

## Computational Physics



# Screeener and enumerator with force-field optimization (SEFFO): Algorithm for searching adsorption sites and configurations on 2D materials <sup>☆</sup>

Leran Lu <sup>\*</sup>, Wei Cao, Romain Botella

Faculty of Science, Nano and Molecular Systems Research Unit, University of Oulu, Fin-90014 Oulu, Finland

## ARTICLE INFO

Dataset link: <https://doi.org/10.5281/zenodo.10804634>

## Keywords:

Adsorption  
Density functional theory  
2D materials  
Selection  
On-the-fly optimization

## ABSTRACT

With the increasing attention to 2D materials for photocatalytic applications, as well as to data science, there is a need for high-throughput computation of adsorption states for experimentally or theoretically discovered structures in order to study (photo-) catalytic mechanism. Despite numerous progresses in high-throughput methods for adsorption study, a general search algorithm is lacking. In this work, SEFFO (Screeener and Enumerator with Force-Field Optimization) algorithm is developed for the automation of adsorption study on 2D material surface. Graph theory is utilized to create the descriptors of the adsorption configurations, which are later input for geometry construction by numerical optimization. The configuration screening process is combining the use of graphs with structural similarity comparison of configurations density functional theory (DFT) produced configurations. The algorithm is validated through four case studies, involving water and carbon dioxide molecules as adsorbates, molybdenum sulfide and carbon nitride as substrate counterparts. The results are consistent with literature while proposing alternative configurations. Additionally, SEFFO can show the evolution between configurations during the process. This method enables the high throughput study of adsorption behavior on 2D materials, and paves the way for future surface studies involving other substrate/adsorbates pairs.

## 1. Introduction

Adsorption is an important process in many surface chemical processes, including the photocatalysis, which is a key technology to mitigate energy and environmental crisis and has been applied in water splitting, CO<sub>2</sub> reduction, and pollutant degradation [1,2]. Efforts have been devoted to finding the candidate materials that are able to improve the photocatalytic performance of photocatalytic reaction. 2D materials stand out, due to their excellent characteristics such as large surface area and the quantum confinement of electrons [3,4]. Their unique properties enable them to show comparable or superior photocatalytic performance to their bulk counterparts [5,6]. Considering the photocatalytic steps where it can hardly be circumvented, the adsorption studies on 2D materials are also important. Furthermore, it is the first step of any photocatalytic reaction. To date, computational study of adsorption is well-known. However, depending on the substrate/adsorbate pair considered, the study of adsorption may vary in its complexity. While it is feasible to go through all the necessary steps for the adsorption study manually with “simple” structures, the computational cost can be exces-

sive when there are too many possible configurations. Therefore, there is a need for a general method to produce reasonable configurations that enables the automation of this process.

In order to develop such an algorithm, it is essential to have a good understanding of the adsorption process. Adsorption comprises two types: physisorption that involves “weak” physical forces such as van der Waals interactions and hydrogen bonds (interaction energy between ca. 0.1 and ca. 1.3 eV) and chemisorption that involves strong electron density sharing between adsorbate and adsorbent (chemical bonds, interaction energy between ca. 11 and ca. 39 eV) [7]. Physisorption commonly happens when an adsorbate approaches a solid surface [8–11]. After being close enough between the adsorbate and adsorbent, a chemisorption may then happen, depending on the degree of unsaturation of the surface and the properties of the adsorbate. In computational study of adsorption, multiple codes are available in order to find proper adsorption configuration. They offer diverse methods to construct the initial geometry of the configuration based on the surface environment. AdsorbateSiteFinder class in pymatgen [12] is the first one that offers the function to detect adsorption sites by using Delaunay triangulation

<sup>☆</sup> The review of this paper was arranged by Prof. W. Jong.

<sup>\*</sup> Corresponding author.

E-mail address: [leran.lu@oulu.fi](mailto:leran.lu@oulu.fi) (L. Lu).

and construct the initial geometry simply by putting the adsorbate above the adsorption sites. To take into consideration the connectivity of the adsorbate with the surface, Catkit [13] developed a method for bulk materials to create configurations that are determined by the innovative adsorption vector and adsorption edges that works with monodentate and bidentate adsorption. The Atomic Surface Adsorbate Structure Provider (ASAP) [14] is another option that enables studies with porous structures. Despite the progresses, the simple construction or algebraic calculation restricted to limited types of adsorptions requires more efforts to approach the adsorption complexity in reality and sometimes produce unrealistic initial structures. Apart from that, the screening process also impacts on the efficiency of theoretical research [15]. A good screening process can save computational resources on calculations, leading to the same results if not a better one. This is especially true for first-principles computation. It has become the workhorse of computational material science due to its general applicability with less empirical interference on the system-specific parameters than others, as well as good compromise between the computational cost and the precision of result. However, it is still time-consuming, and its computational time increases with the size of the system. To date, adsorption on well-defined surfaces of the known 2D materials have been intensively studied, accumulating valuable results to verify incoming and newly developed algorithm. Previous endeavors rely on semi-empirical searching paths where tens or hundreds of configurations are relaxed from the beginning to the very end. This procedure is not resource- or time-efficient for first-principles methods, the use of screening process can alleviate this situation. One example is Surfgraph [16], which screens the adsorption configurations using chemical environment of the adsorbate. Another one is DockOnSurf [17]. It implements an efficient algorithm based on the geometry of the configuration that works well with big molecules or clusters. It is worth noting that Surfgraph uses a construction strategy that is similar to AdsorbateSiteFinder in pymatgen, while DockOnSurf lacks the function to enumerate adsorption sites.

In this work, an algorithm was developed to not only automate the adsorption study on surface of 2D materials, but also to provide an exhaustive list of symmetrically and energetically unique adsorption configurations. This algorithm avoids any bias in the choice of configurations to study. It relies on graph theory for the configuration descriptor and numerical optimization for construction, building on existing works regarding adsorption site search while avoiding irrelevant sites. As for the screening process, it has been built on top of the graph data structure with connectivity created by evaluating similarity. The later is based on the computed values regarding the symmetry of adsorption sites. Besides, the evolution between configurations can be visualized. The validity of the whole process has been tested with water and carbon dioxide molecules on molybdenum sulfide and carbon nitride, in total of four cases. This algorithm enables the high-throughput study of adsorption on 2D materials and is easy to be extended to deal with more complex adsorbent/adsorbate systems.

## 2. Results and discussions

### 2.1. The SEFFO algorithm

Fig. 1a shows the workflow of the designed algorithm. From top to bottom the number of configurations progressively decreases through four steps.

