Accelerating atomic structure search with cluster regularization

K. H. Sørensen,[1] M. S. Jørgensen,[1] A. Bruix,[1] and B. Hammer[1, a)]
Department of Physics and Astronomy, and Interdisciplinary Nanoscience Center (iNANO), Aarhus University, DK-8000 Aarhus C, Denmark.

(Dated: 30 April 2018)

We present a method for accelerating the global structure optimization of atomic compounds. The method is demonstrated to speed up the finding of the anatase $TiO_2(001)$-(1×4) surface reconstruction within a density functional tight-binding theory framework using an evolutionary algorithm (EA). As a key element of the method, we use unsupervised machine learning techniques to categorize atoms present in a diverse set of partially disordered surface structures into clusters of atoms having similar local atomic environments. Analysis of more than 1000 different structures shows that the total energy of the structures correlate with the summed distances of the atomic environments to their respective cluster centers in feature space, where the sum runs over all atoms in each structure. Our method is formulated as a gradient based minimization of this summed cluster distance for a given structure and alternates with a standard gradient based energy minimization. While the latter minimization ensures local relaxation within a given energy basin, the former enables escapes from meta-stable basins and hence increases the overall performance of the global optimization.

I. INTRODUCTION

In computational materials science, knowing the atomic structure of a given molecule, atomic cluster or solid compound is a prerequisite for further prediction of electronic and thermodynamic properties of such a substance. In the emerging fields of combinatorial chemistry and high-throughput computational screening of materials,[1] the use of local relaxation of probed structures is often sufficient since libraries of molecular building blocks and crystal structures can be used to direct the starting points for the searches.[2] However, many problems exist for which the structural motifs in the sought-after structures have no analogues in known structures, and where global optimization must be employed. A prominent example of this is that of surface reconstructions, which often exhibit structural motifs that are unique to a chemical composition, crystaline polymorph, and surface orientation. Monomeric Si adatoms, rest-atoms, and zig-zagging Si-dimers do for example evolve at the Si(111)-(7×7) and Si(001)-c(4×2) surfaces,[3,4] respectively, but are otherwise not present in bulk Si or bulk-truncated Si surfaces. Likewise, rutile and anatase TiO_2 single-crystal surfaces are known to exhibit reconstructions and ad-structures,[5–8] that have no equivalent in any TiO_2 bulk or surface systems.

Many strategies exist for performing global optimization in conjunction with model potentials or first-principles total energy frameworks. Among these, stochastic methods such as random search[9] and basin hopping (BH)[10] are widely used and more advanced methods based on evolutionary algorithms (EA) are becoming increasingly popular.[11–13] Common to these methods is the need for the perturbative update steps followed by local relaxation and evaluation of whether or not to retain the new structure. The nature of the perturbation steps ranges from a mere rattling of the atomic positions to highly advanced cross-over operations in which the atomic structures of several known systems are combined.[14] The exploration[15,16] of configuration space is ensured by the very stochastic nature of these updates, yet updates that are too random will too often lead to rejection of the locally relaxed structure and hence cause slow convergence. The atomic displacement amplitude in a rattling update is a good example of this. The amplitude must be kept small and sometimes only apply to a subset of the atoms, as the new structural candidates otherwise become too unstable.

The present work aims at formulating a means of performing a random update in a way that optimizes the chances of finding more stable structural candidates in subsequent local relaxation steps. The method proposed starts from assigning every atom in a given structure to a cluster of similar atoms present in a reference set of structures. The similarity is measured as the distance in a feature space, chosen in this work sufficiently simple that it can be illustrated. With every atom assigned to a cluster, we evaluate for a given structure the sum of the distances of the atoms from their respective cluster centers in feature space. This structure-specific scalar measure of distance to cluster centers is demonstrated to correlate with the structure stability – the smaller the distance measure, the more likely that the structure is more stable. As a consequence, minimizing the cluster distance measure is likely to bring a structure into a more stable basin, i.e. into a new region of configuration space, that has more stable local minima. The minimization can be done by moving opposite to an analytic gradient and has potential to take a structure out of local energy minima, since the cluster distances are measured in a space that is complementary to that of the energy landscape. We refer to our method as the *cluster regularization method* as it penalizes large distances of cluster members to their cluster centers.

The paper is outlined as follows: In the Method sec-

a)Electronic mail: hammer@phys.au.dk

tion we present the computational setup, outline the reference method used, and describe the required machine learning components of our method, including the choice of feature vector and the clustering technique. In the Feature Space Analysis section we analyse the data that forms the basis for formulating the cluster regularization method in the subsequent Cluster Regularization Method section. In the Global Structure Search Results section, the method is used to conduct a full scale global optimization search for the anatase $TiO_2(001)$-(1×4) surface reconstruction.[6] This section demonstrates the usefulness of the method and further investigates its efficacy as the number of unknown atomic position is increased in the global optimization. The paper ends with a summary section.

