Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

GOCIA: grand canonical Global Optimizer for Clusters, Interfaces, and Adsorbates

Zisheng Zhang,*[abc] Winston Gee,[a] Robert H. Lavroff,[a] and Anastassia N. Alexandrova*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Restructuring of surfaces and interfaces underlie the activation and/or deactivation of a wide spectrum of heterogeneous catalysts and functional materials. The statistical ensemble representation can provide unique atomistic insights into this fluxional and metastable realm, but constructing the ensemble is very challenging, especially for the systems with off-stoichiometric reconstruction and varying coverage of mixed adsorbates. Here we report GOCIA, a general-purpose global optimizer for exploring the chemical space of these systems. It features the grand canonical genetic algorithm (GCGA), which bases the target function on the grand potential and evolves across the compositional space, as well as many useful functionalities and implementation details. GOCIA has been applied to various systems in catalysis, from cluster to surfaces, and from thermal to electro-catalysis.

1 Introduction

Understanding the catalyst's structure under reaction conditions is crucial for deciphering the reaction mechanism and further design or optimization. In the recent decade, with the development of *in situ* and *operando* characterization techniques, many common thermal and electro-catalysts have been found to undergo highly non-trivial restructurings during operation.[1] Moreover, the "restructuring" is not a single transformation but a collective phenomena which involves multiple coexisting catalyst states, pathways, time scales, and intricate interplay with the adsorbates and environments.[2]

Molecular dynamics (MD) based methods, when combined with enhanced sampling techniques[3] and/or machine learning interatomic potentials,[4,5] have become a powerful tool to modeling many dynamical behaviors in catalysis. However, they typically focus on the potential energy landscape of a fixed-composition system and hence are often insufficient in exploring the chemical space of off-stoichiometric restructuring systems with a fluctuating composition and without any well-defined collective variable.

Another approach is the to revise the representation of the catalyst as a statistical ensemble of many catalyst states instead of a single or a few selected structures.[6,7] By extending to a grand canonical (GC) ensemble representation, all reaction-relevant global minimum (GM) and local minimum (LM) catalyst states with varying geometry and composition (including both surface itself and adsorbate/adatom coverage) can be included in the representation, with their individual contributions to reactivity or spectroscopic signals properly evaluated.[8] By probing the response of GC free energetics of the states to external factors (i.e., reaction conditions), the ensemble becomes condition-dependent in nature and can be used to understand and predict structural evolution during operation,[9] or to better simulate spectra by ensemble-averaging.[10]

Despite the simplicity of the ensemble representation theory, obtaining such an ensemble – including the *ab initio* thermodynamics of all relevant surface phase – is rather computationally challenging.[11] The difficulty lies in the exponentially growing chemical space of off-stoichiometric restructuring versus system size and number of elements. Indeed, constructing a realistic ensemble requires inclusion of all relevant states, which means searching extensively the global and local minima on the potential energy surface (PES), for all relevant stoichiometries. Note that the global optimization minima search at density functional theory (DFT) level, even for small clusters with fixed composition, is highly nontrivial.[12,13]

A recently emerging family of GO techniques is to directly use the grand canonical free energy (Ω, also named grand potential), which is a function of system's composition at a given set of chemical potentials, as the target function of the minima search. This allows for GC global optimization, in which the stoichiometry is also treated as a set of discrete variables to optimize. In this way, we do not need to extensively sample each possible stoichiometry in a grid-search fashion, but can efficiently sample into the relevant stoichiometries in the grand canonical free energy surface (FES) and producing a distribution of stoichiometries in the resulted states. GC treatments has recently been successfully ap-

[a] *Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, 90095-1569, USA. Email: ana@chem.ucla.edu*
[b] *Department of Chemical Engineering and SUNCAT Center for Interface Science and Catalysis, Stanford University, Stanford, California 94305, USA. Email: zishengz@stanford.edu*
[c] *SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA.*

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

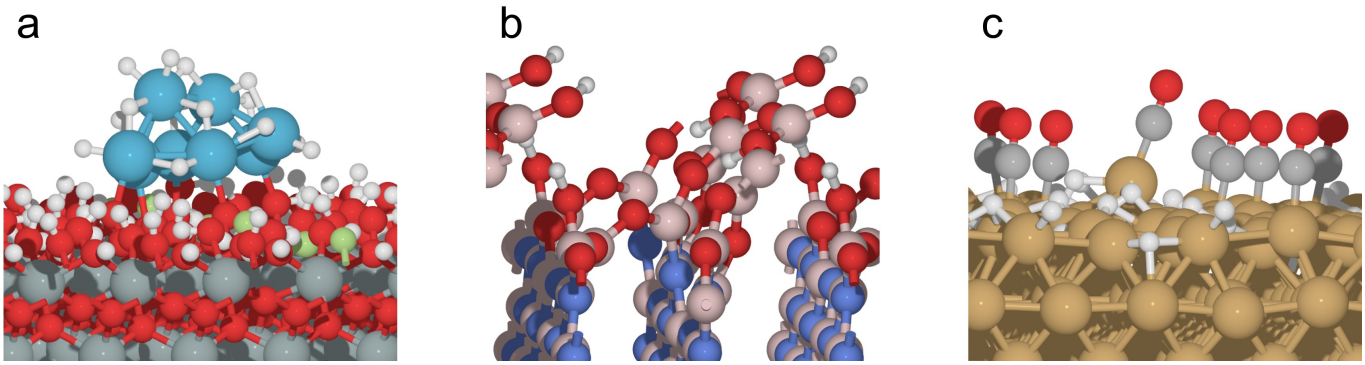

Fig. 1 Examples of previous applications of GOCIA on catalytic systems. (a) H-covered Pt_n clusters supported on hydroxylated F-doped tin oxide in electrochemical conditions. (b) Partially oxidized and hydroxylated over-layer of hexagonal boron nitride. (c) Restructuring of crystalline Cu(100) surface under the coverage of a mixture of H and CO adsorbates.

plied to multiple cluster or surface systems, within algorithms such as GC basin hopping (BH),[14,15] GC Monte Carlo (MC),[16] and GC genetic algorithm (GA).[17,18] However, these algorithms are usually tailored for a specific set of systems, and a general-purpose GC global optimizer has been lacking.

