DockOnSurf: A Python code for the High-Throughput Screening of Flexible Molecules Adsorbed on Surfaces

Carles Martí,*,† Sarah Blanck,‡ Ruben Staub,† Sophie Loehlé,‡ Carine Michel,† and Stephan N. Steinmann*,†

†*Univ Lyon, Ens de Lyon, CNRS UMR 5182, Laboratoire de Chimie, F69342, Lyon, France*

‡*Total Marketing & Services, Chemin du Canal – BP 22, 69360, Solaize, France*

E-mail: carles.marti2@gmail.com; stephan.steinmann@ens-lyon.fr

Abstract

We present the open-source python package DockOnSurf which automates the generation and optimization of low-energy adsorption configurations of molecules on extended surfaces and nanoparticles. DockOnSurf is especially geared towards handling polyfunctional flexible adsorbates. The use of this high-throughput workflow allows to carry out the screening of adsorbate-surface configurations in a systematic, customizable and traceable way, while keeping the focus on the chemically relevant structures. The screening strategy consists on splitting the exploration of the adsorbate-surface configurational space into chemically meaningful domains, i.e., by choosing among different conformers to adsorb, surface adsorption sites, adsorbate anchoring points, orientations and allowing dissociation of (acidic) protons. We demonstrate the performance of the main features based on varying examples, ranging from CO adsorption on a gold nanoparticle to sorbitol adsorption on hematite. Through the use of the

presented program, we aim to foster efficiency, traceability and ease of use in research within tribology, catalysis, nanoscience and surface science in general.

1 Introduction

Computational screening has been proven powerful in chemistry and is not restricted to the field of drug discovery[1–3] but also applied for gas adsorption in metal organic frameworks[4] and for material discovery in photovoltaics,[5] photocatalysis,[6] piezoelectrics,[7] batteries,[8] electrocatalysis[9,10] and heterogeneous catalysis.[11] When dealing with chemistry at solid/gas and solid/liquid interfaces, accounting for the adsorption of molecules on surfaces is key for the properties and reactivity of the device. In this realm, density functional theory (DFT) has become the workhorse, due to its general applicability without system specific parameters and to a favorable compromise between accuracy and computational cost. However, DFT is still too expensive to address the large-scale computational screening of adsorption processes via general-purpose global optimization algorithms like minima hopping,[12] Bayesian optimization,[13] Monte-Carlo[14] or (meta)Dynamics.[15,16] Instead of DFT, model Hamiltonians ranging from simple Lennard-Jones potentials to elaborate cluster-expansions are generally used in conjunction with these techniques.[17–19] In the absence of a global optimization algorithm to be used together with the reliability that DFT offers, the choice of structures for initializing the "local" geometry optimizations has traditionally relied on expert knowledge, i.e. human intuition. This is most problematic for flexible molecules that feature a significant configurational space. This configurational large space combined with the loss of translational and rotational invariance upon adsorption, and with the number of possible adsorption sites yields to too many conformations at the interface to be manually explored. Flexible molecules on surfaces are involved in a large range of applications from modelling lubricant additives,[20] biomass conversion over metal surfaces,[21–23] to the understanding of the oxide support stability and shaping in water in the presence of polyols[24,25] and ligand

arrangements on metal nanoparticles for medical applications,[26] or their shape-control.[27]

The growing concern for making science more reproducible, transparent and systematic[28,29] together with the increase in computational power motivate the development of automatic high-throughput frameworks and workflows. Their applications in the field of interface chemistry span from the calculation of adsorption energies for atoms/small molecules[30] to the force-field parametrization and screening of soft matter,[31] passing by the creation and manipulation of nanoparticle interfaces.[32] Still, we are not aware of a generally applicable screening framework to investigate the (dissociative) adsorption modes of flexible molecules on surfaces and nanoparticles.

Our DockOnSurf package addresses the challenge to efficiently and automatically explore the adsorption modes of flexible molecules on surfaces. DockOnSurf automatically generates structures using different conformers, surface adsorption sites, adsorbate anchoring points, relative orientations and probes the dissociation of protons.

To illustrate the features of DockOnSurf, we show examples of adsorption of small molecules on nanoparticles and various flexible organic molecules on hematite and γ-alumina. The interest in polyfunctional organic molecules on oxide surfaces is motivated by applications in lubrication,[20] where multi-functional polar head-groups are common. Concomitantly, hematite is relevant to photoelectrochemistry,[33] catalysis,[34–36] pollutants removal[37] and energy storage,[38] while γ-alumina is a typical support for metallic catalysts in heterogeneous catalysis and a model for the surface oxide of aluminum during metal working.[39] The adsorption on nanoparticles is, on the other side, inspired by reports pointing towards a strong relation between the adsorption energy and the coordination number of the adsorption site.[40] For this last topic, it is the capability of DockOnSurf to detect the surface normal at each adsorption site that enable systematic and automatic high-throughput computations to probe the accuracy of such structure-property relations.

2 The DockOnSurf package

DockOnSurf[41] is an open-source program written in Python 3. To promote testability, adaptability, extensibility and ease of debugging, DockOnSurf is built in a modular fashion. Furthermore, the algorithms that constitute the code have been implemented following the PEP 8[42] and PEP 257[43] guidelines for code readability and documentation. Moreover, in order to increase the code efficiency and the modularity, the routines composing the program make extensive use of the popular chemistry-related libraries ASE[44] and RDKit.[45] The source code (under MIT license) is available on `http://forge.cbp.ens-lyon.fr/git/dockonsurf`.

A full DockOnSurf execution is composed of three independent steps: (i) isolated, (ii) screening and (iii) refinement. For the sake of flexibility, the input file (see section S2 in the SI) allows to select only one or two of these steps for execution.

The execution of DockOnSurf is highly customizable via the input file. The reproducibility and traceability of the results are guaranteed by recording all actions taken, as well as possible warnings/errors, together with their corresponding timestamp in a log file. For instance, tracking the origin of the atomic configurations produced by DockOnSurf is enabled by keeping track of the parent structure during screening and refinement steps.

2.1 Isolated

The "isolated" step intends to explore the conformational space of the molecule alone, which will be adsorbed on the surface later on. A given number of random conformers are generated exploiting the ETKDG algorithm[46] as implemented in the RDKit library,[45] optionally pre-optimizing their geometry with MMFF94[47–52] or UFF[53] to correct chemically inconsistent structures. These random conformers are then clustered based on their RMSD matrix relying on the HDBSCAN[54] algorithm. For every cluster, an exemplar (i.e. the most representative conformer of the cluster) is chosen using the Affinity Propagation method[55] and its structure

is further optimized using the specified quantum chemistry code (CP2K or VASP).

2.2 Screening

Given a surface and at least one conformer of the molecule to adsorb, the screening step samples the adsorbate-surface configurational space by combining the previously accomplished conformational exploration with an orientational and translational sampling. This is in practice carried out by generating a number of structures resulting from the combination of different possibilities as explained below and schematically summarized in Fig. 1.

