

# Efficient Bayesian Inference of Atomistic Structure in Complex Functional Materials

Milica Todorović<sup>1a</sup>, Michael U. Gutmann<sup>b</sup>, Jukka Corander<sup>c,d</sup>, and Patrick Rinke<sup>a</sup>

<sup>a</sup>Department of Applied Physics, Aalto University, P.O. Box 11100, Aalto, FI-00076, Finland; <sup>b</sup>School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom; <sup>c</sup>Institute of Basic Medical Sciences, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway; <sup>d</sup>Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, Helsinki, FI-00014, Finland

**Tailoring the functional properties of advanced organic/inorganic heterogeneous devices to their intended technological applications requires knowledge and control of the microscopic structure inside the device. Atomistic quantum mechanical simulation methods deliver accurate energies and properties for individual configurations, however, finding the most favourable configurations remains computationally prohibitive. We propose a 'building block'-based Bayesian Optimisation Structure Search (BOSS) approach for addressing extended organic/inorganic interface problems and demonstrate its feasibility in a molecular surface adsorption study. In BOSS, a Bayesian scheme accelerates the identification of material energy landscapes with the number of sampled configurations during active learning, enabling structural inference with high chemical accuracy and featuring large simulation cells. This allowed us to identify several most favourable molecular adsorption configurations for C<sub>60</sub> on the (101) surface of TiO<sub>2</sub> anatase and clarify the key molecule-surface interactions governing structural assembly. Inferred structures were in good agreement with detailed experimental images of this surface adsorbate, demonstrating good predictive power of BOSS and opening the route towards large-scale surface adsorption studies of molecular aggregates and films.**

**F**rontier technologies are increasingly based on functional hybrid materials - engineered blends of organic molecules and inorganic crystals that harness and enhance the functional properties of both substances to perform specific tasks. Organic/inorganic heterostructures and metal-organic frameworks are key components for smart sensors, membranes and coatings, novel optoelectronic and fuel cell technologies, with further applications in data storage, quantum engineering and nanophotonics on the horizon (1–8). Despite outstanding component materials, engineering the microscopic structure of complex heterostructures to tailor their properties towards desired functionality remains a fundamental challenge in physics, chemistry and materials science. It means bypassing the pitfalls of interface artifacts, defects and unfavorable self-assembled structures that lead to poor overall device performance.

Understanding the microscopic structural details of advanced organic/inorganic material blends has emerged as the primary route towards controlling and engineering the functionality of hybrid materials (2, 9). Here computational studies lead the way (10, 11), since nanoscale experimental measurement techniques frequently lack the necessary atomistic detail, and traditional trial-and-error tests are costly and time-consuming. First-principle methods like density functional theory (DFT) are particularly predictive in simulations of hybrid materials because they accurately describe the delicate interplay of microscopic interactions (e.g. electrostatics, dispersion, bond formation and charge transfer) that direct structural assembly (12). DFT maps the atomic structure of a material onto an intrinsic energy, with lower energies indicat-