This article is aimed to introduce our recent efforts in developing a Global Optimizer of Clusters, Interfaces, and Adsorbates (GOCIA)[19] – a versatile Python package featuring GC global optimization of off-stoichiometric restructuring systems – with a detailed dissection of its components, and to showcase its previous successful applications, applicability, and a roadmap to future developments.

2 Overview of features

GOCIA is built to achieve efficient global optimization of periodic systems and can handle internally many nuances that come with the periodic boundary conditions such as collision of atoms and breaking of polyatomic fragments.

The main feature of GOCIA is the grand canonical genetic algorithm (GCGA) which can efficiently explore the relevant regions in the chemical space of varying compositions, by using grand canonical free energy as the search target, and it eliminates the need to grid search for every possible composition. Built on the basis of GA, GCGA can achieve extremely efficient exploration of geometric and compositional space, as compared to MD- or Monte Carlo (MC)-based approaches.

GOCIA was initially built to handle amorphous layers without well-defined bonding modes, where every atom in the sampling region was allowed to form any type of bond (as individual adatoms). A recent update enabled our implementation of GCGA to handle the coverage of polyatomic and mixed adsorbates while maintaining their intactness, which is rather relevant to studying reaction intermediate-relevant surface phenomenon and the complex interplay between surface atoms and multiple types of surface species.

GOCIA's random structure generator, whose primary role is to make the initial population for GCGA, can also work as a good one-shot sampler for the smaller systems such as smaller sub-nanometer clusters supported on surfaces and adsorbate configurations at low coverage.[20]

GOCIA also provides a toolkit and streamlined workflow for grand canonical density functional theory (GCDFT) calculations using the surface charging approach. This is useful for sampling of electrified interfaces, such as those used in electrocatalysis.

Every mentioned component of GOCIA are highly versatile and can be customized to meet a broadness of needs in the areas of catalysis, materials science, surface science, and so on.

GOCIA has been applied to study the structure, reactivity, and spectroscopy of many surface systems ranging from clusters to amorphous over-layers and to reconstruction of crystalline metal electrodes, in thermal- and electro-catalysis.[21] A few representative systems shown in Fig. 1a-c are: fluorine-doped tin oxide (FTO) supported Pt_n (n=1-8) clusters under varying H coverage during electrochemical hydrogen evolution reaction (HER);[10] partial boron oxide/hydroxide over-layer formed on hexagonal boron nitride (hBN) in conditions of oxidative dehydrogenation of propane (ODHP);[22,23] restructuring of crystalline Cu facets induced by H and CO coverage in CO_2 reduction reaction (CO_2RR) conditions.[24] Other notable applications include restructuring of Cu in acidic HER conditions,[9,25] metal-support contact angle of small nano-particles (NPs),[20] and the structure of amorphous nickel oxide/hydroxide on Pt surface.[26]

3 Code architecture

3.1 The Interface class

Central to GOCIA is the Interface class which is a representation of the system of study.

The Interface class is based on the Atoms class (from the ASE module[27]) with some additional structure-related metadata as is illustrated in Figure 2. There are two atomistic parts within an Interface object, a constrained region and a relaxed region. The constrained region is usually the bottom few layers of the slab and can mimic the behaviour of the bulk. The relaxed region is the part of the surface that can interact with the external environment but cannot change its own composition, usually the top few layers of the slab or supported surface species such as subnanometer clusters or adatoms.

The user would also need to define a rectangular sampling box

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(by the coordinates of its vertices) which intersects with the top few layers of the relaxed region. Compositional changes are only allowed within the sampling box.

In the case of sampling polyatomic adsorbates, one would also need to supply a list of atomic indices for each adsorbate, so that GOCIA can keep track of the connectivity and make sure every adsorbate is intact during the local and global optimizations, with a similar practice to ref[13].

A number of useful functions are built-in under the Interface class for easy access, modification, and geometric analysis of each individual component.

3.2 Data structure

During the global optimization, a large number of structures are generated, and each must be fully optimized to a local minimum before it can be added to the ensemble. GOCIA will make a dedicated sub-directory to each structure, so that the local optimization jobs would be performed in separate sub-directories and not interfere with each other. After a local optimization job finishes, the results will be updated to the project database file in the main directory.

The project database file (a SQL database in ASE format) stores all optimized structures along with their metadata (calculator, energy, magnetic moments, fragment lists, labels, population information, etc.), to allow for easy query and manipulation.

All structures in a global optimization search share the same definition of the constrained region, the relaxed region, and the sampling box. These information are stored in a *substrate.vasp* file (it can be in any format that supports periodic structure with constraints) which is one of the required input files.

The other variables needed to set up a global optimization run, such as the dictionary of chemical potentials, control parameters of GA, and paths/commands to initiate software, can be provided as a separate *input.py* file in the main directory or included in the main "manager" script (*vide infra*).

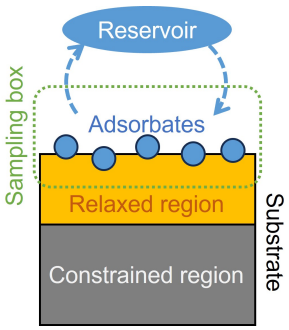

Fig. 2 Schematic of the components of the Interface class.

3.3 Parallel scheme

The overall parallelization efficiency of the global optimization depends on two factors. (i) The scaling performance of the local optimization calculation: For most electronic structure codes, the scaling performance versus the number of nodes is rather poor, and the optimal parallelism setting is usually within 20 nodes *per* instance.[28] (ii) The population updating of GA: To avoid too drastic a change of the population, it is more beneficial to add new structures to the population one-by-one or in small batches (similar to the population size), instead of in large batches.

Depending on the job requirements and queuing policy of the high performance computer (HPC), GOCIA users can choose from two different workflows:

(i) If the HPC allows submission of a large number of small jobs from a single user: submit a manager job of long wall time, as a single-core process on the login node or interactive session (the manager sleeps periodically and is not resource intensive at all). The manager job will automatically make and submit many worker jobs, each performing a series of local optimization calculations on a structure to which the worker is assigned. The manager will check the queue constantly and resubmit a new worker job if an old one has finished.