II. METHOD

A. Density functional tight-binding theory

Density functional tight-binding (DFTB) theory calculations were performed for TiO_2 structures using parameters from ref.[17] Self-consistent charge DFTB calculations for TiO_2 systems have previously been conducted successfully in studies where either the system size or the number of calculations is too large for a full first principles density functional theory (DFT) study.[18,19] In this study, we choose not to include the self-consistent charge correction in order to further increase the computational speed. The loss of accuracy in using DFTB over DFT is not of significant importance because our goal is not to find exact geometries, but to enhance global structure optimization independently of the level of theory used to calculate the energy of the system. In the present work, we employ 2D periodic super cells accomodating the anatase $TiO_2(001)$ surface with (1×4) periodicity with lattice parameters $a = 3.94$ Å and $c = 9.47$ Å. A mesh of 2×1 **k**-points is used to sample the Brillouin zone. Slabs of three different thicknesses were considered, containing two, three, or four layers of atoms. The three-layer system is illustrated in Fig. 1, showing the template of four static TiO_2 units (a), examples of two disordered (b,c) and one ordered (d) structure with nine extra TiO_2 units added, the latter being the global minimum energy structure with three well ordered atomic layers and one extra row of TiO_2 protruding out of the surface.[6]

B. Evolutionary algorithm

The disordered structures in Figs 1b-c are snapshots from search runs performed with the evolutionary algorithm (EA) available in the Atomic Simulation Environment (ASE).[20] The EA iteratively improves the positions of the atoms atop the template structure (Fig. 1a) using cross-over operations combining two parent structures

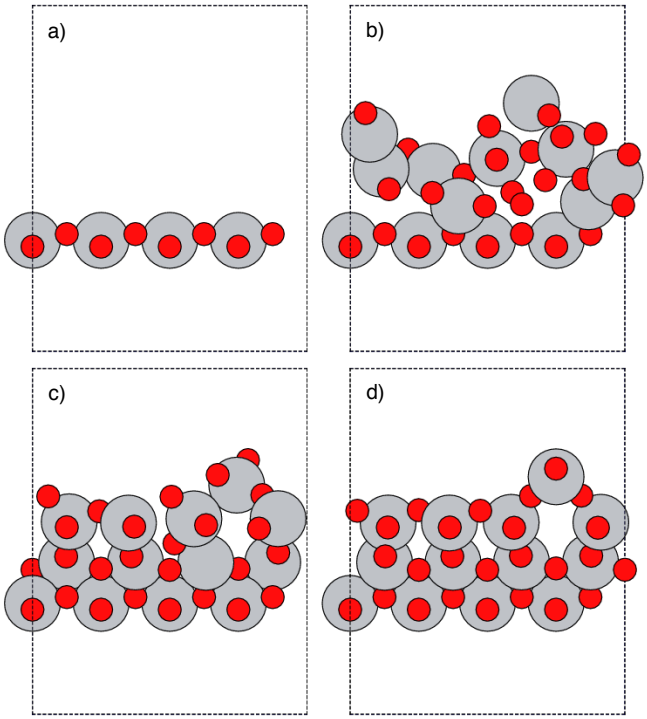

FIG. 1. Side views of computational cells with Ti_nO_{2n} structures. Small red spheres: Oxygen, large grey spheres: Ti. a) The static template of four TiO_2 units. b-c) Two examples of intermediate structures from an evolutionary search. Nine TiO_2 units have been placed on the template. d) The sought-after "3-layer" global minimum structure for the anatase $TiO_2(001)$-(1×4) surface.

or mutation operations applied to single parent structures. The resulting offspring structures undergo local relaxation within DFTB following the atomic forces (i.e., the negative of the energy gradient). A population of $N_{\mathrm{pop}} = 20$ parent structures is maintained, and new candidates are tested one at a time with the possibility of updating the population after each test. The relaxed offspring structures are included in the population according to their fitness evaluated as the negative of the DFTB energy. More details on the EA are given in Refs.[21,22]

The cluster regularization method proposed in this work is implemented as a cross-over operation and its usefulness is gauged by comparing an EA search where this operation is used either exclusively or occasionally to a benchmark EA search, where it is not used at all. We also test the method in a basin hopping (BH) framework,[10] in which it serves as a perturbing update step that alternates with an energy relaxation step.

C. Machine learning techniques

Machine learning techniques are becoming popular tools for various tasks in chemical physics.[23–32] Our method uses two machine learning techniques: the rep-

resentation of the data with a feature vector, and the classification of the data utilizing a clustering method.

1. Feature vector

Feature vector representations of the atomic structure of molecules, clusters, and solids serve the general purpose of introducing invariance with respect to symmetries obeyed by the total energy operator, i.e. the Hamiltonian.[24] The symmetries are those of translation or rotation of the entire system or the permutation of the order of atoms with the same chemical identity. Without a symmetry invariant representation, e.g. using Cartesian coordinates, a certain structure will have a finite distance to for instance a translated copy of itself, even though the two systems will have identical physical properties. However, when measured in a proper feature space, two such structures may have a zero distance.

A large and increasing number of feature vector formulations appear in the chemical physics literature. Global feature vectors of entire systems include the bag-of-bonds[33] and fingerprint[34,35] feature vectors, that are composed of either a sorted list or a histogram of all interatomic distances in the compounds. More sophisticated methods such as the Coulomb matrix[36,37] have been proposed.

Atom-specific feature vectors representing the local chemical environments of the atoms in a compound structure are particularly abundant in the literature. Such feature vectors are usually formulated in internal coordinates and therefore have the translational and rotational invariance conveniently build in by construction. The Behler-Parrinello (BP) symmetry function formulation[38] stands out as a seminal contribution, yet one with rather many parameters to be chosen by the user. With BP symmetry functions, each atom is represented by a selection of 2-body distance terms and 3-body angular terms. All terms are attenuated with a cutoff function that ensures a smooth decay to zero at a set cutoff distance. Recently, more advanced atom-specific feature vectors have been developed for which higher-order interactions (e.g. 4-atom torsion angles) are included. Such feature vectors have proven superior in reproducing energies, HOMO-LUMO gaps, polarizations, and a number of other properties for a diverse set of molecules.[39,40]