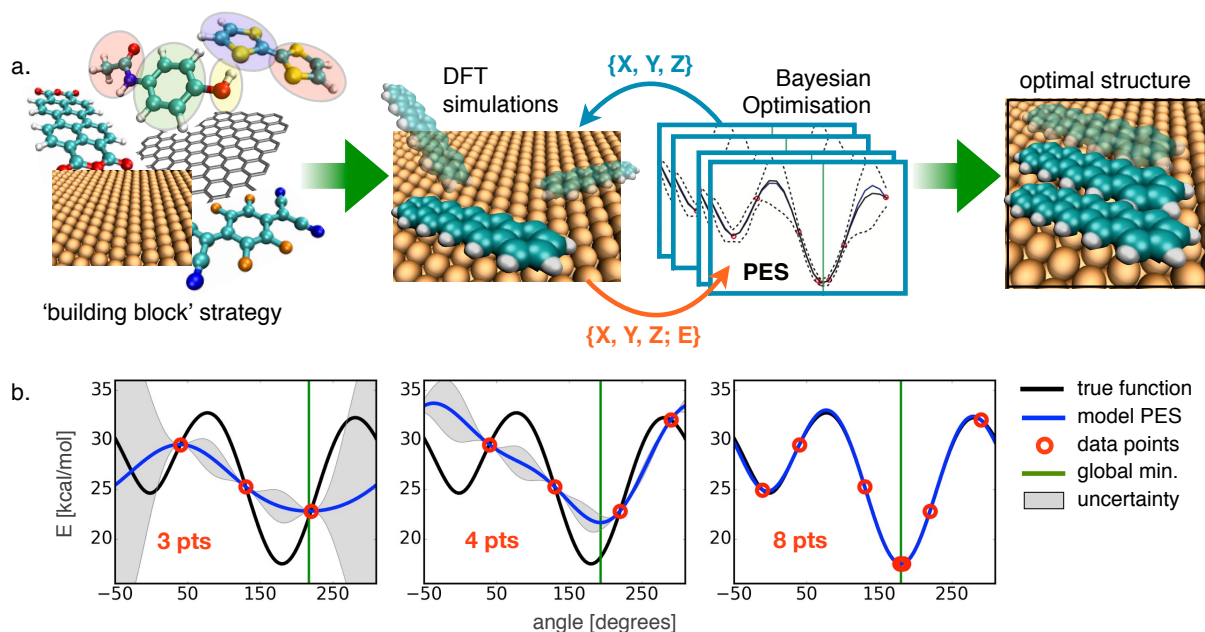
ing more stable material polymorphs. Theoretical structure prediction methods focus on exploring the resulting configurational phase-space, the potential energy surface (PES) (13, 14). Extensive DFT sampling is computationally prohibitive and reduced to comparing several 'most-likely' structures proposed by chemical intuition, which is unreliable in complex materials. For this reason, hybrid organic/inorganic interfaces present a special challenge for structure search methods. Their PES is complicated by the variety of different morphologies that molecular films can adopt against the solid material. Moreover, the large size of functional molecules means that extensive simulation cells (large lengthscales) are needed to describe molecular film morphologies, making computations particularly expensive.

Here, we propose a structure search scheme based on machine learning (ML) that is capable of accelerated and unbiased PES refinement across large length-scales, while also minimizing the amount of configurational sampling. Recently, algorithms from machine learning (ML) were coupled with DFT to approximate the PES (15–17) or improve sampling and accelerate structure prediction in single material clusters and solids (18–23). Their application to heterostructures is not straightforward, and they may not scale up to required sizes. In some cases, framework setup and the choice of ML parameters was found to affect the results (15, 24). Many schemes rely on large data sets with 1,000-10,000 sampled points (25), which are costly to produce. Our ideal method would need to be (i) efficient (minimal sampling costs), (ii) accurate (both in robust model convergence and DFT chemical accuracy), (iii) comprehensive (delivering the entire PES information of global and local minima), (iv) transferable (minimal dependence on ML parameters), (v) versatile (adaptable to targeting properties, structural prescreening, etc.), (vi) flexible (easily combined with other schemes) and (vii) truly multi-scale in its scope.

## Significance Statement

Computational materials design with accurate but costly simulation methods stands to benefit greatly from the machine learning techniques designed for minimal sampling. Our BOSS framework combines cutting edge machine learning with quantum mechanics. It is a general structure search method adapted to infer the microscopic structure of organic/inorganic interfaces with the aim to enhance the functionality of prospective devices. Here, we demonstrate the feasibility, robustness and predictive power of the BOSS approach in predicting the structure of single molecular surface adsorbates.

<sup>1</sup>To whom correspondence should be addressed. E-mail: milica.todorovic@aalto.fi



**Fig. 1.** a. Illustration of key steps in BOSS application to structure search at the inorganic surface: from the choice of materials and building blocks, through selection of the BO degrees of freedom and the iterative optimization, towards the inferred individual adsorbate and thin-film structures; b. Example of BOSS iterative inference of a simple 1-dimensional PES featuring a global and local minimum. The global minimum location is learned after six sampled configurations and the entire PES is learned in four acquisitions more.

To address the structure search challenge at organic/inorganic interfaces, we have conceived the Bayesian Optimisation Structure Search (BOSS) method. It delivers on all fronts thanks to its key ingredients: 1) state-of-the-art DFT or quantum chemistry treatment, 2) the Bayesian Optimisation (BO) machine-learning method and 3) the “building block” approach.

