuration contains ξ_1 instances of cluster 1 (which could, e.g., be a single-body term), ξ_2 instances of cluster 2, etc. To determine the energy contribution of each of these clusters, we need to compute the energy of several configurations (e.g., from DFT) and solve an inverse problem to obtain $\varepsilon_1, \varepsilon_2, \ldots$. Let us thus consider a set C of such configurations having energies $E_1, E_2, \ldots, E_{|C|}$ (recall that $|C|$ is the cardinality of set C, i.e., the number of configurations). Since we use "cluster" and "pattern" as synonyms herein, we call the number of distinct clusters |P| as stated in the context of Eq. (11). Let us further denote with $\xi_{c\kappa}$ the number of instances of pattern κ in configuration c. The energies $E_1, E_2, \ldots, E_{|C|}$ will then be given by Eq. (15),

$$\begin{cases} \xi_{\mathbf{11}}\varepsilon_1 + \xi_{\mathbf{12}}\varepsilon_2 + \cdots + \xi_{1|P|}\varepsilon_{|P|} = E_1 \\ \xi_{\mathbf{21}}\varepsilon_1 + \xi_{\mathbf{22}}\varepsilon_2 + \cdots + \xi_{2|P|}\varepsilon_{|P|} = E_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \xi_{\mathbf{c1}}\varepsilon_1 + \xi_{\mathbf{c2}}\varepsilon_2 + \cdots + \xi_{c|P|}\varepsilon_{|P|} = E_c \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \xi_{|C|\mathbf{1}}\varepsilon_1 + \xi_{|C|2}\varepsilon_2 + \cdots + \xi_{|C||P|}\varepsilon_{|P|} = E_{|C|} \end{cases}. \tag{15}$$

This system of equations is more conveniently casted in the matrix form

$$\boldsymbol{E} = \boldsymbol{\Xi} \cdot \boldsymbol{\varepsilon}, \tag{16}$$

where Ξ is a $|C| \times |P|$ matrix; its entry $\xi_{c\kappa}$ is the occurrence of pattern κ in configuration c.

Although |C| and |P| can be chosen to be equal, the Ξ matrix is in general a rectangular matrix. Whenever there are less patterns than configurations, the system is *overdetermined* and invertible "in the sense of the least squares." Operatively, systems are commonly inverted by the most general definition of matrix inverse, that is, the *Moore-Penrose pseudoinverse*.[48,49] The left pseudoinverse of the occurrence matrix is written as

$$\boldsymbol{\Xi}^{+} = \left(\boldsymbol{\Xi}^T \cdot \boldsymbol{\Xi} \right)^{-1} \cdot \boldsymbol{\Xi}^T, \tag{17}$$

multiplying Eq. (16) by the right-hand side of Eq. (17) and simplifying terms, the following solution for the pattern coefficients is obtained in the form of Eq. (18),

$$\boldsymbol{\varepsilon} = \boldsymbol{\Xi}^{+} \cdot \boldsymbol{E}. \tag{18}$$

The solution of Eq. (18) gives a set of coefficients to represent Hamiltonians of adsorbed layers.

The cardinality of the set P could be, in principle, very large; this has the effect of requiring an even larger set of configurations to be computed. Cluster expansions are, however, known to converge fast so that the expansion can be truncated at a finite distance and at low orders of many-body patterns.[50–53] Additionally, the system of equations is likely to contain linear dependencies: for instance, there could exist a configuration whose pattern count $\boldsymbol{\xi}$ is a linear combination of the counts of two other configurations. This may, in principle, be tackled by avoiding superposition of configurations while sampling. However, in terms of statistics and recalling that we map a continuous space (adsorbates on the surface) into a discrete space (lattice occupations), removing such "redundancies" would preclude to being able to assess the soundness (introduced) error due to this discretization. The other source derives from the truncation of the expansion: removal of long-range pattern may in fact generate identical (short-range) configurations. Since no *a priori* knowledge of the cutoff or of the expansion order is available for an arbitrary catalytic system, this second source of linear dependencies can hardly be avoided. A *singular value decomposition* (SVD) during the Moore-Penrose pseudo-inversion is capable of circumventing the issues related to the linear dependencies.[54] SVD (with eigenvectors corresponding to an eigenvalue below machine precision being removed) was employed for all of the selected applications discussed in Sec. V.

While the Moore-Penrose pseudoinverse identifies a unique matrix corresponding to the solution with the least squares of the residuals ($\boldsymbol{E} - \boldsymbol{\Xi} \cdot \boldsymbol{\varepsilon}$), one would like to find that the parameters have chemically reasonable values, since it would give confidence that the parameters are "transferable" to new configurations. In particular, we expect significant stabilizing single body patterns for chemisorbed molecules and less important but repulsive two-body patterns. For this reason, we will also report maximum and minimum values for the one- and two-body contributions. In our experience, this expectation is satisfied for systems of equations that have a rank greater or equal to |P|.

A first step towards assessing the reliability or rather the trustworthiness of a given least squares fit can be done "internally." Since, as mentioned above, the "sampling problem" does not have an *a priori* solution, the importance of each point/configuration for the fit is an important measure. For example, if one parameter of the fit is connected to only one configuration, then this configuration is highly important (in statistics we would call it an outlier) for the fit and the value of this parameter is uncertain. The only way to properly deal with such a situation is to increase the sampling set to include more configurations, allowing us to define this specific parameter. *Cook's distance*[55,56] of a configuration is convenient to measure its influence on the fit. Cook's distance can readily be obtained from the Moore-Penrose pseudoinverse. First, the so-called *hat matrix* of the system $\boldsymbol{H}$ is defined as

$$\boldsymbol{H} = \boldsymbol{\Xi} \cdot \boldsymbol{\Xi}^{+}. \tag{19}$$

The diagonal element of the hat matrix H_{ii} is known as the *leverage* of point i; it measures how far a configuration stands from the average pattern occurrence. The leverage is used to standardize the differences between the configurational energy predicted by the model and the ones known from raw data; these normalized differences are known as *standardized residuals* e_i and are defined by the following equation:

$$e_i = \frac{E_i - \sum_j \xi_{ij} p_j}{\sqrt{1 + H_{ii}}}. \tag{20}$$

Cook's distance D_i of a given configuration is obtained as

$$D_i = \frac{H_{ii} e_i^2}{|P|\, \sigma^2 (1 + H_{ii})^2}, \tag{21}$$

where σ^2 is the mean squared prediction/raw data residual. Cook's distance is a tool to highlight the need for a more thorough acquisition of data/sampling. Rigorously, one should perform an F-test to identify the influential points and then analyze these points in order to decide whether or not they are to be included in the final analysis. While some authors suggest that 4/n (with n being the number of samples) should be the maximum Cook's distance,[57] others suggest to simply remove samples with Cook's distances above one.[58] In the following, we adopt the later cutoff. Since our data points are obtained from optimized DFT geometries, high Cook's distances cannot be equated to regular outliers: they still are physically meaningful. However, they can be considered outliers with respect to the training set. Hence the user has the choice to either increase the sampling around these points or to remove them, if he considers them irrelevant for his specific application.