In the present work we adopt a simple atom-specific feature vector with only three components:

$$\mathbf{f}_i = \left[Z_i, \rho_i^{Z^{\mathrm{O}}}, \rho_i^{Z^{\mathrm{Ti}}} \right] \tag{1}$$

where Z_i is the atomic number of atom i, while $\rho_i^{Z^{\mathrm{O}}}$ and $\rho_i^{Z^{\mathrm{Ti}}}$ are measures for atom i of the density of neighboring oxygen and titanium atoms, respectively. These densities are inspired by the radial symmetry functions of Behler *et al.*[38] For a discussion of these functions see the work of Botu and Ramprasad.[27] In our work we use:

$$\rho_i^Z = \sum_{j \neq i, Z_j = Z} e^{-r_{ij}/\lambda} g_c(r_{ij}) \tag{2}$$

where Z is either Z^{O} or Z^{Ti}, j runs over all atomic indices, Z_j is the atomic number of the j'th atom, r_{ij} is the distance $|\mathbf{r}_{ij}| = |\mathbf{r}_j - \mathbf{r}_i|$ from the i'th to the j'th atom, and $\lambda = 1$ Å is a chosen length scale. g_c is a cut-off function given by:

$$g_c(r) = \begin{cases} \frac{1}{2}\cos(\pi \frac{r}{r_c}) + \frac{1}{2}, & r < r_c \\ 0, & r \geq r_c \end{cases} \tag{3}$$

which ensures that the densities vanish algebraically at the cut-off radius, r_c. For r_c we use 11.9 Å which was determined in a parameter search detailed below. When used with periodic boundary conditions as in the present work, it is important that the sum in Eq. (2) runs over all replicas of given atoms in neighboring unit cells that are within the cut-off.

2. Clustering

Molecular and chemical physicists are often faced with vast configuration spaces when studying the structures of chemical compounds. Clustering techniques have proven valuable tools when trying to find recurring structural moieties,[41] especially in combination with molecular dynamics where thousands of structure trajectories are generated.[42,43] Clustering has also previously been used to aid global structure optimization.[44]

The clustering is done in feature space. Given N_S different structures, each having N atoms, the $N_S N$ different feature vectors are clustered in N_c clusters. Unless otherwise stated, we use a fixed number of clusters, $N_c = 5$, since this emerged from a parameter optimization presented below. The specific clustering method used in this study is k-means with two modifications. The first is a modification to avoid generating empty clusters[45] and the second modification is the use of k-means++ cluster initialization.[46]

Whenever clustering of atom-specific feature vectors has been performed, a set of resulting cluster centers, $\{\mathbf{c}_k\}$, will be known. Representing by $k(i)$ the index of the cluster center that a given atom i belongs to, we can now evaluate the sum of all cluster distances for a given structure according to:

$$D_S = \sum_{i \in S} \left| \mathbf{f}_i - \mathbf{c}_{k(i)} \right| \tag{4}$$

where S is the structure (a list of atoms). Note that D_S is a scalar number, which in principle exists for any conceivable set of N atomic positions. As such, a continuous "cluster-distance"-landscape exists in $\mathbb{R}^{3N}$, and

the atomic positions may be changed towards an overall smaller cluster distance by following the negative of the gradient of the cluster distance with respect to the atomic positions, $\mathbf{R}$:

$$-\nabla_{\mathbf{R}} D_S = -\frac{1}{2} \sum_{Z \in \left\{ \substack{Z^{\mathrm{O}} \\ Z^{\mathrm{Ti}}} \right\}} \sum_{i \in S} \left(\nabla_{\mathbf{R}} \rho_i^Z \right) \frac{\partial}{\partial \rho_i^Z} \left| \mathbf{f}_i - \mathbf{c}_{k(i)} \right| \tag{5}$$

where the cluster centers are kept fixed. This is analogous to the atomic forces directing the minimization of the energy in the energy landscape. The initial factor $\frac{1}{2}$ originates from the double sums over all atoms appearing when Eq. (2) is inserted in Eq. (5).

III. FEATURE SPACE ANALYSIS

Owing to its low dimensionality, the atom-specific feature vector can be illustrated in two-dimensional scatter plots. This is done in Figs. 2 and 3 where features pertaining to oxygen and titanium atoms are shown in red and grey color, respectively. The two axes in the plots are the $\rho^{Z^{\mathrm{O}}}$ and $\rho^{Z^{\mathrm{Ti}}}$ densities. In Fig. 2, the atomic structure considered is that of the three-layer global minimum energy structure introduced in Fig. 1d. The figure shows that strong correlations exist for the features. Features for oxygen atoms lie on one line, while features for titanium atoms lie on another. Within each line the 26 data points for oxygen atoms and 13 data points for titanium atoms further appear in a number of clusters. The lines reflect that in this global minimum energy structure the atoms have a nearly constant ratio of neighboring atoms of the two possible types.

Figure 3 presents the distribution of atomic features for the 26+13 atoms in about 1000 different structures of the three-layer system. The structures are taken from several EA searches and have been filtered so that they can be considered distinct structures. The wide distribution of the feature vectors for general structures as opposed to the highly ordered features for the global minimum energy structure indicates that despite its simplicity the feature vector holds great potential to capture variations in local structure present in the disordered structures. The feature vector is clearly not unique as for example an atom with a single close neighbor may have the same feature vector as an atom with two more distant neighbors. However, in the current context, this is not of importance, since the rich variety in feature vectors over a real data set as seen in Fig. 3 is the desired property.