In the first step, the structures of the adsorbate and the substrate are collected, along with specifications about the anchored atoms. From the slab that contains the substrate, all the adsorption sites of top, bridge and hollow type as illustrated in Fig. 1b can be found by using Delauney triangulation [18]. These configurations assume that the adsorbate is first chemisorbed to the surface, which is a valid starting point for any adsorption study. While this process already produces satisfying results, a more refined filter can be applied to eliminate some unwanted sites,

especially the hollow sites. These sites can be located at the centers of obtuse triangles close to a nearby top site, but also at the of triangles close to the center of polygons they are in, at their centers or at their edges' center. Any adsorbate placed in this region will experience strong repulsion that will most likely end up in the adsorbate being "optimized" away from the surface, or in the final configuration having high energy (local minimum). The AdsorbateSiteFinder class implemented in Pymatgen ignores the hollow sites in obtuse triangles, but some unwanted sites are still kept in the final result. A filter is further developed in this work to remove these unwanted sites. It is based on the creation and modification of the basic graph data structure [19]. A graph  $D = (V_D, E_D)$  named site graph that stores the connectivity of all adsorption sites is created.  $V_D$  is its set of vertices, in which each vertex corresponds to an adsorption site, and  $E_D$  is its set of edges. All notations employed in the current study and their meanings are illustrated at the end of the paper (section 5). By computing the distances between each pair, edges in  $E_D$  are created, in condition that each site (excluding top sites) is only connected to its nearest top site

$$E_D = \bigcup_{u \in V_D - T} \left\{ \{u, v\} \mid \forall v \in T, \left| d_{uv} - \min_{v' \in T} d_{uv'} \right| \leq \epsilon_D \right\} \quad (1)$$

$T$  is the set of top sites, and  $d_{uv}$  ( $d_{uv'}$ ) are the distances between the atomic positions of sites  $u$  and  $v$  ( $v'$ ) respectively.  $\epsilon_D$  is the error allowed when determining the nearest top site. It is illustrated in Fig. S1. The filter is executed on  $V_D$ , yielding a filtered  $V_D$

$$V_D \rightarrow \{v \mid \forall v \in V_D, \deg v > 1\} \quad (2)$$

and the corresponding edges in  $E_D$  that connects to sites that are not in the filtered  $V_D$  will also be removed.

The second step aims to find the descriptors of the possible configurations based on the choice of anchors (adatom in Fig. 1b) and sites. To anchor atoms to the surface as close as possible while not producing contacts that leads to high energy but still in the chemisorption range, the anchors are "shifted" from their initially computed positions. These site shifts are computed using the site positions as well as the covalent radii of the anchored atoms and the surface atoms that connect to the corresponding sites. (see section "Computing the shifted sites" in Supporting Information). The three basic cases about computing the shifts are illustrated in the lower part of Fig. 1b. More complex ones can be achieved by enumerating on the surface atoms that connect the sites of interest until a position and orientation is found where the anchored atom has no overlap with any of them. Afterwards, the shifted sites are used for enumerating the configurations. In the case where there is more than one anchor (multidentate adsorption), the same number of sites will be chosen, and each of them will be paired with an anchor. Since in the current stage only small molecules (ambient gas molecules) are considered, the adsorbates are assumed to be rigid, i.e. the bond lengths and angles are fixed. Therefore, if the distances between anchors and the distances between corresponding sites differ respectively within a specified precision  $\epsilon_E$ , then this combination will be regarded as valid.

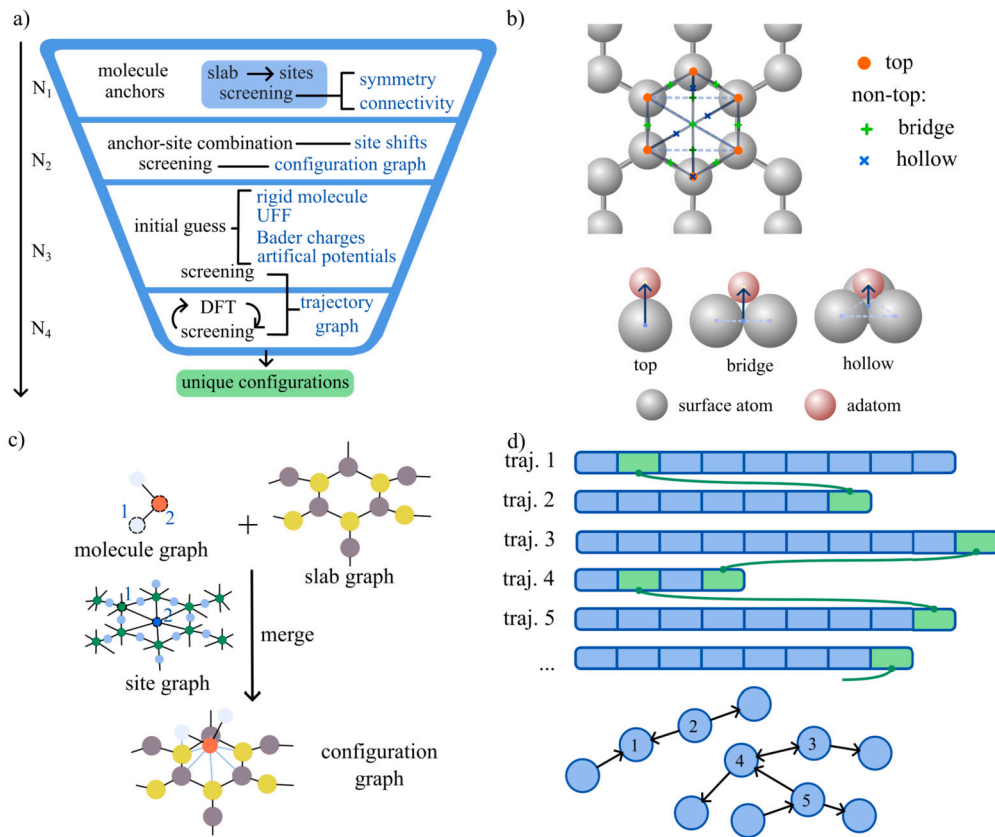
A screening process based on graph theory is then executed. As illustrated in Fig. 1c, the graphs of molecule and the substrate are merged to create a configuration graph associated with the adsorption configuration. It is achieved by connecting the anchors in  $V_M$  with the surface atoms in  $V_S$  that are connected to the corresponding sites in  $V_D$  by using the connectivity in  $E_D$ . Hence, the configuration graph is below

$$C = (V_C, E_C) \quad (3)$$

$$V_C = V_S \cup V_M \quad (4)$$

$$E_C = E_S \cup E_M \bigcup_{i=1}^n \{ \{a_i, v\} \mid \forall v \in \text{csa}(p_i), a_i \in A, p_i \in P \} \quad (5)$$

where  $C = (V_C, E_C)$  is the configuration graph.  $S = (V_S, E_S)$  and  $M = (V_M, E_M)$  are the slab graph and molecular graph. They are respectively



**Fig. 1.** Illustration of SEFFO. a) The overview of the algorithm, where  $N_1$ ,  $N_2$ ,  $N_3$ ,  $N_4$  are the number of possible configurations. b) Illustration of different types of sites and the shifts represented as arrows in solid line. A hollow site is defined as the center of more than two neighboring surface atoms. c) Diagram showing the creation of configuration graph, where each vertex in molecule graph, slab graph, and configuration graph represents an atom, and in site graph green ones are surface atoms/top sites and blue ones are bridge or hollow sites. d) Schematic representation of how trajectories are connected (above) and its graph representation (below). The arrows indicate the evolution direction from one trajectory to another. Double arrows mean the two trajectories connected evolve into the same pattern at the end. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

the graph representation of the substrate and the molecule.  $csa(v)$  is the function that gets all the surface atoms that are connected to site  $v$ . It ensures that in the later procedure the anchored atoms are correctly placed on the surface, with a possibility of making bonds with the neighboring atoms on surface in case of chemisorption. The construction of the  $S$ ,  $M$ , and the definition of  $csa(v)$  are elaborated in section “Construction of graphs” in Supporting Information.  $A = (a_1, a_2, \dots, a_n)$  and  $P = (p_1, p_2, \dots, p_n)$  are the sequences of anchors and corresponding sites respectively. Two configuration graphs  $C_1 = (V_{C1}, E_{C1})$  and  $C_2 = (V_{C2}, E_{C2})$  are isomorphic if there is a mapping function  $f(v)$  so that