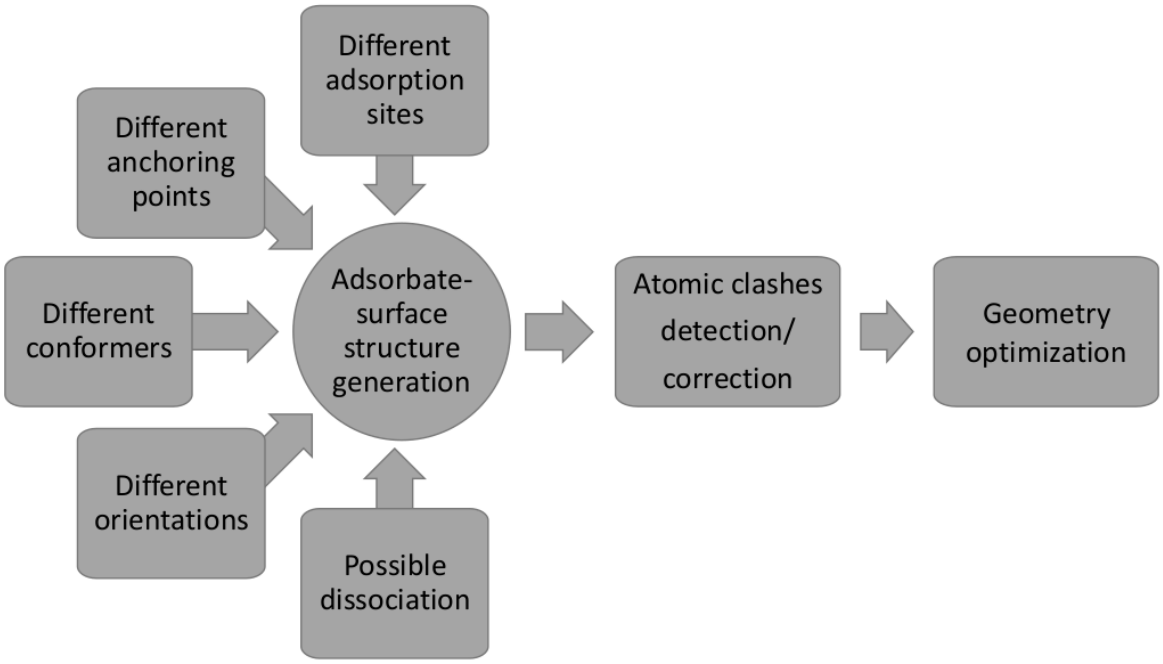

Figure 1: Schematic overview for the screening (main) step of DockOnSurf.

The division of the entire configurational space into smaller chemically meaningful degrees of freedom allows some degree of understanding of the system. More importantly, the division allows a fine tuning of the exploration according to the user's needs and the particularities of the studied system to increase overall efficiency and reduce "useless" computations.

Distinct conformers generated in the "isolated" step can be selected for screening according to the value of certain quantities (e.g. their moment of inertia or energy). Specially

interesting are the conformers with the largest moments of inertia, as they tend to expose their functional groups "outwards", so that they are available for interaction with the surface.

Each adsorption site is defined in the input file as a group of surface-atom indices for which the arithmetic mean of their atomic coordinates is computed. For example, an atop site is defined as only one index, a bridge as a group of two and a hollow as a group of three. This approach also allows out-of-standard site definition granting the user more possibilities specially useful for irregular surfaces.

For a multi-functional molecule, a list of several "anchoring points" should be provided. The adsorption screening (see Fig. 1) loops over this list by placing the anchoring points on top of the adsorption sites. The anchoring points are defined in analogy to the adsorption site. Thus, an anchoring point can be a single atom (e.g., N of NH_3), the midpoint of two atoms (e.g., of a C≡C bond or the center of a benzene ring), etc.

The different orientations of a conformer adsorbed on a surface site via a given anchoring point can be explored by defining both the set of angles to be scanned and the number of points per angle to be sampled. While Euler angles easily allow to explore the entire rotational space, they are unadapted to chemistry and, thus, often lead to unphysical adsorption modes. Therefore, we have also implemented "internal angles" (see Figure 2), which are intuitive to the chemist and allow a more relevant orientational sampling. The use of internal angles for adsorption on surfaces is also available as a standalone package.[56]

The sample size is specified in a per-angle basis, i.e., by defining that n rotations are to be performed for each angle. This results in a total of $n^3 - n(n-1)$ unique total rotations.[1]

Stepped and kinked surfaces or even nanoparticles present a very interesting and rich surface chemistry, but their surface normal has to be defined locally. Hence, we enable the use of DockOnSurf for adsorption sites that differ in orientation. For every adsorption

[1] All possible combinations of the three different angles are generated, leading to n^3 raw configurations. However, for both, the Euler (in its x-convention) and the internal set of angles, redundant configurations are generated with a dependence on the rotations per angle (n) equal to $n(n-1)$: In the x-convention of Euler angles, $(\alpha, 0, 0) \equiv (0, 0, \alpha)$. In the internal angles, when the bond angle is flat, all the bond-dihedral angle rotations (center panel in Figure 2) are ineffective. These redundant configurations generated both in the Euler and Internal angles are algorithmically filtered out.

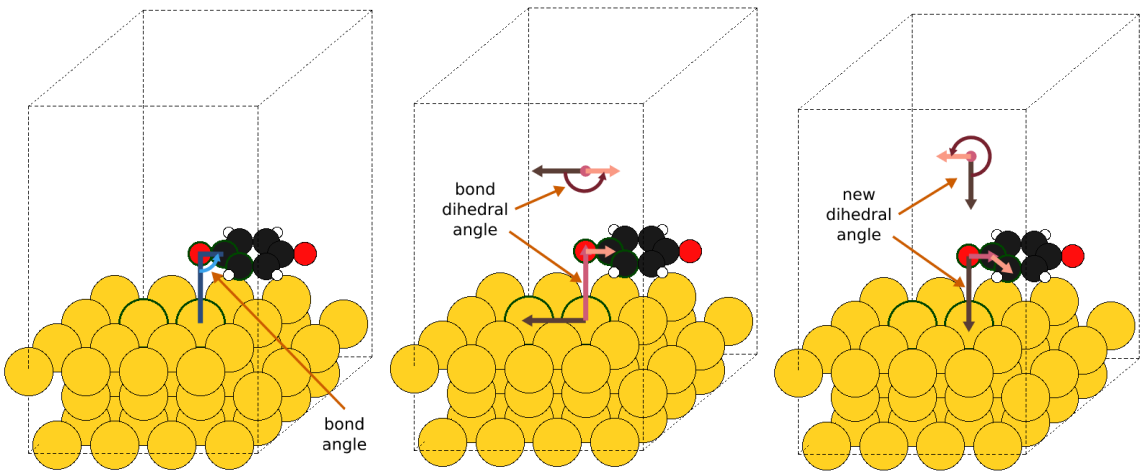

Figure 2: The internal set of angles

site, a normal vector to the surface is either provided by the user, or, more conveniently, automatically determined using the ASANN[57] algorithm.

Dissociative adsorption is a very common phenomenon, particularly for molecules with acidic protons on basic or noble-metal surfaces. Therefore, based on a flexible user-definition of both acidic protons and basic surface atoms, DockOnSurf detects protons that could be dissociated and generates the corresponding atomic configurations by displacing the proton from the adsorbate to the target surface atom.

