

# Determining Surface Phase Diagrams Including Anharmonic Effects

Yuanyuan Zhou, Matthias Scheffler, and Luca M. Ghiringhelli  
*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin-Dahlem, Germany*

(Dated: August 30, 2019)

We introduce a massively parallel replica-exchange grand-canonical sampling algorithm to simulate materials at realistic conditions, in particular surfaces and clusters in reactive atmospheres. Its purpose is to determine in an automated fashion equilibrium phase diagrams for a given potential-energy surface (PES) and for any observable sampled in the grand-canonical ensemble. The approach enables an unbiased sampling of the phase space and is embarrassingly parallel. It is demonstrated for a model of Lennard-Jones system describing a surface in contact with a gas phase. Furthermore, the algorithm is applied to  $\text{Si}_M$  clusters ( $M = 2, 4$ ) in contact with an  $\text{H}_2$  atmosphere, with all interactions described at the *ab initio* level, i.e., via density-functional theory, with the PBE gradient-corrected exchange-correlation functional. We identify the most thermodynamically stable phases at finite  $T, p(\text{H}_2)$  conditions.

## I. INTRODUCTION

A prerequisite for analyzing and understanding the electronic properties and the function of surfaces is the detailed knowledge of the surface composition and atomistic geometry under realistic conditions. The structure of a surface at thermodynamic equilibrium with its environment is in fact a configurational statistical average over adsorption, desorption, and diffusion processes.

A temperature-pressure phase diagram describes the composition and structure of a system at thermal equilibrium and is an essential tool for understanding material properties. The *ab initio* atomistic thermodynamics (aiAT) approach<sup>1-5</sup> has been very successful in predicting phase diagrams for surfaces<sup>6,7</sup> and gas-phase clusters<sup>8-10</sup> at realistic  $T, p$  conditions. The key assumption is, however, that *all* relevant local minima of the potential energy surface (PES) of a given system are enumerated, a (strong) limitation in case of unexpected surface stoichiometries or geometries. Such limitation can only be overcome by an unbiased sampling of configurational and compositional space. A further assumption in most work has been that the vibrational contributions to the change of the free energy are largely canceled and can be neglected. We will see below that this is not always justified.

In this paper, we introduce a Replica-Exchange (RE) Grand-Canonical (GC) Monte-Carlo (MC)/Molecular-Dynamics (MD) algorithm, that enables the efficient calculation of complete temperature-pressure phase diagrams of surfaces, nanoparticles, or clusters in contact with reactive gas atmospheres. The RE and GC steps of the algorithm are formulated in the Metropolis MC framework, while the canonical sampling of configurations (diffusion) is supported via both MC and MD. In the case of surface in contact with a gas phase reservoir, the gas molecules can physi-/chemisorb on the surface, while adsorbed molecules or single atoms can desorb from the surface to the gas phase. At thermodynamic equilibrium, the number of desorbed molecules/atoms balances the adsorbed one, so that on average a constant number of molecules/atoms is present on the surface. We specifi-

cally target thermodynamically open systems in the GC ensemble, aiming at describing (nano)structured surfaces in a reactive atmosphere at realistic  $T, p$  condition, so that the surfaces can exchange particles with the gas reservoir. The initial idea of RE<sup>11-14</sup> is to allow for an efficient sampling of the configurational space by shuttling configurations from regions of low  $T$  to regions of high  $T$ . Later, de Pablo *et al.*<sup>15,16</sup> extended the concept to other intensive thermodynamic variables, such as the chemical potential ( $\mu$ ) in order to simulate the phase equilibria of Lennard-Jones (LJ) systems. This allows systems with different number of particles (the conjugate variable of  $\mu$ ) to be shuttled across different values of  $\mu$ , thus enhancing the sampling, following the same spirit of the temperature replicas in traditional RE. By combining advantages of both GC and RE, our massively parallel algorithm requires no prior knowledge of the phase diagram and takes only the potential energy function together with the desired  $\mu$  and  $T$  ranges as inputs. The partition function is estimated using the output of the simulation, thus calculating thermodynamic observables is straightforward.

The structure of this paper is as follows. In Sec. II the method and implementation of our REGC algorithm will be discussed in details. In section III we show two applications of the REGC method. The first, in section III A, proof-of-concept application is the determination of the  $p$ - $T$  phase diagram of a system composed of a LJ (frozen) surface in contact with a LJ gas phase. Next, in section III B, we address the calculation of the phase diagram of the  $\text{Si}_2$  dimer and  $\text{Si}_4$  cluster in a reactive atmosphere of  $\text{H}_2$  molecules by performing REGC with aiMD using Perdew-Burke-Ernzerhof (PBE)<sup>17</sup> xc approximation. During the last several decades, silicon hydrides have attracted a lot of attention because of their potential applications in semiconductors, optoelectronics, and surface growth processes.<sup>18-21</sup> The binary clusters of silicon and hydrogen play key roles in the chemical vapor deposition of thin films, and photoluminescence of porous silicon. However, most of the previous research on silicon hydrides focused on the search of global minima structures, but the decisive issue of stability and metastability of silicon hydrides at realistic conditions (exchange of

atoms with an environment) has not been addressed so far. The purpose of this application is to investigate the phase diagrams of silicon hydrides in reactive hydrogen atmosphere. In the outlook section (IV), the capabilities and current limitations of our REGC method will be discussed.

## II. METHOD AND IMPLEMENTATION

The sampling of complex systems, e.g., thermodynamically open systems, composed of many atoms arranged in molecules, clusters, condensed phases, etc., remains a challenge. The main factors that limit sampling efficiency is (i) that systems' configurations get easily trapped — especially at low temperatures — in local minima and (ii) the inherently long characteristic relaxation times in complex many-molecules systems (e.g., atoms' diffusion that require collective motions involving several degrees of freedom). During the last decades, many powerful methods have been developed to deal with the first difficulty, e.g., J-walking<sup>22,23</sup>, multicanonical sampling<sup>24,25</sup>, nested sampling<sup>26</sup>, simple tempering<sup>12,27</sup>, 1/k sampling<sup>28</sup>, expanded ensembles<sup>29</sup>, and parallel tempering<sup>11,14</sup>. While these methods are effective in overcoming kinetic barriers, they do little to accelerate the slow relaxation at low temperatures.

Open ensembles, described at equilibrium by the grand-canonical-ensemble formalism, provide an effective mean to overcome slow-relaxation problems: atoms can get in and out of a system, effectively generating thermodynamically possible defects, along unphysical pathways (e.g., atoms' insertion or removal), thereby circumventing diffusional bottlenecks by disentangling degrees of freedom. We took advantage of both the replica-exchange and grand-canonical-ensemble concepts to design an algorithm that alleviate both kinetic trapping and slow phase space diffusion. In Sec. IIA, we describe our replica-exchange grand-canonical algorithm. Later, in Sec. IIB we describe how to use the results from replica-exchange grand-canonical simulations to calculate phase diagrams and free energy surfaces.

### A. Replica-Exchange Grand-Canonical Monte Carlo / Molecular Dynamics

Our Replica-Exchange Grand-Canonical Monte Carlo or Molecular-Dynamics approach is outlined in Fig. 1. In a REGCMC or REGCMD simulation,  $S$  replicas of the original system of interest are considered, each evolving in a different thermodynamical states ( $T_i$ ,  $\mu_i$ , where  $i$  is the index of the replica). During the simulation, first the system has a probability  $x_0$  ( $0 \leq x_0 \leq 1$ ) to attempt exchanging a particle with the reservoir and probability  $(1 - x_0)$  to perform a replica-exchange move (see below). After the particle/replica-exchange attempt,  $S$  parallel molecular dynamics or Monte Carlo runs follow,

to diffuse the system in the canonical ensemble. i.e., at temperature  $T_i$ , with fixed number of particles  $N$  and volume  $V$  of the system ( $NVT$  ensemble). Then, the procedure is iterated until convergence of defined quantities is achieved. See further for the convergence criterion we adopted.

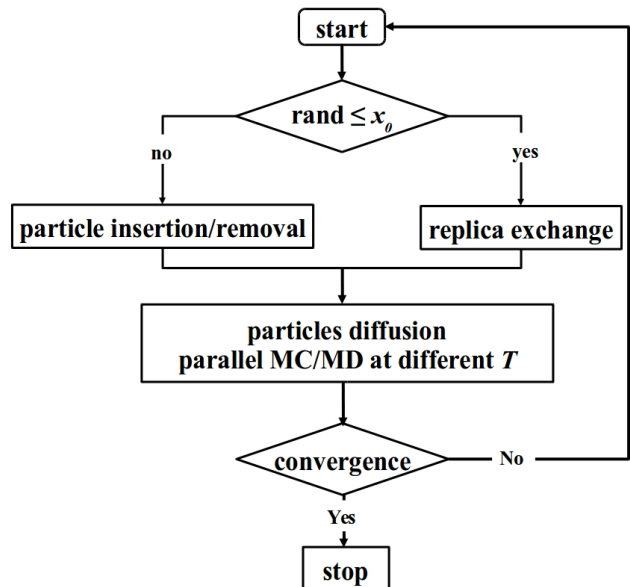


FIG. 1. The flow chart of Replica-Exchange Grand-Canonical Monte Carlo/Molecular Dynamics algorithm. Here  $\text{rand}$  is a (pseudo) random number generated uniformly distributed between 0 and 1.