1. On vertex level, nodes  $f(v_1) = v_2$  satisfy that  $v_1$  and  $v_2$  are of same element and same layer index according to the slab graph created, i.e.  $elem(v_1) = elem(v_2)$  and  $lyr(v_1) = lyr(v_2)$ , where  $elem(v)$  returns the chemical element of the atom  $v$ , and  $lyr(v)$  returns the index of layer where the atom  $v$  is;
2. On edge level, edges  $\{u_1, v_1\}, \{u_2, v_2\}$  that have relations  $f(u_1) = u_2, f(v_1) = v_2$  satisfies that
  - (a) If  $\{u_1, v_1\}$  and  $\{u_2, v_2\}$  are both in the union  $E_S \cup E_M$ , i.e.  $\{u_1, v_1\} \in E_S \cup E_M \wedge \{u_2, v_2\} \in E_S \cup E_M$ , then  $|d_{u_1 v_1} - d_{u_2 v_2}| \leq \epsilon_C$  where  $\epsilon_C$  is the precision,
  - (b) Otherwise, neither of  $\{u_1, v_1\}$  and  $\{u_2, v_2\}$  is in the union  $E_S \cup E_M$ , i.e.  $\{u_1, v_1\} \notin E_S \cup E_M \wedge \{u_2, v_2\} \notin E_S \cup E_M$ ,

where  $u_1, v_1 \in V_{C1}, u_2, v_2 \in V_{C2}, \{u_1, v_1\} \in E_{C1}, \{u_2, v_2\} \in E_{C2}$ . Isomorphic configuration graphs mean the descriptors lead to the same geom-

etry when constructing the adsorption configurations, therefore, only one is kept to avoid redundancy.

In step three, the descriptors from step two are used for constructing the geometry of the structure. It minimizes the system energy  $E$  through the following step

$$\begin{aligned} \min_{\text{pos}(v), v \in V_M} E = & \min_{\text{pos}(v), v \in V_M} \frac{1}{N_S N_A} \sum_{u \in V_S} \sum_{v \in V_M} 4\epsilon_{uv} \left[ \left( \frac{\sigma_{uv}}{d_{uv}} \right)^{12} - \left( \frac{\sigma_{uv}}{d_{uv}} \right)^6 \right] \\ & + \frac{1}{N_S N_A} \sum_{u \in V_S} \sum_{v \in V_M} \frac{q_u q_v}{4\pi\epsilon_0 d_{uv}} \\ & + \frac{1}{n} \sum_{i=1}^n \frac{1}{2} k_{as} d_{a_i, \text{shift}(p_i; a_i)}^2 \\ & - \frac{1}{N_S (N_A - n)} \sum_{u \in V_S} \sum_{v \in V_M - A} \frac{1}{2} C_u k_{as} d_{uv}^2 \end{aligned} \quad (6)$$

subject to  $d_{uv} = d_{uv}|_{t=0}, \forall \{u, v\} \in E_M$

$$\begin{aligned} \angle_{uvw} = \angle_{uvw}|_{t=0}, \forall \{u, v\}, \{v, w\} \in E_M \\ 0 < \text{pos}(v) < 1, \forall v \in V_M \end{aligned} \quad (7)$$

where  $N_S$  and  $N_A$  are the numbers of atoms in substrate and adsorbate respectively. The first two terms introduce physical effects in the optimization. They are the main intermolecular interaction. The first term is the Lennard-Jones 12-6 van der Waals potentials.  $\epsilon_{uv}$  and  $\sigma_{uv}$  are obtained by using Lorentz-Berthelot mixing rules [20,21]. The second term

is the Coulomb potential between the adsorbate and the substrate.  $q_u$  and  $q_v$  are Bader charges. The last two terms are artificial elastic potentials. The first one creates an attractive force field between the anchored atoms and the position of sites, with an elastic constant  $k_{as}$ , while the second one is repulsive, with an elastic constant  $C_u k_{as}$  in which  $C_u$  is a coefficient.  $\text{shift}(u; v)$  is the function that maps the site  $u \in V_D$  to the one with shifts  $u' \in V_D^v$  according to the anchor  $v \in V_M$ . Their goal is to constraint the molecule to the space close to the absorption sites while weakening the interaction between the non-anchored atoms and the surface as much as possible. This constraint aims at enabling triggering the chemisorption at varying degrees to significantly probe the possibility of bond forming with the surface. If despite this constraint, the adsorbate does not form any chemical bond with the surface, then the adsorbate is most likely physisorbed. Their existence can also be regarded as the chemical bonding between the molecule and the surface depending on the strength of  $k_{as}$ . In the force-field optimization, no change of the bond lengths and bond angles are allowed, which corresponds to the first and second constraints, as well as the fractional coordinates of molecular atoms remain within the cell, which is the third constraint, where  $\text{pos}(v)$  is the fractional coordinates of atom  $v$ .

A screening process is run on all constructed configurations. A feature vector is generated from each structure that stores the information about atomic position and symmetry. A score based on the differences between these feature vectors will be computed to evaluate their similarity. To achieve this, the following objective about the likeness  $L(V_M, V'_M)$  is optimized

$$\min_{f: V_M \rightarrow V'_M} L = \min_{f: V_M \rightarrow V'_M} \frac{1}{N_A N_D} \sum_{u \in V_M} \sum_{j=1}^{N_D} (d_{uj} - d_{f(u)j}) \quad (8)$$

$$\text{subject to } \forall f(u) = u', \text{elem}(u) = \text{elem}(u')$$

$$\forall f(u_1) = u'_1, f(u_2) = u'_2, u_1 \neq u_2 \wedge u'_1 \neq u'_2 \quad (9)$$

where  $V_M$  and  $V'_M$  are the set of vertices of molecule graph from two configurations respectively. They differ in the atomic positions of the molecule.  $d_{uj}$  ( $d_{f(u)j}$ ) are the grouped and sorted distances between each pair of an atom  $u$  ( $u' = f(u)$ ) in the molecule and an adsorption site indexed  $j$  (see Supporting Information).  $N_D$  is the total number of adsorption sites.  $f: V_M \rightarrow V'_M$  is a mapping which is sought to minimize the value of the objective, i.e. likeness  $L(V_M, V'_M)$ . Two structures will be deemed similar if the likeness is small, e.g., below the covalent radius of a hydrogen atom. This choice of criterion is due to the fact that it is enough for the algorithm to recognize what type of adsorption site a smallest atom is at. More details on this process can be found in section ‘‘Screening in trajectory graph’’ in Supporting Information.

In the fourth step, the loop of DFT geometry optimization and screening starts. Using the trajectories of partially or fully DFT-optimized structures, the configurations that end in the same geometry can be merged into one by employing the same screening procedure with an extension. To alleviate the risk of retaining an outlying configuration instead of a valid one, whole sequences of the consecutive geometries, instead of a single one, are considered. They form a matched pattern. If the matched pattern of a trajectory is found in the middle or end of another, then it indicates that the former will eventually evolve into the latter. In the event of the matched pattern being found at the end of the trajectory, it is called an *end pattern*. Consequently, the connectivity between trajectories can be created as in Fig. 1d, and the arrows indicate the direction of evolution. Text can be drawn on edges in the format of  $m@n$  or  $m$ , where  $m$  is the length of the matched pattern, and  $n$  is the index of the last image of the matched pattern in the trajectory the arrow points to. The former, which is drawn on simple arrows, means that the matched pattern of a trajectory is in another trajectory before its end. The latter, which is drawn on double arrows, means they have the same end pattern. They play a central role in eliminating the configurations that evolve into the same ones. In this generated graph  $J = (V_J, E_J)$ , namely

trajectory graph, the strongly connected subgraphs are replaced by configurations that leads to the minimum energies respectively, then all vertices that satisfy

$$\text{deg}^- v = 0 \quad (10)$$

are removed, where  $v \in V_J$ . This procedure will be repeated. After several rounds, the number of configurations remaining can be decreased. Meanwhile, the precision of DFT calculation can be promoted gradually. Each isolated subgraph in the trajectory graph  $J$  is denoted in format of  $J_v$ , where  $v$  is the vertex that satisfies  $\text{deg}^+ v = 0$ , i.e. no edge from this vertex. In case there are multiple vertices that meets this condition, the corresponding subgraph can be denoted as  $J_{v_1-v_2-\dots}$  and  $\text{deg}^+ v_i = 0, i = 1, 2, \dots$