One of the bottlenecks of screening adsorption modes of molecules stems from the possibility of unphysical adsorption modes, where atoms collide and lead to unintended chemical reactions or numerical artifacts. The internal set of angles implicitly prevents many of the adsorbate-surface collisions. Nevertheless, colliding atoms cannot be excluded when adsorbing multi-functional molecules on surfaces that might be intrinsically rough or which contain pre-adsorbed species. To prevent these unphysical initial structures, DockOnSurf implements two methods to detect (and later try to correct) such atomic clashes. (i) A fast method, useful for flat surfaces, based on the definition of a height threshold; whenever a configuration is generated that leads to atoms below this threshold, action is taken. (ii) A slow but more flexible approach based on the detection of overlapping spheres, where the radius of the sphere corresponds to the covalent radii of the corresponding atom as implemented in ASE,

multiplied by a user-defined factor. In order to produce the largest number of physically meaningful structures possible, a small angular correction is applied to configurations where atomic clashes occur to try to avoid the collision. The maximum angular correction applied for every angle is half the size of its sampling range, thus avoiding duplicate structures to be generated. Furthermore, for increased flexibility, the user can opt-out the molecular anchoring point from the collision detection.

Depending on the computational resources available and the combination of parameters set in the input file, the number of configurations generated can easily lead to an overwhelming number of computations. The user may, however, specify a maximum number of generated structures. In this case the corresponding number of configurations are randomly selected for the DFT computations. Performing a dry run, where all the configurations are generated but no computations are performed, is also possible and useful in this and other cases.

Once all the structures have been generated, a (DFT based) geometry optimization for each adsorption configuration is performed.

2.3 Refinement

The last step of DockOnSurf is the refinement. In this phase, the results from the screening are read and all structures corresponding to relative energies below a specified energy with respect to the most stable structure, are recomputed at a higher level of theory or accuracy.

2.4 Supported software

The actual quantum chemical computations are performed using the user-provided CP2K or VASP input files. These computations can either be run locally or in a computing center by using a batch queuing system. For the moment, SLURM, LSF and SGE are supported. The user can also specify a maximum number of simultaneous jobs to be either on queue, running or the sum of both.

The magnetism of hematite (one of our test systems) plays a fundamental role in the correct description of the surface properties.[58,59] In order to properly deal with the anti-ferromagnetic coupling, the definition of two kinds of Fe atoms with different spin-orbital occupations is needed. In practice, such a definition comes, necessarily, with a different labeling of the two Fe atom types. Taking into account these out-of-the-periodic-table chemical symbols could cause problems in the underneath libraries that DockOnSurf relies on (i.e. ASE and RDKit). A specific option has, thus, been implemented allowing the definition of special kinds of atoms in the input file by relating them to, and taking the properties of, the traditional ones present in the periodic table.

3 Computational details

For the tests performed in Section 4, four different surfaces were used.

To demonstrate the screening on nanoparticles, a stable, non-crystalline Au_{38} nanoparticle was used. The generalized coordination numbers of their surface atoms ranged from 1.6-5, making it an ideal test-case, as already observed in our previous study on coordination numbers.[57]

For the identification of adsorption modes on corrugated surfaces, we chose a 3×3 supercell of the (0001) facet of hematite in its Fe-O_3-Fe- termination when hydrated by one dissociated water molecule per unit cell.[60] Its lowest four atomic layers were kept frozen in its bulk arrangement. The corresponding dehydrated surface was used to compare the efficiency of internal angles for adsorption compared to Euler angles. Finally to illustrate the capacity to deal with multiple adsorption sites and anchoring points, the (100) facet of γ-alumina was used,[20] which exposes three inequivalent Al atoms and three inequivalent O atoms.

All computations have been performed at the DFT level as implemented in CP2K, using the hybrid GPW basis set. For the plane waves basis set, an energy cutoff of 400 Ry was

chosen, while for Gaussians the MOLOPT double-ζ basis set was used. Integration was performed using a multigrid method composed of 5 grids and the relative energy cutoff among grids was set to 50 Ry. Finally the screening of overlapping Gaussians was controlled by setting an EPS_DEFAULT parameter of 10^{-9}.

In the screening stage, the target accuracy for the SCF convergence was set to 10^{-5} while the geometry optimization criterion for the maximum force component was set to 0.01 Hartree/bohr. In the Refinement stage instead, the target accuracy for the SCF convergence was set to 10^{-7} and the default geometry optimization criteria of CP2K were used. All the computations were driven by DockOnSurf. The relevant files for setting up each DockOnSurf execution are available in the SI. The CP2K input files are also provided as examples and to clarify the specific technical settings.

4 Results and discussion

4.1 Internal coordinates for adsorption to increase efficiency

When searching for the most stable structure of a molecule adsorbed over a surface, the sampling process becomes capital. An algorithm that explores the adsorbate-surface configurational space efficiently by naturally avoiding nonphysical atomic configurations results in a larger effective number of configurations to explore and is, thus, more likely to identify the most stable structure.

As mentioned in Section 2, DockOnSurf splits the adsorbate-surface configurational space by sampling separately the conformational space of the molecule and the orientational space upon adsorption. The latter is explored by treating the given conformer as a 3D rigid body and rotating it over a set of angles. Two sets of angles are implemented (i) the Euler angles, in its x-convention,[61] and the internal angles[56] shown in Figure 2.

The sampling efficiency of both sets of angles were compared by generating a number of adsorbate-surface structures with each set and counting how many physically meaningful

structures (i.e., avoiding clashes) each of the set of angles produced. In other words, in this section we performed a "dry run" of the screening step. The most stable conformer of triethanolamine in gas-phase was chosen as an exemplar of voluminous flexible molecule while the surface of choice was the (0001) facet of hematite.

One of the oxygen atoms of triethanolamine was chosen as the anchoring point of the molecule. This maximizes the moment of inertia when rotating over it and thus increases the probability of collision. The adsorption site was defined as one of the uppermost (equivalent) Fe atoms as it is the only undercoordinated atom on the hematite (0001) surface.

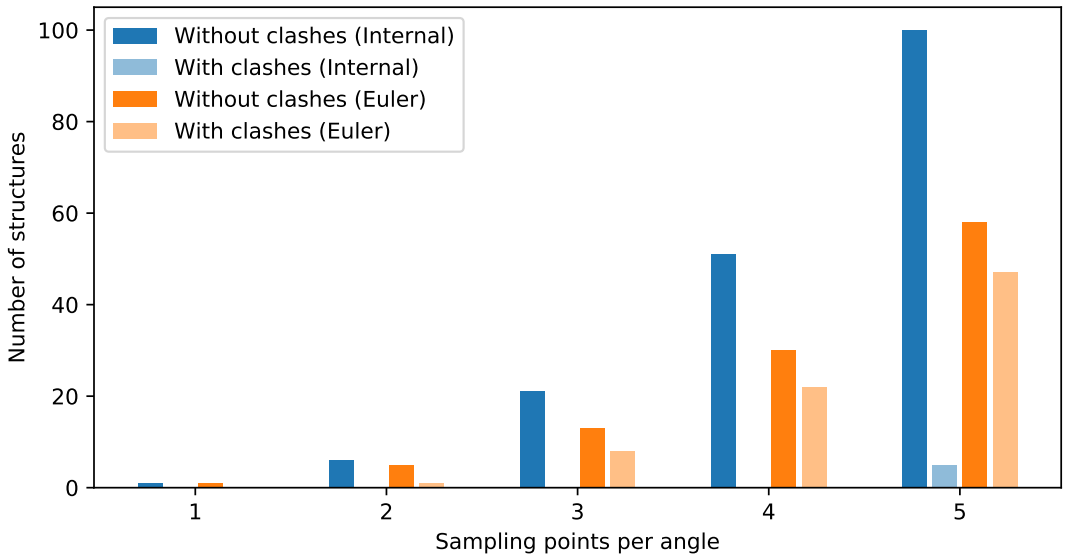

Figure 3: Summary of the efficiency of the two available sets of angles for the exploration of the orientational space.