Cook's distances measure the sparseness of the available sample. A more frequently resorted tool to analyze interpolation reliability is cross-validation.[40] Cross validation schemes partition the sample into a training and a validation set; regression is then performed using data contained into the former, while the latter is employed to obtain a residual (error estimate). The scheme is iterated as to cover all the possible training/validation partitions of the sample, and residuals are summed as to get validation scores for each of the model's coefficients. Although no cross validation score was calculated in the present work, Cook's distances and cross-validation scores are complementary to address the robustness of models; assuming that clusters with high uncertainties and/or close to zero mean values during cross-validation should be either refined or removed, this would, indeed, help "compress" the cluster expansions and making them more robust/transferable. Furthermore, this could be exploited to identify the configurations that require higher body patterns to be well described.

Pattern detection and determination of the cluster energies as describe above are implemented in an in-house FORTRAN90 code interfaced with LAPACK[59] and will be made public in the near future.

V. APPLICATION TO THE ETHYLENE-PALLADIUM SYSTEM

In Secs. II–IV, a complete framework to recognize patterns and parameterize a cluster expansion for adsorbate layers was proposed. In this section, this scheme is applied to ethylene adsorption on the Pd(111) surface.

Palladium is well-known in the catalysis community for its hydrogenation capacity, e.g., for fatty acid hydrogenation.[60] The hydrogenation of ethylene over Pd is a widely studied model system.[61–63] Previous theoretical studies have identified two single-molecule adsorption modes: The most stable mode is the di-σ mode in which the molecule is bonded to two palladium atoms, distorting its geometry quite strongly according to the Dewar-Chatt-Duncanson model.[64,65] The latter model accounts for hydrocarbon-metal interactions in which a fraction of metal electron occupies the anti-bonding orbitals of the molecule.[66] The carbon-carbon bond distance extends by 0.12 Å (9%) and the hydrogen atoms are tilted by 9° instead of being co-planar. The other adsorption mode is the π mode. This mode induces no significant change in the molecular structure. The molecule merely interacts with a single palladium atom around its center of mass.

A. The training set

Out of these two basic modes, we have constructed a diverse training set of 32 configurations on a p(4 × 4) slab supercell of Pd(111) that span different coverages (1/4 monolayer, ML to 1/16 ML) and different relative orientations. The most important ones are represented on Fig. 5. We started with the low-coverage (1/16 ML) situation, where the di-σ and π configurations are isolated in the surface unit cell (see configuration 7 and 29 in Fig. 5). Then, we also included three configurations representative of the high coverage situation (1/4 ML): 4 π bound, 4 di-σ bound, and 2 π together with 2 di-σ bound molecules on the p(4 × 4) cell. 31 configurations representative of intermediate coverage (1/8 ML) were constructed based on various combinations of di-σ and π coordination and neighboring adsorption sites. Additionally, 10 configurations (3/16 ML) were constructed based on the 1/8 ML configurations by adding one molecule wherever possible. Note that the construction of this training set is merely exemplary and is neither complete nor ideal but just serves to illustrate the application of the framework outlined above. In other words, some of the configurations included might rarely occur in reality while others, in particular high coverage structures, might be missing. We consider the question of the construction and validation of the training set as a separate topic, which will be addressed in the future. All configurations were optimized at the DFT level exploiting the Projector Augmented Wave (PAW) formalism.[67,68] All computations were performed with VASP 5.3.3.[69,70] The functional of Perdew, Burke, and Ernzerhof (PBE)[71] was used, with the dispersion correction of Steinmann and Corminboeuf (dDsC).[72,73] The (111) surface was modeled by a p(4 × 4) unit cell with 6 metallic layers, 2 of which held fixed to simulate bulk properties. A vacuum layer of 15 Å was used. The plane wave basis set was chosen to have a cut-off energy of 400 eV. Brillouin zone integration was performed by a 3 × 3 × 1 Monkhorst-Pack[74] k-points grid and a Methfessel-Paxton smearing[75] of 0.2 eV. The wavefunction and geometric gradient were converged to 10^{-6} eV and 5×10^{-2} eV/A, respectively. As for our computational tools, van der Waals radii were taken from the work of Bondi,[46] while covalent radii came from Pyykkö and Atsumi.[47]

Figure 5 presents the adsorption energy per ethylene molecule for all investigated configurations ordered by increasing energy (configurations 1 and 32 are most and least stable, respectively) per ethylene molecule. The structures of the high (1/4 ML) and low (1/16 ML) coverage configurations are depicted as well. As expected, the di-σ mode (configuration 7; E_{ads} = −1.28 eV) is significantly more stable than the π mode (configuration 29; E_{ads} = −1.14 eV). In general, the through-surface and through-space lateral interactions between co-adsorbed molecules should have a destabilizing effect either due to Pauli repulsion or by limiting electron donation to or from the surface. We therefore expect that the di-σ mode at low coverage (1/16 ML) exhibits the lowest adsorption energy per molecule. Configuration 7 in Fig. 5 corresponding to that mode is, however, surpassed by six other modes lower in energy. Considering configurations 1 and 2 that correspond to the high coverage (1/4 ML) di-σ mode, we recognized slightly attractive lateral interactions as being responsible (in the order of 0.01 eV per molecule). The small stabilization originates from a subtle balance between repulsive and attractive London dispersion interactions. Indeed, we have checked that the results at the PBE level, i.e., without the