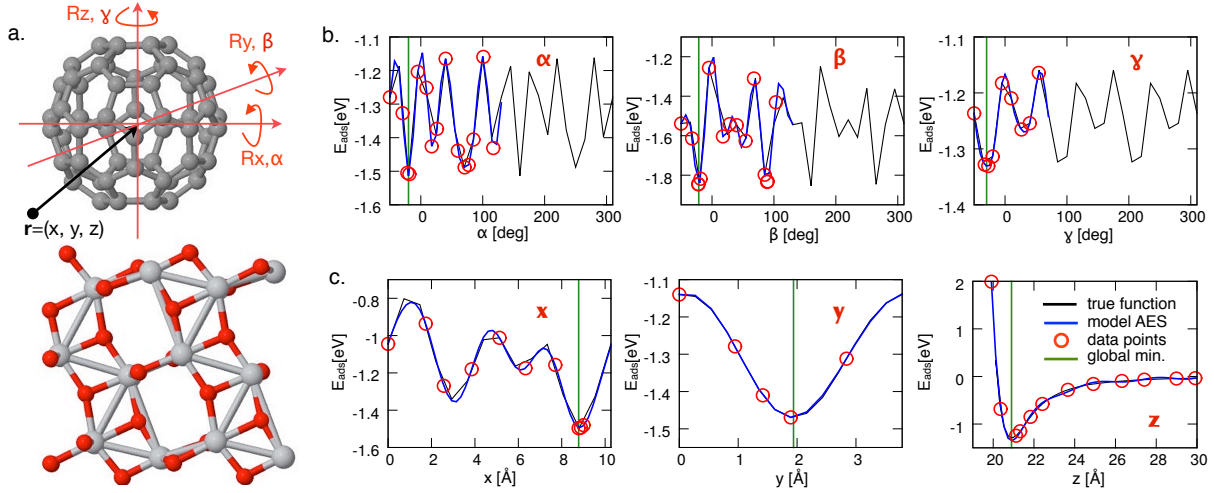
Efficient sampling underpins the BOSS approach to allow accurate DFT-based predictions despite their computational cost. Approximate Bayesian Computation<sup>(26)</sup> is a class of likelihood-free inference (LFI) methods where data sampling involves complex evaluation. It has recently been combined with BO<sup>(27)</sup> to accelerate model prediction where the evaluation is also costly. Here, we adapted the resulting BOLFI scheme<sup>(28)</sup> to search for minima of the PES in an arbitrary phase space using simple models. The Bayesian algorithm for acquiring new energy points balances exploration with exploitation of accumulated configurational data to quickly determine the PES global minimum. The PES function, including global and local minima as well as the barriers between them, is actively learned during the sampling procedure. The framework features only two hyperparameters, which are also learned on-the-fly.

BOSS utilizes an advanced DFT framework designed for efficient multi-scale materials simulations on supercomputer infrastructures<sup>(29)</sup>. To expedite structure search over large surface areas, we fix the internal components of the material that tend to maintain their structural integrity (e.g. aromatic rings, functional groups, or entire molecules)<sup>(30, 31)</sup>. Treating parts of the material as mobile “building blocks” accelerates the search by focusing on important regions of a lower-dimensional phase space, and eliminates wasteful sampling by removing irrelevant degrees of freedom. This has already allowed us to overcome the size limitations of similar schemes<sup>(32)</sup> and

apply our method to large surface adsorbates.

Our long-term goal is to predict the structure of organic/inorganic interfaces. In this article, we will focus on the first necessary step: the efficient structure prediction of a single adsorbed organic molecule. While some methods acquire single adsorption configurations by intuition and focus on complex lattice-based film morphology search<sup>(33, 34)</sup>, we aim to treat both the molecular adsorbates and aggregates within the BOSS framework by increasing search degrees of freedom. Employing BOSS to learn the individual molecule-surface interactions and structure efficiently is a key step that will later allow us to extend the search to molecular aggregates, monolayers and films. We focus on the adsorption geometry of fullerene molecules on the (101) surface of TiO<sub>2</sub> anatase. Both materials are frequently employed in organic optoelectronics: C<sub>60</sub> as an optically active electron donor, and TiO<sub>2</sub> anatase as a high conductivity dielectric buffer layer<sup>(35–37)</sup>. High-resolution atomic force microscopy (AFM) images available for this functional surface exhibit sub-molecular resolution<sup>(38)</sup>, which would allow us to verify our findings.

In this manuscript, we present the BOSS technique for large-scale structure search in complex heterostructures. We describe the main features of the machine learning approach and its practical implementation alongside atomistic simulations. BOSS was employed to infer the microscopic details of C<sub>60</sub> adsorption on TiO<sub>2</sub> anatase. We sought to explore its efficiency and accuracy as a function of increasing dimensionality of the search, and evaluate the quality of the ‘building block’ approximation. Understanding the advantages and limitations of the present approach will allow us to modify the scheme as we upscale the structure search towards more complex problems and more realistic interface geometries.