Results reported on Figure 3 depict the number of configurations generated after the eventual "correction" that DockOnSurf carries out when clashes occur. If the structure generation is successful, the corresponding structure is counted as acceptable (blue). Otherwise, it is rejected due to clashes (pink). The superior performance of internal with respect to the Euler angles for producing meaningful structures is evident starting from three points per angle, where already 40% of the Euler-generated structures have to be discarded. For internal angles, clashes only occur when 5 points per angle are sampled.

In summary, Euler angles allow an unbiased exploration in case the user has no idea how his molecules adsorb on the surface. The numerical test confirm, however, that the use of internal angles allows to exploit chemical intuition (expert knowledge) for generating physically meaningful structures. Therefore, we recommend the use of the more efficient internal angles in general.

4.2 Adsorption on irregular shaped objects

DockOnSurf can automatically guess the local surface normal using the ASANN algorithm.[57] To illustrate this capability, we have defined all surface atoms of a low-energy non-crystaline Au_{38} nanoparticle as adsorption sites to be screened. The CO molecule was chosen as a typical and simple adsorbate. A single adsorbate orientation per site was specified, where the neighbor atom (O) to the anchoring point (C) points away from the surface. A superposition of all the generated configurations is presented on Figure 4. From this picture, where all CO molecules are placed as expected, it becomes clear that the surface normal has been successfully identified even for this highly irregular object.

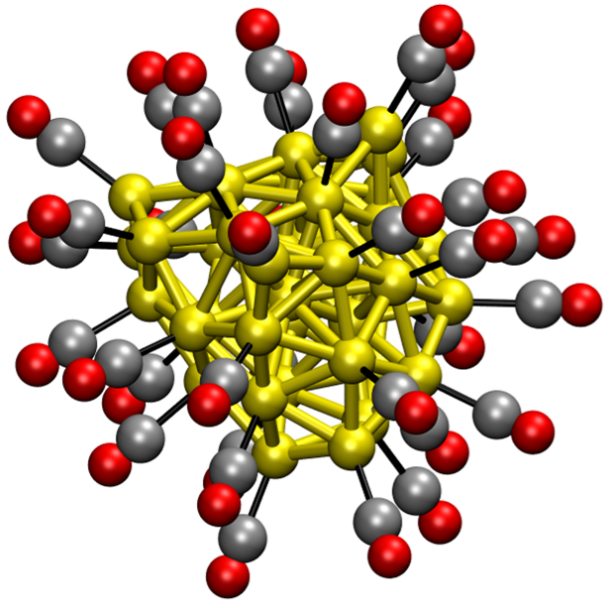

Figure 4: Superposition of all the CO-Au_{38} nanoparticle configurations generated by DockOnSurf using the ASANN method to identify the surface normal.

All these initial structures have been optimized at the DFT level and Figure 5 reports the adsorption energy of CO as a function of the generalized coordination number (GCN)[40] of the corresponding adsorption site. This small and rather soft gold nanoparticle shows a slight tendency for reconstructions upon adsorption (root mean square deviations up to 0.6 Å). This might be responsible for the significant spread in data. Nevertheless, there is a rough trend that CO binds somewhat stronger to the undercoordinated (edge) adsorption sites (low GCN) than to the gold atoms that are more embedded in the nanoparticle.

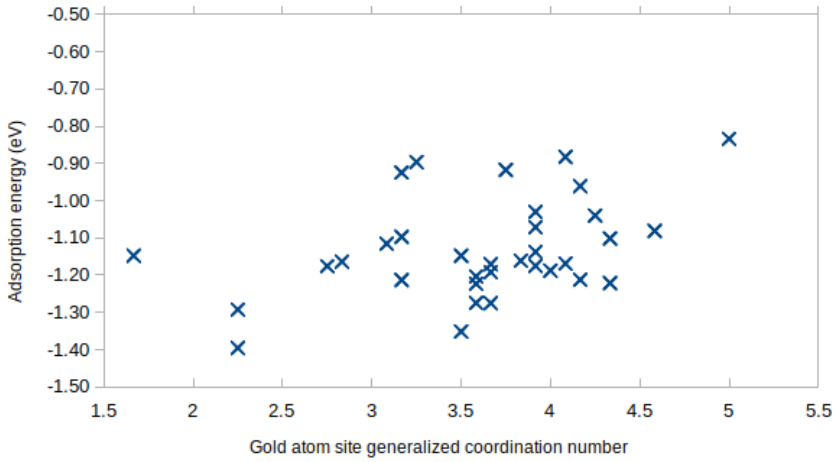

Figure 5: Adsorption Energy of CO adsorbed over the Au_{38} nanoparticle against the generalized coordination number of the gold atom used as adsorption site.

This example, where a large number of diverse adsorption configurations has been generated and optimized automatically illustrates the high-throughput capabilities of DockOnSurf. To test the hypothesis of adsorption energies that depend on the GCN, one could easily scan the surface sites of nanoparticles with different shapes and sizes with minimal human effort. This is, however, beyond the scope of this report.

4.3 Adsorption modes on pre-covered, corrugated surfaces

While flat surfaces are common in surface science and are straight forward to deal with, corrugated surfaces are also important applications for DockOnSurf. The corrugation of surfaces can either have its origin in the atomic structure of the pristine surface (e.g., a surface

of a zeolite[62]) or become corrugated when pre-adsorbed species are present. As a typical example of the second category, we here investigate adsorption on the Fe-O_3-Fe- termination of the (0001) facet of hematite when it is hydrated by one (dissociated) water molecule per unit cell. This surface state reflects the presence of a chemisorbed water molecule on the hematite surface under ambient humidity.[60] Our test case consists in finding the most stable adsorption mode for acetonitrile on a 3×3 supercell of the hydrated hematite (see right panel in Figure 6). Indeed, the dissociated water molecules, partially covering the surface, still allow a small molecules such as acetonitrile to reach the surface. Nevertheless, many of its orientations are impossible in the sense that they lead to clashes with the surface OH groups. Therefore, this system illustrates very well the capability of DockOnSurf to handle these constraints imposed by the surface roughness. At the same time, if the filtering of generated structures is too aggressive, relevant adsorption modes might be missed.