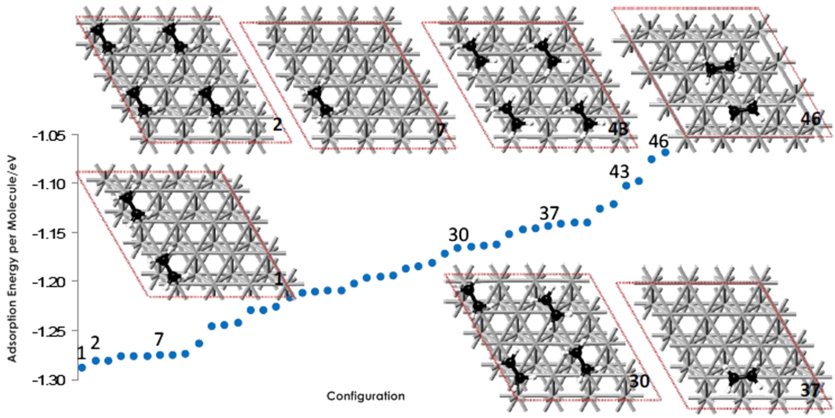

FIG. 5. Adsorption energies per molecule in eV for all configurations considered. Geometries are given for the low (7 and 37) and high (2, 30, 43) coverage limits and the most and least stable binding structure (1, 46). The unit cell is given by red dotted lines, Pd (in grey) atoms of the first two layers are shown. Carbon is depicted in black and hydrogen in white. The structural representation has been produced by the MAPS software.[76]

explicit inclusion of dispersion interactions, are consistent with the expectation of purely repulsive lateral interactions. In other words, since the adsorbates in the high coverage regime not only experience repulsion but also are stabilized by dispersion interactions, predicting the stability of high coverage structures is sensitive to the use of a dispersion correction.

B. Pattern recognition

Pattern recognition is first carried out for the two low coverage (1/16 ML) adsorption modes (di-σ and π in Fig. 6). Clusters are shown as a function of the size of the atoms, expressed as the barycenter function of R_{cov} and R_{vdw} [$x \cdot R_{vdw} + (1 - x) \cdot R_{cov}$], and the different sites distinguished by the graphical lattice. As mentioned in Sec. III, considering H adsorption as a limiting case, we can derive a minimal fraction of the vdW radius to be included so that H is large enough to occupy the smallest site, as well as a maximal fraction, assuming that a H atom at a given site should not be recognized as occupying a neighboring site as well. Therefore, we show the results for the x_{min} and the x_{max} in Fig. 6. The first line is a lattice only accounting for top sites and is denoted by T. Adding the 3-fold hollow sites (H) or two-fold bridge sites, (B) represents a more detailed lattice (TH and TB lattice, respectively). Finally, combining all three sites together yields the TBH lattice in the bottom line. In these pictures, the position of the C and H atomic nuclei is given by black dots, the size of the atom by a black circle with the corresponding atomic radius and the occupied sites (overlap > 50%) are given by colored circles (T, B, and H are red, blue, and green, respectively). For the coarsest T lattice, the two adsorption modes (di-σ and π) are readily distinguished as occupying two sites and one site, respectively, see Fig. 6, top row. This accounts well for the expected bi- and mono-dentate adsorption mode and the dependence thereof on the chosen atomic radii is negligible.

This significantly differs from the more detailed TH lattice: At x_{min} (30% R_{vdw}), a "top-top" mode is obtained for the di-σ configuration, which intuitively corresponds to an ethylene π bond being transformed into two C–Pd bonds. Accordingly, for the π configuration, a top mode is obtained. At x_{max} (70% R_{vdw}), however, the strongly overlapping sites, characteristic for the bond and represented by the top sites previously recognized, are now surrounded by a crown of (hollow) sites. Recognizing that these sites are occupied by ethylene is