#### 1. Grand-Canonical Monte Carlo

The particle insertion/removal step is handled by applying the formalism of the grand-canonical ensemble, where the subsystem of our interest (e.g., a surface or a cluster in contact with a gas phase), defined in a volume ( $V$ ), is in equilibrium with a reservoir at given temperature ( $T$ ), and chemical potential ( $\mu$ ) of one species (or more species, each with its own chemical potential). In practice, the reservoir is modeled as an ideal gas and  $\mu$  depends on  $T$  and the pressure  $p$ , as will be specified in the application cases. The number of atoms or molecules in the subsystem is a fluctuating variable, determined by specifying the chemical potential and temperature of the reservoir of (ideal) gas-phase atoms or molecules. The probability density of a grand-canonical ensemble of identical particles is:<sup>30</sup>

$$\mathcal{N}_{\mu,V,T}(\mathbf{R}; N) \propto \frac{e^{(\beta\mu N)} V^N}{\Lambda^{3N} N!} e^{-\beta E(\mathbf{R}; N)} \quad (1)$$

where  $\beta = 1/k_B T$ ,  $\Lambda = h/\sqrt{2\pi m k_B T}$  is the thermal wavelength of a particle of mass  $m$ , and  $E(\mathbf{R})$  is the potential energy of a configuration  $\mathbf{R}$  of the  $N$ -particle

system. The GCMC algorithm consists of the following MC moves: 1) insertion of a gas atom/molecule into the system at a random position, 2) removal of a randomly selected gas atom/molecule from the system, 3) displacement of a gas atom to a new random position in the system to sample the potential energy surface (PES). In our algorithm, the displacement (diffusion) is taken care of separately (see section II A 3) and can be done via either Monte Carlo or MD. Here, we consider the insertion and removal moves, where microscopic reversibility (also called ‘detailed balance’, a *sufficient* condition for an MC scheme to converge the evaluation of observable properties in the desired ensemble<sup>30</sup>) is ensured by having equal number of insertion and removal attempts, for all particles described by the given chemical potential. In practice, we first randomly select if a particle will be inserted or removed, i.e., by generating a (pseudo)random number  $y_1^{\text{GC}}$  uniformly distributed between 0 and 1 and performing a removal if  $y_1^{\text{GC}} < 0.5$ .

For a removal, a particle (an atom or a molecule) is selected at random (by generating a new random number  $y_2^{\text{GC}}$  and selecting particle  $i$  if  $(i-1)/N \leq y_2^{\text{GC}} < i/N$ ). In order to fulfill detailed balance, a possible (and common) choice for accepting the removal of the selected particle is with probability<sup>30</sup>:

$$P_{(N \rightarrow N-1)} = \min \left[ 1, \frac{\Lambda^3 N}{V} e^{-\beta[\mu + E_{N-1} - E_N]} \right] \quad (2)$$

where  $N$  is the number of atoms (or molecules) for which a reservoir at given temperature  $T$  and chemical potential  $\mu$  is defined, and which are in the system before the attempted removal.  $E_N$  is the energy of the system of  $N$  particles,  $E_{N-1}$  is the energy of the same system, without the selected particle, and  $V$  is the system volume, which is fixed during the simulation. According to this formula, if the change in energy due to the particle removal is similar in value to  $\mu$ , there is a high probability that the removal is accepted.

For the insertion, first a location is randomly chosen, uniformly in the simulation volume (in a rectangular cell, by driving three independent uniformly distributed random numbers, one for each Cartesian coordinate). Then, a particle is positioned in the selected location and its insertion is accepted with probability<sup>30</sup>:

$$P_{(N \rightarrow N+1)} = \min \left[ 1, \frac{V}{\Lambda^3 (N+1)} e^{-\beta[\mu - E_{N+1} - E_N]} \right] \quad (3)$$

The probability of accepting an insertion can be low in dense systems as random locations will have high probability to end up too close to already-present particles, henceforth yielding large  $E_{N+1} - E_N$  and consequent rejection of the insertion. Since we are modeling adsorption on surfaces or clusters in contact with a gas phase, we have a relatively rarefied system, especially if the considered volume of particle insertion (and removal) does not include the subsurface (see further).

## 2. Replica Exchange in the Grand-Canonical ensemble

We define an extended ensemble that is the collection of  $S = L \times M$  replica of a given system, arranged in  $L$  values of temperature and  $M$  values of the chemical potential, as illustrated in Fig. 2a. In this paper, we consider only one species that exchange particles with the reservoir, hence, one chemical potential. The *partition function* of this extended ensemble is the product of the partition functions of the individual  $(\mu_m, V, T_l)$  ensembles, where  $l = 1, 2, \dots, L$  and  $m = 1, 2, \dots, M$ :

$$Q_{\text{extended}} = \prod_{l=1}^L \prod_{m=1}^M \frac{e^{\beta_l \mu_m N_{l,m}} V^{N_{l,m}}}{\Lambda_l N_{l,m}!} \int d\mathbf{R} e^{-\beta_l E(\mathbf{R}; N_{l,m})} \quad (4)$$

In the following, we label the temperature indifferently by  $T_l$  or  $\beta_l = 1/k_B T_l$ . The key observation is that taken one configuration along the evolution of a replica at given  $(\mu_m, V, T_l)$ , statistical mechanics allows us to write a well defined probability that the same configuration belongs to the another state  $(\mu_o, V, T_k)$ . We now randomly select a pair of replicas. The replica at state  $(\mu_m, V, T_l)$  is in configuration  $\mathbf{R}_i$  (e.g., represented by the  $3 \times N_{l,m}$  matrix of coordinates) and the replica at state  $(\mu_o, V, T_k)$  is in configuration  $\mathbf{R}_j$ . We then aim at defining a rule for accepting the swap of the configurations between the two replicas, in order to satisfy the detailed balance in the extended ensemble. To the purpose, one has to impose the following equality:

$$\begin{aligned} & \mathcal{N}_{(\beta_l, \mu_m, \mathbf{R}_i)} \mathcal{N}_{(\beta_k, \mu_o, \mathbf{R}_j)} \\ & \times P_{[(\beta_l, \mu_m, \mathbf{R}_i), (\beta_k, \mu_o, \mathbf{R}_j) \rightarrow (\beta_l, \mu_m, \mathbf{R}_j), (\beta_k, \mu_o, \mathbf{R}_i)]} \\ & = \mathcal{N}_{(\beta_l, \mu_m, \mathbf{R}_j)} \mathcal{N}_{(\beta_k, \mu_o, \mathbf{R}_i)} \\ & \times P_{[(\beta_l, \mu_m, \mathbf{R}_j), (\beta_k, \mu_o, \mathbf{R}_i) \rightarrow (\beta_l, \mu_m, \mathbf{R}_i), (\beta_k, \mu_o, \mathbf{R}_j)]} \quad (5) \end{aligned}$$

where  $\mathcal{N}$  is the probability density in the grand-canonical ensemble (Eq. 1), and  $P$  is the probability to swap configurations. Our choice of  $P$  that satisfies the detailed balance is:

$$\begin{aligned} & P_{[(\beta_l, \mu_m, \mathbf{R}_i), (\beta_k, \mu_o, \mathbf{R}_j) \rightarrow (\beta_l, \mu_m, \mathbf{R}_j), (\beta_k, \mu_o, \mathbf{R}_i)]} \\ & = \min \left[ 1, \left( \frac{\beta_l}{\beta_k} \right)^{\frac{3}{2}(N_{l,m} - N_{k,o})} \times \right. \\ & \left. e^{[-(\beta_l - \beta_k)(E(\mathbf{R}_j) - E(\mathbf{R}_i)) + (\beta_l \mu_m - \beta_k \mu_o)(N_{l,m} - N_{k,o})]} \right] \quad (6) \end{aligned}$$

A similar swap-acceptance probability has been proposed in Refs. 15 and 16, but we include a factor  $(\frac{\beta_l}{\beta_k})^{\frac{3}{2}(N_{l,m} - N_{k,o})}$  that is probably neglected in those papers. Furthermore, our scheme adopts a two-dimensional grid of values of temperatures and chemical potentials, while in Refs. 15 and 16 the values of  $T$  and  $\mu$  are constrained to be along a phase boundary of the studied system (vapor-fluid coexistence for the LJ system), therefore being a uni-modal scheme, i.e., one-dimensional in practice.

It is clear from Eq. 6 that swap trial moves are more likely to be accepted the larger the overlap between the energy distributions of the two replicas. A large overlap of energy distribution is verified if the values of the thermodynamic variables  $(\mu, T)$  defining the two replicas are not too dissimilar. In traditional one-dimensional RE, swap moves are attempted only between neighbor replicas. In that case, each replica has two neighbors (or one, for the largest and smallest values of the chosen replicated thermodynamic variable, typically  $T$ ). In our two-dimensional scheme (Fig. 2), each replica has between 3 and 8 neighbors, thus enhancing the possibility for configurations to “diffuse” across replicas. We adopted a “collective” scheme for the attempted swaps that involves the definition of four different types of neighboring swaps, as illustrated in Fig. 2. At each RE move, one type of swaps is selected at random (each with probability 1/4). This choice has the advantage to involve all replicas (when the number of  $T$ -replicas and  $\mu$ -replicas is even) in one attempted swap. An alternative scheme could be to select randomly one replica and independently one neighbor to perform the attempted swap, then to repeat until no replica has an unselected neighbor. We are exploring this scheme for higher-dimensional settings (e.g.,  $T$  and more than one  $\mu$  for more than one type of particles that are exchanged with the reservoir).