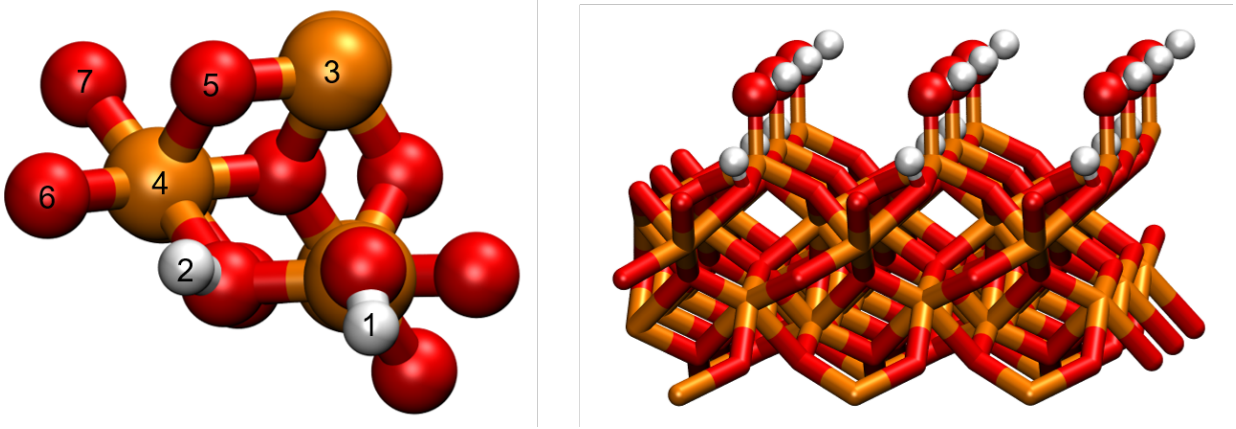

Figure 6: The (0001) facet of hematite in its Fe-O_3-Fe termination hydrated by a pre-adsorbed, dissociated water molecule leading to non-negligible surface corrugation. Left and right show the top view with labels of adsorption sites and a side view, respectively.

For a full exploration, seven distinct adsorption sites have been defined. Adsorption sites 1 and 2 in left panel of Figure 6 are located on top of the hydration layer. Direct bonding to the "original" surface is represented by adsorption sites 3 to 7. It is to be noted that adsorption site number 7 has not been defined as on top of the underneath O atom, which is placed 2 Å below the surface. Rather, it was specified as the baricenter of 3 of the uppermost

oxygen atoms. The anchoring point of the molecule (N atom) has been placed at a distance of 1.5 Å from the adsorption site. Furthermore, the anchoring point was opt-out of the collision evaluation to allow it to be placed in direct contact with the surface Fe atoms.

One hundred and forty-seven raw configurations were generated by combining the 7 different sites of the hydrated hematite surface with 21 orientations of the acetonitrile molecule obtained by rotating over the Euler angles which are adapted to this linear molecule. Due to the corrugation of the surface, the collision detector for spheres method was chosen to identify nonphysical configurations. Among the initial raw configurations, 66 were discarded due to collisions that could not be avoided. The screening stage was thus continued with the remaining 81 structures, for which a low-precision geometry optimization was performed. The geometry optimization was successful on all 81 structures, meaning that the clashes detection was effective and no unphysical structure was generated. Out of the 81 optimized geometries, 44 of were found within 0.5 eV of the most stable configuration, which was the specified energy threshold to select the configurations for the refinement stage.

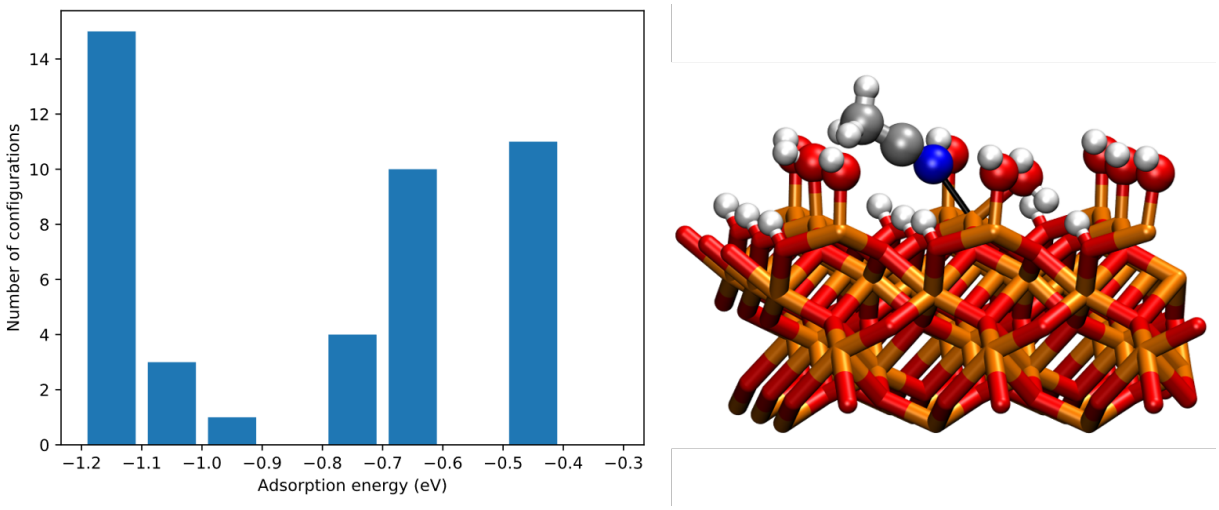

Figure 7: Histogram of structures obtained with DockOnSurf, after the refinement step, as function of their adsorption energies (left) and most stable structure (right) for acetonitrile adsorbed over a hydrated hematite surface.

The histogram of the structures obtained after the refinement step of DockOnSurf (see left panel of Figure 7) shows a trimodal distribution with respect to the adsorption energy.

A detailed visual inspection revealed three binding modes. In the first and most stable configuration, the N atom of acetonitrile directly binds to a surface Fe atom, displacing the corresponding OH. This OH recombines with a proton to form water (see right panel of Figure 7). In the second mode, with adsorption energies in the range -0.6 to -0.75 eV, the N atom interacts with the dissociated proton (i.e. the one directly bonded to the surface O, site 2 in Figure 6). The third mode encompasses structures loosely interacting with the OH atoms of the hydration layer in a variety of arrangements. This example also illustrates the ability of DockOnSurf to sample a variety of adsorption modes.

4.4 Screening multiple anchoring points on multiple adsorption sites

The presence of multiple possible anchoring points renders the identification of preferred adsorption modes cumbersome for multifunctional molecules. This is exacerbated when the molecule is adsorbed on surfaces which feature multiple potential adsorption sites. To assess the ability of DockOnSurf to deal with such systems, we here investigate the adsorption of triethanolamine on the γ-alumina (100) surface, a non-corrugated multi-site surface.

Triethanolamine presents four Lewis-basic anchoring points: one nitrogen atom and three hydroxyl groups. Once adsorbed on the surface and depending on the acidity of the adsorption site, the OH groups can dissociate. γ-alumina (100) exposes four Lewis-acidic aluminum atoms. They are truncated octahedral, coordinated to five oxygen atoms. Among these four surface aluminum atoms, Al_{Vb} and $Al_{Vb'}$ are equivalent, leaving three distinct possible adsorption sites on the surface: Al_{Va}, Al_{Vb} and Al_{Vc}. Our test consists in docking the most stable gas-phase conformer of triethanolamine on the three non-equivalent adsorption sites of a 4×3 supercell of the (100) γ-alumina surface via its four non-equivalent anchoring points to find the optimal structure.

To find the most stable gas-phase structure of the flexible triethanolamine, 2000 conformers were generated and 42 of them were optimized at the DFT level after the clustering step.

We selected the most stable conformer for screening its adsorption modes. 720 adsorption structures were generated by DockOnsurf, which combine the three possible adsorption sites on the surface with the four anchoring points of the molecule and tests the possibility of dissociating adsorbed OH groups. Due to the near spherical shape of the molecule, the rotations of the adsorbate over the surface were performed using the Euler set of angles. Among these initial adsorption structures, 509 calculations have converged during the screening part and 17 of them were then selected (< 0.5 eV above the minimum structure) to be further optimized at the refinement step.