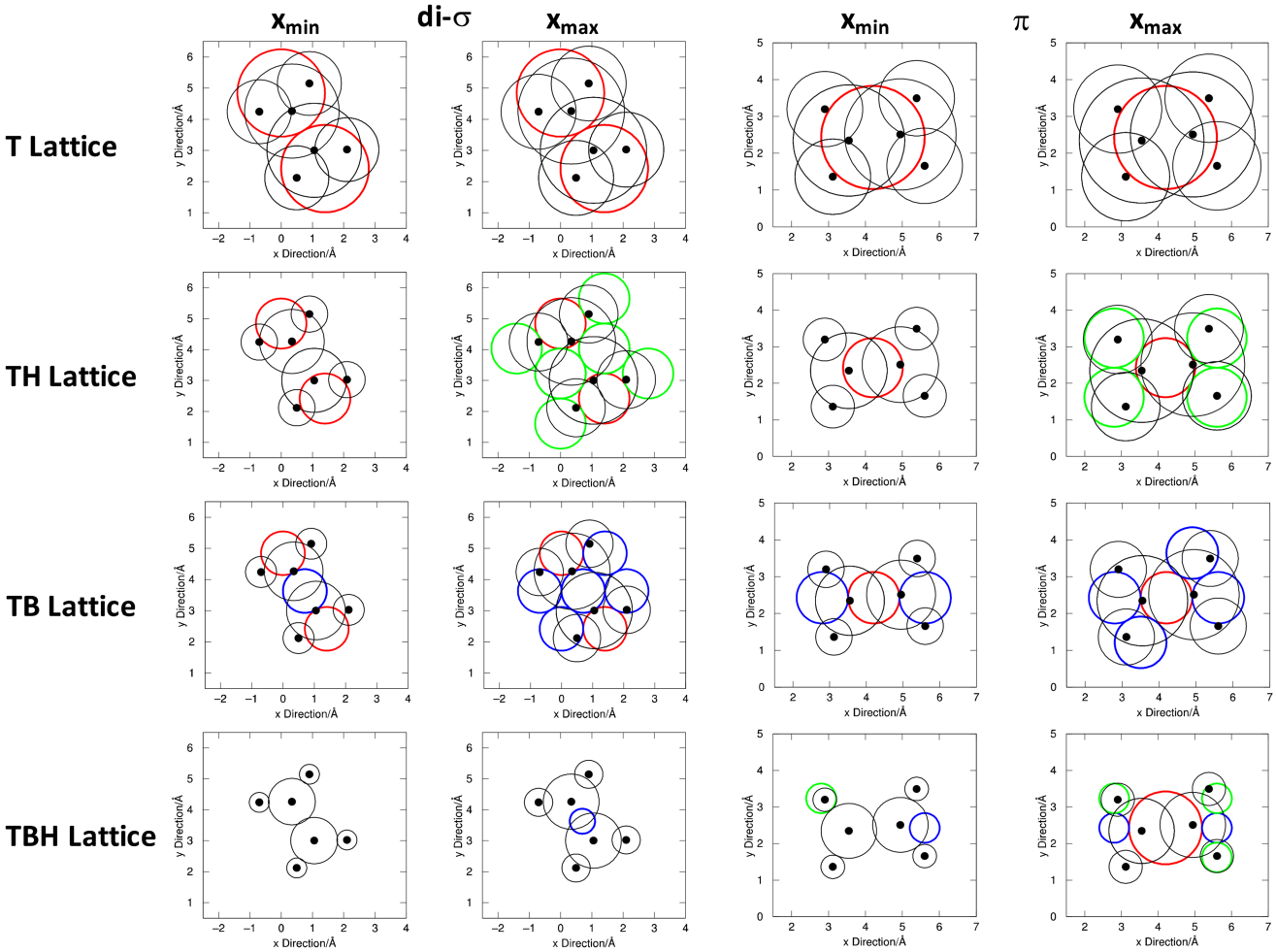

FIG. 6. Recognition of the di-σ (left, configuration 7) and π (right, configuration 37) as a function of the atomic radii [$x \cdot R_{vdw} + (1 - x) \cdot R_{cov}$] and the four different lattices. C and H atomic nuclei and radii are indicated as black dots and circles, respectively. Recognized top, bridge, and hollow sites are given by red, blue, and green circles, respectively. X_{min} and x_{max} are 0.8 and 1.0 for the T lattice, 0.3 and 0.7 for the TH lattice, 0.2 and 0.5 for the TB lattice, and 0 and 0.15 for the TBH lattice, respectively.

important, in the sense that it excludes other molecules to adsorb on these sites. In the case of x_{max}, this excludes, for instance, that a second di-σ mode adsorbs in the immediate neighborhood without distortion (the top site in the empty corner delimited by the green hollow sites cannot be occupied without an overlap with the already occupied H sites). The TB lattice, which in general is less relevant than the TH lattice, as several small adsorbates such as H or CO strongly interact with the hollow sites, overall provides a similar description, with patterns that account well for the elongated shape of the molecule. In contrast to the TH lattice, at x_{min} (20% R_{vdw}), the bridge sites already play a role and are recognized as being occupied by the molecule, which makes, of course, perfect sense, since they are in the middle of the C–C and between the H atoms of the CH_2 group for the di-σ and π configuration, respectively. The TBH lattice turns out to be least useful for our purpose mostly because of the large imbalance of the size of different sites: since the smallest sites have a radius of only 0.4 Å, x_{min} and x_{max} are 0 and 0.15, respectively, which does not allow for a reliable recognition of the different adsorption modes. Therefore, we will not discuss results for this lattice in Sec. V C. In conclusion, the chemically most intuitive patterns are obtained on the T lattice, independent on the atomic radii. On the TH lattice, only with the larger radii (70% R_{vdw}), reasonable patterns are obtained that reflect well the space occupied by ethylene and the role hollow sites could play in this case. However, in this case adsorption on sites close by is not possible without distortion, a topic on which we will come back in Sec. V C. On the TB lattice, already at x_{min}, the picture is "reasonable," although just like on the top lattice, the different adsorbates can in principle approach each other very closely. Finally, the sites in the TBH lattice are so small that the recognition is very sensitive to the atomic radii and no particularly suitable patterns are obtained.

Having discussed the low coverage patterns, we now address the following question: Do lateral interactions lead to detectable distortions? To do so, Fig. 7 addresses the extreme case of configuration 46, which is the least bound configuration at the per ethylene molecule basis and is therefore subject to the most significant lateral interactions. While the T lattice remains insensitive to the atomic radii and simply identifies two π modes, the TH and TB lattices clearly illustrate the distortion of the p mode due to lateral interactions (compared to the two left columns of Fig. 6): instead of simply a top site or a linear bridge-top-bridge pattern obtained at low coverage for x_{min}, the lateral interactions induce a distortion that leads to bent, asymmetric structures. In fact, the distortions due to the lateral interactions are actually seen both by the loss of some sites and the gain of others. In other words, distortion arises only when molecules are arranged in such a way that their van der Waals contours overlap; consequently, their graphical representation is reshaped. This implies that the distorted single-body patterns should not occur "alone," but only in the vicinity of other patterns. Hence, we can expect that there are some two-molecule patterns that should be treated as a single entity, as opposed to two independent molecules, simply because the distorted patterns cannot exist "alone." This is also the only way of dealing with situations where neighbouring