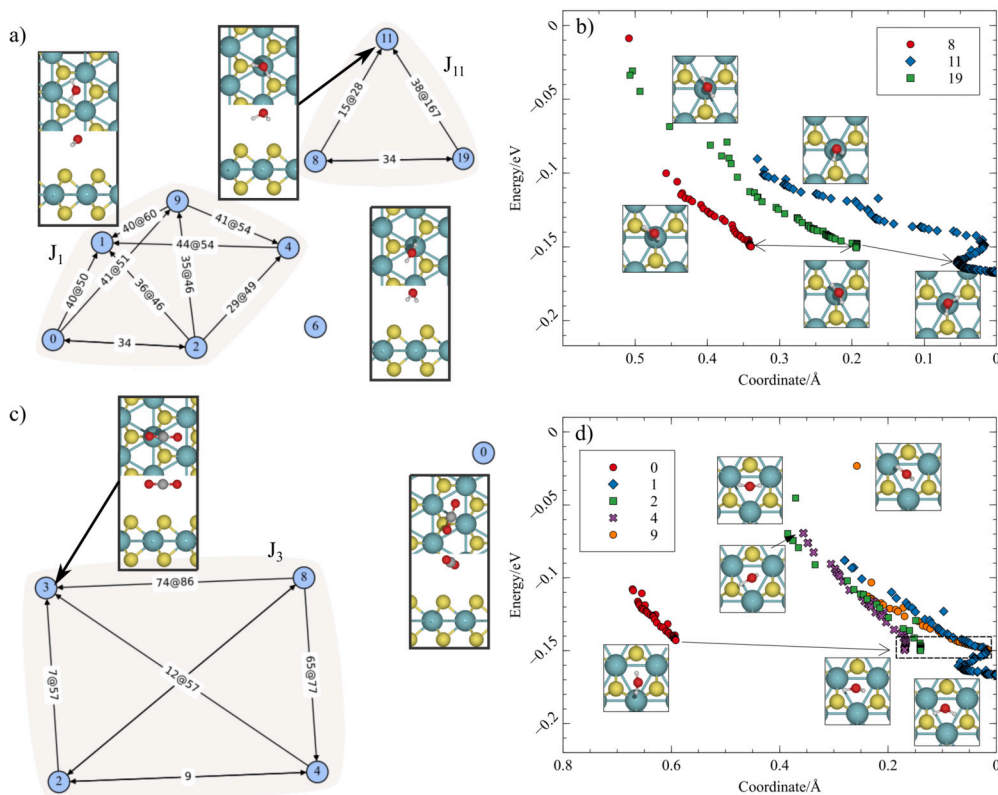
The remaining configurations are the unique ones. By comparing the adsorption energy, the most preferable configuration can be determined, as others are only metastable states corresponding to local minima on the potential energy surface. It is worth mentioning that despite effective in justification of lowest-energy configuration through the current algorithm, further investigation needs to be carried out to define the nature of the adsorption. This is due to the fact that, in some cases, the adsorption energy is not sufficient to indicate chemisorption. The repulsive Hartree and Madelung interaction may have a significant reducing effect on the adsorption energy, therefore, leading to a contradiction with what is observed from charge density difference [22].

In this screening process, the reconstruction of the surface is not considered, since the reconstructed surface has a different structure. In case of that, the substrate needs to be relaxed first before running with adsorbates in SEFFO.

## 2.2. Case studies

The above algorithm is applied in four case studies, water and  $\text{CO}_2$  as adsorbates,  $\text{MoS}_2$  and  $\text{C}_3\text{N}_4$  as adsorbents. The choice of adsorbate enables the study of the Lewis acido-basicity of the surface by providing a Lewis base  $\text{H}_2\text{O}$  and a Lewis acid  $\text{CO}_2$  that can chemisorb at the surface. The adsorbates, as well as the substrates chosen have been well studied and references can be found for comparison with the results in this work. While  $\text{MoS}_2$  has a simpler surface structure, the case studies with  $\text{C}_3\text{N}_4$  demonstrate SEFFO’s capability of working with more complex surfaces.

Fig. 2a and Fig. 2c show the trajectory graphs of water and  $\text{CO}_2$  adsorption on  $\text{MoS}_2$  monolayer respectively. The graphs show they are both adsorbed since the numbers of configurations for both systems are greatly decreased after the 3rd round of screening. Most of the connections already appears after the second round of the loop. The starting nine configurations of water merge into three unique ones at the end of the process. Among the nine configurations, five evolved into configuration 1 alone. The adsorption energies of the final structures are listed in Table 1. The vdW corrections were also considered to evaluate their impacts onto the adsorption configuration. A PBE-D3 correction [23,24] was employed to calculate the adsorption energies. Results with and without the PBE-D3 are tabulated in the two columns of energies. Configuration 11 is the most energetically preferable, while the value of configuration 1 is almost isoenergetic. In configuration 1, the water molecule is at a hollow site of  $\text{MoS}_2$ , while in configuration 11, on the top of a molybdenum atom. The former is the same as the most stable one found in references [25] and [26]. Fig. 2b and Fig. 2d show the evolution of adsorption energy, i.e., the ionic optimization process of the configurations that connects to 11 and 1 respectively. The distance from the center of mass of the adsorbate at any position of the trajectory (in this case water) to its final position was adopted as the coordinate. As reflected in the inset pictures of structure near the curves, all water molecules tend to evolve into the same configuration despite their different initial orientations. Since the designed loop of calculation and screening always leads to the geometry of lower energy, the converged



**Fig. 2.** The trajectory graph of a) water and c)  $\text{CO}_2$  adsorption on  $\text{MoS}_2$  monolayer. Text on arrows in format  $m@n$  means  $m$ -length end pattern of a trajectory is found in the pointed trajectory at index  $n$ , while in format  $m$  means they have the same end pattern of  $m$ . b) and d) plot adsorption energy vs center of mass of adsorbate of configurations in cluster  $J_{11}$  and  $J_1$  respectively in a). The arrows and rectangles drawn in dashed line imply the merging of multiple trajectories. The inset images are the snapshots of the structure of the data point nearby.