The results of the refinement step lead to different adsorbate-surface configurations with adsorption energies ranging from -2.01 to -2.50 eV (Fig. 8a). The most stable structure obtained from the DockOnSurf execution, shown on (Fig. 8b), consists in a bidentate adsorption of two oxygen atoms of the trimethanolamine molecule on two different aluminum sites of the alumina surface. This structure is furthermore stabilized by two intramolecular hydrogen bonds and another one between the adsorbate and the surface. While carefully checked, the dissociation of hydroxyls were not found to be stabilizing.

This example demonstrates how DockOnSurf efficiently manages the sampling of a large configurational space, by keeping the focus on the chemically interesting/meaningful structures without the need for human intervention.

4.5 Conformational complexity: From gas-phase to the adsorbed state

The choice of the conformer, with which to carry out the adsorption, is decisive for the screening of adsorbate-surface structures. From an operational point of view, choosing the most stable gas-phase conformer is simple and seems intuitively well justified. However, in flexible, multi-functional molecules, the different functional groups often interact strongly with each other in the most stable gas-phase conformer (e.g., via intramolecular hydrogen bonds). These interactions lead to "folded" geometries, where the functional groups are

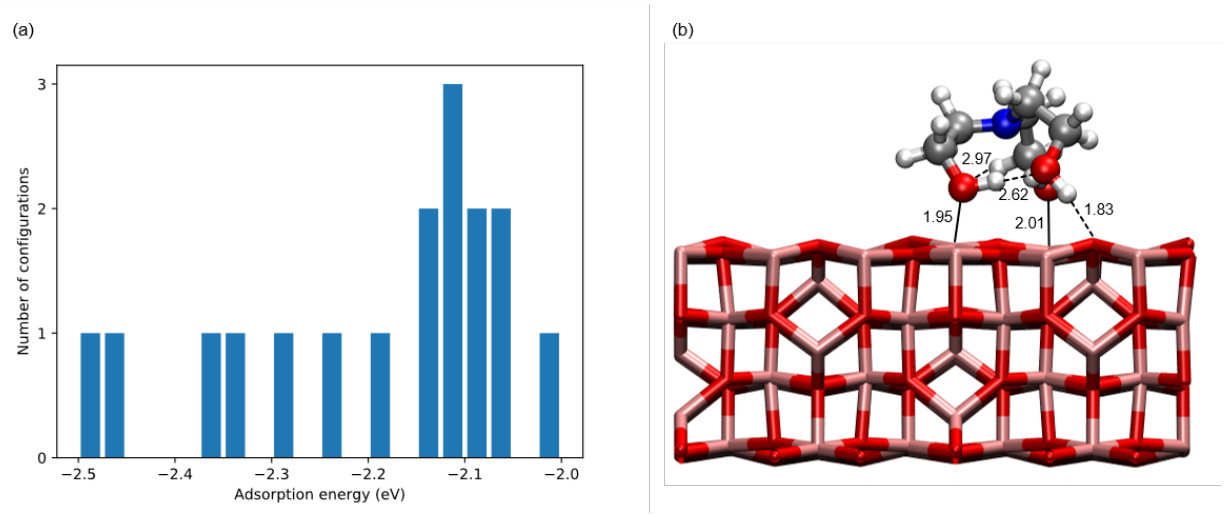

Figure 8: (a) Histogram of the distribution of adsorbate-surface structures obtained for triethanolamine adsorption on alumina after the refinement step as function of their adsorption energies. (b) Most stable structure found for the adsorption of triethanolamine on alumina. Relevant distances are indicated on the figure, for both bonds between the adsorbate and the surface (black lines) and hydrogen bonds (dashed lines).

unavailable for strong interactions with the adsorption sites. In contrast, higher-energy conformers where the functional groups are freely available, i.e., exposed to the gas-phase might be better candidates to identify the most stable adsorption mode. Since functional groups in flexible molecules are mainly formed by heteroatoms (i.e. atoms, heavier than the carbon and hydrogen atoms composing the backbone), a simple method for choosing wide-open conformers is to select the ones with large values of the trace of the moment of inertia tensor ($\mathrm{tr}(\mathbf{I})$).

To check the hypothesis that large moments of inertia are leading to stable adsorbate configurations, we docked sorbitol on bare hematite. To have a large conformational sampling, DockOnSurf was set to generate 500 raw conformers which were preoptimized using the MMFF forcefield before clustering them. The 11 conformers resulting from the clustering process were further optimized at the DFT level. The quantity for the selection of conformers for proceeding to the adsorbate screening phase was set to the total energy and the moment of inertia ($tr(\mathbf{I})$) of the optimized conformers. Three conformers per quantity were set to be selected, i.e. the ones having the largest, the smallest and the median values of the energy

and $tr(\mathbf{I})$, respectively. This resulted in a selection of five conformers, since the conformer with largest moment of inertia was also the one with highest total energy. For each of these five conformers, the orientational sampling was performed using the internal set of angles with two points per angle. Monodissociation of protons linked to O atoms in sorbitol was allowed, letting O atoms of the hematite surface act as H acceptors. This resulted in a total of 97 adsorbate-surface atomic configurations after removing eventual clashes. An energy threshold of 1.0 eV was set for the refinement of structures obtained at the screening stage, which lead to only 20 structures being fully optimized.

Since DockOnSurf keeps track of the parent structure all along the different steps, we can easily identify the origin of each of the 20 final adsorption modes (see Table S1 in the SI). The most stable adsorbate structure of sorbitol on hematite (see left panel of Figure 9) was obtained after the optimization of a structure built using the conformer with the largest moment of inertia. Indeed, it was the only structure able to bind to a total four different Fe surface atoms. The second most stable ($\Delta_{rel} = 0.69$ eV) adsorbed structure (see central panel of Figure 9), is derived from the conformer with the second largest $tr(\mathbf{I})$ among the 5 selected. The somewhat naive choice of using the most stable conformer in gas-phase resulted in an adsorbed conformer with a relative energy of 0.88 eV. It is, nevertheless, the third most stable adsorption mode.

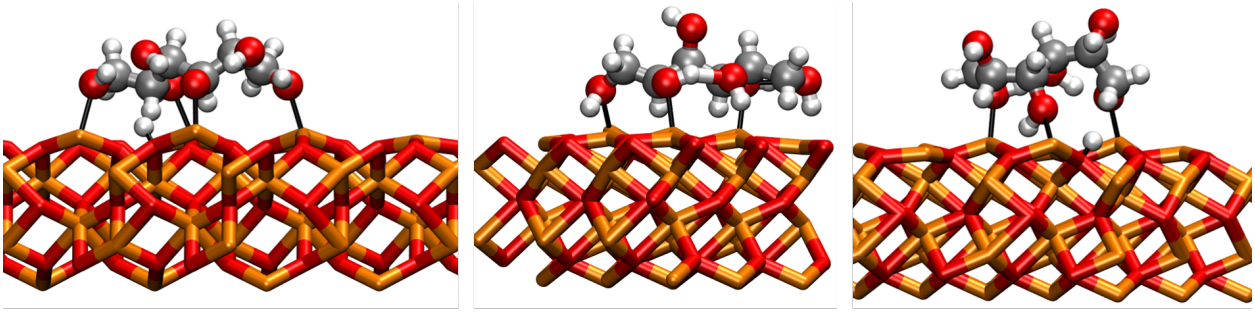

Figure 9: Top-3 most stable Sorbitol-Hematite structures, each originating from a different conformer. The adsorption energy is given with respect to the most stable gas-phase conformer.