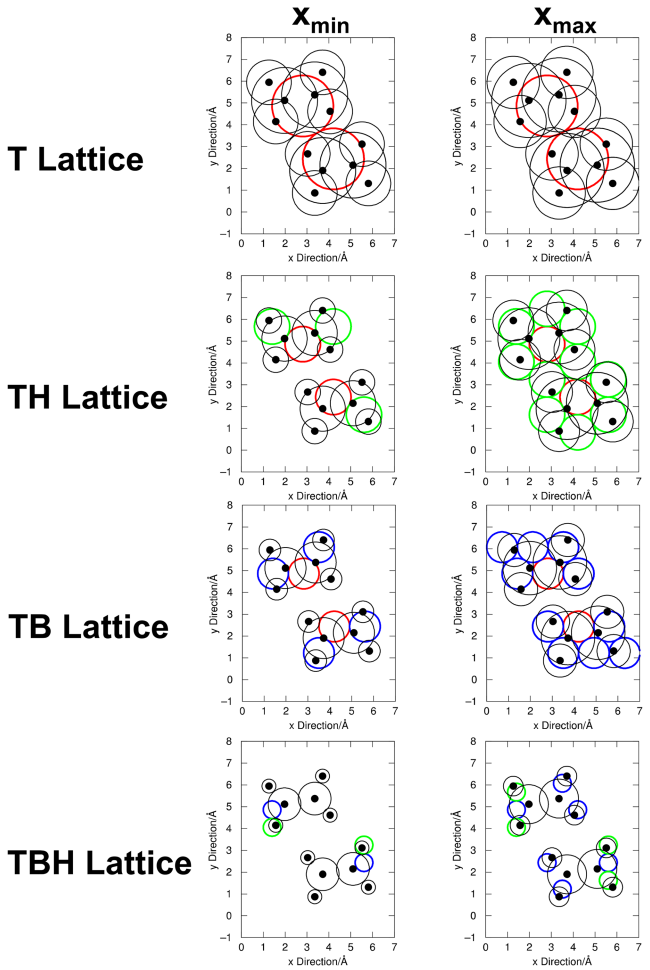

FIG. 7. Illustration of the pattern recognition mode (configuration 46) as a function of the atomic radii [x*R_{vdw} + (1 − x)*R_{cov}] and the four different lattices. Atomic nuclei and radii are indicated as black dots and circles, respectively. Recognized top, bridge, and hollow sites are given by red, blue, and green circles, respectively. X_{min} and x_{max} are 0.8 and 1.0 for the T lattice, 0.3 and 0.7 for the TH lattice, 0.2 and 0.5 for the TB lattice, and 0 and 0.15 for the TBH lattice, respectively.

molecules start to actually overlap, since this is unacceptable for the graph theoretical representation unless they are merged into a single entity.

Overall, we find that rather small atomic radii are most promising, although they can lead to adsorptions on the graphical lattice, which are "too close," i.e., leading to excessive lateral interactions. These situations can, however, be taken care of separately by introducing the corresponding prohibitive penalty terms.

C. Cluster expansion of the interaction energy

We have parameterized a variety of cluster expansions (CEs) as a function of the lattice type and of the atomic radii and nearest neighbour interactions. For each value of the atomic radii and each 46 configurations, we first detect the patterns for the molecule/surface interaction as explained above in Sec. III and with results discussed in Sec. V B. 1-Body patterns are associated with a single molecule (or a "super molecule" composed of molecules that would overlap otherwise) while patterns corresponding to the interaction of a given pair of

TABLE I. Summary of cluster expansions using all 46 configurations as a function of the graphical lattice, the size of the atoms, and the inclusion of interactions with neighbors (2 body patterns: NN: nearest neighbors). The number of unique patterns and their size is reported, and the minimum and maximum of single and two body pattern energies, together with the RMS and maximum error and the maximum Cook's distance. The value in square brackets indicates the number of linear dependencies, which is roughly equal to the number of patterns that should treat two molecules in one 1B pattern. A star (*) is used to indicate that linear dependencies between 1B and 2B patterns prohibit the definition of 2B interaction energies.

Lattice	%van der Waals	Neighbours included	No. of patterns	No. of sites min, max	Min, max E(1B)	Min, max E(2B)	Error (RMS, MAX) (eV)	Max Cook's D
T	80	None	2	1, 2	−1.25 −1.11	N.A.	0.07, 0.22	0.22
T	100	None	2	1, 2	−1.24 −1.12	N.A.	0.10, 0.33	0.26
TH	30	None	7	1, 4	−1.28 −1.05	N.A.	0.03, 0.12	>1000
TH	70	None	11 [1]	5, 8	−1.28 −0.99	N.A.	0.02, 0.05	>1000
TB	20	None	8	2, 4	−1.27 −1.06	N.A.	0.05, 0.19	>1000
TB	50	None	17 [1]	5, 7	−1.28 −0.97	N.A.	0.02, 0.10	>1000
T	80	NN	14	1, 4	−1.28 −1.12	−0.01 0.24	0.03, 0.13	>1000
TH	30	NN	27 [4]	1, 7	−1.28 −1.05	−0.03 0.16	0.01, 0.05	>1000
TB	20	NN	13 [3]	2, 8	−1.27 −0.91	N.A.*	0.04, 0.19	>1000

molecules are called 2-body patterns. The energy contributions for these 1-body or 2-body patterns are then calculated from the energy of the configurations by applying the inversion procedure of Sec. IV.

Table I provides a summary of the parameterizations, which describes all 46 configurations. Depending on the case, a rank deficient system of equations is obtained. This can happen when, for instance, two 1-body patterns contained in a 2-body pattern are heavily distorted and, therefore, do not occur in any other configuration. The same can, actually, already occur at the 1-body expansion level. Mathematically, our inversion procedure using SVD still allows solving the system of equations, even though it is effectively rank deficient. However, it is impossible to separately calculate the energy contributions of the individual, linearly dependent patterns, since they will always appear together in any configuration. From a chemical physics point of view, this means that the corresponding arrangement should be treated as a single entity of two molecules: its energy is well defined. Furthermore, in this case, the corresponding distorted single-molecule patterns are not supposed to occur alone but only in the neighbourhood of another (distorted) molecule. Hence, the procedure in which "primitive patterns" that are linearly dependent on each other are summed together in larger clusters (single body patterns of two molecules) makes both mathematical and chemical sense. The number of such patterns is indicated in square brackets in Table I.

We measure the quality of a given cluster expansion mainly by the root mean square error (RMSE) and the maximum error. In addition, Fig. 8 shows selected parity plots.

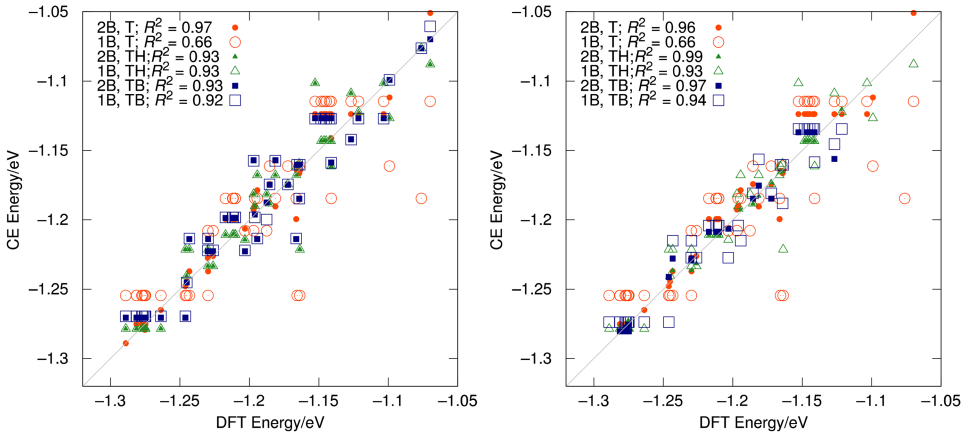

FIG. 8. Parity plots of the interaction energy per molecule given by the cluster expansion (CE) compared to DFT values for four typical models of Table I (left), which describe all 46 configurations and of Table II (right), which account only for a subset of configurations. Single body (1B) expansions and 2-body (2B) expansions that take nearest neighbors (NN) into account are distinguished. The percentage of vdW radii and the R^2 is given before and after the semi-colon, respectively. The bisector is indicated by a thin grey line.