A strong correlation between the total cluster distance, D_S, and the total DFTB energy, E_S^{DFTB}, is revealed when clustering of the atomic feature vectors presented in Fig. 3 is performed. The correlation illustrated in Fig. 4 may in part be explained by the general observation that nature often favors high symmetry,[9] i.e. structures with low energy often have recurring atomic motifs. A single crystal is an obvious example of this, since all the atoms

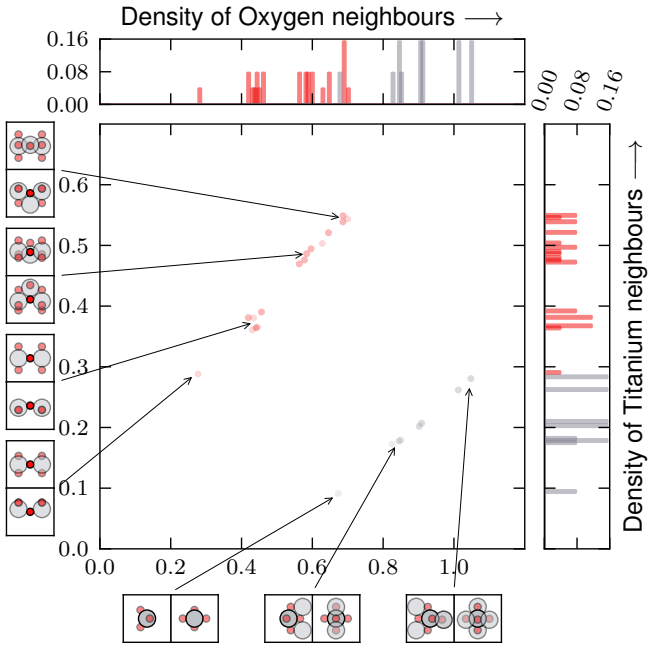

FIG. 2. Visualization of global minimum for three-layer TiO_2 structures ($Ti_{13}O_{26}$) in feature space. Feature vectors are shown in the $\left(\rho_i^{Z^{\mathrm{O}}}, \rho_i^{Z^{\mathrm{Ti}}}\right)$-plane and the Z_i-dimension which can only assume two values, either Z^{O} or Z^{Ti} is color coded, red meaning Z^{O} (oxygen), grey meaning Z^{Ti} (titanium). The insets are representative local structures at given positions in feature space.

assume the same local environments. In fact, for bulk anatase TiO_2, the scatter plot similar to Figs. 2 and 3 only contains just one point for each of the two chemical elements, oxygen and titanium. Hence, with two cluster centers, the total cluster distance for bulk anatase TiO_2 would be zero. In our case, the slab geometry used is causing the presence of two surfaces that act to give several possible motifs, as evidenced for the global minimum energy structure by Fig. 2, where now the total cluster distance becomes finite, reflecting the surface energy cost. For any disordered slab structure, the feature vectors scatter even more and the cluster distance increases further at the same time as the total energy of the structure rises above that of the global minimum energy structure (Fig. 4).

IV. CLUSTER REGULARIZATION METHOD

The observed correlation between energy and cluster distance forms the basis for our new method for perturbing an atomic structure in the search for the global minimum energy structure. The method simply performs a local minimization of the cluster distance for fixed cluster centers. I.e. structures subjected to the method undergo regularization with respect to their cluster distance measure. The regularization takes place as a gradient

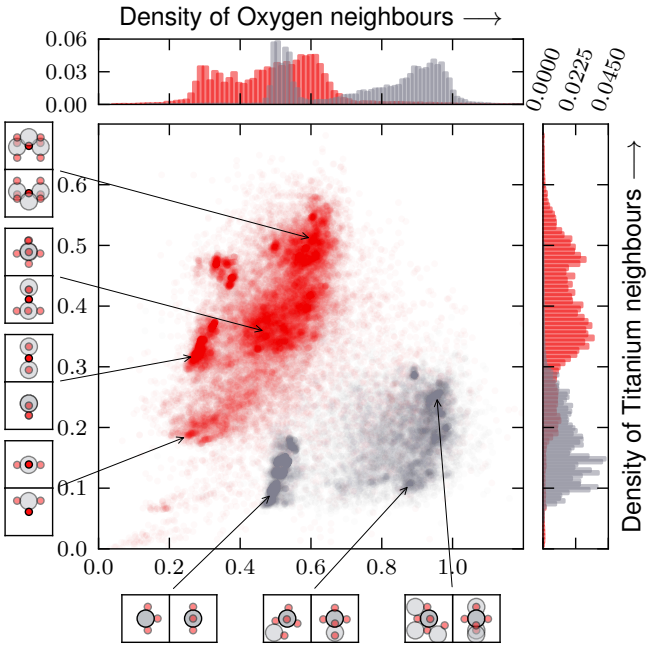

FIG. 3. Visualization of atoms from 1019 $Ti_{13}O_{26}$ structures in feature space. Each structure contributes $N = 39$ separate spots that are colored as in Fig. 2. Every vector is marked with a transparent symbol. In regions with many vectors, the symbols add up to a darker contrast.