### 3. Atoms' displacement

At each cycle of our REGC scheme, after the RE or GC move has been performed, the atoms in each replica perform in parallel a sampling of the canonical (fixed  $N$ , fixed  $V$ , fixed  $T$ ) ensemble. This is achieved with the standard Metropolis MC or with MD.

According to MC, one atom-displacement step requires to select at random one atom and assigning to it a random displacement, typically uniformly distribute in a cube or sphere of size comparable with the typical inter-atomic distances at equilibrium. The move is accepted with probability<sup>30</sup>:

$$P_{(\mathbf{r} \rightarrow \mathbf{r} + \Delta \mathbf{r})} = \min \left[ 1, e^{-\beta[E(\mathbf{r} + \Delta \mathbf{r}, \mathbf{r}^{N-1}) - E(\mathbf{r}, \mathbf{r}^{N-1})]} \right] \quad (7)$$

where  $\mathbf{r}$  is the position before the random displacement  $\Delta \mathbf{r}$  of the selected atom and  $[E(\mathbf{r} + \Delta \mathbf{r}, \mathbf{r}^{N-1}) - E(\mathbf{r}, \mathbf{r}^{N-1})]$  is the potential-energy difference between the system with one atom displaced and all the other  $N - 1$  atoms kept in place, and the system before displacement. In MC schemes, one cycle is the application of the attempted displacement  $N$  times, so that on average each atom is attempted to be displacement once.

According to MD, the forces among atoms are calculated and the Newton equation is numerically integrated in order to obtain one displacement step for all atoms<sup>30</sup>. This scheme samples the constant energy, constant  $V$ , constant  $N$  ensemble (microcanonical). In order to sample the canonical ensemble, the velocities of the

atoms need to be modified in order to obey the Maxwell-Boltzmann distribution at the desired  $T$ . This is achieved via numerical thermostats<sup>30</sup>.

The choice between the two schemes, MC or MD, for the canonical sampling step of our REGMC or REGCMD algorithm is dictated only by convenience. In both case our choice is to perform few (about 10) MD steps or MC cycles between two applications of the REMC step, in order to take full advantage of the enhanced sampling allowed by the REGC accepted moves.

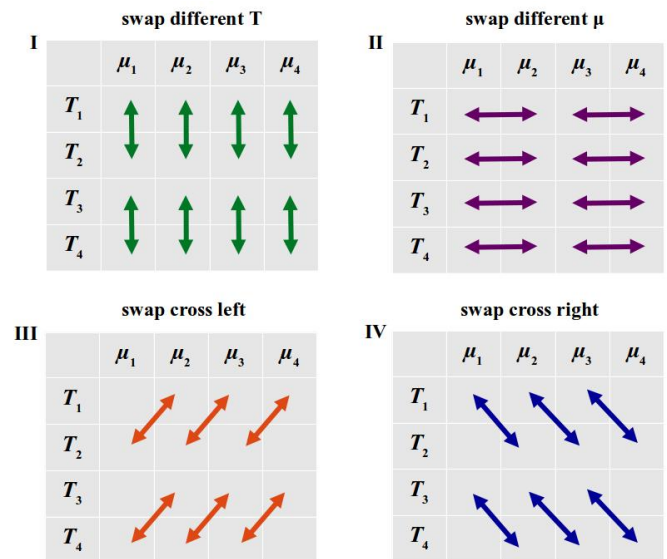


FIG. 2. The 2D schematic of of Replica-Exchange Grand-Canonical method.

### 4. Implementation

Due to the inherently parallel nature of replica exchange, the REGC method is particularly suitable to implement on super computers in parallel. MD or MC simulation of each replica at different  $T$  are performed simultaneously and independently for the same time steps/MC moves. The whole computation resources are propotional to the number of replicas  $S$ , e.g., if each replica requires  $q$  cores, in total,  $S \times q$  cores are assigned to this REGC simulation.

## B. Calculating Phase Diagrams

After a REGC simulation, we obtain  $\Omega_{l,m}$  equilibrium samples from each of the  $S = L \times M$  thermodynamic states  $(\mu_m, V, T_l)$  within the grand-canonical ensemble. For each sample, a wide range of observable values can be collected, starting from the potential energy, the number of particles, and going to properties that are not related to the sampling rules. For instance, structural quantities

like the radial distribution function or electronic properties such as the HOMO-LUMO gap of the system. In order to construct a phase diagram for the studied system, one has first to define which *phases* are of interest. For instance, we can define as one phase all samples with the same number of particles  $N$ . The task is then to evaluate the free energy  $f_i(\mu, T)$  of phase  $i$ , as function of  $\mu$  and  $T$ , and for each value of  $(\mu, T)$  the most stable phase is the one with lowest free energy. From textbook statistical mechanics, the free energy is related to the probability  $p_i$  to find the sampled system in a certain phase (i.e., having a certain value of an observable quantity) as follows:

$$\begin{aligned} f_i(\mu, T) &= -k_B T \ln p_i(\mu, T) \\ &= -k_B T \ln \frac{\int_{\Gamma} d\mathbf{R} \chi_i(\mathbf{R}) q(\mathbf{R}; \mu, \beta)}{\int_{\Gamma} d\mathbf{R} q(\mathbf{R}; \mu, \beta)} \end{aligned} \quad (8)$$

where  $\mathbf{R}$  denotes the configuration of the system,  $\chi_i$  is the indicator function for the state  $i$  — e.g., equal to 1 when  $N$  is a given  $N^*$  and 0 otherwise —, and  $q(\mathbf{R}; \mu, \beta)$  is the density function for the specific statistical ensemble. The integrals are over the whole configuration space  $\Gamma$ .

The normalization term at the denominator of Eq. 8 is known as the *partition function*,  $c(\mu, \beta)$ . Once  $q(\mathbf{R}; \mu, \beta)$  is defined for the sampled ensemble (see further), the nontrivial task is to estimate  $c(\mu, \beta)$ , in order to evaluate the free energy and find its minimum.

To efficiently estimate the partition function from our REGC sampling, we adopted the multistate Bennett acceptance ratio (MBAR)<sup>31</sup> approach, as implemented in the pymbar code (<https://github.com/choderalab/pymbar>). The MBAR method starts from defining the reduced potential function for the grand-canonical ensemble  $U(\mathbf{R}; \mu, \beta)$  for state  $(\mu, \beta)$ <sup>31</sup>:

$$U(\mathbf{R}; \mu, \beta) = \beta \left[ E(\mathbf{R}) - \mu N(\mathbf{R}) \right] \quad (9)$$

where  $N(\mathbf{R})$  is the number of particles for the considered configuration. We note that there is a sign mistake in front of  $\mu N$  for the corresponding formula in the original MBAR paper<sup>32</sup>. The grand-canonical density function is then  $q(\mathbf{R}; \mu, \beta) = \exp[-U(\mathbf{R}; \mu, \beta)]$ .

The MBAR approach provides the lowest-variance estimator for  $c(\mu, \beta)$ , first by determining its value over the set of actually sampled states, via the set of coupled nonlinear equations<sup>31</sup>:

$$\hat{c}_{l,m} = \frac{\sum_{l=1}^L \sum_{m=1}^M \sum_{i=1}^{\Omega_{l,m}} q(\mathbf{R}_{i,l,m}; \mu_m, \beta_l)}{\sum_{l=1}^L \sum_{m=1}^M \Omega_{l,m} \hat{c}_{l,m}^{-1} q(\mathbf{R}_{i,l,m}; \mu_m, \beta_l)} \quad (10)$$

where the index  $i$  runs over all the samples in one state. Crucially, all samples enter the estimator for  $\hat{c}_{l,m}$ , at state  $(l, m)$ , irrespective of the state they were sampled in. Once the set of equations for the  $L \times M$   $\hat{c}_{l,m}$ 's is solved,  $c(\mu, \beta)$  can be estimated for any new state  $(\mu, \beta)$  via the same formula, with the observation that the  $\hat{c}_{l,m}$ 's at the denominator are now known.

Next, Eq. 8 can be evaluated. Following the example where the phase  $i$  is identified by the number of particles in the system, the values of  $N$  that minimizes  $f_i(\mu, \beta)$  is the stable phase at the particular value of  $(\mu, \beta)$ . Graphically, one can assign a color to each value of  $N$  and, for each  $(\mu_i, \beta_j)$  on a grid, the color is assigned to a pixel of size  $(\delta\mu, \delta\beta)$  centered at  $(\mu_i, \beta_j)$  (see Fig. 3).