As mentioned above, we also report Cook's distances, which indicate the presence (or not) of influential configurations. As it can be seen, there are many cluster expansions for which some configurations have Cook's distances above 1, which is indicative of highly influential points. Indeed, patterns that are defined by only one configuration have automatically an infinite Cook's distance, which simply reflects the fact that a given cluster is not present more than once in the training set. Since, however, all 46 configurations correspond to local minima on the DFT surface, these cluster energies are still physically meaningful. Nevertheless, we have also established a second family of cluster expansions (see Table II), where we have adopted the common practice[58] to exclude configurations that lead to Cook's distances greater than 1.0, as these points clearly are highly influential for the least squares fit. Of course, it would be preferable to increase the training set further, in order to probe similar configurations. But as the optimal construction of a training and validation set is out of the scope of the present work and is likely to depend on the target (e.g., low vs. high coverage), we do not follow this computationally much more expensive strategy. Exclusion of configurations necessarily means that the training set is less diverse and that some of the configurations can no longer be described. The parity plots (Fig. 8), however, demonstrate that we still span a similar range of adsorption energies per molecule by excluding between 1 and 20 configurations. In Fig. 8 we also see that the R^2 values are barely affected by this procedure suggesting that our training set is reasonably diverse.

Starting with the simplest approximation, the single body expansions on a T lattice where only di-σ and π modes are recognized, we find large maximum errors (>0.2 eV), given the range of interaction energies. Without much surprise, the simple picture with only two modes and no lateral interactions is insufficient. Reassuringly, however, the di-σ and π modes have an energy of −1.25 and −1.11 eV, respectively, which compares very well to −1.28 and −1.14 eV for the low coverage (1/16 ML) configurations. This observation actually holds for all CEs presented, i.e., the minimal (π) and maximal (di-σ) energy coefficients obtained for the single body patterns (6th column of Table I) are consistent with the corresponding low-coverage adsorption energies. Increasing the atomic radii to 100% R_{vdw} can lead to "miss-assignments" on the T lattice (e.g., π becomes di-σ), which explains the larger RMSE. Adding the 2-fold bridge or 3-fold hollow sites in the pattern recognition procedure increases the number of unique 1-body patterns and allows to improve the quality of the cluster expansion: these topological features allow for a detailed description of single modes even at small intermolecular distances, accounting for the deformation energy induced by the lateral interactions (without explicitly considering these interactions in the expansion). Indeed, the TH and TB lattices are characterized by 7 and 8 single body patterns at x_{min}. This more detailed description slightly reduces the RMSE from 0.07 to 0.03–0.05 eV and the maximum error lies now around 0.1 eV. However, without much surprise, some of the heavily distorted modes only occur in one (or few) configuration(s). Hence, Cook's distance reaches very high levels. Folding the lateral interactions into 1-body energy terms for slightly deformed adsorption modes is, apparently, a successful way to improve the fit with 1-body patterns exclusively. It is, however, unclear whether this would lead to a good energy prediction of configurations outside the training set; this question will be addressed in future work.

Including nearest neighbour interactions (2-body patterns) reduces the RMS and maximum errors from 0.07 and 0.2 eV to 0.03 and 0.1 eV, respectively, for the coarse T lattice. These two body patterns are repulsive by up to 0.2 eV and visibly increase the accuracy of the cluster expansion significantly, which is also seen in the parity plot of Fig. 8 (left hand side). On the other hand, Cook's distances identify highly influential configurations. For the more detailed TH and TB lattice, on the other hand, the inclusion of nearest neighbour interactions is of minor importance, although small improvements can still be seen in the overall performance. This slightly increased accuracy comes, however, at the cost of linear dependencies, which means that combinations of two single body patterns into one single body pattern describing two ethylene

TABLE II. Summary of cluster expansions using a subset of configurations as a function of the graphical lattice, the size of the atoms, and the inclusion of interactions with neighbors (2 body patterns: NN: nearest neighbors). The number of unique patterns and their size is reported, and the minimum and maximum of single and two body pattern energies as determined by the least squares fit together with the RMS and maximum error. Configurations leading to Cook's distances larger than 1 have been excluded and the resulting number of configurations used in the fit is given in the last column.

Lattice	%van der Waals	Neighbours included	No. of patterns	No. of sites min, max	Min, max E(1B)	Min, max E(2B)	Error (RMS, MAX) (eV)	No. of configurations
TH	30	None	6	1, 3	−1.28 −1.06	N.A.	0.03, 0.12	45
TB	20	None	4	2, 3	−1.27 −1.06	N.A.	0.03, 0.07	38
T	80	NN	10	1, 4	−1.28 −1.12	−0.01 0.24	0.03, 0.13	41
TH	30	NN	9 [1]	1, 7	−1.28 −1.05	−0.03 0.16	0.01, 0.02	29
TB	20	NN	5	2, 7	−1.27 −1.14	0.01 0.08	0.02, 0.05	27

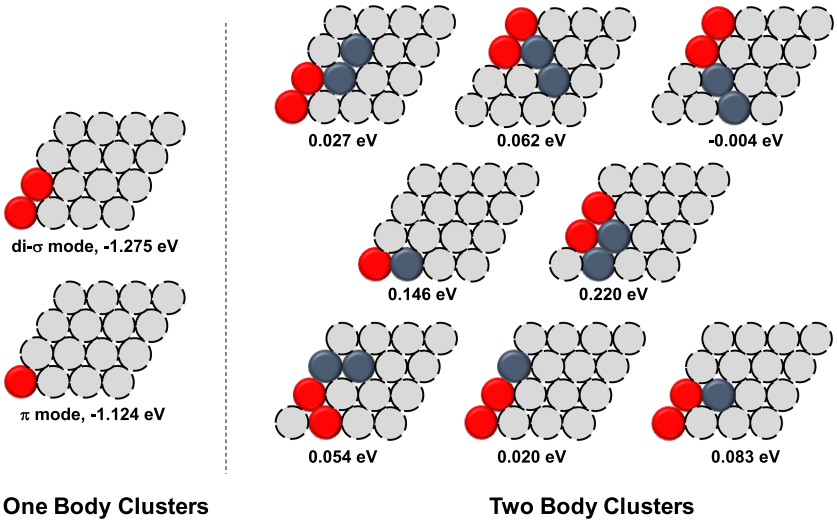

FIG. 9. The cluster expansion of ethylene on Pd(111) as recognized with a criterion of 80% van der Waals radii.