**Fig. 2.** a. Atomistic model of  $C_{60}$  adsorption on the (101) surface  $TiO_2$  anatase in the reference configuration, with the energetically dominant degrees of freedom for the molecule indicated in black (translational motion) and red (rotational motion); b. Comparison of the converged 1D AES with the true function for all rotational variables; c. Comparison of the converged 1D AES with the true function for all translational degrees of freedom. Learning in b. and c. was initialized with 5 quasi-random points and the models converged in 5-10 BO acquisitions.

## Methods

**Active PES learning with BOSS.** The BOSS strategy for identifying molecular surface adsorption structures is illustrated in Fig. 1a. It is first necessary to identify the simulation 'building block' segments that will be kept fixed. This step determines the dimensionality of the PES search. The BO algorithm then commences structural sampling over this model, an iterative process during which the global minimum structure and the PES model are inferred. The model PES is represented by a Gaussian Process (GP) (39), characterized by the GP posterior mean and the variance functions. The posterior variance describes the degree of belief in the GP mean, and supplies a useful measure of uncertainty which is minimized as the PES model is refined. To keep the PES smooth and continuous in all dimensions and encode the periodic boundaries of our atomistic simulations, we employed a non-isotropic standard periodic GP kernel. It features two hyperparameters: lengthscale  $l$  of PES variation in phase space and the standard deviation  $\sigma$  that describes PES magnitude. These hyperparameters are learned iteratively alongside the GP prior by maximizing the GP likelihood function.

Active PES learning with BOSS is illustrated in 1 dimension (1D) in Fig. 1b in several snapshots. A few initial data points were selected as a Sobol quasi-random sequence. Next, we sampled the phase space deterministically by minimizing the lower confidence bound (LCB) selection criterion (27). To avoid excessive data exploitation and local minima traps, we modified the acquisition function into an exploratory LCB (e-LCB) version with a pure exploratory step if the uncertainty on the PES guess becomes too small. Consequently, data was sampled both near the minima (exploitation) and in regions of high uncertainty (exploration). The acquired energy points were used to update the GP model via the Bayes' theorem, and the procedure was repeated until convergence. The model evolution shown in Fig. 1b. is typical of 1D functions: five data acquisitions suffice to pinpoint the global minimum, and with five more, the PES model converges to the true function.

**BOSS application to the  $C_{60}/TiO_2$  interface.** The (101) facet of the  $TiO_2$  anatase surface and the  $C_{60}$  cage exhibit minimal deformation upon interaction with other materials and are both treated as rigid building blocks. Fullerene adsorption configuration at the (101) facet of  $TiO_2$  anatase depends on both the position of the molecule above the surface and its orientation with respect to the surface. The position of  $C_{60}$  was supplied by the radius vector of the molecular center of mass  $r=[x, y, z]$ . The molecular orientation was described by angles of rotation  $\alpha$ ,  $\beta$  and  $\gamma$  with respect to Cartesian axes of rotation  $R_x$ ,  $R_y$  and  $R_z$ , respectively. The total

interaction energy  $E_{tot}$  between the two building blocks is thus a 6-dimensional (6D) function  $E_{tot}=E(x, y, z, \alpha, \beta, \gamma)$ . The BOSS search for the lowest energy adsorption structure proceeds by finding the minima of the adsorption energy surface (AES) with respect to this vector of variables. The adsorption energy  $E_{ads}$  was computed as  $E_{ads}=E_{tot}-(E_{surf}+E_{mol})$ , where  $E_{surf}$  and  $E_{mol}$  are the total energies of the clean surface and isolated molecule, respectively.

A reference configuration, illustrated in Fig. 2a, was employed to initialize the BOSS search. We used a local minimum structure to set the reference molecule position  $r_0=[x_0, y_0, z_0]$ . The reference  $[\alpha, \beta, \gamma]=[0, 0, 0]$  orientation features hexagonal  $C_{60}$  facets in the XY plane terminating the molecule at the top and bottom, both pierced through their center by the Z-axis. The high symmetry of the  $C_{60}$  cage was broken by the asymmetric surface features, allowing us to take limited advantage of molecular symmetry.

**Computational details.** BOLFI based on the *gpml* package (39) was implemented in a serial MATLAB code, which can be interfaced with any total energy simulation method. The GP model and its hyperparameters were updated every 10 acquisitions until convergence. We performed all configurational sampling with the all-electron DFT code FHI-aims (29). Simulations were carried out with converged Tier 2 basis sets free of g and h functions, and the PBE exchange-correlation functional (40) augmented with van der Waals correction terms (41). We utilized relativistic corrections to account for the heavy elements. *Light* grids with  $\Gamma$ -point reciprocal space sampling was employed to build the PES model, but global minima structures were refined with *tight* grids and a  $2 \times 2 \times 1$  k-point mesh.