In order to obtain a more familiar  $(p, T)$  phase diagram from the evaluated  $(\mu, \beta)$ , we use the relationship  $\mu(p, T) = k_B T \ln(p/p_0)$ , where  $p_0$  is chosen such that  $-k_B T \ln(p_0)$  summarizes all the pressure-independent components of  $\mu$ , i.e., translational, rotational, etc. degrees of freedom.<sup>6,33,34</sup>

We now turn our attention to evaluating the ensemble-averaged value of some property, at a given state point  $(\mu, \beta)$ . To give a concrete example for which we actually give results in section III A 2, let's consider the radial distribution function  $g(r)$ , i.e., the probability to find a particle at a given distance  $r$  from any selected particles, averaged over all particles and samples. Here, we are in particular interested in the average (or expected) value of a property like  $g(r)$  when the system is in a given phase, e.g., has a certain number of particles  $N$ . The ensemble average value of  $g(r)$  at a given  $r$  and given state point  $(\mu, \beta)$ , and phase  $i$  is:

$$\langle g(r) \rangle_{\mu, \beta, i} = \frac{\int_{\Gamma} d\mathbf{R} \chi_i(\mathbf{R}) g(r; \mathbf{R}) q(\mathbf{R}; \mu, \beta)}{\int_{\Gamma} d\mathbf{R} q(\mathbf{R}; \mu, \beta)} \quad (11)$$

where the function  $g(r; \mathbf{R})$  at any given  $r$  depends on the whole configuration  $\mathbf{R}$ . In the MBAR formalism, the integrals are estimated over the sampled points via:

$$\langle g(r) \rangle_{\mu, \beta, i} = \sum_{n=1}^{\Omega_i} \frac{g(r; \mathbf{R}_n) c_{\mu, \beta}^{-1} q(\mathbf{R}_n; \mu, \beta)}{\sum_{l,m} \Omega_{l,m} \Omega_{l,m,i} c_{\mu_m, \beta_l}^{-1} q(\mathbf{R}_{l,m,i}; \mu_m, \beta_l)} \quad (12)$$

where  $\Omega_i$  is the number of samples in phase  $i$  and therefore the sum over  $n$  runs over all samples belonging to phase  $i$ . Similarly,  $\Omega_{l,m,i}$  is the number of samples in phase  $i$  in each sampled state point  $(m, l)$ . In practice,  $g(r)$  is discretized into a histogram, in which bin  $k$  counts how many particles are found between distance  $r_{k-1}$  and  $r_k$  (see section section III A 2 for more details). One should note that the average value of each bin in the histogram is evaluated independently by MBAR.

### III. RESULTS

#### A. Lennard-Jones surface

As first example, we applied our REGC algorithm to a two-species Lennard-Jones (LJ) system, consisting of a fcc(111) frozen surface of species A, in contact with a gas phase of species-B particles. Details on the interactions between BB and AB LJ particles are given in the Appendix, here we mention that we chose them so that AB

interactions are much stronger than BB (being the A particles frozen, there is no interaction defined among them). The equilibrium distances  $d_{ij}^{\text{eq}}$  are mismatched such that  $d_{AB}^{\text{eq}} > d_{BB}^{\text{eq}}$ , and both are shorter than the fixed AA first-neighbor distances. Other choices are possible, but here we focus on only one choice, in order to show in depth the type of *a posteriori* analysis an REGC run allows for. The sub-system labeled as  $A_{18}$  is a 2-layer slab with a  $3 \times 3$  lateral supercell (i.e., 18 A atoms), periodically replicated in the  $x$  and  $y$  direction, while the  $z$  direction is aligned with the [111] direction of the slab. The gas particle B is only allowed to insert in the “surface” zone. We defined the “surface” zone as a slab of height 48.0 Å above (i.e., in the positive  $z$  direction), starting from the  $z$  position of the topmost atoms of  $A_{18}$ . At the same time, particles B are inserted at all  $x$  and  $y$  coordinates, uniformly. Insertion and deletion attempts have been performed with equal probabilities. Ten sequential MC moves are performed after each particle/replica exchange attempt. In the calculations, 160 replicas are defined i.e., 10 temperatures ranging from 200 to 650 K, with an interval of 50 K, and 16 chemical potentials ranging from -2.4 to -0.9 eV, with an interval of 0.1 eV. The range of chemical potentials is selected such that the lowest value of  $\mu$  is comparable to and slightly lower than the adsorption energy of one B particle on  $A_{18}$ , in order to assure that the sampling includes states where zero or few particles are adsorbed (in order to have the pristine surface appearing in the phase diagram). The highest value of  $\mu$  is ideally always close to zero, in order to scan up to the condensation of B particles and formation of a bulk B phase. The range of temperature was chosen to be slightly lower than the solid/liquid/gas triple point of the B particles and ranging to few times (here, four) its critical temperature<sup>35</sup>. In practice, pre-knowledge of the studied system can be applied in order to frame a suitable  $(\mu, T)$  window containing phases of interest. The spacing between  $T$  and  $\mu$  values is more difficult to estimate *a priori*. During the simulation, one has to check that the acceptance ratio of RE attempted moves is not too low, in order to ensure a proper diffusion of replicas in the  $(\mu, T)$  window. For instance, the present choice ensured an acceptance ratio of about 25%. Configuration swaps were attempted every 100 REGC steps, and  $x_0$  was set equal to 0.99; a total of  $1.2 \times 10^5$  REGC steps were performed to reach convergence, that is, there was no change in the density of reduced-energy states  $\rho(U)$ , with increasing simulation steps. The density  $\rho(U)$  is sampled by binning the sampled configurations according to their value of  $U$ .

### 1. Phase diagram

The phase diagram shown in Fig. 3(a) is constructed by using MBAR and shows the  $(p_B, T)$  regions where different number of adsorbed B particles are in thermodynamic equilibrium with their gas phase. The B

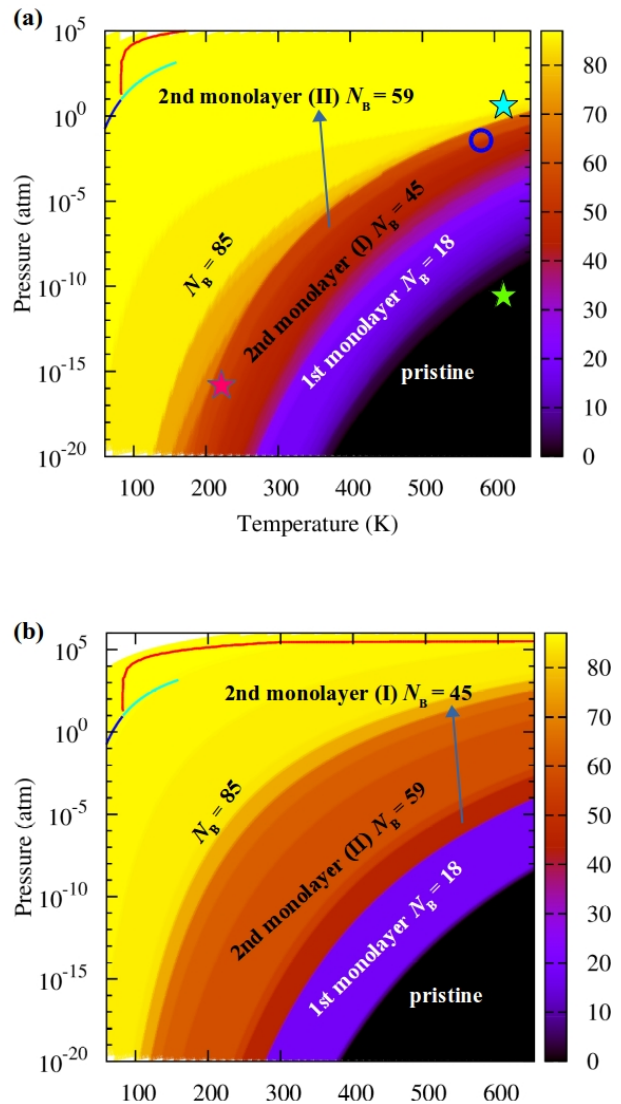


FIG. 3. Phase diagrams of a LJ gas-phase (particles B) in contact with a frozen fcc(111) LJ frozen surface calculated via by MBAR from the REGC sampling (panel a) and aiAT (panel b) at  $(p_B, T)$  conditions corresponding to a range from zero adsorbed particles (all in gas phase, region labeled as “pristine”, referred to the surface) to the deposition of the LJ B particles into a bulk solid. The red line is the melting line for the LJ B particles, the sublimation line is blue, and the vaporization line is cyan. The cyan, green and pink stars correspond to the “corner” states for the REGC sampling: (650 K, -0.9 eV), (650 K, -2.4 eV) and (200 K, -0.9 eV), respectively. The fourth corner, (200 K, -2.4 eV) falls outside the  $(p, T)$  window shown in the plot. The blue circle indicates (600 K,  $8.89 \times 10^{-2}$  atm) and (200 K,  $2.03 \times 10^{-17}$  atm) is exactly the pink star, corresponding to two states in Fig. 6c and Fig. 6b, respectively.

reservoir is assumed to be an ideal gas, so the chemical potential of the reference state is defined as  $\mu_{id.gas}^0 \equiv k_B T \ln(\Lambda^3)$ . The relationship between pressure  $p_{id.gas}$  in the reservoir and the chemical potential  $\mu$  is  $\beta\mu \equiv$

$\beta\mu_{id,gas}^0 + \ln(\beta p_{id,gas})$ , that is  $p_0 = (k_B T)^{\frac{5}{2}} (\frac{2\pi m}{h^2})^{\frac{3}{2}}$ . The whole output data of REGCMC is sub-sampled every 100 REGC steps, that is, recording data after every attempted replica exchange, to remove correlations in the sampled quantities.