molecules would be not only more efficient but also necessary from a mathematical point of view. Excluding highly influential configurations changes the overall picture only moderately (see Table II and Fig. 8, right): for the 1-body expansions up to 8 out of the 46 configurations need to be excluded, which does not significantly affect the maximum and minimum cluster energies. The situation for the 2-body expansions is somewhat different, since, except for the T lattice, more configurations have to be excluded in order to have a well-balanced training set. Nevertheless, even in these cases, we obtain CEs that span a reasonable range of configurations and adsorption energies.

In summary, we find that a 2-body expansion (with about 10 clusters as displayed in Fig. 9) on the T lattice is probably the most convenient parametrization for ethylene adsorption on Pd(111). However, in hydrogenation reactions, for instance, hollow sites are necessarily included in order to describe the adsorbed hydrogen atoms. Hence, expansions on the TH lattice also need to be considered. In this case, we find that in the case of ethylene on Pd(111), lateral interactions tend to be small enough to be incorporated in slightly deformed single body patterns or can be accounted for explicitly. The ultimate comparison of the two possibilities in order to decide which approach is more effective in the determination of the ground-state geometry would require the application of the CEs to larger unit cells and validation of the obtained results against DFT.

VI. CONCLUSIONS

This paper has proposed a set of machine-learning tools to produce unbiased cluster expansions of atomic and molecular adsorbed layers on a graphical lattice. A measure of site recognition was proposed by associating circles of maximum, non-overlapping radii to lattice points, and projecting the molecular surface based on atomic radii onto the lattice plane. A molecule is assigned to a graphical site from an overlap of at least 50% between the atoms of a molecule and a site. Mapping the recognized patterns into the configurational energy of the system finally retrieves the expanded Hamiltonian to be calculated by pseudo-inversion.

The proposed scheme has been applied to the adsorption of ethylene on Pd(111). The cluster expansions reproduced the known adsorption features in an accurate way, with both di-σ and π modes being clearly identified. Lattices, which are more detailed than just atop sites, were found capable of resolving important parts of interactions at short contacts, since the repulsion induces measurable distortions at the DFT level. 2-body patterns, on the other hand, are also an efficient approach to increase the accuracy of the cluster expansions. This contribution has not addressed the optimal construction of a training set and validation procedure, although we consider that these tasks are highly important and human-time intensive. However, for a given set of DFT configurations, we have proposed a clear protocol to establish cluster expansions based on a set of configurations and their energy without any user input and without ambiguity. The later point is the more relevant since adsorption modes of somewhat flexible molecules such as ethylene are not trivially deduced from DFT computations at short contacts, which are typical for high-coverage situations. We suggest that around 30 configurations are sufficient to have a statistically relevant sampling of the different adsorption modes and nearest neighbor interactions. Having this tool in hand allows one to construct cluster expansions for arbitrary systems in a black-box manner based on the corresponding DFT configurations. Hence, a lattice based kinetic Monte Carlo model for selective hydrogenation on Pd(111) of acetylene including one intermediate could, by now, be constructed with about 300 DFT computations and minimal time for the determination of the cluster expansions. It also opens up the possibility to construct cluster expansions for more complex surfaces such as alloys, where the identification of the different adsorption modes is even more challenging.

SUPPLEMENTARY MATERIAL

See supplementary material for all *ab initio* computed energies and geometries.

[1] J. M. Thomas and J. W. Thomas, *Principles and Practice of Heterogeneous Catalysis*, 1st ed. (Wiley-VCH, Weinheim, 1996).
[2] *Computational Methods in Catalysis and Materials Science*, edited by R. A. van Santen and P. Sautet (Wiley-VCH, Weinheim, 2009).