**Table 1**

Adsorption energy of the final configurations. The first column of energy are values with DFT-D3 corrections, and the second column without. The part of  $\text{CO}_2@C_3N_4$  is put in Supporting Information as Table S1.

| Adsorbate@substrate         | Configuration | Energy/eV·particle <sup>-1</sup> |        |
|-----------------------------|---------------|----------------------------------|--------|
|                             |               | PBE-D3                           | PBE    |
| $\text{H}_2\text{O}@MoS_2$  | 1             | -0.167                           | -0.075 |
|                             | 6             | -0.142                           | -0.056 |
|                             | 11            | -0.167                           | -0.081 |
| $\text{CO}_2@MoS_2$         | 0             | -0.166                           | -0.039 |
|                             | 3             | -0.183                           | -0.040 |
|                             | 10            | -0.461                           | -0.323 |
|                             | 18            | -0.220                           | -0.141 |
|                             | 24            | -0.319                           | -0.272 |
| $\text{H}_2\text{O}@C_3N_4$ | 33            | -0.184                           | -0.116 |
|                             | 41            | -0.304                           | -0.309 |
|                             | 47            | -0.220                           | -0.140 |

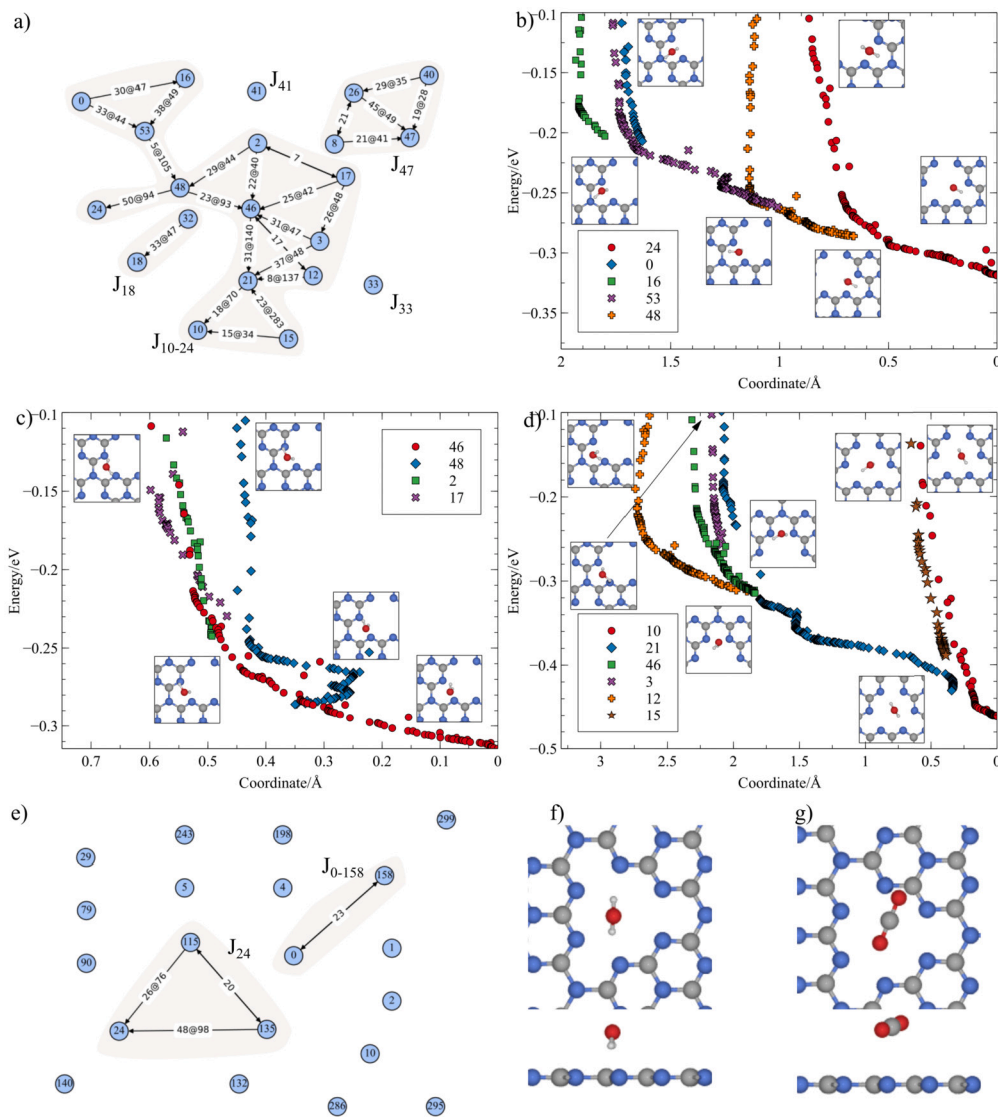
optimized structures stay at local minima on the potential energy surface. Therefore, the screening process merges the trajectories that end in the same minimum according to the specified precision of structural difference, to prevent the calculation of the duplicate structures. This is true especially for configuration 1, where 4 curves are already close to each other when coordinates are in range 0–0.2 Å.

For the adsorption of  $\text{CO}_2$  on  $\text{MoS}_2$  in Fig. 2c, the configurations that lead to 3 are visually similar after several ionic steps. As the most stable one, configuration 3 is consistent with previous work [27]. Therefore, many overlaps were found between their trajectories, similar to the case of water adsorption on  $\text{MoS}_2$ . This comes from the use of a high value as the maximum number of ionic steps in DFT calculation. It can be solved

by using a small number of ionic steps, and in accordance, the number of rounds should be increased.

Fig. 3a shows the trajectory graph for water adsorption on  $C_3N_4$ . The starting number of configurations is 21, which is twice the number of configurations found on  $\text{MoS}_2$ . This is due to the fact that  $C_3N_4$  has a more complex surface, hence, more possible configurations. Correspondingly, to speed up the calculation and decrease the overlaps between trajectories, the number of ionic steps was reduced. Most of the connections were made during the second round of optimization. This indicates the capability of  $C_3N_4$  to adsorb the water molecule, similar to the case of adsorbates on  $\text{MoS}_2$ . After five rounds of screening, only 6 configurations remain. Among them, configuration 10 exhibits the lowest energy as shown in Fig. 3f. Therein the water molecule is at the center of the big hollow site of  $C_3N_4$  with the HOH plane perpendicular to the  $C_3N_4$  surface with the oxygen atom being the furthest from it. This is consistent with the result in the reference [28]. Besides, other configurations in their work can also be found among the remaining final configurations.

Fig. 3b-d show the evolution of configurations that leads to configurations 10 and 24. Because the paths from each configuration to configuration 10 are long, it is divided into 3 parts, one to configuration 24, one to configuration 46 and one to configuration 10. There are more than one evolution from configuration 48. It can evolve into either configuration 24 or configuration 46. Even the coordinates of water molecule are close, the orientation is different. Configuration 46 is an important node, where the water molecule is at the bridge site between two nitrogen atoms. Many configurations where the water molecule is at the adsorption sites other than the big hollow site evolved into it. It then joined configuration 21, where the similar structures are found and, after geometry optimization, the water molecule moves toward the hollow site and merges into configuration 10 and 15, where it is at the hollow site. These observations clearly show how each configuration



**Fig. 3.** The trajectory graph of a) water and e)  $\text{CO}_2$  adsorption on  $\text{C}_3\text{N}_4$ . Text on arrows in format  $m@n$  means  $m$ -length end pattern of a trajectory is found in the pointed trajectory at index  $n$ , while in format  $m$  means they have the same end pattern of  $m$ . b), c) and d) plot adsorption energy vs center of mass of adsorbate of configurations in clusters  $J_{10-24}$  a). The inset images are the snapshots of the structure of the data point nearby. f) and g) show the most preferable configurations for water and  $\text{CO}_2$  adsorption after optimization respectively.

evolved into configuration 10 after geometry optimization even when their starting point is different.