The MBAR@REGC phase diagram is compared to the aiAT@REGC phase diagram (Fig. 3(b)), which is calculated via the following steps: (i) For each observed number  $N_B$  of adsorbed (B) particles in the REGCMC sampling, the lowest energy configuration is selected. We note that identifying phases (the phase is identified by  $N_B$ ) via grand-canonical sampling is not the usual strategy for aiAT. Typically phases are enumerated on the basis of pre-knowledge and local minimization (at fixed number of adsorbed particles. In other words, the aiAT study presented in this paragraph is already richer than usual due to the unbiased structure sampling. (ii) The formation Gibbs free energy for each of these phases is calculated via:

$$\Delta G_{N_B}^f(T, p_B) = F_{N_B} - F_{A_{18}} - N_B \mu(T, p_B) \quad (13)$$

Here, the free energy of  $F_{N_B}$  of the system  $A_{18}B_{N_B}$  and  $F_{A_{18}}$  of the pristine  $A_{18}$  slab is approximated by the LJ energies of the two systems, i.e., all the vibrational contributions to the free energy are assumed to cancel out. This is often a justified assumption for systems studies via aiAT<sup>6</sup>. As we will see, it is not a good approximation for this LJ system, at least at larger  $N_B$ . (iii) As for MBAR@REGC, at each  $(T_i, \mu_j)$  on a grid the phase with the lowest  $\Delta G^f$  determines the color of the pixel of size  $(\delta T, \delta \mu)$  centered at  $(T_i, \mu_j)$ . This aiAT@REGC approach, used here only for comparing to MBAR@REGC in order single out the role of the vibrational contribution to free energy, including anharmonic effects, is similar to the method recently proposed in Ref. 36. There, the configurations are sampled by means of an approximated GC scheme at one temperature only and without replica exchange for either temperature or chemical potential. The effect of the reservoir to the free energy is taken care of by an expression similar to Eq. 13.

By comparing the two panels of Fig. 3, we note that up to  $N_B = 18$ , the two phase diagrams almost coincide, especially at lower temperatures (in the Suppl. Material, we show a zoom-in of the region between 60 and 350 K). There are, however, significant differences at larger  $N_B$ : There are many more phases in Fig. 3(a) that are missing in Fig. 3(b) for  $N_B > 18$  and the region of stability of larger coverages is shifted to higher temperatures and lower pressures. This can be understood as due to increasingly larger vibrational contributions, especially in the direction  $z$ , perpendicular to the slab, while at low coverage the free energy is indeed essentially given by the LJ energy. We come back to this in the next section, after analyzing the structural properties of the different phases.

The analysis of the phase diagram Fig. 3(b) reveals that for many values of number of adsorbed B particles,  $N_B$ , there is a region of stability in the phase diagram,

however, for some specific values of  $N_B$  larger stability areas are found. Besides  $N_B = 0$  (the pristine surface), we recognize  $N_B = 18$  as the first complete mono-layer,  $N_B = 45$  as the addition of a second complete mono-layer, plus a third phase,  $N_B = 59$  with a thicker second monolayer (see further). We also identify a large-coverage phase,  $N_B = 85$  which can be described by the formation of a “third” layer around 1.9 Å, but in this case the particle distribution does not go completely to zero between second and third layer as it does between first and second, as shown in Fig. 4. The diagram extends till the melting (red), vaporization (cyan), and sublimation (blue) line for bulk B particles. The phase transition curves are derived from the published equations of state for the LJ system.<sup>37–43</sup> We underline that the phase diagram outside the  $(p, T)$  region sampled directly via the REGC run is not *extrapolated*. It is obtained as for all the diagram by Boltzmann re-sampling the configurations actually visited, using the measured (reduced) potential energies.

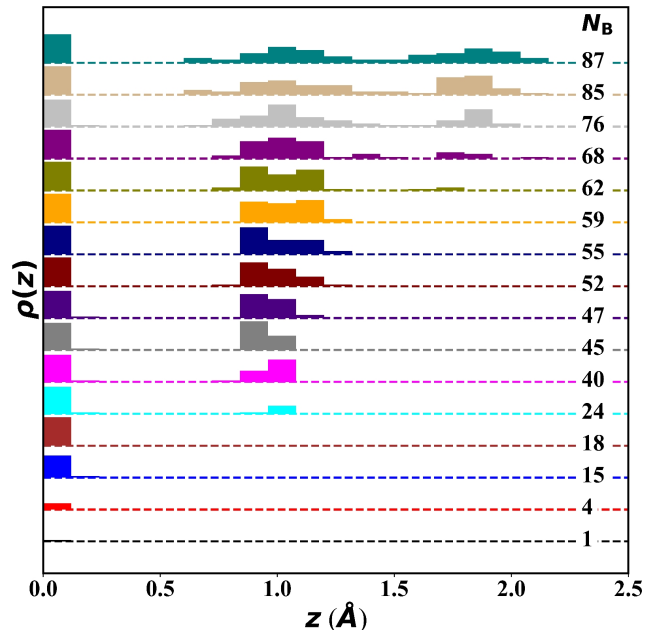


FIG. 4. Axial distribution function of adsorbed particles for each  $N_B$  composition generated in REGC sampling. The curves are displaced by 20 units and each dash line is a zero reference line for the curve with the same color.

## 2. Structural properties

The REGC sampling allows for much deeper analysis than the evaluation of the phase diagram. For instance, the structural properties of the adsorbed phases can be characterized in a statistical way. The axial distribution function  $\rho(z)$  was calculated by dividing the cell into slabs of width 0.12 Å, parallel to the surface, and collecting a

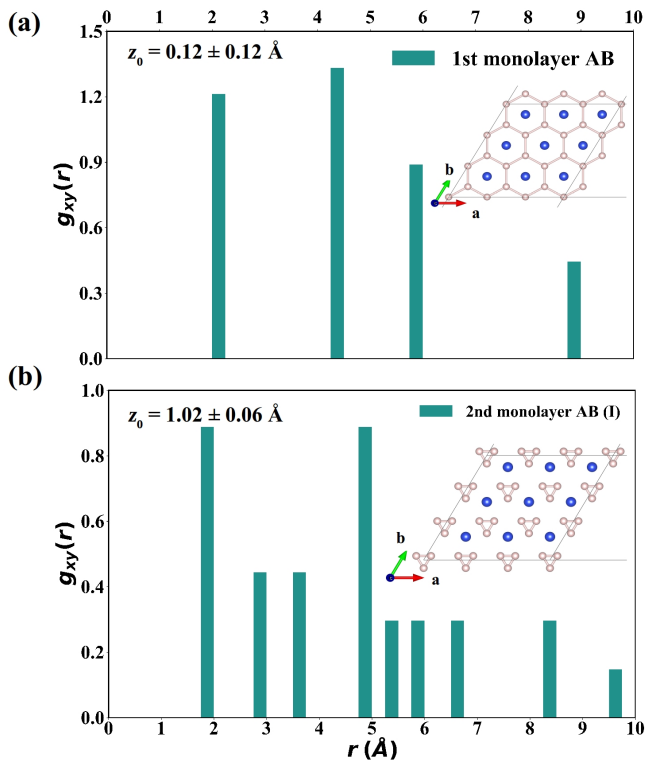


FIG. 5. Lateral radial distribution functions  $g_{xy}(r)$  for (a) first monolayer, and (b) second monolayer (I), respectively. The blue and pink balls in the insets indicate A and B particles, respectively.

histogram of the number of particles in each slab along the REGC sampling. As shown in Fig. 4, the adsorbate has a clear layered structure up to the second layer. For larger  $N_B$ , i.e.,  $N_B > 59$ , there are more and more particles adsorbed in the range  $1.2 \leq z \leq 1.8 \text{ \AA}$ , though another noticeable peak around  $1.9 \text{ \AA}$  occurs. As intuitively predictable, the first layer consists of 18 B particles located in all the hollow sites of the  $3 \times 3$  surface. When the second full monolayer  $N_B = 45$  is stable, the B particles occupy the 27 bridge sites of the  $A_9$  surface layer.