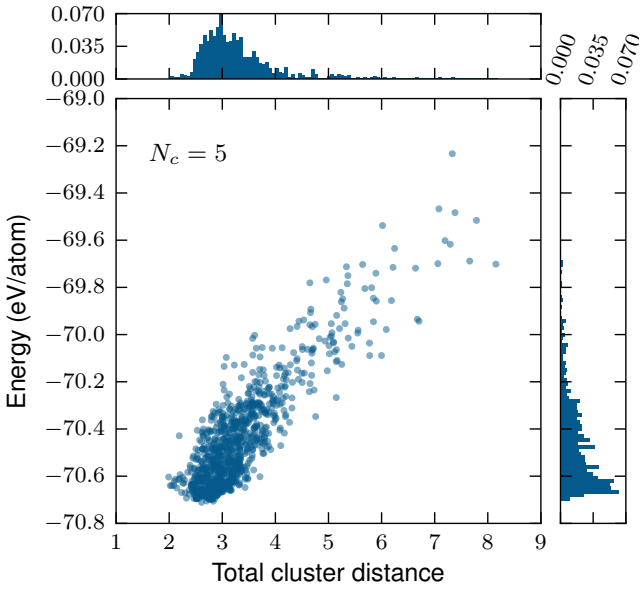

FIG. 4. Correlation of energy (in eV/atom) with cluster distance according to Eq. (4). Each dot represents one of the $N_S = 1019$ structures that form the basis for Fig. 3. The $N_c = 5$ cluster centers used in Eq. (4) were: $(22, 0.89, 0.19)$, $(22, 0.53, 0.14)$, $(8, 0.62, 0.44)$, $(8, 0.46, 0.38)$, and $(8, 0.31, 0.27)$.

based optimization using standard techniques. There is no guarantee or even expectation that the cluster regularization will directly perturb a given structure towards a lower energy structure, but the cluster distance-energy correlation in Fig. 4 is suggestive that when moved according to (minus) the cluster distance gradient, Eq. (5), a given structure may leave a meta-stable basin region of the energy landscape and move into a more favorable basin in the energy landscape.

The principle of the method is illustrated schematically in Fig. 5a, where transformations between real space and feature space enable the structure to escape local minima. Note that a local minimum in real space is only escaped if the local minimum in feature space takes the structure to another energy basin in real space. If this is not the case, the optimization might enter a dynamical equilibrium state between real space and feature space, and thus the structure may become trapped in one or more energy basins (Figs. 5b and c). The stochastic nature of the k-means method may, however, help the optimization procedure escape from trapped states. In the following, we demonstrate the usefulness of the method in two different setups, either as the perturbation step in a basin hopping (BH) framework or as a cross-over operation in an evolutionary search.

3. Perturbation step in BH

Figure 6 illustrates the outcome of the simplest conceivable way of using the cluster regularization, namely as the perturbing step in a basin hopping setup. In such a setup, the cluster centers entering Eq. (4) are obtained by clustering the atomic features present in the latest structure of the search only (i.e. here $N_S = 1$ is used). Figures 6a-c illustrate how alternating cluster regularization steps (red color) and local relaxation, i.e. energy minimization, steps (blue color) lead to progressively more favorable structures. Figures 6a and b show the evolution of the energy and the cluster distance, respectively, during the updates. The starting point is a fully relaxed random structure which through the cluster regularization perturbations is taken out of many different local energy minima, thereby gaining about 0.5 eV/atom over the course of 25 updates. Figure 6c combines the two measures, cluster distance and energy, giving an overview of how the structure develops. We observe (not shown) that many BH runs end up in the dynamic equilibrium illustrated in Fig. 5b. However, occasionally, the BH runs succeed in identifying the global energy minimum structure, as illustrated with the penultimate and ultimate optimization steps of such BH runs in Figs. 6e-f.

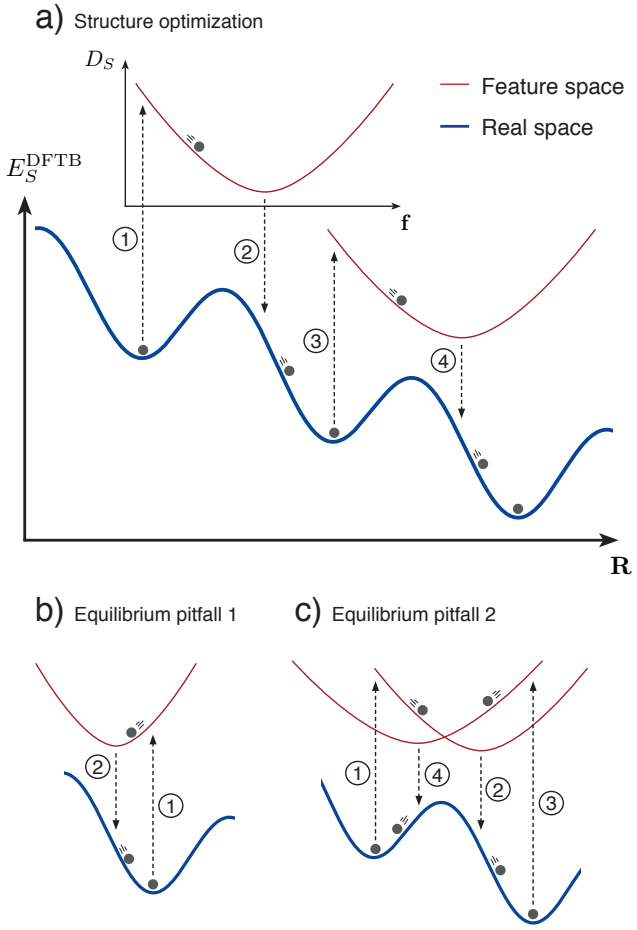

FIG. 5. a) Structure optimization via alternation between minimization in real space and feature space. b+c) Situations where the structure is stuck in one or more real space local minima due to a dynamic equilibrium between relaxations in real space and feature space.

4. Cross-over operation in EA

The cluster regularization method lends itself particularly well for use as a cross-over operation in connection with an evolutionary algorithm (EA). It serves in a manner similar to that in a basin hopping setup, but owing to the stochastic elements in the EA, the risk of being trapped in a dynamic equilibrium is largely reduced.