(ii) If the HPC strongly encourages large job by measures such as limiting the wall time of smaller jobs: Use the multiprocessing module of Python to maintain a pool of many worker processes. The main script will automatically spawn a new worker process to the idle nodes whenever an old one has finished. This should be submitted as a single large bundled job.

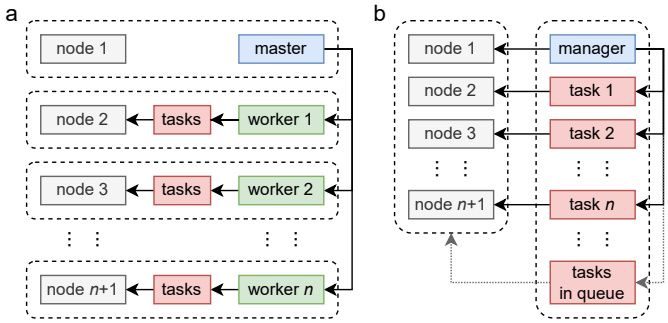

Fig. 3 Parallel scheme of GOCIA on computing clusters. (a) The distributed scheme where each master or worker job are submitted as separate jobs on each allocated node. (b) The bundled scheme where one master job manages all tasks within a large bundled job on all allocated nodes.

3.4 Extensibility

GOCIA currently supports VASP the best, covering all functionalities described in this article. In principle, GOCIA can interface with any code *via* the ASE Calculator class to perform the core functionalities. Note that, although the ASE Calculator class interface is easy to use, it comes with some compromise in computational efficiency (charge density and/or wavefunction IO from the use of a Python wrapper *per* force call) and some advanced functionalities (iterative local optimization with fragment information and GCDFT). A workaround is to define the calculator such that it runs a local optimization internally using the code's own optimizer, and then GOCIA calls it for a single point, which conserves the conveniences of using the ASE calculator class while suffering no IO bottleneck. Since GC global optimization is a highly computationally intensive task, we plan to ultimately make an individual optimized interface for each popular periodic DFT and

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

semi-empirical code.

4 Grand canonical genetic algorithm

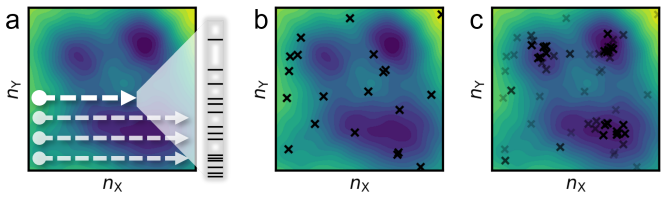

Fig. 4 Schematic comparison of different approaches to explore off-stoichiometric restructuring involving elements X and Y. The grand canonical free energy landscape is shown as a contour plot depending on the number of X and Y atoms. (a) Grid search within a defined range of compositions, performing a canonical global optimization at each grid. The inset bar shows the energetic distribution of states of the same composition. (b) Stochastic one-shot sampling, with × representing the samples. (c) Grand canonical global optimization with an iterative scheme. Lighter and deeper colors represent samples in earlier and later iterations, respectively.

Before going into the details of the GCGA, we first discuss the challenges in exploring the off-stoichiometric restructuring. In the context of thermal and electro-catalytic surfaces, we assume that the system is always in the electronic ground state for a given set of nuclear positions. Finding the stable and metastable structures of a certain stoichiometry is then equal to locating the global minimum and local minima of the ground state potential energy surface (PES) defined by a non-convex function $E(\mathbf{r})$, where $\mathbf{r}$ is the atomic coordinate. For a system containing N atoms, there are $3N$ variables, spanning a vast high-dimensional space. Moreover, there is no analytical expression of $E(\mathbf{r})$ due to its quantum mechanical nature, and all values (energy) and gradients (forces) needs to be computed numerically, which is extremely resource-intensive.

Abstraction, such as treating surface adsorption configurations as lattices of graphs, could help reduce the dimensionality of the problem. However, these abstraction will only hold when the surface itself is relatively rigid regardless of the adsorbate/adatoms on it. In other words, the coupling between surface species coverage/configuration and arrangement of surface atoms are negligible. This might be actually the case for some systems, but it is quite dangerous to assume so universally, with the growing collection of reports on non-trivial restructurings of surface and clusters.[29] For the latter, there exists no shortcut.

The picture further complicates when we allow the composition to vary —- the system becomes a collection of many constant-composition potential energy hyper-surfaces, each with different dependence on external factors/conditions. Let us consider a system where the number of X and Y atoms, n_X and n_Y, can vary. In the discrete compositional space, each grid point defined by (n_X, n_Y) entails a full PES.

The most straightforward approach to explore this chemical space is the grid search (Fig. 4a) —- performing a canonical global optimization on the corresponding constant-composition PES of each (n_X, n_Y). This approach would in theory yield the most uniform sampling distribution over the whole chemical space, however, it is extremely inefficient as the vast majority of the (n_X, n_Y) grids are in the irrelevant regime to the ambient or operating condition of the catalyst. In addition, the compositional space is infinite, and the initial definition of the grid (i.e., the upper and lower bounds) is arbitrary.

Stochastic sampling into random compositional grids can provide a bird's-eye view of the GC free energy landscape at a very low cost (Fig. 4b). For smaller systems, the one-shot samples may sometimes even suffice as a (sub-)ensemble. However, for larger systems, it is as inefficient as the grid search approach, because, again, the majority of the grid points are catalytically irrelevant.

To guide the search towards those relevant regions in the compositional space, one can adopt the GC free energy Ω, within the GC ensemble (μVT), as the basis of the target function. The composition is then treated as an additional set of variables to optimize. In a typical iterative GC global optimization search, the initial stochastic samples inform the searcher about the "promising" regions, and the search direction is adaptively updated throughout the search to sample denser and denser into the relevant minima regions (Fig. 4c).