Compared to the other three case studies already shown, the  $\text{CO}_2$  adsorption on  $\text{C}_3\text{N}_4$  is quite different, since there are only four edges in the trajectory graph, and after 5 rounds of geometry optimization/screening, only three were merged into others. Its trajectory graph is presented in Fig. 3e and the most preferable one is configuration 135, shown in Fig. 3g, where the  $\text{CO}_2$  molecule is near the big hollow site with the O-C-O axis pointing to a carbon atom of  $\text{C}_3\text{N}_4$ . This position is between the two most preferred sites found in [29], rather than the most preferred one where the molecule is above a nitrogen atom near the hollow site. The reason behind this inconsistency may be the flatness of the potential energy surface of  $\text{CO}_2$  adsorption on  $\text{C}_3\text{N}_4$ , which is indicated by the non-reducibility of the trajectory graph. It is also why the initialization of  $\text{CO}_2$  positions and orientations does not change much after geometry optimization. The possibility of the initial position of  $\text{CO}_2$  being too far from the substrate can be excluded because in the first few steps, the  $\text{CO}_2$  molecule is pushed away from the surface due to repulsion from the surface (steric hindrance). Therefore, we can assert that the adsorption of  $\text{CO}_2$  is weaker than the adsorption of  $\text{H}_2\text{O}$  according to the results of

the present algorithm. This can be understood in terms of the electrostatic forces between the molecules and the surface. The two hydrogen atoms of the water molecule facing downward suggest attractive forces with the surrounding surface nitrogen atoms, which stabilizes its adsorption on surface. The two oxygen atoms in  $\text{CO}_2$  carrying negative charges point repulsive forces with the nitrogen atoms, therefore the molecule is stirred away from the surface. It is worth mentioning that the interaction is stronger when water interacts with the surface of  $\text{C}_3\text{N}_4$  than when it is interacting with  $\text{MoS}_2$ , since the charges have greater absolute value in  $\text{C}_3\text{N}_4$  than that in  $\text{MoS}_2$  (Bader charges are  $\sim -0.6$  for sulfur in  $\text{MoS}_2$  and  $\sim -1.0$  for nitrogen in  $\text{C}_3\text{N}_4$  respectively).

After studying the four substrate/adsorbate pairs, only physisorption has been obtained. This is also verified by the charge density differences shown in Fig. S5-8. It is observed that the charge transfer are all too weak to form chemical bonding. Usually, chemisorption happens at a shorter distance from the surface and its strength is higher than physisorption. Considering the adsorption process, chemisorption frequently needs an appreciable activation energy while physisorption does not. Hence, a preceding physisorption process is often preferred in chemisorption [8-10,30,11]. The difficulty of chemisorption to re-

alize chemical bond formation with the surface comes from the stable monolayers and the closed shell adsorbate considered. The monolayers are stable because they do not show any unsaturations (e.g. defects) as opposed to surfaces created from bulk cleavage that necessitates bond breaking. Hence, the way to using this algorithm for chemisorption studies will be to create unsaturations in the monolayer, then posing new challenges such as stability of the monolayer itself [31]. In terms of adsorbate, the molecules of water and carbon dioxide both have all of their electrons associated in chemical bonds and/or electron doublets, therefore presenting an important degree of stability. The electron doublet (in water for example) confers this molecule a Lewis basicity. However, water could not chemisorb on MoS<sub>2</sub> due to the metal center being surrounded by sulfur atoms. The introduction of a radical (such as ·OH·), is likely to enable the chemisorption of unsaturated monolayers as it has already been seen elsewhere [32–34].

The comparison between various packages for adsorption study is shown in Table S2. It shows the function of this work covers the whole lifecycle of adsorption study. While site enumeration has been implemented in most of them, the construction of the structure based on numerical optimization gives its strength to be more flexible and extendable by simply introducing more robust molecular constraints. The loop of DFT and screening not only decreases the time for calculations with duplicates, but also additionally gives evolutionary information of trajectories.

### 3. Conclusion

In conclusion, SEFFO is developed for automating the adsorption study on 2D materials based on graph theory for descriptor and screening, and numerical optimization for construction. It was tested on four systems using H<sub>2</sub>O and CO<sub>2</sub> as the adsorbate and MoS<sub>2</sub> and C<sub>3</sub>N<sub>4</sub> as the substrate for its validation. The results are consistent with the literature except the one about CO<sub>2</sub> on C<sub>3</sub>N<sub>4</sub>, which shows its relatively weak adsorption behavior that is indicated from hardly reduced trajectory graph. The creation of trajectory graph reflects the evolutionary relationships between configurations and offers a way to reduce the computational time during the automation through comparing the likeness. Next objectives will be to use this algorithm on new, more challenging systems (larger adsorbates) as well as open the possibility for chemisorption, by vacancy creation and the use of radicals that play a central role in photocatalytic reaction mechanisms.

### 4. Computational details

Four cases were tested using the algorithm SEFFO developed, water and carbon dioxide molecules as the adsorbates, molybdenum sulfide and carbon nitride as substrates. Their monolayer structures were downloaded from the C2DB [35] and relaxed first before use. 5 × 5 supercell of MoS<sub>2</sub> and 3 × 3 supercell of C<sub>3</sub>N<sub>4</sub> were used, with the topmost atoms within 0.9 Å selected as the surface atoms. Atoms in adsorbates were enumerated and combined as a set of anchors of different sizes. The radii of atoms were scaled with a factor of 1.2 based on their covalent radii, and the distance between atoms were truncated to 1 Å if it is less during the process of enumerating anchor-site combinations. In the construction step, the parameters for the LJ 12-6 potentials were from the Universal force field on OpenKIM [36–42].  $\eta$  and  $\xi$  were set to 1.5 and 1 respectively, and the elastic constants were 10 and 2 in the first and second numerical minimization of the objective, and the ratio of the repulsive potentials to the attractive ones is 0.1. After the construction, the screening process was run with precision of structural difference 0.15 Å and precision of energy 0.05 eV. During the loop of DFT calculation and screening, the precision of structural difference is 0.3 Å and that of the energy difference is 0.05 eV, with 5 images as the minimum length of the matched sequence of structure that are allowed to merge. For MoS<sub>2</sub>, three rounds were run, in which the maximum number of ionic steps in DFT calculation was 100, while for C<sub>3</sub>N<sub>4</sub>, it is 5 rounds and each run for

50 ionic steps at maximum. The criteria of convergence were improved after each round.

Atomic Simulation Package is used for the operation on atomic structures [43,44]. The part that involves the manipulation of graph data structure is written with NetworkX [45]. The vf2 algorithm [46] was used for checking graph isomorphism and the strongly connected components in the graph were found by Tarjan's algorithm [47] with Nuutila's modifications [48]. The DFT calculations were all done in Vienna Atomic Simulation Package (VASP) [49–51] with PBE functionals [52,53]. Multiple Python packages [54–57] were used to do the data analysis.

### 5. Notations

| Description                                   | Symbol                         |
|---|--------------------------------|
| Slab graph                                    | $S = (V_S, E_S)$               |
| Molecule graph                                | $M = (V_M, E_M)$               |
| Site graph                                    | $D = (V_D, E_D)$               |
| No. vertices in slab graph                    | $N_S$                          |
| No. vertices in molecule graph                | $N_A$                          |
| No. vertices in site graph                    | $N_D$                          |
| Set of top sites                              | $T, T \subset V_D$             |
| Configuration graph                           | $C = (V_C, E_C)$               |
| Set of sites for adsorption                   | $P = (p_i)_{i=1,2,\dots,n}$    |
| Set of anchors for adsorption                 | $A = (a_i)_{i=1,2,\dots,n}$    |
| Trajectory graph                              | $J = (V_J, E_J)$               |
| Isolated subgraph in $J$                      | $J_v, J_{v_1-v_2-\dots}$       |
| Fractional position of vertex $v$             | $\text{pos}(v)$                |
| Distance between vertices $u$ and $v$         | $d_{uv}$                       |
| Angle between edges $\{u, v\}$ and $\{v, w\}$ | $\angle_{uvw}$                 |
| Surface atoms                                 | $F \subset V_S$                |
| Surface atoms connected to site $v \in V_D$   | $\text{csa}(v)$                |
| Atomic type $e$ of vertex $v$                 | $\text{elem}(v) \rightarrow e$ |
| Layer index $n$ of vertex $v$                 | $\text{lyr}(v) \rightarrow n$  |
| Final energy $g$ of trajectory $v \in V_J$    | $\text{fen}(v) \rightarrow g$  |

### CRedit authorship contribution statement

**Leran Lu:** Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Wei Cao:** Writing – review & editing, Supervision, Funding acquisition. **Romain Botella:** Writing – review & editing, Supervision, Data curation.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 101002219). Financial supports from Jane ja Aatos Erkon Säätiö (JAES) and Tiina ja Antti Herlinin säätiö (TAHS) on Advanced Steels for Green Planet (AS4G) are also acknowledged. Authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources and Marko Huttula and Samuli Urpelainen for the management of AS4G project.