Furthermore, among the ten most stable adsorption modes, four and three originated from the conformer with the largest and second largest $tr(\mathbf{I})$, respectively. In contrast, only one adsorption mode came from the most stable conformer in gas phase. This illustrates the benefit of using the moment of inertia at least as a secondary metric to choose the gas-phase conformers which should be adsorbed on the surface.

5 Conclusions

The reproducible, but customizable, screening strategy for finding low-energy adsorbate-surface structures implemented in DockOnSurf is especially designed for polyfunctional flexible molecules and ensures a maximum of chemically meaningful final geometries. The key element of the strategy is a combination of different chemically familiar concepts like surface adsorption sites, adsorbate's conformers, anchoring points and orientations as well as dissociation of protons. The generated geometries are optimized in a two-step procedure (screening followed by refinement), typically using a density-functional theory as implemented in CP2K or VASP. The traceability of the results is guaranteed through the detailed log file, recording all steps between generation of gas-phase conformers and final, refined adsorption geometries. DockOnSurf features three major innovations: (i) the use of internal angles rather than Euler angles for defining the relative orientation of the molecule on the surface, which prevents steric clashes with $> 95\%$ efficiency compared to $> 55\%$ efficiency for Euler angles. (ii) Convenient definitions of adsorption modes on irregular surfaces and nanoparticles is enabled through a locally defined surface normal, relying on the robust ASANN algorithm. (iii) The conformation of the adsorbate to be docked on the surface can not only by determined by its (relative) stability in the gas-phase, but also by the magnitude of the moments of inertia in order to select structurally diverse conformations. The implemented algorithms have been illustrated on five different test cases. In particular, we have performed an automatic, high-throughput screening of the CO adsorption energy as a function of the generalized co-

ordination number on a low-energy, non-crystalline, Au_{38} nanoparticle. Furthermore, the capability to deal with corrugated surfaces has been illustrated by the identification of three major adsorption modes of acetonitrile over hydrated hematite. Finally the importance of the choice of a suitable gas-phase conformer for finding low-energy adsorption modes has been demonstrated with the case of sorbitol ($C_6H_{14}O_6$) adsorption on hematite. In this case, using a conformation with a high moment of inertia, rather than the lowest gas-phase energy, has been key, as such conformations expose the functional groups and thus facilitate the interaction with the surface. In summary, we have demonstrated that the DockOnSurf package simplifies finding low-energy adsorbate-surface structures and is applicable to a very broad range of adsorbate/surface systems.

Acknowledgement

S. B. gratefully acknowledges Total MS and the ANRT for her PhD fellowship. The authors thank the SYSPROD project and AXELERA Pôle de Compétitivité for financial support (PSMN Data Center). This work was also granted access to the HPC resources of CINES and IDRIS under the allocation 0800609 made by GENCI. We thank Paul Clabaut for helpful discussions.

Supporting Information Available

Archive with DockOnSurf input files (input_files.zip) and additional tables.

References

(1) Talele, T. T.; Khedkar, S. A.; Rigby, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Current topics in medicinal chemistry* **2010**, *10*, 127–141.

(2) Van Drie, J. H. Computer-aided drug design: the next 20 years. *Journal of computer-aided molecular design* **2007**, *21*, 591–601.

(3) Vijayakrishnan, R. Structure-based drug design and modern medicine. *Journal of Postgraduate Medicine* **2009**, *55*, 301–304.

(4) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry* **2012**, *4*, 83.

(5) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters* **2011**, *2*, 2241–2251.

(6) Singh, A. K.; Mathew, K.; Zhuang, H. L.; Hennig, R. G. Computational screening of 2D materials for photocatalysis. *The journal of physical chemistry letters* **2015**, *6*, 1087–1098.

(7) Armiento, R.; Kozinsky, B.; Fornari, M.; Ceder, G. Screening for high-performance piezoelectrics using high-throughput density functional theory. *Physical Review B* **2011**, *84*, 014103.

(8) Yao, Z.; Hegde, V. I.; Aspuru-Guzik, A.; Wolverton, C. Discovery of Calcium-Metal Alloy Anodes for Reversible Ca-Ion Batteries. *Advanced Energy Materials* **2019**, *9*, 1802994, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aenm.201802994.

(9) Cheng, L.; Assary, R. S.; Qu, X.; Jain, A.; Ong, S. P.; Rajput, N. N.; Persson, K.; Curtiss, L. A. Accelerating electrolyte discovery for energy storage with high-throughput screening. *The journal of physical chemistry letters* **2015**, *6*, 283–291.

(10) Greeley, J.; Nørskov, J. K. Combinatorial density functional theory-based screening of surface alloys for the oxygen reduction reaction. *The Journal of Physical Chemistry C* **2009**, *113*, 4932–4939.

(11) Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Nørskov, J. K. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *Journal of Catalysis* **2006**, *239*, 501–506.

(12) Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of chemical physics* **2004**, *120*, 9911–9917.

(13) Carr, S.; Garnett, R.; Lo, C. BASC: applying Bayesian optimization to the search for global minima on potential energy surfaces. International Conference on Machine Learning. 2016; pp 898–907.

(14) Wales, D. J.; Scheraga, H. A. Global Optimization of Clusters, Crystals, and Biomolecules. *Science* **1999**, *285*, 1368–1372, Publisher: American Association for the Advancement of Science Section: Review.

(15) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141 – 151.

(16) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(17) Chen, W.; Schmidt, D.; Schneider, W. F.; Wolverton, C. Ordering and Oxygen Adsorption in Au–Pt/Pt(111) Surface Alloys. *J. Phys. Chem. C* **2011**, *115*, 17915–17924.

(18) Zhu, B.; Creuze, J.; Mottet, C.; Legrand, B.; Guesmi, H. CO Adsorption-Induced Surface Segregation and Formation of Pd Chains on AuPd(100) Alloy: Density Functional

Theory Based Ising Model and Monte Carlo Simulations. *J. Phys. Chem. C* **2016**, *120*, 350–359.

(19) Vignola, E.; Steinmann, S. N.; Le Mapihan, K.; Vandegehuchte, B. D.; Curulla, D.; Sautet, P. Acetylene Adsorption on Pd–Ag Alloys: Evidence for Limited Island Formation and Strong Reverse Segregation from Monte Carlo Simulations. *J. Phys. Chem. C* **2018**, *122*, 15456–15463.

(20) Blanck, S.; Loehlé, S.; Steinmann, S. N.; Michel, C. Adhesion of lubricant on aluminium through adsorption of additive head-groups on γ-alumina: A DFT study. *Tribology International* **2020**, *145*, 106140.

(21) Steinmann, S. N.; Sautet, P.; Michel, C. Solvation free energies for periodic surfaces: comparison of implicit and explicit solvation models. *Physical Chemistry Chemical Physics* **2016**, *18*, 31850–31861.

(22) Gu, G. H.; Schweitzer, B.; Michel, C.; Steinmann, S. N.; Sautet, P.; Vlachos, D. G. Group Additivity for Aqueous Phase Thermochemical Properties of Alcohols on Pt(111). *J Phys Chem C* **2017**, *121*, 21510–21519.