4.1 Calculation of the grand canonical free energy

Now we introduce the calculation of the main thermodynamic metric used in GC global optimization, GC free energy Ω. In the context of off-stoichiometric surface restructuring under a certain reaction condition, we divide the atoms into two groups: group A includes species (blue spheres in Figure 2) that the system can freely exchange with the reservoir, such as adatoms and adsorbates; and group B are atoms in the substrate (relaxed and constrained regions in Figure 2). The whole system is labeled as AB. The number of atoms in group B is constant, while those in the group A can fluctuate. The GC free energy of a certain AB configuration with respect to the group A species can then be written as:

$$\Omega_A = U_{AB} - TS_{AB} - \sum_A \mu_i N_i - \sum_B \mu_j N_j \tag{1}$$

Because the number of group B atoms does not change, the fourth term is a constant for all states in the ensemble and does not influence the relative energetics. Here we take the bare surface as a reference state for group B atoms and set value of $\sum_B \mu_i N_i$ as the electronic energy of a bare surface slab, E_B.

$$\Omega_A = U_{AB} - TS_{AB} - \sum_A \mu_i N_i - E_B \tag{2}$$

In a strict sense, the calculation of U_{AB} and TS_{AB} terms requires vibrational analysis, which is unaffordable in the context of *ab initio* global optimization involving tens of thousands of configurations. Hence, we approximate the value of $U_{AB} - TS_{AB}$ to the electronic energy of the whole system, E_{AB}. The lost thermal correction terms related to group A species are then absorbed into the chemical potential as a new μ' term. The GC free energy with respect to group A species can then be expressed as:

$$\Omega_A \approx E_{AB} - E_B - \sum_A (\mu_i - \delta E_i) N_i = E_{AB} - E_B - \sum_A \mu_i' N_i \tag{3}$$

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Here δE denotes the thermal correction terms to the free energy related to the group A species, including the zero point energy, constant pressure heat capacity, and vibrational entropy. Note that, for consideration of cost, we assume that any group A species in any configuration has the same δE to avoid explicit vibrational analysis for every configuration.

The μ is a function of reaction conditions such as temperature, partial pressure, concentration, pH, and electrode potential. For example, the corrected chemical potential of H, μ'_{H}, can be expressed as:

$$\mu'_{\mathrm{H}} = \frac{1}{2}E^{gas}_{\mathrm{H_2}} + \delta E^{\mathrm{gas}}_{\mathrm{H}} - \ln(10)k_{\mathrm{B}}T\mathrm{pH} - |e|U_{\mathrm{SHE}} - \delta E^{\mathrm{ads}}_{\mathrm{H}} \quad (4)$$

The $E^{\mathrm{gas}}_{\mathrm{H_2}}$ is the electronic energy of an optimized gas phase H_2 molecule. The δE^{gas} can be obtained from vibrational analysis of the gas phase H_2 molecule and thermochemistry calculations. pH effect is incorporated using the Nernst equation, and the electrode potential effect is included using the computational hydrogen electrode model. The $\delta E^{\mathrm{ads}}_{\mathrm{H}}$ can be obtained from vibrational analysis and thermochemistry calculations on one or a set of relevant H adsorption configurations.

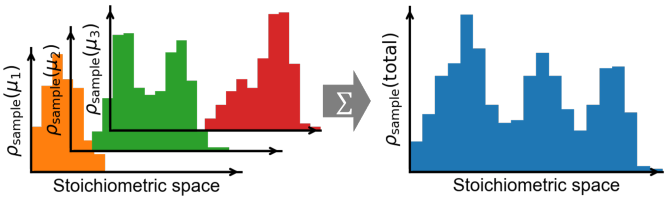

Fig. 5 The recommended practice for constructing a well-sampled GC ensemble. Multiple GC global optimizations are performed at a series of chemical potentials (μ_n, $n = 1, 2, 3, ...$). The samples (sub-ensembles) from multiple runs are then merged in to a total ensemble. If the sample distribution is continuous over the stoichiometric space of interest, the merged ensemble can be used in the interpolated μ range among the μ_n values.

Note that, the calculation of μ for some elements or species can be less straightforward for a lack of appropriate reference state and/or the limitation of electronic structure method. The calculated μ can be off by up to a few hundred meV from the realistic condition, and in some cases, one may only be able to estimate a relevant range of μ for a specific species. In those cases, it is advised to perform multiple searches at various μ values in the relevant range, so as to gain a broader distribution of stoichiometry. If there are prior experimental information on surface composition or adsorbate coverage, one may also vary the μ on a sub-ensemble (from one-shot sampling or an unfinished search) and probe the response of the GM stoichiometry by using the ensemble analysis functions provided by GOCIA (*vide infra*). This will help narrow down the μ window relevant to the experiment.

Each GC global optimization run would yield likely a multimodal distribution of stoichiometries (Fig. 5, left panel). The number of modes and the width of the distribution can be highly system dependent, so it is recommended to always check the stoichiometric distribution in the final ensemble merged from multiple searches —- they should ideally join and have a more or less uniform density over the stoichiometric space of interest (Fig. 5, right panel). If there is any discontinuity, then more sampling is deserved at its corresponding μ values. After sufficient sampling, the final merged ensemble can be used for further analysis at any μ within the interpolated range among the μ values used in sampling .

4.2 Random structure generation

GOCIA offers three types of structure generation methods from a base surface:
(i) Growth sampling. It first randomly selects an existing atom from the relaxed region. A random unit vector will be generated to be the direction of the "growth". The adatom/adsorbate is then aligned to the "growth" direction and placed along it, with the selected surface atom as the starting point. The distance between the adatom/adsorbate and the selected surface atom is then sampled from the bond length distribution algorithm (BLDA),[30] based on the covalent bond radii of the two atoms that should form the surface-adsorbate bond. This methods can generate new structures with the most reasonable interatomic distances with high efficiency, but it may fail for some corner cases where the growth direction is ambiguous, such as the interface between a large cluster and the surface, or when the surface is already quite crowded with adsorbates.
(ii) Box sampling. It directly makes attempts to place adatoms/adsorbates into the sampling box with random positions and orientations. Since it is less dependent on the surface structure, it works well on cases with irregular shapes and morphology, non-directional and multi-center bonds, as well as very crowded surfaces. Note that this method can also be used to generate molecular packing structures, such as micro-solvation slab,[31] by applying connectivity constraints and expanding the sampling box.
(iii) Graph sampling. This method constructs a connectivity graph of the top surface layer, and then identify the atop, bridge, and hollow sites using the NetworkX module.[32] Adsorbates are then added to the identified sites with random rotations. Note that this method expects well-defined lattices and works the best for exploring adsorption on unrestructured surfaces or just to enrich the initial population.