### Appendix A. Supplementary material

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.cpc.2024.109440>.

## Data availability

The data generated and analyzed in this article are available on Zenodo at <https://doi.org/10.5281/zenodo.10804634> or upon request.

## References

- [1] A. Habibi-Yangjeh, S. Asadzadeh-Khaneghah, S. Feizpoor, A. Rouhi, Review on heterogeneous photocatalytic disinfection of waterborne, airborne, and foodborne viruses: can we win against pathogenic viruses?, *J. Colloid Interface Sci.* 580 (2020) 503–514, <https://doi.org/10.1016/j.jcis.2020.07.047>.
- [2] S.B. Torrisi, A.K. Singh, J.H. Montoya, T. Biswas, K.A. Persson, Two-dimensional forms of robust CO<sub>2</sub> reduction photocatalysts, *npj 2D Mater. Appl.* 4 (1) (Jul. 2020), <https://doi.org/10.1038/s41699-020-0154-y>, times Cited in Web of Science Core Collection: 10 Total Times Cited: 10.
- [3] S. Feng, B. Li, B. Xu, Z. Wang, Hybrid perovskites and 2D materials in optoelectronic and photocatalytic applications, *Crystals* 13 (11) (2023) 1566, <https://doi.org/10.3390/cryst13111566>.
- [4] F. Yang, P. Hu, F. Yang, X.-J. Hua, B. Chen, L. Gao, K.-S. Wang, Photocatalytic applications and modification methods of two-dimensional nanomaterials: a review, *Tungsten* 6 (1) (2024) 77–113, <https://doi.org/10.1007/s42864-023-00229-x>.
- [5] Y. Lin, G. Yuan, R. Liu, S. Zhou, S.W. Sheehan, D. Wang, Semiconductor nanostructure-based photoelectrochemical water splitting: a brief review, *Chem. Phys. Lett.* 507 (4) (2011) 209–215, <https://doi.org/10.1016/j.cplett.2011.03.074>.
- [6] A.B. Murphy, P.R.F. Barnes, L.K. Randeniya, I.C. Plumb, I.E. Grey, M.D. Horne, J.A. Glasscock, Efficiency of solar water splitting using semiconductor electrodes, *Int. J. Hydrog. Energy* 31 (14) (2006) 1999–2017, <https://doi.org/10.1016/j.ijhydene.2006.01.014>.
- [7] J. Echeverría, S. Alvarez, The borderless world of chemical bonding across the van der Waals crust and the valence region, *Chem. Sci.* 14 (42) (2023) 11647–11688, <https://doi.org/10.1039/D3SC02238B>.
- [8] P. Atkins, J. De Paula, *Physical Chemistry*, vol. 1, Macmillan, 2006.
- [9] P.L. Houston, *Chemical Kinetics and Reaction Dynamics*, Courier Corporation, 2012.
- [10] F. Huber, J. Berwanger, S. Polesya, S. Mankovsky, H. Ebert, F.J. Giessibl, Chemical bond formation showing a transition from physisorption to chemisorption, *Science* 366 (6462) (2019) 235–238, <https://doi.org/10.1126/science.aay3444>.
- [11] J.E. Lennard-Jones, Processes of adsorption and diffusion on solid surfaces, *Trans. Faraday Soc.* 28 (1932) 333–359.
- [12] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, *Comput. Mater. Sci.* 68 (2013) 314–319, <https://doi.org/10.1016/j.commatsci.2012.10.028>.
- [13] J.R. Boes, O. Mamun, K. Winther, T. Bligaard, Graph theory approach to high-throughput surface adsorption structure generation, *J. Phys. Chem. A* 123 (11) (2019) 2281–2285, <https://doi.org/10.1021/acs.jpca.9b00311>.
- [14] S.A. Wilson, C.L. Muhich, Fast identification, and construction of adsorbate-adsorbent geometries for high throughput computational applications: the automatic surface adsorbate structure provider (ASAP) algorithm, *Comput. Theor. Chem.* 1216 (2022) 113830, <https://doi.org/10.1016/j.comptc.2022.113830>.
- [15] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal-organic frameworks, *Nat. Chem.* 4 (2) (2012) 83–89, <https://doi.org/10.1038/nchem.1192>.
- [16] S. Deshpande, T. Maxson, J. Greeley, Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis, *npj Comput. Mater.* 6 (1) (2020) 1–6, <https://doi.org/10.1038/s41524-020-0345-2>.
- [17] C. Martí, S. Blanck, R. Staub, S. Loehlé, C. Michel, S.N. Steinmann, DockOnSurf: a Python code for the high-throughput screening of flexible molecules adsorbed on surfaces, *J. Chem. Inf. Model.* 61 (7) (2021) 3386–3396, <https://doi.org/10.1021/acs.jcim.1c00256>.
- [18] J.H. Montoya, K.A. Persson, A high-throughput framework for determining adsorption energies on solid surfaces, *npj Comput. Mater.* 3 (1) (2017) 1–4, <https://doi.org/10.1038/s41524-017-0017-z>.
- [19] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, *Introduction to Algorithms*, fourth edition, The MIT Press, Cambridge, Massachusetts, 2022.
- [20] R.J. Good, C.J. Hope, New combining rule for intermolecular distances in intermolecular potential functions, *J. Chem. Phys.* 53 (2) (1970) 540–543, <https://doi.org/10.1063/1.1674022>.
- [21] D.R. Pesuit, Model-calculated combining rules for distance force constants and pseudocritical volumes, *J. Chem. Phys.* 68 (7) (1978) 3149–3151, <https://doi.org/10.1063/1.436157>.
- [22] S.B. Mishra, B.R.K. Nanda, Facet dependent catalytic activities of anatase TiO<sub>2</sub> for CO<sub>2</sub> adsorption and conversion, *Appl. Surf. Sci.* 531 (2020) 147330, <https://doi.org/10.1016/j.apsusc.2020.147330>.
- [23] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, *J. Chem. Phys.* 132 (15) (2010) 154104, <https://doi.org/10.1063/1.3382344>.
- [24] S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory, *J. Comput. Chem.* 32 (7) (2011) 1456–1465, <https://doi.org/10.1002/jcc.21759>.
- [25] F. Ferreira, A. Carvalho, Í.J.M. Moura, J. Coutinho, R.M. Ribeiro, Adsorption of H<sub>2</sub>, O<sub>2</sub>, H<sub>2</sub>O, OH and H on monolayer MoS<sub>2</sub>, *J. Phys. Condens. Matter* 30 (3) (2018) 035003, <https://doi.org/10.1088/1361-648X/aaa03f>.
- [26] N.S. Bobbitt, M. Chandross, Interactions of water with pristine and defective MoS<sub>2</sub>, *Langmuir* 38 (34) (2022) 10419–10429, <https://doi.org/10.1021/acs.langmuir.2c01057>.
- [27] S. Zhao, J. Xue, W. Kang, Gas adsorption on MoS<sub>2</sub> monolayer from first-principles calculations, *Chem. Phys. Lett.* 595–596 (2014) 35–42, <https://doi.org/10.1016/j.cplett.2014.01.043>.
- [28] X. Zhou, C. Zhao, C. Chen, J. Chen, X. Chen, The interaction of H<sub>2</sub>O, O<sub>2</sub> and H<sub>2</sub>O + O<sub>2</sub> molecules with g-C<sub>3</sub>N<sub>4</sub> surface: a first-principle study, *Diam. Relat. Mater.* 125 (2022) 108995, <https://doi.org/10.1016/j.diamond.2022.108995>.
- [29] B. Zhu, L. Zhang, D. Xu, B. Cheng, J. Yu, Adsorption investigation of CO<sub>2</sub> on g-C<sub>3</sub>N<sub>4</sub> surface by DFT calculation, *J. CO<sub>2</sub> Util.* 21 (2017) 327–335, <https://doi.org/10.1016/j.jcou.2017.07.021>.
- [30] K.J. Laidler, J.H. Meiser, B. Ramachandran, *Solutions Manual, Physical Chemistry*, Houghton Mifflin Co., Boston, MA, 1999.
- [31] L. Lu, R. Botella, W. Cao, Theoretical study of stability of halogen-defective trihalide monolayers: cases of AlI<sub>3</sub>, AsI<sub>3</sub>, and IrBr<sub>3</sub>, *Phys. Status Solidi (b)* 260 (9) (2023) 2300001, <https://doi.org/10.1002/pssb.202300001>.
- [32] H. Li, L. Xu, X. Huang, J. Ou-Yang, M. Chen, Y. Zhang, S. Tang, K. Dong, L.-L. Wang, Two-dimensional C<sub>3</sub>N/WS<sub>2</sub> vdW heterojunction for direct Z-scheme photocatalytic overall water splitting, *Int. J. Hydrog. Energy* 48 (6) (2023) 2186–2199, <https://doi.org/10.1016/j.ijhydene.2022.10.102>.
- [33] L. Luan, L. Han, D. Zhang, K. Bai, K. Sun, C. Xu, L. Li, L. Duan, AlSb/ZrS<sub>2</sub> heterojunction: a direct Z-scheme photocatalyst with high solar to hydrogen conversion efficiency and catalytic activity across entire PH range, *Int. J. Hydrog. Energy* (Oct. 2023), <https://doi.org/10.1016/j.ijhydene.2023.09.156>.
- [34] Y. Zhang, H. Qiao, Z.-H. Yan, L. Duan, L. Ni, J.-B. Fan, PtS<sub>2</sub>/g-C<sub>3</sub>N<sub>4</sub> van der Waals heterostructure: a direct Z-scheme photocatalyst with high optical absorption, solar-to-hydrogen efficiency and catalytic activity, *Int. J. Hydrog. Energy* 48 (39) (2023) 14659–14669, <https://doi.org/10.1016/j.ijhydene.2022.12.329>.
- [35] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P.S. Schmidt, N.F. Hinsche, M.N. Gjerding, D. Torelli, P.M. Larsen, A.C. Riis-Jensen, J. Gath, K.W. Jacobsen, J. Jørgen Mortensen, T. Olsen, K.S. Thygesen, The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals, *2D Mater.* 5 (4) (2018) 042002, <https://doi.org/10.1088/2053-1583/aacfc1>.
- [36] J.E. Jones, On the determination of molecular fields. I. From the variation of the viscosity of a gas with temperature, *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.* 106 (738) (1924) 441–462, <https://doi.org/10.1098/rspa.1924.0081>.
- [37] J.E. Jones, On the determination of molecular fields. II. From the equation of state of a gas, *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.* 106 (738) (1924) 463–477, <https://doi.org/10.1098/rspa.1924.0082>.
- [38] J.E. Lennard-Jones, On the forces between atoms and ions, *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.* 109 (752) (1925) 584–597, <https://doi.org/10.1098/rspa.1925.0147>.
- [39] R.S. Elliott, Efficient ‘universal’ shifted Lennard-Jones model for all KIM API supported species developed by Elliott and Akerson (2015) v003, OpenKIM (2018), <https://doi.org/10.25950/962b4967>.
- [40] R.S. Elliott, Efficient multi-species Lennard-Jones model with truncated or shifted cutoff v003, OpenKIM (2018), <https://doi.org/10.25950/ac258694>.
- [41] E.B. Tadmor, R.S. Elliott, J.P. Sethna, R.E. Miller, C.A. Becker, The potential of atomistic simulations and the knowledgebase of interatomic models, *JOM* 63 (7) (2011) 17, <https://doi.org/10.1007/s11837-011-0102-6>.
- [42] R.S. Elliott, E.B. Tadmor, Knowledgebase of Interatomic Models (KIM) application programming interface (API), 2011.
- [43] S.R. Bahn, K.W. Jacobsen, An object-oriented scripting interface to a legacy electronic structure code, *Comput. Sci. Eng.* 4 (3) (2002-05/2002-06) 56–66, <https://doi.org/10.1109/5992.998641>.
- [44] A.H. Larsen, J.J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, P.C. Jennings, P.B. Jensen, J. Kermode, J.R. Kitchin, E.L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J.B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K.S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K.W. Jacobsen, The atomic simulation environment—a Python library for working with atoms, *J. Phys. Condens. Matter* 29 (27) (2017) 273002.
- [45] A.A. Hagberg, D.A. Schult, P.J. Swart, Exploring network structure, dynamics, and function using NetworkX, in: G. Varoquaux, T. Vaught, J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference*, Pasadena, CA USA, 2008, pp. 11–15.
- [46] P. Foggia, C. Sansone, M. Vento, An improved algorithm for matching large graphs, in: *3rd IAPR-TC15 Workshop on Graph-Based Representations in Pattern Recognition*, 2001, pp. 149–159.
- [47] R. Tarjan, Depth-first search and linear graph algorithms, *SIAM J. Comput.* 1 (2) (1972) 146–160, <https://doi.org/10.1137/0201010>.
- [48] E. Nuutila, E. Soisalon-Soininen, On finding the strongly connected components in a directed graph, *Inf. Process. Lett.* 49 (1) (1994) 9–14, [https://doi.org/10.1016/0020-0190\(94\)90047-7](https://doi.org/10.1016/0020-0190(94)90047-7).
- [49] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* 6 (1) (1996) 15–50, [https://doi.org/10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).