To better characterize the structure of the adsorbate layers, in Figs. 5a–b and Figs. 6a–c we show the  $g_{xy}(r)$ , i.e., the radial distribution functions (RDF) in the  $xy$ -plane for the different adsorbate layers (i.e., for B-particles in a slab  $z_0 \pm \delta z_0$  as specified in each panel). The structures shown in Fig. 6b–c are obtained via MBAR by evaluating Eq. 12. We observe that the first monolayer and second monolayer (I)  $N_B = 45$  have a  $g_{xy}(r)$  characteristic of the solid phase with well-defined peaks and long-range order, whereas for the second monolayer (II)  $N_B = 59$ , the  $g_{xy}(r)$  is more disordered. In the relaxed structure of  $A_9B_{59}$  (Fig. 6a), B particles occupy approximately both hollow and bridge sites, relative to the top  $A_9$  layer and form a ring-like structure around the projection of the A particles. At (200 K,  $2.03 \times 10^{-17}$  atm), the average radial distribution function  $\langle g_{xy}(r) \rangle$  of this

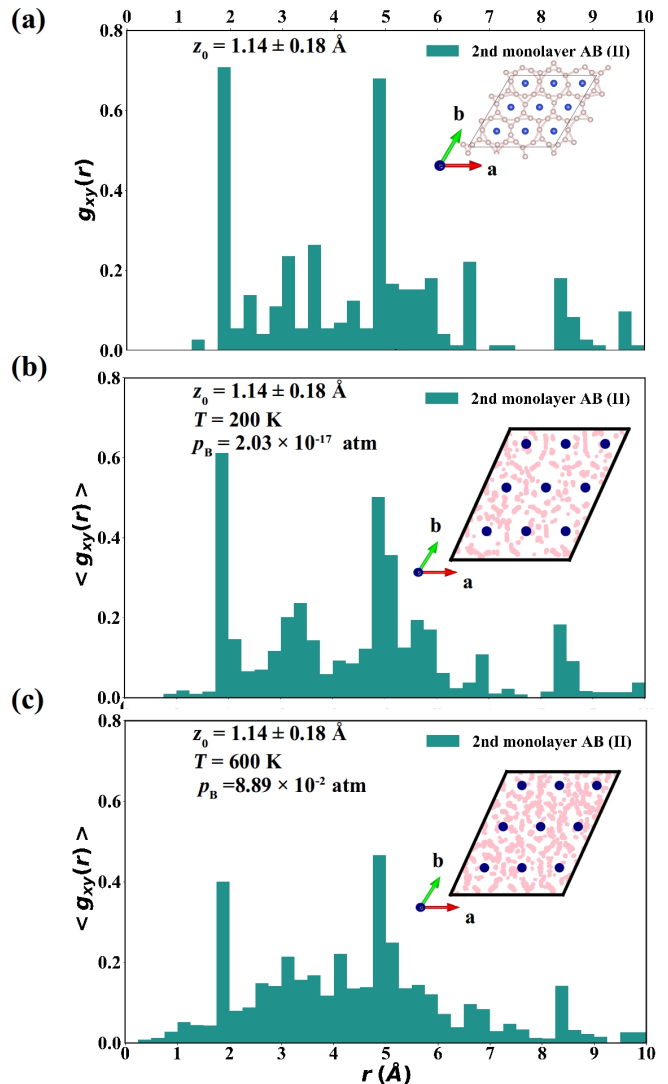


FIG. 6. Lateral radial distribution function  $g_{xy}(r)$  for (a) relaxed second monolayer (II)  $A_9B_{59}$ , average distribution function  $\langle g_{xy}(r) \rangle$  at (200 K,  $2.03 \times 10^{-17}$  atm) state (b), and at (600 K,  $8.89 \times 10^{-2}$  atm) (c) for the same composition. The blue and pink balls in the insets indicate A and B particles, respectively.

phase shares some similar peak positions with that of its lowest-energy isomer. It is clear that the ring structure formed by B particles can be still found in the average adsorbate structure though there are a few B particles diffusing around the projection of the A particles. At (600 K,  $8.89 \times 10^{-2}$  atm), more and more B particles diffuse and the ring structure is not as noticeable as before. Consistently, the  $\langle g_{xy}(r) \rangle$  shares a few major peaks with that of its lowest-energy isomer, but they appear more smeared.

The example of  $\langle g_{xy}(r) \rangle$  at the two state points was selected in order to demonstrate the power of the REGC sampling to reveal detailed thermodynamic information on the simulated system. A crucial observation is that



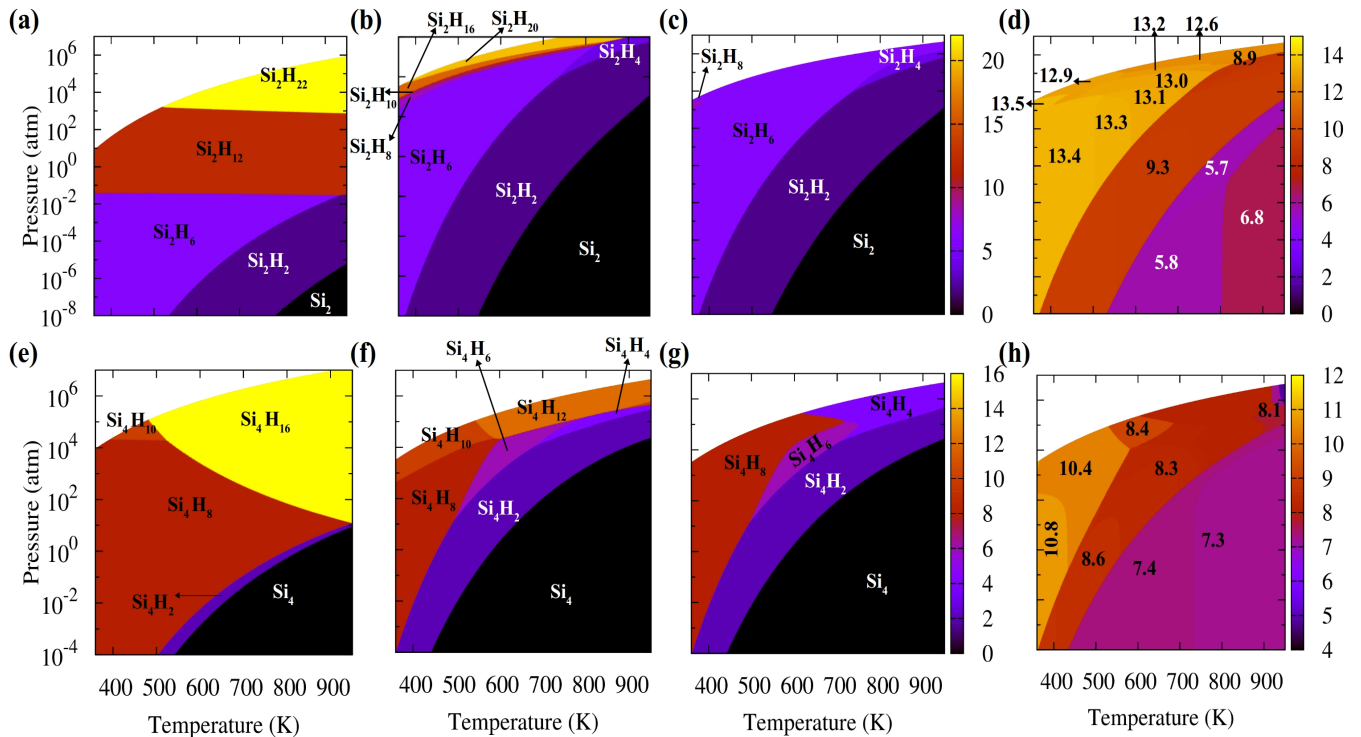


FIG. 7. Phase diagrams of  $\text{Si}_2$  with  $\text{H}_2$  reactive gas phase calculated by (a) aiAT@REGC (b) MBAR@REGC. MBAR@REGC phase diagrams of (c) chemisorbed  $\text{Si}_2\text{H}_N$  and (d) HOMO-LUMO gap of  $\text{Si}_2\text{H}_N$ . Phase diagrams of  $\text{Si}_4$  with  $\text{H}_2$  reactive gas phase calculated by (e) aiAT@REGC and (f) MBAR@REGC. MBAR@REGC phase diagrams of (g) chemisorbed  $\text{Si}_4\text{H}_N$  and (h) HOMO-LUMO gap of  $\text{Si}_4\text{H}_N$  at PBE0 level. HOMO-LUMO gaps in panels (d) and (h) are in eV.

such information is already contained in the REGC sampling, no further simulation is needed, only post-processing statistical analysis of the sampled data points is required.

Coming back to the differences between aiAT@REGC and MBAR@REGC phase diagrams (Fig. 3), we observe, in Fig. 4 that up to the complete first monolayer ( $N_B = 18$ ), the adsorbed particles have essentially no freedom to move in the  $z$  direction. As soon as the second monolayer is established, the adsorbed particles display a broader and broader distribution along the  $z$  direction. The distribution becomes even bimodal for  $N_B \geq 62$ . This enhanced configurational freedom creates a large, negative, vibrational free energy contribution that stabilizes the higher coverages compared to when only the energetic contribution is taken into account (as in the aiAT@REGC phase diagram).

## B. *Ab initio* $\text{Si}_2\text{H}_N$ and $\text{Si}_4\text{H}_N$ clusters

The REGC algorithm coupled to *ab initio* MD was applied to identify the thermodynamically stable and metastable compositions and structures of  $\text{Si}_M\text{H}_N$  ( $M=2, 4$ ) clusters at realistic temperatures and pressure of the molecular hydrogen gas.