In all three methods, the interatomic distances of attempted geometry are checked to avoid bad contacts. The user may also opt to check the similarity of a new structure with already generated structures to prevent duplicates in the very beginning. GOCIA also offers many user-defined constraints such as bonds that must (or must not) form, upper and lower limits of the coordination number, whether the added adatom/adsorbate can incorporate into the relaxed region or must stay above. If multiple types of adatoms/adsorbates are to be added, the list can be randomly shuffled before addition to prevent biases from the original ordering. The process iterates until all adatoms/adsorbates have been added to the sampling box while satisfying all geometric constraints.

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

4.3 Pre-optimization & iterative optimization

To ensure an aggressive sampling, which underlies extensive and delocalized exploration of the chemical space, oftentimes one would allow some unphysical connectivity or interatomic distances to form in the random structure generation. This may cause slow-down (or even failure) of the self-consistent field (SCF) or force convergence to the initial steps in the local optimization.

A remedy to this problem is to perform a pre-optimization at a lower level of theory before the structure is fed to the electronic structure calculator. GOCIA currently supports Hookean and Lennard-Jones potential as the calculator for the preoptimization. Any code can be interfaced to GOCIA as the pre-optimizer *via* the ASE Calculator class.

To reduce the overall computational expense, we adopt a multi-stage local optimization strategy (Figure 6, left), where each stage has a different level of precision and convergence criteria, from computationally cheaper to more expensive. In this way, we can rationalize the structure in earlier and cheaper stages and bring the structure closer to its local optimum, before the final stage of higher precision for production.

Since electronic structure codes do not intrinsically constrain connectivity (bonds are determined quantum mechanically), some unwanted motifs or bonds may form during the local optimizations. GOCIA also offers an iterative local optimization scheme which checks the geometry for undesired connectivity after each stage. If any unwanted substructure is detected, GOCIA would modify the structure to meet the constraint and call for another multi-stage local optimization. This again goes iteratively until convergence of the connectivity (Figure 6, right). Currently, GOCIA supports the following connectivity constraints: (1) make sure where is no desorbed species that is not connected to the slab; (2) remove any atom that is outside the sampling region; (3) force all adsorbates to directly form bonds with the surface; (4) remove fragments that are not intact; (5) prevent bonds between fragments; (6) remove atoms that are not involved in a specific type of bonds. Each connectivity constrained can be switched on and off or modified easily. Users can also define their own constraints (geometric or compositional) inside the worker script to archive the unwanted structure, terminate the job, or to modify the structure and send it back for re-optimization.

4.4 Crossover, mutation, and selection

The crossover, mutation, and selection process largely follows the original genetic algorithm proposed by ref[33] and the gradient-embedded genetic algorithm by ref[12]. Here we only highlight a few notable modifications and additions in Figure 7.

In the crossover process, the parent structures are split-and-spliced along the same cutting plane. In case of any bad atomic contact, the one whose center is closer to the cutting plane would be preserved, while the farther one removed. In case of poly-atomic adsorbate, the bridle atom (*via* which the adsorbate is supposed to bind to the substrate) would be viewed as the center of the adsorbate.

In the mutation process, GOCIA offers the following operators:

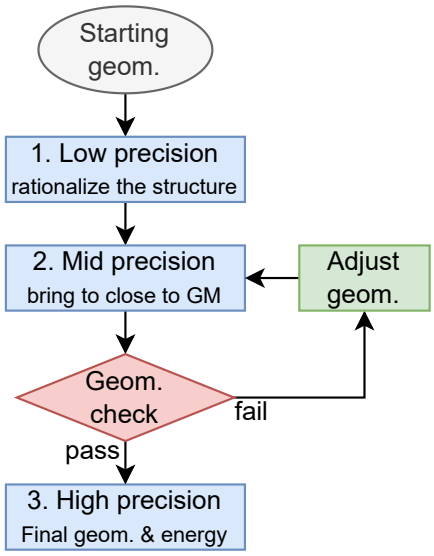

Fig. 6 The workflow of the iterative multi-stage local optimization process in GOCIA.

(i) adding an atom/fragment, (ii) deleting an atom/fragment; (iii) moving a random atom/fragment to a random empty site; (iv) rattling the surface atoms along random vectors drawn from a normal distribution; (v) translating the buffer slab along x or y axis by a fraction of the cell length; (vi) permuting a random fraction of the buffer slab. If an offspring is too similar to its parent, its mutation rate is raised to 100% to avoid recalculating the same structure.

In the selection process, an over-mating penalty factor of $1+(N_{\mathrm{mate}})^{-3/4}$ is multiplied to the grand canonical free energy-based fitness factor. Here N_{mate} is the mating counts, and it penalizes the candidates that have mated too many times to diversify the population. Similarity checks against the current population are performed before adding any new candidate to remove duplicates. Adopted mutation operations include: Upon the addition of each offspring to the population, the candidate with the lowest fitness is archived to maintain the population size.

4.5 Filtering and sorting the ensemble

It is important to avoid or prevent duplicate structures during the global optimization or final analysis of the ensemble. GOCIA adopts an adapted version of the similarity checker proposed by ref[34], which considers both energetic and structural aspects.

After duplicate removal, the unique structures in the ensemble would be sorted and written to a new database which contains all essential metadata from the search. The database file can be used for statistical analysis or computing ensemble-average properties. GOCIA would also report an oversampling ratio which reflects how extensively the chemical space has been sampled. A low oversampling ratio suggests that the sampling is far from extensive, while a high oversampling ratio often means that the search is extensive enough.