We introduced the cluster regularization as an EA cross-over operation in the following way:

1. Draw randomly N_{par} parents from the population.
2. Calculate atomic features for all parents.
3. Cluster them and find cluster centers.
4. Minimize the cluster distance for the lowest energy parent.
5. Return the resulting structure (the child).

Note that when used with $N_{par} = 1$, the method reduces to a mutation operation. When the method returns the child, the EA will make an energy relaxation and we have a cycle similar to that of the BH setup illustrated in Fig. 6. However, for the EA cycle, the random drawing of parents will ensure a more stochastic nature of the search for two reasons: i) The cluster centers found in step 3 will vary depending on the parents and the initialization of the k-means search, and ii) the lowest energy parent picked in step 4 will vary as well.

The optimum values of the free parameters in the new cross-over operation were determined with Bayesian hyper-parameter optimization (BHO) test runs on the three-layer system. We found $N_c = 5$ clusters, $N_{\mathrm{par}} = 3$ parents, and a cut-off radius of $r_c = 11.9$ Å as the optimum values. Next, a separate BHO search was done to identify the optimum frequency by which the new method should be used. The search gave an optimum ratio of 70% cluster regularization, 28% cut-and-splice, and 2% rattle mutation when cluster regularization was included, which compares to a BHO result of 59% cut-and-splice and 41% rattle mutation in the absence of the cluster regularization. The latter parameters were used for benchmark runs. In order to gauge the efficiency of the cluster regularization method on its own we further did EA runs using that method exclusively (i.e. having 100% cluster regularization and 0% cut-and-splice, and 0% rattle mutation).

V. GLOBAL STRUCTURE SEARCH RESULTS

The three EA settings – benchmark, all-cluster regularization, and combined – were tested for two-, three-, and four-layer TiO_2 systems. The results are presented in Fig. 7 which plots the accumulated success rates of 300, 300, and 600 restarts of the EA for the three system sizes, respectively. Note that the number of required attempts (measured on the x-axis) changes as the system size increases.

For all system sizes, it is seen that the all-cluster regularization method (orange curve) starts finding the global minimum energy structure much sooner than the benchmark method (yellow curve). For the two-layer system, the all-cluster regularization method levels off (i.e. stagnates) at around 60-70% success rate, which is smaller than what is obtained in the benchmark runs. The larger systems were not run long enough to see if a similar effect would occur for these systems. We speculate that the observed stagnation of the all-cluster regularization can be explained as a set of dynamic equilibria having been reached, now involving a large number of structures (essentially most of the population members).

By resorting to the combined method, where the cluster regularization method is used with its optimum 70% likelihood, the overall best performance is reached for all three system sizes (red curves). The combined method still exhibits a much sooner onset of finding the global en-