[3] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nat. Chem. **1**, 37 (2009).
[4] W. Guo, M. Stamatakis, and D. G. Vlachos, ACS Catal. **3**, 2248 (2013).
[5] K. J. Andersson, F. Calle-Vallejo, J. Rossmeisl, and I. Chorkendorff, J. Am. Chem. Soc. **131**, 2404 (2009).
[6] F. Calle-Vallejo, J. Jakub Tymoczko, V. Colic, Q. Huy Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann, and A. S. Bandarenka, Science **350**, 185 (2015).
[7] O. Sinanoğlu and K. S. Pitzer, J. Chem. Phys. **32**, 1279 (1960).
[8] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).
[9] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).
[10] M. A. van Hove and S. Y. Tong, *Surface Crystallography by LEED* (Springer-Verlag, Berlin, Heidelberg, 1979).
[11] A. P. van Bavel, D. Curulla Ferré, and J. W. Niemantsverdriet, Chem. Phys. Lett. **407**, 227 (2005).
[12] K. Reuter, in *Modeling and Simulation of Heterogeneous Catalysis Reactions: From Molecular Process to the Technical System*, edited by O. Deutschmann (Wiley-VCH, Weinheim, 2011), pp. 370–392.
[13] A. P. J. Jansen, *An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions* (Springer-Verlag, Berlin, 2012).
[14] M. Stamatakis and D. G. Vlachos, ACS Catal. **2**, 2648 (2012).
[15] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, London, 1982).
[16] G. Mussardo, *Statistical Field Theory. An Introduction to Exactly Solved Models in Statistical Physics* (Oxford University Press, New York, 2010).
[17] D. de Fontaine, in *Solid State Physics*, edited by H. Ehrenreich and D. Turnbull (Academic Press, New York, 1994), Vol. 47, pp. 33–176.
[18] G. Inden, *Materials Science and Technology* (Wiley-VCH, Weinheim, 2013), pp. 519–581.
[19] T. D. Lee and C. N. Yang, Phys. Rev. **87**, 410 (1952).
[20] S. M. Bhattacharjee and A. Khare, Curr. Sci. **69**, 816 (1995), http://www.jstor.org/stable/24097007.
[21] A. van de Walle and G. Ceder, J. Phase Equilib. **23**, 348 (2002).
[22] A. van de Walle, M. Asta, and G. Ceder, CALPHAD **26**, 539 (2002).
[23] J. W. D. Connolly and A. R. Williams, Phys. Rev. B **27**, 5169 (1983).
[24] M. Stamatakis, Y. Chen, and D. G. Vlachos, J. Phys. Chem. C **115**, 24750 (2011).
[25] M. Stamatakis, M. A. Christiansen, D. G. Vlachos, and G. Mpourmpakis, Nano Lett. **12**, 3621 (2012).
[26] M. D. Marcinkowski, A. D. Jewell, M. Stamatakis, M. B. Boucher, E. A. Lewis, C. J. Murphy, G. Kyriakou, E. Charles, and H. Sykes, Nat. Mater. **12**, 523 (2013).
[27] S. Piccinin and M. Stamatakis, ACS Catal. **4**, 2143 (2014).
[28] M. Stamatakis and S. Piccinin, ACS Catal. **6**, 2105 (2016).
[29] K. Reuter and M. Scheffler, Phys. Rev. B **73**, 45433 (2006).
[30] D. J. Liu and J. W. Evans, Prog. Surf. Sci. **88**, 393 (2013).
[31] A. Bajpai, K. Frey, and W. F. Schneider, J. Phys. Chem. C **121**, 7344–7354 (2017).
[32] S. D. Miller and J. R. Kitchin, Mol. Simul. **35**, 920 (2009).
[33] K. Frey, D. J. Schmidt, C. Wolverton, and W. F. Schneider, Catal. Sci. Technol. **4**, 4356 (2014).
[34] D. J. Schmidt, W. Chen, C. Wolverton, and W. F. Schneider, J. Chem. Theory Comput. **8**, 264 (2012).
[35] C. Wu, D. J. Schmidt, C. Wolverton, and W. F. Schneider, J. Catal. **286**, 88 (2012).
[36] M. Borg, C. Stampfl, A. Mikkelsen, J. Gustafson, E. Lundgren, M. Scheffler, and J. N. Andersen, ChemPhysChem **6**, 1923 (2005).
[37] C. Stampfl, Phase Transitions **80**, 311 (2007).
[38] M. Stamatakis and D. G. Vlachos, J. Chem. Phys. **134**, 214115 (2011).
[39] J. Nielsen, M. d'Avezac, J. Hetherington, and M. Stamatakis, J. Chem. Phys. **139**, 224706 (2013).
[40] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
[41] N. L. Biggs, *Interaction Models* (Cambridge University Press, New York, 1976).
[42] A. J. Stone, *The Theory of Intermolecular Forces*, 2nd ed. (Oxford University Press, Oxford, United Kingdom, 2013).
[43] T. L. Hill, *An Introduction to Statistical Thermodynamics* (Dover Publications, New York, 1960).
[44] V. P. Zhdanov, J. Chem. Phys. **110**, 8748 (1999).
[45] J. F. Gonthier, S. N. Steinmann, M. D. Wodrich, and C. Corminboeuf, Chem. Soc. Rev. **41**, 4671 (2012).
[46] A. Bondi, J. Phys. Chem. **68**, 441 (1964).
[47] P. Pyykkö and M. Atsumi, Chem. - Eur. J. **15**, 186 (2009).
[48] R. Penrose, Math. Proc. Cambridge Philos. Soc. **51**, 406 (1955).
[49] J. C. A. Barata and M. S. Hussein, Braz. J. Phys. **42**, 146 (2012).
[50] G. Ceder and G. D. Garbulsky, Phys. Rev. B **51**, 57 (1995).
[51] N. A. Zarkevich and D. D. Johnson, Phys. Rev. Lett. **92**, 255702 (2004).
[52] A. Seko and I. Tanaka, Phys. Rev. B **83**, 224111 (2011).
[53] A. Seko, K. Shitara, and I. Tanaka, Phys. Rev. B **90**, 174104 (2014).
[54] G. Golub and W. Kahan, J. Soc. Ind. Appl. Math., Ser. B, Numer. Anal. **2**, 205 (1965).
[55] R. D. Cook, Technometrics **19**, 15 (1977).
[56] R. D. Cook, J. Am. Stat. Assoc. **74**(365), 169 (1979).
[57] K. A. Bollen and R. W. Jackman, in *Modern Methods of Data Analysis* (Sage, Newbury Park, 1990), pp. 257–291.
[58] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*, 4th ed. (Wiley Interscience, Hoboken, 2006).
[59] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed. (SIAM, Philadelphia, PA, 1999).
[60] S. J. Eun, Y. J. Mun, and D. B. Min, Compr. Rev. Food Sci. Food Saf. **1**, 22 (2005).
[61] D. Stacchiola, S. Azad, L. Burkholder, and W. T. Tysoe, J. Phys. Chem. B **105**, 11233 (2001).
[62] M. Neurock and R. A. Van Santen, J. Phys. Chem. B **104**, 11127 (2000).
[63] E. W. Hansen and M. Neurock, J. Catal. **196**, 241 (2000).
[64] Y.-T. Wong and R. Hoffmann, J. Phys. Chem. **95**, 859 (1991).
[65] P. Sautet and J. F. Paul, Catal. Lett. **9**, 245 (1991).
[66] D. M. P. Mingos, J. Organomet. Chem. **635**, 1 (2001).
[67] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[68] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[69] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[70] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[71] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[72] S. N. Steinmann and C. Corminboeuf, J. Chem. Theory Comput. **7**, 3567 (2011).
[73] S. Gautier, S. Steinmann, C. Michel, P. Fleurat-Lessard, and P. Sautet, Phys. Chem. Chem. Phys. **17**, 28921 (2015).
[74] J. D. Pack and H. J. Monkhorst, Phys. Rev. B **13**, 5188 (1977).
[75] M. Methfessel and A. T. Paxton, Phys. Rev. B **40**, 3616 (1989).
[76] Scienomics, MAPS platform, version 4.0, 2016, France, http://scienomics.com.