### 1. Phase diagram

*a.  $\text{Si}_2$*  Twenty replicas of  $\text{Si}_2$  are selected in contact with different thermodynamic states, that is, with temperatures of 500, 650, 800, and 950 K and  $\text{H}_2$  chemical potentials of -0.2, -0.16, -0.12, -0.08, and -0.05 eV. The selection of the temperature range is made according to the experimental deposition temperature of chemical vapor deposited silicon films<sup>44,45</sup>, which starts from around 600 K. Ideally, the lowest  $\mu_{\text{H}}$  should be around  $-1.2$  eV, which is the half adsorption energy of  $\text{H}_2$  on  $\text{Si}_2$ , according to our DFT calculations (see details in appendix). However, in order to focus the sampling on a more interesting region, where more H atoms are adsorbed, we started from a much higher minimum  $\mu_{\text{H}_2}$ . The studied  $\text{Si}_{2,4}\text{H}_N$  systems are confined in a sphere with radius 4 Å, by applying reflecting boundaries. This avoids that H atoms diffuse at arbitrary distance from the  $\text{Si}_M$  cluster, without perturbing the statistics as the cutoff distance is such that the H atoms are not any more interacting with the Si cluster. *Ab initio* molecular dynamics is performed for each system after exchanging particle with the reservoir or swapping with neighboring replicas. For this REGCMD study,  $x_0$  is chosen as 0.9. For comparison, we analyzed the stability of  $\text{Si}_2\text{H}_N$  clusters using *ab initio* atomistic thermodynamics (aT) in Fig. 7a. For

each number of adsorbed hydrogens  $N_H$ , the lowest DFT energy isomer is identified among all the configurations obtained along the REGC *ab initio* MD sampling. The Gibbs free energy of each phase is calculated as:

$$\Delta G_f(T, p_{H_2}) = F_{Si_{2,4}H_N} - F_{Si_{2,4}} - N\mu_H(T, p_{H_2}) \quad (14)$$

Here,  $F_{Si_{2,4}H_N}$  and  $F_{Si_{2,4}}$  are the Helmholtz free energies of the  $Si_{2,4}H_N$  and the pristine  $Si_{2,4}$  cluster (at their configurational ground state), respectively.  $\mu_{H_2}$  is the chemical potential of the hydrogen molecule.  $F_{Si_{2,4}H_N}$  and  $F_{Si_{2,4}}$  are calculated using DFT information and are expressed as the sum of DFT total energy, DFT vibrational free energy in the quasi-harmonic approximation, as well as translational, and rotational free-energy contributions. The dependence of  $\mu_{H_2}$  on  $T$  and  $p_{H_2}$  is calculated using the ideal (diatomic) gas approximation with the same DFT functional as for the clusters.<sup>33,34,46</sup> So  $p_0$  here is calculated as follows:

$$p_0 = \left[ \left( \frac{2\pi m}{h^2} \right)^{\frac{3}{2}} (k_B T)^{\frac{5}{2}} \left( \frac{8\pi^2 I_A k_B T}{h^2} \right) \frac{e^{\left( \frac{k_B T}{E_{DFT}} \right)}}{e^{\left( \frac{h\nu_{HH}}{k_B T} \right) - 1}} \right] \quad (15)$$

$E_{DFT}$  is the DFT total energy,  $m$  is the mass,  $I_A$  is the inertia moments,  $\nu_{HH}$  is the H-H stretching frequency of  $3080 \text{ cm}^{-1}$ , and  $E_{DFT}$  of  $-31.74 \text{ eV}$ . The  $(p_{H_2}, T)$  phase diagram of  $Si_2H_N$  cluster is also constructed via the MBAR@REGC method. As shown in Fig. 7b, besides  $Si_2$ ,  $Si_2H_2$ , and  $Si_2H_6$ , which have their wide stability regions revealed in both phase diagrams, there is a narrow  $(T, p_{H_2})$  stability domain for  $Si_2H_4$ , which is only revealed by the MBAR@REGC phase diagram that includes without approximation all the anharmonic contributions to the free energy. Another difference between two phase diagrams is that the stable  $(p_{H_2}, T)$  range of each phase is quite different. The  $Si_2H_N$  phases in Fig. 7b include not only chemically adsorbed H atom, but also  $H_2$  molecule or isolated H atoms. In order to further investigate the chemisorbed phase stability, we construct the phase diagram (Fig. 7c) for a new observable: the number of adsorbed H atoms. A H atom is considered adsorbed on the Si cluster when the distance to the closest Si is smaller than  $1.7 \text{ \AA}$ .

*b.  $Si_4$*  Twenty thermodynamic states for the  $Si_4H_N$  system are selected, with temperature of 560, 685, 810, and 935 K, and chemical potentials of -0.3, -0.2, -0.17, -0.14, and -0.11 eV. The lowest value of  $\mu_H$  is selected as a bit larger than the half adsorption energy (-0.6 eV) of  $H_2$  on  $Si_4$ . The other settings are the same as in  $Si_2$  simulation. As for the  $Si_2H_N$  case, we construct both the aiAT@REGC and MBAR@REGC phase diagram, for comparison, plus the MBAR@REGC phase diagram for the adsorbed H atoms. In Fig 7e and 7f, the results indicate that two stable  $Si_4H_4$  and  $Si_4H_6$  are missing in aT phase diagram.  $Si_4H_4$  and  $Si_4H_6$  have considerable larger stable range in chemisorbed phase diagram shown in Fig 7g than in both physi- and chemisorbed one. Besides, the stable  $(p_{H_2}, T)$  range of each phase transitions are quite different in phase diagrams calculated by two method.

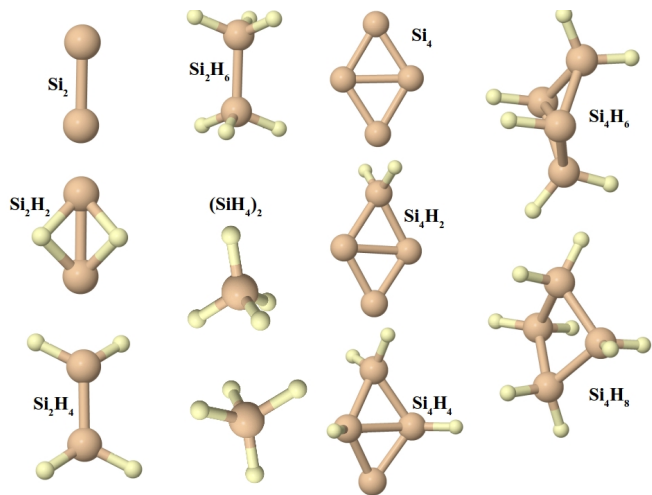


FIG. 8. Structures of  $Si_2H_N$  and  $Si_4H_N$ , found by the REGC sampling, that have a region of thermodynamic stability in the phase diagrams of Fig. 7.

## 2. Structural and electronic properties of silicon hydrides

In Fig. 8, we show the structures of each thermodynamically stable cluster size appearing in the phase diagrams. All previously reported structures are found in our REGC *ab initio* MD simulations and illustrated in Fig. S2. Besides, we identified many other isomers at each composition, via the REGC *ab initio* MD sampling, as shown in Fig. S2.

The HOMO-LUMO gap  $E_g$  is also chosen as further observable for the evaluation of phase diagrams for  $Si_2H_N$  Fig. 7d and  $Si_4H_N$  Fig. 7h.  $E_g$  is evaluated as the difference between the vertical electron affinity (VEA) and vertical ionization potential (VIP). The VEA (VIP) is evaluated —via the PBE0 hybrid<sup>47</sup> xc functional, with the Tkatchenko-Scheffler<sup>48</sup> pairwise vdW correction— as the energy difference between the neutral cluster and its monovalent anion (cation), at fixed geometry of the neutral species. It has been clearly shown Fig. 7d and 7h that the HOMO-LUMO gap increases with increasing  $N_H$  for both  $Si_2H_N$  and  $Si_4H_N$ , as the VEA decreases with increasing  $N_H$  (Fig. S1b and S1e) while VIP increases (Fig. S1c and S1f). This electronic-structure phase diagram can be used to provide guidance to synthesize the material with desired electronic properties, by tuning the environmental conditions, i.e, the temperature and pressure of reactive gas phase.

## IV. CONCLUSION AND OUTLOOK

In summary, we have developed a massively parallel Replica-Exchange Grand-Canonical Monte Carlo/*ab initio* Molecular-Dynamics (REGCMC/MD) algorithm to perform simulations on surfaces/nanoclusters in contact with reactive  $(T, p)$  gas and demonstrated how it

can be used, in combination with the multistate-Bennett-acceptance-ratio (MBAR) reweighting approach to determine  $(T, p)$  phase diagrams. This massively parallel algorithm requires no prior knowledge of the phase diagram and takes only the potential energy function together with the desired  $\mu$  and  $T$  ranges as inputs. The particle insertion/removal Monte Carlo move, which implements the GC sampling, together with the exchange of configurations among thermodynamic states introduced by RE, allows for an efficient sampling of the configurational space. The approach is applied to a model surface described by the Lennard-Jones empirical force-fields and small Si clusters in reactive  $H_2$  atmosphere described at the *ab initio* DFT level. Besides free-energy  $(T, p)$  phase diagrams, the combination of the REGC sampling and *a posteriori* analysis via MBAR allows for the determination of phase diagrams for any (atom position dependent) observables, therefore indicating how to tune the environmental condition ( $T$  and  $p$ ) to get a material with desired properties. It can therefore be applied to a wide range of practical issues, e.g., dopant profiles, surface segregation, crystal growth and more.