The evolution trajectory of a GCGA run, although carrying no physical meaning in a strict sense, contains many useful informa-

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

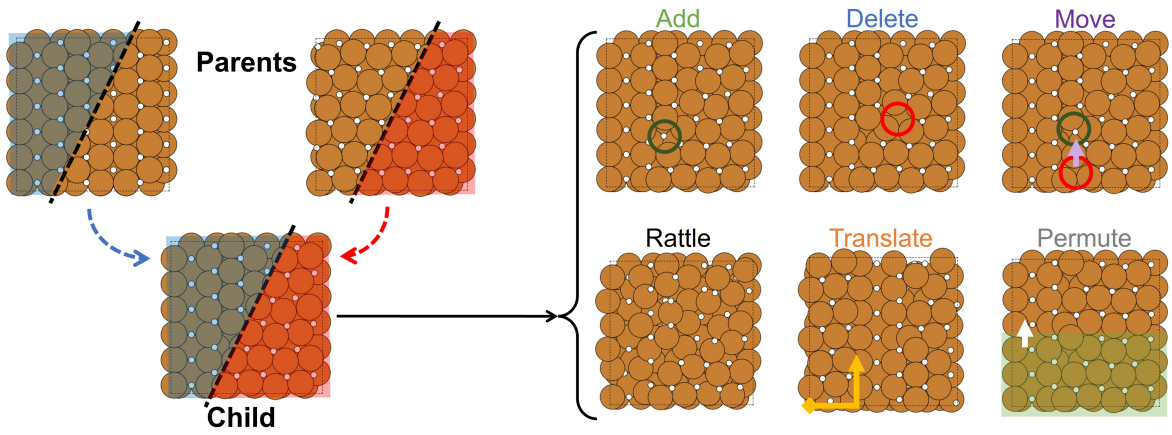

Fig. 7 Crossover of two parent structures to produce a child structure, with an illustration of possible mutation operations.

tion. GOCIA offers scripts that can be used to track the progress of GCGA by plotting the Ω versus number of samples on-the-fly. It can easily visualize the key new GM's in the search history and if there is a good sign of convergence. It can also inform if there is any sign of significantly restructuring, usually characterized by a apparent dive of the population's Ω to a much lower value and staying there, without needing to inspect each structure in the trajectory.

GOCIA also stores the inheritance information of each candidate in the database. To be specific, the identity of each candidate's parents and the type of mutation (if any) that it went through. GOCIA offers scripts that can track the lineage of any candidate and plot it family tree. This can inform putative pathways *via* which the restructured GM may arise from pristine structures, and which mutation operations are the most effective for the system of study.

4.6 Ensemble analysis and beyond

The filter and sorted ensemble of unique minima structures well covers the GM and relevant LMs to a specific condition defined by the used μ. By merging multiple ensembles from searches at different sets of μ (followed by filtering and sorting again), a more complete GC ensemble is yielded and can be applied to all interpolated μ values among the sampled ones.

GOCIA offers a GCE class for *ab initio* thermodynamic analysis of the GC ensemble database. But before anything, an important thing to check is the distribution of stoichiometries. The GCE class offer functions that can cluster the minima into separate groups of the same stoichiometry. By plotting statistical histograms, one can learn about the density (counts) of samples for each stoichiometry, which informs whether the samples cover a continuous range in the chemical space which is the prerequisite for further analysis with interpolated μ values. By calculating the structural similarity metric (the same as in Section 4.5) with respect to a few reference structures, one can also group the samples by their restructuring patterns and check their sampling density.

Within each group, it is straight forward to extract the low-energy local minima (LELMs) as a relevant sub-ensemble, which can be used for further refinement at a higher level of theory or with additional treatments such as solvation and GCDFT. A recommended energy cutoff relative to the GM of each group is $10k_BT$, however, one should always check if the relative energies of the LELMs would reorder at a different level of theory, and there may be a need to use a higher cutoff.

The GCE class offers functions for easy calculation of Ω and Boltzmann population, p, of any states within the ensemble at a specific μ or a series of μ values (Fig. 8a-c) by:

$$p_i(\mu) = \frac{e^{-\Omega_i(\mu)/k_BT}}{\sum_j^N e^{-\Omega_j(\mu)/k_BT}} \tag{5}$$

The μ-dependent populations can then be used to calculate the GC ensemble average of a specific function X (Fig. 8d) by:

$$\langle X \rangle = \sum_i^N p_i(\mu) F_i \tag{6}$$

Here X can be a single-value property (activation energy, adsorbate coverage, etc.) or an array (simulated microscopy image, spectrum, etc.). In this way, we can obtain the ensemble averaged X as a function of any reaction condition within the μ range of sampling.

In the cases where the Boltzmann statistics fail, the ensemble can still serve as an *ab initio* thermodynamics database for kinetics simulations, as it well covers the relevant LELMs. The combination of global optimization and quasi-kinetic MC simulation has been used to study the off-equilibrium structural evolution such as Ostwald ripening of sub-nano clusters[35] and surface roughening of Cu electrodes during CO_2RR.[24]

5 Grand canonical density functional theory

GOCIA also supports GCDFT calculations using the surface charging approach.[36] Specifically, the potential-dependent grand canonical electronic free energy, $\Omega_{el}(\phi)$, of a charged electrode/electrolyte interface at a constant potential (i.e., a constant μ_e), is approximated by an effective capacitor model with con-

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

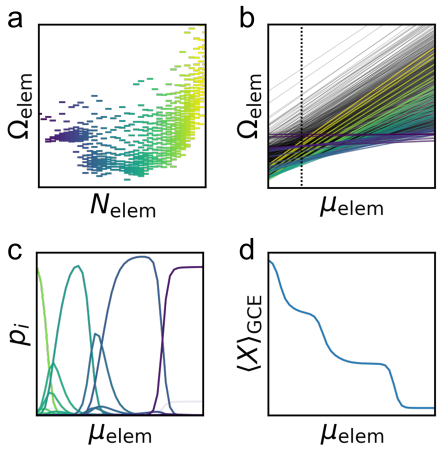

Fig. 8 A typical analysis of the grand canonical ensemble. (a) Computing the grand canonical free energy Ω of all states within the ensemble with respect to some elements at a given set of μ. Each bar represents a unique state. (b) Computing Ω on a series of μ values to generate a condition dependent phase diagram. Each line represents a unique state. Slicing at the dotted line would yield panel a. (c) Calculate the Boltzmann population p_i of each state as a function of μ. (d) Use the p_i to calculate μ-dependent ensemble averaged property or spectra. All steps shown here are straightforward by using the functions within the GCE class of GOCIA.

stant capacitance:

$$\Omega_{el}(\phi) = E(\phi) - q(\phi) \cdot F\phi \approx -\frac{1}{2} C_{eff}(\phi - \phi_0)^2 \quad (7)$$

Here, $E(\phi)$ is the electronic energy of the surface under the a potential ϕ that is calculated by referencing the Fermi level of the system against the vacuum level. $q(\phi)$ is the surface charge difference referenced against the neutral system, and F is the Faraday constant. ϕ_0 is the potential of zero free charge (PZFC) of the system, and C_{eff} is the effective capacitance of the interface. The linearized Poisson-Boltzmann model as implemented in VASPsol[37] is used to represented the polarizable electrolyte region. By varying the number of electrons (N_{el}) in the system, the surface is charged/discharged, and the electrolyte is polarized. The center of the empty region in the cell (vacuum filled with implicit solvation) is then used as the reference energy level to track the change in the Fermi level of the system. By sampling a series of q (through varying N_{el}), we can obtain a data set of $E(\phi)$ and their corresponding ϕ, which can then be used to fit the quadratic relation (Eqn. 7).