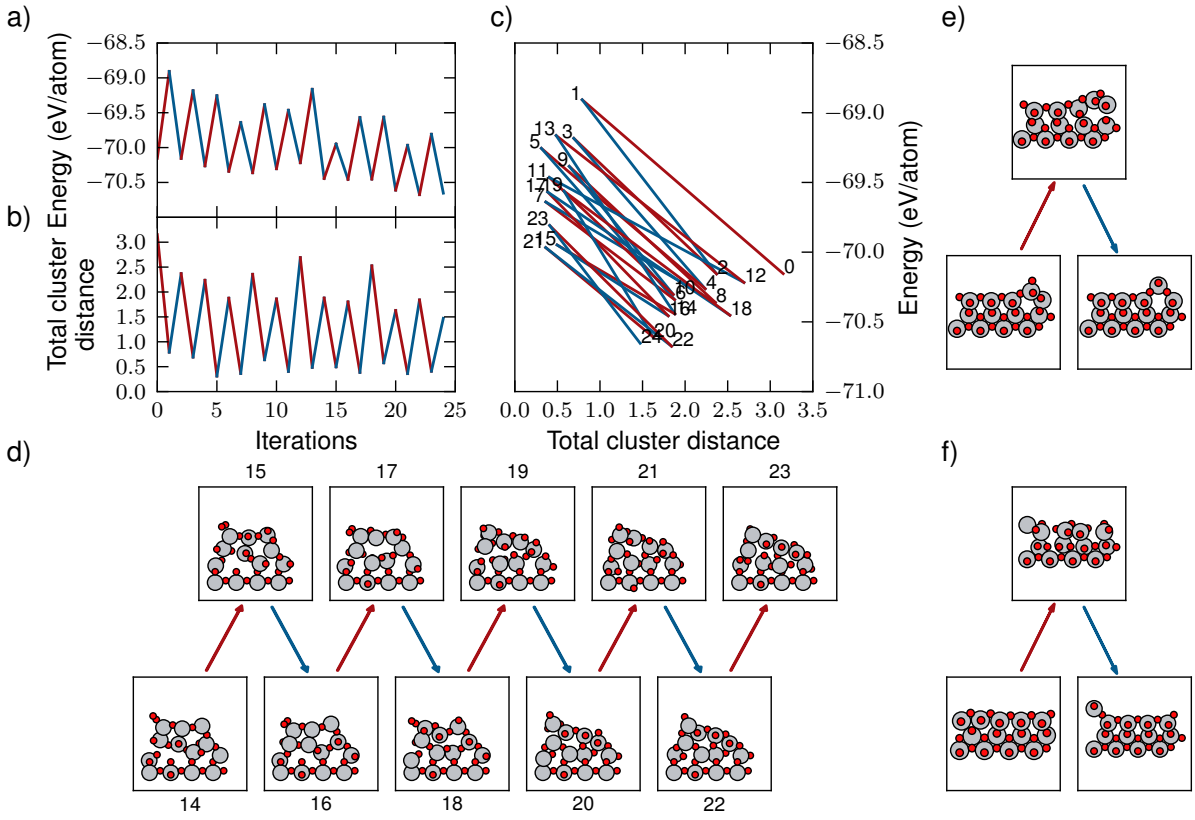

FIG. 6. Evolution of energy (in eV/atom), cluster distance, and structures during basin hopping updates using the cluster distance regularization method as perturbation in the structural updates. Red lines and arrows indicate cluster distance minimization steps, while blue lines and arrows indicate energy relaxations. a-c) Data for the same 25 iterations in one basin hopping run. d) Structural illustrations of 10 selected states from the same run. The progress in the individual minimization steps is small and sometimes hard to appreciate. However, from state 16 to 17 the cluster regularization acts to fill the vacancy in the second layer and from state 18 to 19 it splits up the two oxygen atoms present in the top left corner. e-f) Structural illustrations of the second-last and last states in two different basin hopping runs that led to successful finding of the global minimum structure.

ergy minimum structure than the benchmark runs, and a small degree of stagnation is seen.

It is interesting to compare the results across the three different system sizes. First of all, it is seen that as the complexity of the problem increases from optimizing the positions within the 2-layer system (12 static + 15 optimized atoms) to the 3-layer and 4-layer systems (12 static + 27 and 39 optimized atoms), the overall efficiency of the EA in its benchmark form goes down. After 1000 attempts, the benchmark method finds the global minimum in ∼80, ∼10, and ∼1 % of the EA restarts. The longer runs with more attempts for the 3-layer and 4-layer systems bring the success rates up to ∼33 and ∼3 % after 2000 and 4000 attempts, respectively. Importantly, these levels are reached with either the all-cluster regularization method or the combined method after 5-600 and 1200-1500 attempts for the two sizes, respectively. This can be phrased as a more than three-fold speedup. We thus conclude that our new method shows great promise for accelerating global structure optimization. We do, however, also acknowledge that even with this improvement in speed, we are still facing challenges with the overall scaling of the search methods as system sizes are increased. The remedy of this important issue would involve the development of scalable search methods.

VI. SUMMARY

In this work, we have used machine learning techniques to characterize local atomic environments in structures found during structural optimization of the anatase $TiO_2(001)$ surface in search for the global minimum energy structure, the (1×4) ridge reconstruction. Upon clustering according to feature vectors describing the local atom environments, we find a correlation between the sum of distances of the feature vectors to their respec-

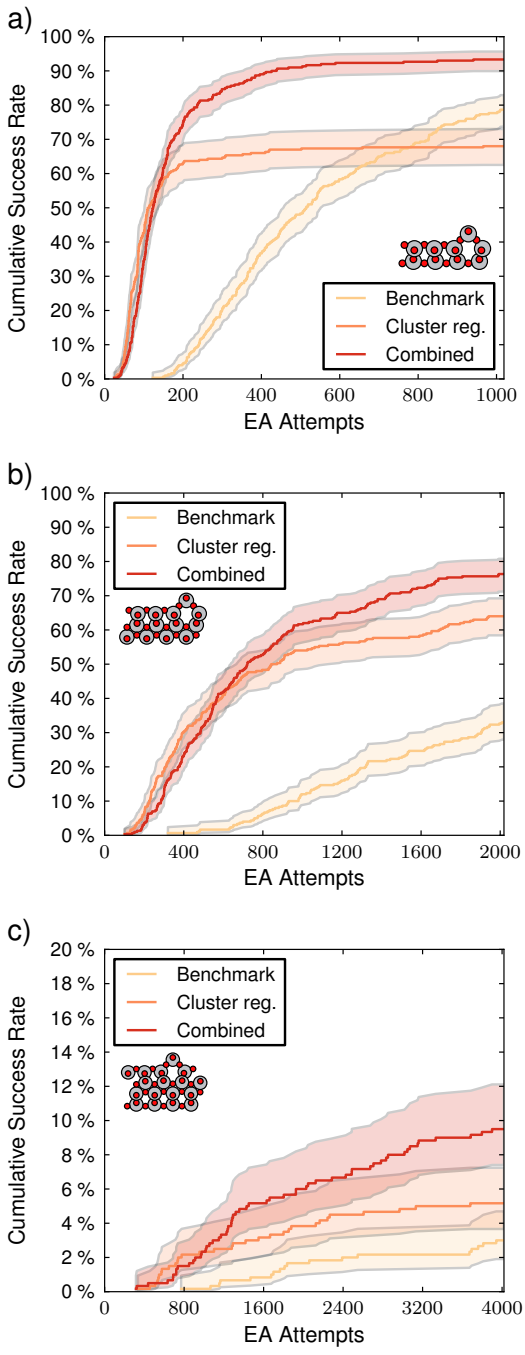

FIG. 7. Cumulative success rates for a) 2-layer, b) 3-layer, and c) 4-layer TiO_2 systems. Benchmark runs (yellow) use 59% cut-and-splice and 41% rattle mutation. All-cluster regularization runs ("Cluster reg.", orange) use 100% cluster regularization cross-over. Combined runs (red) use 70% cluster regularization, 28% cut-and-splice, and 2% rattle mutation in the GA updates. The shaded regions indicate 95% confidence intervals.

tive cluster centers for individual structures to the density functional tight binding energies for these structures. The correlation enables the formulation of a perturbation operation for a basin hopping search or a cross-over operation for an evolutionary algorithm search that have a bias toward taking structures into lower energy basins. The cross-over operation uses gradient based methods to minimize for one parent the sum of the distances to the cluster center as defined by an ensemble of parents. The usefulness of the cross-over method is demonstrated in evolutionary searches for the anatase $TiO_2(001)$ (1×4) ridge reconstruction for three different system sizes.