- [50] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B* 54 (16) (1996) 11169–11186, <https://doi.org/10.1103/PhysRevB.54.11169>.
- [51] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, *Phys. Rev. B* 47 (1) (1993) 558–561, <https://doi.org/10.1103/PhysRevB.47.558>.
- [52] G. Kresse, J. Hafner, Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements, *J. Phys. Condens. Matter* 6 (40) (1994) 8245, <https://doi.org/10.1088/0953-8984/6/40/015>.
- [53] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B* 59 (3) (1999) 1758–1775, <https://doi.org/10.1103/PhysRevB.59.1758>.
- [54] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, *Nature* 585 (7825) (2020) 357–362, <https://doi.org/10.1038/s41586-020-2649-2>.
- [55] J.D. Hunter, Matplotlib: a 2D graphics environment, *Comput. Sci. Eng.* 9 (3) (2007) 90–95, <https://doi.org/10.1109/MCSE.2007.55>.
- [56] W. McKinney, Data structures for statistical computing in Python, in: *Proceedings of the 9th Python in Science Conference*, Jan. 2010.
- [57] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, Í. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, S. Contributors, SciPy 1.0—fundamental algorithms for scientific computing in Python, *Nat. Methods* 17 (3) (2020) 261–272, <https://doi.org/10.1038/s41592-019-0686-2>, comment: Article source data is available here: <https://github.com/scipy/scipy-articles>, arXiv:1907.10121.