We can then replace the electronic energy terms (E_{AB} and E_B in Eqn. 3) with the resulted $\Omega_{el}(\phi)$. In this way, we can eventually obtain the potential-dependent total GC free energy, Ω_{tot}, with respect to all adatoms/adsorbates as well as electrons:

$$\Omega_{tot}(\phi) \approx \Omega_{el,AB}(\phi) - \Omega_{el,B}(\phi) - \sum_A \mu_i' N_i \quad (8)$$

5.1 Slab symmetrization

Symmetrized slabs are recommended for constant-potential calculations. GOCIA can construct a symmetrized slab using mirror and center symmetry operations from an asymmetric slab. This operation only requires a few structural parameters and can be easily applied to a large number of structures within the same ensemble. The user can also make customized operations that combines multiple symmetry operations and atoms addition/removal for slabs with unusual stacking or chirality.

5.2 Automated surface charging workflow

GOCIA provides a wrapper for easy surface charging calculations. The user only needs to provide a list of numbers of fractional electrons that needs to be added/removed from the system, and GOCIA would calculate the corresponding N_{el} and make the input files. A separate job sub-directory will be made for each N_{el}, and it again can be run in a serial or parallel way. After jobs corresponding to all N_{el} values converge, GOCIA can automatically parse the output files, extract the key results, and then fit and report the Ω_{el} – ϕ relation. After all GCDFT calculations converge, GOCIA can extract the fitting parameters and write them into the database file for further data query and analysis (similar to in 4.5).

6 Comments & perspectives

We would like to note that GOCIA is not a black box, but rather a open toolbox with many tunable parts and options. The user should be prepared to make customization according to the nature of the system to study, especially what to do with each individual component. Otherwise the sampling could go off to unwanted configurational subspace and waste a lot of computational resource.

Future developments of the GOCIA would include: (i) Varying the chemical potentials (corresponding to reaction conditions) during the search. The "scan rate" can be adaptive and depend on how extensive the local chemical potential regime has been sampled. This can be useful in identifying the critical conditions where there is a switch in thermodynamic global minimum. (ii) Symmetry-based operations and substructure representation, which may accelerate the convergence for some systems where the bonding is more directional and coordination patterns are more well-defined. (iii) Motif-based operations, which can keep track of energetically favorable structural motifs during the search and include them in later structure generation steps, similar to ref[13] but covering periodic and multi-component cases. (iv) Sampling of explicit solvation layers. Some key goals are determining electrolyte hydration structures, and building micro-solvation models for surface species in a more adaptive and efficient way.

Machine learning (ML) models, especially the interatomic potentials, have undergone impressive development over the recent decade.[38–41] However, in our opinion, there are still two obstacles in applying them to global optimizations: (i) Overall cost. The computational cost for generating the training data for making a good model that well covers the corner cases would be comparable to, if not larger than, that of a direct global optimization approach. (ii) Force accuracy. Unlike the case of MD, global optimizations requires very accurate force (at the magnitude of a few meV/Å) to ensure that the final ensemble contains only minima states and exclude saddle points or other structures on flat local

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

regions of the PES. We look forward to further advances in ML model architectures that can enable more accurate force predictions and surpass the limitations discussed afore. At that time, GOCIA would still serve as a excellent generator of diverse and off-equilibrium training dataset – or it can be incorporated as an on-the-fly component into active learning workflows.

7 Conclusions

Herein, we report GOCIA, an open-source Python package for general-purpose global optimization of various off-stoichiometric restructuring systems. GOCIA has proven efficient and successful in a wide range of applications involving adatoms, clusters, crystalline surfaces, amorphous over-layers, and/or adsorbate coverage.

This manuscript covers the main features of GOCIA, with detailed descriptions of its code structure and the grand canonical genetic algorithm. The relevant theories are explained, and other key functionalities are introduced.

GOCIA is a highly versatile and extendable code, and it can be potentially customized to study many other systems beyond heterogeneous catalysis, such as plasma chemistry, metallurgy, batteries, environmental chemistry, and functional materials. GOCIA is an ongoing effort and is open to comments and contributions from researchers in all afore mentioned areas, and we hope to continue the development and implementation of community-needed features in the future.

Author Contributions

Zisheng Zhang: Conceptualization, Investigation, Methodology, Software, Supervision, Visualization, Writing – Original Draft, Writing – Review & Editing. Winston Gee: Methodology, Software, Writing – Review & Editing. Robert H. Lavroff: Software, Writing – Review & Editing. Anastassia N. Alexandrova: Conceptualization, Supervision, Writing – Review & Editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by National Science Foundation CBET grant 2103116 and U.S. Department of Energy, Office of Science, Basic Energy Science Program, grant DE-SC0020125 and DE-SC0019152. ZZ was support by Edwin W. Pauley Fellowship and Dissertation Year Award at UCLA, and currently by Stanford Energy Fellowship at Stanford. The computational resource used for development and application of GOCIA includes: Hoffman2 the UCLA-shared cluster; Cori and Perlmutter of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract DE-AC02-05CH11231; Theta of the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program at the Argonne Leadership Computing Facility, a U.S. Department of Energy Office of Science User Facility operated under Contract DE-AC02-06CH11357.

Notes and references

1 Z. Zhang, B. Zandkarimi and A. N. Alexandrova, *Accounts of chemical research*, 2020, **53**, 447–458.

2 R. H. Lavroff, H. W. Morgan, Z. Zhang, P. Poths and A. N. Alexandrova, *Chemical science*, 2022, **13**, 8003–8016.

3 G. Piccini, M.-S. Lee, S. F. Yuk, D. Zhang, G. Collinge, L. Kollias, M.-T. Nguyen, V.-A. Glezakou and R. Rousseau, *Catalysis Science & Technology*, 2022, **12**, 12–37.