VII. ACKNOWLEDGEMENTS

We acknowledge support from the Danish Council for Independent Research — Natural Science (grant no. 0602-02566B) and from VILLUM FONDEN (Investigator grant, project no. 16562).

[1] S. Curtarolo *et al.*, Nature Mater **12**, 191-201 (2013).
[2] J. Hafner, C. Wolverton, and G. Ceder, MRS Bulletin **31**, 659-668 (2006).
[3] K. Takayanagi, Y. Tanishiro, S. Takahashi, and M. Takahashi, Surf. Sci. **164**, 367 (1985).
[4] K. C. Low and C. K. Ong, Phys. Rev. B **50**, 5352-5357 (1994).
[5] U. Diebold, Surf. Sci. Rep. **48**, 53-229 (2002).
[6] M. Lazzeri and A. Selloni, Phys. Rev. Lett. **87**, 266105 (2001).
[7] X. Gong, N. Khorshidi, A. Stierle, V. Vonk, C. Ellinger, H. Dosch, H. Cheng, A. Selloni, Y. He, O. Dulub, U. Diebold, Surf. Sci. **603** 138-144 (2009).
[8] R. Bechstein, H. H. Kristoffersen, L. B. Vilhelmsen, F. Rieboldt, J. Stausholm-Møller, S. Wendt, B. Hammer, and F. Besenbacher, Phys. Rev. Lett. **108**, 236103 (2012).
[9] C. J. Pickard and R. J. Needs. J. Phys.: Condens. Matter **23** (2011) 053201 (23pp) DOI:10.1088/0953-8984/23/5/053201
[10] D. J. Wales, J. Phys. Chem. A **101**, 5111 (1997).
[11] R. L. Johnston, Dalton Trans. **22**, 4193 (2003).
[12] A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[13] S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, Phys. Rev. Lett. **111**, 135501 (2013).
[14] D. M. Deaven and K. M. Ho, Phys. Rev. Lett. **75**, 288 (1995).
[15] M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, and B. Hammer, J. Phys. Chem. A, **122**, 1504 (2018).
[16] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, Phys. Rev. Mater. **2**, 013803 (2018).
[17] G. Dolgonos, B. Aradi, N. H. Moreira, and T. Frauenheim, J. Chem. Theory Comput. **6**, 266-278 (2010).
[18] D. Selli, G. Fazio, G. Seifert, and C. D. Valentin, J. Chem. Theory Comput. **13**, 3862-3873 (2017).
[19] D. Selli, G. Fazio, and C. D. Valentin, J. Chem. Phys. **147**, 164701 (2017).
[20] A. H. Larsen *et al.*, J. Phys. Condens. Matter **29**, 273002 (2017).
[21] L. B. Vilhelmsen and B. Hammer, Phys. Rev. Lett. **108**, 126101 (2012).
[22] L. B. Vilhelmsen and B. Hammer, J. Chem. Phys. **141**, 044711 (2014).
[23] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[24] A. P. Bartok, R. Kondor, and G. Csanyi, Phys. Rev. B **87**, 184114 (2013).
[25] K. Hansen *et al.*, J. Chem. Theory Comput. **9**, 2404 (2013).
[26] Z. Li, J. R. Kermode, and A. De Vita, Phys. Rev. Lett. **114**, 096405 (2015).
[27] V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1075-18083 (2015). DOI:10.1002/qua.24836
[28] A. Khorshidi and A. A. Peterson, Comput. Phys. Commun. **207**, 310 (2016).

[29]H. Zhai and A. N. Alexandrova, J. Chem. Theory Comput. **12**, 6213-6226 (2016).
[30]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[31]K. Yao, J. E. Herr, S. N. Brown, and J. Parkhill, J. Phys. Chem. Lett. **8**, 2689 (2017).
[32]T. K. Patra, V. Meenakshisundaram, J.-H. Hung and D. S. Simmons, ACS Comb. Sci. **19**, 96 (2017).
[33]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, J. Phys. Chem. Lett. **6**, 23262331 (2015).
[34]A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006)
[35]T. L. Jacobsen, M. S. Jørgensen, and B. Hammer, Phys. Rev. Lett. **120** 026102 (2018).
[36]J. E. Moussa, Phys. Rev. Lett. **109**, 059801 (2012).
[37]M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, Phys. Rev. Lett. **109**, 059802 (2012).
[38]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[39]B. Huang and O. A. von Lilienfeld, J. Chem. Phys. **145**, 161102 (2016).
[40]F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, J. Chem. Theory Comput. **13**, 5255-5264 (2017)
[41]K. M. Gilbert and C. A. Venanzi, J. Comput. Aided Mol. Des. **20**, 209-225 (2006).
[42]L. L. Duan, Y. Mei, D. Zhang, Q. G. Zhang, and J. Z. H. Zhang, J. Am. Chem. Soc. **132**, 11159-11164 (2010).
[43]D. Chema and A. Goldblum, J. Chem. Inf. Comput. Sci. **43**, 208-217 (2003).
[44]M. S. Jørgensen, M. N. Groves, B. Hammer, J. Chem. Theory Comput. **13**, 1486-1493 (2017).
[45]A Modified k-means Algorithm to Avoid Empty Clusters Malay K. Pakhira, International Journal of Recent Trends in Engineering, Vol 1, No 1, May 2009
[46]D. Arthur and S. Vassilvitskii, *K-means++: the advantages of careful seeding* in In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, (2007).