(23) Ould Hamou, C. A.; Réocreux, R.; Sautet, P.; Michel*, C.; Giorgi*, J. B. Adsorption and Decomposition of a Lignin β-O-4 Linkage Model, 2-Phenoxyethanol, on Pt (111): Combination of Experiments and First-Principles Calculations. *The Journal of Physical Chemistry C* **2017**, *121*, 9889–9900, Times Cited: 17.

(24) Chiche, D.; Chizallet, C.; Durupthy, O.; Chanéac, C.; Revel, R.; Raybaud, P.; Jolivet, J.-P. Growth of boehmite particles in the presence of xylitol: morphology oriented by the nest effect of hydrogen bonding. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11310–11323, Publisher: The Royal Society of Chemistry.

(25) Réocreux, R.; Girel, É.; Clabaut, P.; Tuel, A.; Besson, M.; Chaumonnot, A.; Cabiac, A.;

Sautet, P.; Michel, C. Reactivity of shape-controlled crystals and metadynamics simulations locate the weak spots of alumina in water. *Nat Commun* **2019**, *10*, 3139.

(26) Chan, C.-H.; Poignant, F.; Beuve, M.; Dumont, E.; Loffreda, D. Effect of the Ligand Binding Strength on the Morphology of Functionalized Gold Nanoparticles. *J. Phys. Chem. Lett.* **2020**, *11*, 2717–2723, Publisher: American Chemical Society.

(27) Aït Atmane, K.; Michel, C.; Piquemal, J.-Y.; Sautet, P.; Beaunier, P.; Giraud, M.; Sicard, M.; Nowak, S.; Losno, R.; Viau, G. Control of the anisotropic shape of cobalt nanorods in the liquid phase: from experiment to theory... and back. *Nanoscale* **2014**, *6*, 2682–2692, Publisher: Royal Society of Chemistry.

(28) Baker, M. 1,500 scientists lift the lid on reproducibility. **2016**,

(29) National Academies of Sciences Engineering and Medicine and others, *Reproducibility and replicability in science*; National Academies Press, 2019.

(30) Montoya, J. H.; Persson, K. A. A high-throughput framework for determining adsorption energies on solid surfaces. *npj Computational Materials* **2017**, *3*, 1–4.

(31) Summers, A. Z.; Gilmer, J. B.; Iacovella, C. R.; Cummings, P. T.; McCabe, C. MoSDeF, a Python Framework Enabling Large-Scale Computational Screening of Soft Matter: Application to Chemistry-Property Relationships in Lubricating Monolayer Films. *Journal of Chemical Theory and Computation* **2020**, *16*, 1779–1793.

(32) Mathew, K.; Singh, A. K.; Gabriel, J. J.; Choudhary, K.; Sinnott, S. B.; Davydov, A. V.; Tavazza, F.; Hennig, R. G. MPInterfaces: A Materials Project based Python tool for high-throughput computational screening of interfacial systems. *Computational Materials Science* **2016**, *122*, 183 – 190.

(33) Podsiadły-Paszkowska, A.; Tranca, I.; Szyja, B. M. Tuning the Hematite (110) Surface

Properties To Enhance Its Efficiency in Photoelectrochemistry. *The Journal of Physical Chemistry C* **2019**, *123*, 5401–5410.

(34) Shekhah, O.; Ranke, W.; Schlögl, R. Styrene synthesis: in situ characterization and reactivity studies of unpromoted and potassium-promoted iron oxide model catalysts. *Journal of Catalysis* **2004**, *225*, 56 – 68.

(35) He, Y.; Guo, F.; Yang, K. R.; Heinlein, J. A.; Bamonte, S. M.; Fee, J. J.; Hu, S.; Suib, S. L.; Haller, G. L.; Batista, V. S.; Pfefferle, L. D. In Situ Identification of Reaction Intermediates and Mechanistic Understandings of Methane Oxidation over Hematite: A Combined Experimental and Theoretical Study. *Journal of the American Chemical Society* **2020**, *142*, 17119–17130.

(36) Li, P.; Miser, D. E.; Rabiei, S.; Yadav, R. T.; Hajaligol, M. R. The removal of carbon monoxide by iron oxide nanoparticles. *Applied Catalysis B: Environmental* **2003**, *43*, 151 – 162.

(37) Weng, Y.; Vekeman, J.; Zhang, H.; Chou, L.; Elskens, M.; Tielens, F. Unravelling phosphate adsorption on hydrous ferric oxide surfaces at the molecular level. *Chemosphere* **2020**, *261*, 127776.

(38) Parkinson, G. S. Iron oxide surfaces. *Surface Science Reports* **2016**, *71*, 272 – 365.

(39) Jeurgens, L. P. H.; Sloof, W. G.; Tichelaar, F. D.; Mittemeijer, E. J. Structure and morphology of aluminium-oxide films formed by thermal oxidation of aluminium. *Thin Solid Films* **2002**, *418*, 89–101.

(40) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angewandte Chemie International Edition* **2014**, *53*, 8316–8319.

(41) Martí, C. DockOnSurf. `http://forge.cbp.ens-lyon.fr/redmine/projects/dockonsurf`, 2020.

(42) van Rossum, G.; Warsaw, B.; Coghlan, N. *Style Guide for Python Code*; PEP 8, 2001.

(43) Goodger, D.; van Rossum, G. *Docstring Conventions*; PEP 257, 2001.

(44) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C., et al. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(45) RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

(46) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574, PMID: 26575315.

(47) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **1996**, *17*, 490–519.

(48) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry* **1996**, *17*, 520–552.

(49) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *Journal of computational chemistry* **1996**, *17*, 553–586.

(50) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. Conformational energies and geometries for MMFF94. *Journal of computational chemistry* **1996**, *17*, 587–615.

(51) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **1996**, *17*, 616–641.

(52) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(53) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(54) McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2017**, *2*.

(55) Frey, B. J.; Dueck, D. Clustering by passing messages between data points. *science* **2007**, *315*, 972–976.

(56) ChemCat. `https://github.com/RubenStaub/ChemCat`, Accessed: 2020-12-01.

(57) Staub, R.; Steinmann, S. N. Parameter-free coordination numbers for solutions and interfaces. *The Journal of Chemical Physics* **2020**, *152*, 024124.

(58) Huang, X.; Ramadugu, S. K.; Mason, S. E. Surface-Specific DFT + U Approach Applied to α-Fe2O3(0001). *The Journal of Physical Chemistry C* **2016**, *120*, 4919–4930.

(59) Wang, R. B.; Hellman, A. Surface terminations of hematite (α-Fe2O3) exposed to oxygen, hydrogen, or water: dependence on the density functional theory methodology. *Journal of Physics: Condensed Matter* **2018**, *30*, 275002.

(60) Souvi, S. M.; Badawi, M.; Paul, J.-F.; Cristol, S.; Cantrel, L. A DFT study of the hematite surface state in the presence of H2, H2O and O2. *Surface Science* **2013**, *610*, 7 – 15.

(61) Herbert Goldstein, J. L. S., Charles P. Poole *Classical Mechanics (3rd Edition)*, 3rd ed.; Addison Wesley, 2001.

(62) Treps, L.; Gomez, A.; Bruin, T. d.; Chizallet, C. Environment, Stability and Acidity of External Surface Sites of Silicalite-1 and ZSM-5 Micro and Nano Slabs, Sheets, and Crystals. *ACS Catal.* **2020**, *10*, 3297–3312, Publisher: American Chemical Society.

Graphical TOC Entry