4 J.-C. Liu, L. Luo, H. Xiao, J. Zhu, Y. He and J. Li, *Journal of the American Chemical Society*, 2022, **144**, 20601–20609.

5 J. S. Lim, J. Vandermause, M. A. Van Spronsen, A. Musaelian, Y. Xie, L. Sun, C. R. O'Connor, T. Egle, N. Molinari, J. Florian *et al.*, *Journal of the American Chemical Society*, 2020, **142**, 15907–15916.

6 H. Zhai and A. N. Alexandrova, *Fluxionality of catalytic clusters: when it matters and how to address it*, 2017.

7 B. Zandkarimi and A. N. Alexandrova, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2019, **9**, e1420.

8 Z. Zhang, E. Jimenez-Izal, I. Hermans and A. N. Alexandrova, *The journal of physical chemistry letters*, 2018, **10**, 20–25.

9 Z. Zhang, Z. Wei, P. Sautet and A. N. Alexandrova, *Journal of the American Chemical Society*, 2022, **144**, 19284–19293.

10 Z. Zhang, T. Masubuchi, P. Sautet, S. L. Anderson and A. N. Alexandrova, *Angewandte Chemie*, 2023, **135**, e202218210.

11 T. Lee and A. Soon, *Nature Catalysis*, 2024, **7**, 4–6.

12 A. N. Alexandrova and A. I. Boldyrev, *Journal of chemical theory and computation*, 2005, **1**, 566–580.

13 A. N. Alexandrova, *The Journal of Physical Chemistry A*, 2010, **114**, 12591–12599.

14 F. Calvo, D. Schebarchov and D. Wales, *Journal of Chemical Theory and Computation*, 2016, **12**, 902–909.

15 G. Sun, A. N. Alexandrova and P. Sautet, *ACS Catalysis*, 2020, **10**, 5309–5317.

16 R. B. Wexler, T. Qiu and A. M. Rappe, *The Journal of Physical Chemistry C*, 2019, **123**, 2321–2328.

17 G. Sun, A. N. Alexandrova and P. Sautet, *The Journal of chemical physics*, 2019, **151**, 194703.

18 B. C. Revard, W. W. Tipton, A. Yesypenko and R. G. Hennig, *Physical Review B*, 2016, **93**, 054117.

19 Z. Zhang, *GOCIA: Global Optimizer for Clusters, Interfaces, and Adsorbates*, `https://github.com/zishengz/gocia`.

20 A. Salcedo, D. Zengel, F. Maurer, M. Casapu, J.-D. Grunwaldt, C. Michel and D. Loffreda, *Small*, 2023, **19**, 2300945.

21 Z. Zhang, *PhD thesis*, University of California, Los Angeles, 2024.

22 Z. Zhang, I. Hermans and A. N. Alexandrova, *Journal of the American Chemical Society*, 2023, **145**, 17265–17273.

23 M. C. Cendejas, O. A. Paredes Mellone, U. Kurumbail, Z. Zhang, J. H. Jansen, F. Ibrahim, S. Dong, J. Vinson, A. N. Alexandrova, D. Sokaras *et al.*, *Journal of the American Chemical Society*, 2023, **145**, 25686–25694.

24 Z. Zhang, W. Gee, P. Sautet and A. N. Alexandrova, *Journal of the American Chemical Society*, 2024.

25 D. Cheng, Z. Wei, Z. Zhang, P. Broekmann, A. N. Alexandrova and P. Sautet, *Angewandte Chemie*, 2023, **135**, e202218575.

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

26 C. Wan, Z. Zhang, J. Dong, M. Xu, H. Pu, D. Baumann, Z. Lin, S. Wang, J. Huang, A. H. Shah *et al.*, *Nature Materials*, 2023, **22**, 1022–1029.

27 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus *et al.*, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.

28 F. Wende, M. Marsman, Z. Zhao and J. Kim, International Workshop on OpenMP, 2017.

29 J. Zhang and V.-A. Glezakou, *International Journal of Quantum Chemistry*, 2021, **121**, e26553.

30 H. Zhai and A. N. Alexandrova, *Journal of chemical theory and computation*, 2016, **12**, 6213–6226.

31 A. H. Shah, Z. Zhang, Z. Huang, S. Wang, G. Zhong, C. Wan, A. N. Alexandrova, Y. Huang and X. Duan, *Nature Catalysis*, 2022, **5**, 923–933.

32 A. Hagberg, P. J. Swart and D. A. Schult, *Exploring network structure, dynamics, and function using NetworkX*, Los alamos national laboratory (lanl), los alamos, nm (united states) technical report, 2008.

33 D. M. Deaven and K. M. Ho, *Physical Review Letters*, 1995, **75**, 288–291.

34 L. B. Vilhelmsen and B. Hammer, *The Journal of Chemical Physics*, 2014, **141**, 044711.

35 B. Zandkarimi, P. Poths and A. N. Alexandrova, *Angewandte Chemie*, 2021, **133**, 12080–12089.

36 S. N. Steinmann and P. Sautet, *Journal of Physical Chemistry C*, 2016, **120**, 5619–5623.

37 K. Mathew, V. S. Kolluru, S. Mula, S. N. Steinmann and R. G. Hennig, *Journal of Chemical Physics*, 2019, **151**, 234101.

38 D. Tang, R. Ketkaew and S. Luber, *Chemistry–A European Journal*, 2024, e202401148.

39 K. Wan, J. He and X. Shi, *Advanced Materials*, 2024, **36**, 2305758.

40 S. Bae, D. Shin, H. Kim, J. W. Han and J. M. Lee, *Journal of Chemical Theory and Computation*, 2024, **20**, 2284–2296.

41 H. Jung, L. Sauerland, S. Stocker, K. Reuter and J. T. Margraf, *npj Computational Materials*, 2023, **9**, 114.

https://doi.org/10.26434/chemrxiv-2024-cw1tt-v2 **ORCID:** https://orcid.org/0000-0002-4370-4038 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0