Such as an undertaking has its limitation in the cost of *ab initio* molecular dynamics needed for the REGC sampling. However, its embarrassingly parallel nature makes our approach “towards exascale” friendly, and can be regarded as a very efficient and internally consistent high-throughput approach. An obvious and indeed currently investigated generalization of the method is to consider more than one reactive gas in the so-called “constrained equilibrium”<sup>6,7</sup> (different species do not react in the gas phase, but only at the surface). In order to avoid a dimensional explosion, an algorithm with an adaptive  $\mu_i$  grid is under development.

## V. ACKNOWLEDGEMENTS

We thank Fawzi R. Mohamed for crucial help in the parallel implementation of the REGC algorithm and helpful discussions. We thank Chunye Zhu for helpful discussions and a critical reading of the manuscript.

- 
- <sup>1</sup> C. Weinert and M. Scheffler, in *Materials Science Forum*, Vol. 10 (Trans Tech Publ, 1986) pp. 25–30.
  - <sup>2</sup> M. Scheffler, *Physics of solid surfaces-1987*, edited by J. Koukal (Elsevier, Amsterdam, 1988).
  - <sup>3</sup> E. Kaxiras, Y. Bar-Yam, J. D. Joannopoulos, and K. C. Pandey, *Phys. Rev. B* **35**, 9625 (1987).
  - <sup>4</sup> G.-X. Qian, R. M. Martin, and D. J. Chadi, *Phys. Rev. B* **38**, 7649 (1988).
  - <sup>5</sup> N. Moll, M. Scheffler, and E. Pehlke, *Phys. Rev. B* **58**, 4566 (1998).
  - <sup>6</sup> K. Reuter and M. Scheffler, *Phys. Rev. B* **68**, 045407 (2003).
  - <sup>7</sup> K. Reuter and M. Scheffler, *Phys. Rev. Lett.* **90**, 046103 (2003).
  - <sup>8</sup> E. C. Beret, M. M. van Wijk, and L. M. Ghiringhelli, *International Journal of Quantum Chemistry* **114**, 57 (2014).
  - <sup>9</sup> S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, *Phys. Rev. Lett.* **111**, 135501 (2013).
  - <sup>10</sup> S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, *New Journal of Physics* **16**, 123016 (2014).
  - <sup>11</sup> R. H. Swendsen and J.-S. Wang, *Phys. Rev. Lett.* **57**, 2607 (1986).
  - <sup>12</sup> E. Marinari and G. Parisi, *EPL (Europhysics Letters)* **19**, 451 (1992).
  - <sup>13</sup> C. J. Geyer and E. A. Thompson, *Journal of the American Statistical Association* **90**, 909 (1995).
  - <sup>14</sup> Y. Sugita and Y. Okamoto, *Chemical Physics Letters* **314**, 141 (1999).
  - <sup>15</sup> Q. Yan and J. J. de Pablo, *The Journal of Chemical Physics* **111**, 9509 (1999).
  - <sup>16</sup> R. Faller, Q. Yan, and J. J. de Pablo, *The Journal of Chemical Physics* **116**, 5419 (2002).
  - <sup>17</sup> J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **78**, 1396 (1997).
  - <sup>18</sup> L. Sari, M. McCarthy, H. F. Schaefer, and P. Thaddeus, *Journal of the American Chemical Society* **125**, 11409 (2003).
  - <sup>19</sup> W.-N. Wang, H.-R. Tang, K.-N. Fan, and S. Iwata, *The Journal of Chemical Physics* **114**, 1278 (2001).
  - <sup>20</sup> C. Xu, T. R. Taylor, G. R. Burton, and D. M. Neumark, *The Journal of chemical physics* **108**, 7645 (1998).
  - <sup>21</sup> A. Kasdan, E. Herbst, and W. Lineberger, *The Journal of Chemical Physics* **62**, 541 (1975).
  - <sup>22</sup> D. D. Frantz, D. L. Freeman, and J. D. Doll, *The Journal of Chemical Physics* **93**, 2769 (1990).
  - <sup>23</sup> W. Ortiz, A. Perlloni, and G. E. López, *Chemical Physics Letters* **298**, 66 (1998).
  - <sup>24</sup> B. A. Berg and T. Neuhaus, *Physics Letters B* **267**, 249 (1991).
  - <sup>25</sup> B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68**, 9 (1992).
  - <sup>26</sup> R. J. N. Baldock, L. B. Pártay, A. P. Bartók, M. C. Payne, and G. Csányi, *Phys. Rev. B* **93**, 174108 (2016).
  - <sup>27</sup> A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, *The Journal of Chemical Physics* **96**, 1776 (1992).
  - <sup>28</sup> B. Hesselbo and R. B. Stinchcombe, *Phys. Rev. Lett.* **74**, 2151 (1995).
  - <sup>29</sup> F. A. Escobedo and J. J. de Pablo, *The Journal of Chemical Physics* **105**, 4391 (1996).
  - <sup>30</sup> D. Frenkel and B. Smit, *Understanding molecular simulation: From algorithms to applications* (Elsevier, 2002).
  - <sup>31</sup> M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics* **129**, 124105 (2008).
  - <sup>32</sup> Private communication with M.R. Shirts.
  - <sup>33</sup> E. C. Beret, M. M. Wijk, and L. M. Ghiringhelli, *International Journal of Quantum Chemistry* **114**, 57 (2014).
  - <sup>34</sup> S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, *New Journal of Physics* **16**, 123016 (2014).

- <sup>35</sup> L. Rowley, D. Nicholson, and N. Parsonage, *Molecular Physics* **31**, 365 (1976).
- <sup>36</sup> R. B. Wexler, T. Qiu, and A. M. Rappe, *The Journal of Physical Chemistry C* **123**, 2321 (2019).
- <sup>37</sup> J. Nicolas, K. Gubbins, W. Streett, and D. Tildesley, *Molecular Physics* **37**, 1429 (1979).
- <sup>38</sup> J. K. Johnson, J. A. Zollweg, and K. E. Gubbins, *Molecular Physics* **78**, 591 (1993).
- <sup>39</sup> J. Kolafa and I. Nezbeda, *Fluid Phase Equilibria* **100**, 1 (1994).
- <sup>40</sup> Y. Tang and B. C.-Y. Lu, *Fluid phase equilibria* **165**, 183 (1999).
- <sup>41</sup> W. Okrasinski, M. Parra, and F. Cuadros, *Physics Letters A* **282**, 36 (2001).
- <sup>42</sup> M. A. Van der Hoef, *The Journal of Chemical Physics* **113**, 8142 (2000).
- <sup>43</sup> M. A. van der Hoef, *The Journal of chemical physics* **117**, 5092 (2002).
- <sup>44</sup> K. Nakazawa, *Journal of Applied Physics* **69**, 1703 (1991).
- <sup>45</sup> P. A. Breddels, H. Kanoh, O. Sugiura, and M. Matsumura, *Japanese Journal of Applied Physics* **30**, 233 (1991).
- <sup>46</sup> S. Bhattacharya, S. V. Levchenko, L. M. Ghiringhelli, and M. Scheffler, *Physical review letters* **111**, 135501 (2013).
- <sup>47</sup> C. Adamo and V. Barone, *The Journal of Chemical Physics* **110**, 6158 (1999).
- <sup>48</sup> A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.* **102**, 073005 (2009).
- <sup>49</sup> V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, *Computer Physics Communications* **180**, 2175 (2009).
- <sup>50</sup> L. Verlet, *Phys. Rev.* **159**, 98 (1967).
- <sup>51</sup> G. Bussi, D. Donadio, and M. Parrinello, *The Journal of Chemical Physics* **126**, 014101 (2007).
- <sup>52</sup> M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli,

F. Pietrucci, R. A. Broglia, and M. Parrinello, *Computer Physics Communications* **180**, 1961 (2009).

## VI. APPENDIX

*a. Force-field calculations* The interaction between particles in the surface and gas phase was taken to be a Lennard-Jones 12-6 potentials  $\phi(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ . The parameters  $\epsilon_{AB}$ , and  $\epsilon_{BB}$  are 0.66 and 0.01 eV, respectively. The  $\sigma_{AA}$ ,  $\sigma_{AB}$ , and  $\sigma_{BB}$  are 2.5, 1.91 and 1.2 Å. The length of the lattice vectors of this 2D hexagonal supercell is 11.489 Å.

*b. First-principles calculations* All DFT calculations were performed with the all-electron, full-potential electronic-structure code package FHI-aims<sup>49</sup>. We used the Perdew-Burke-Ernzerhof (PBE)<sup>17</sup> exchange-correlation functional, with a tail correction for the van der Waals interactions (vdW), computed using the Tkatchenko-Scheffler scheme<sup>48</sup>. A “tier 1” basis for both Si and H with “light” numerical settings were employed. All AIMD (Born–Oppenheimer) trajectories between REGC attempted moves (0.02 ps each) are performed in the *NVT* ensemble. The equations of motion were integrated with a time step of 1 fs using the velocity-Verlet algorithm<sup>50</sup>. The stochastic velocity rescaling thermostat was adopted, with a decay-time parameter  $\tau = 0.02$  ps, to sample the canonical ensemble<sup>51</sup>. The reflecting conditions to confine the system in a sphere of radius 4 Å are imposed via PLUMED<sup>52</sup> interfaced with FHI-aims, by applying a repulsive polynomial potential of order 4.