The (101)  $TiO_2$  anatase surface in a slab configuration featured three typical trilayers in a  $10.27\text{\AA} \times 11.36\text{\AA} \times 52.77\text{\AA}$  periodic unit cell, exposing a  $1 \times 3$  unit cell surface area. We found that molecular adsorption energies converged with three trilayers; the lowest two trilayers were kept fixed during full structural optimisations. To define the boundaries of BOSS search phase space, we relied on the surface and molecule symmetry and periodicity. Molecular registry search space was limited to the smallest periodically repeating surface unit  $10.27\text{\AA} \times 3.78\text{\AA}$  and informed by this periodicity. The non-periodic z variable search was conducted  $10\text{\AA}$  in height from the  $1.5\text{\AA}$  closest surface approach. Molecular orientation search was conducted in minimal unique periods of 180deg. for  $\alpha$  and  $\beta$  angles, and 120deg. for the  $\gamma$  angle, exploiting the symmetry of the  $C_{60}$  cage. With the efficient parallelisation of FHI-aims (42), a single point data acquisition calculation on 168 atoms required 10min on 120CPUs.



## Results and Discussion

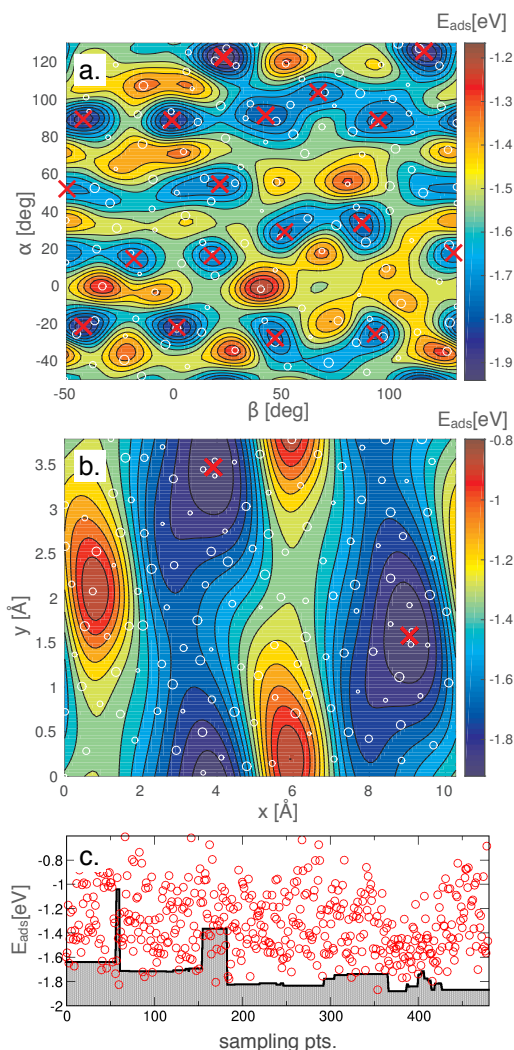
**BOSS search of optimal surface adsorption structure.** We applied the BOSS framework to the atomistic structure search of the surface adsorption model presented in Fig. 2a. To monitor the increase in search complexity with the number of dimensions, we opted to incrementally build up to the full dimensionality of the problem. To begin with, we explored the 1D AES of each of the translational  $x, y$  and  $z$  and rotational molecular variables  $\alpha, \beta$  and  $\gamma$  independently, while keeping all other variables fixed to reference values. The results are presented in Figs. 2b and 2c. Despite the simplicity of low-dimensional BOSS sampling, we obtained a wealth of information about the binding and structure at the  $C_{60}/TiO_2$  interface.

We discovered molecular rotation to be the dominant energetic factor in the molecule-surface interactions. Owing to the high symmetry of the  $C_{60}$  cage and its repeating units, molecular rotation variables produce complex fast-varying AES curves with multiple deep minima (Fig. 2b.). The  $\beta$  angle rotation commanded an AES energy variation of 0.6eV with a global minimum at -1.85eV. The lowest energy recorded for the  $\alpha$  variable AES was -1.5eV. To gain physical and chemical insight behind the BOSS-inferred minima, we analyzed the structures corresponding to these molecular rotations. This allowed us to identify the  $C_{60}$  hexagonal facet and the C-C bond shared by two hexagonal facets ( $C_h-C_h$ ) as the key reactive sites of the molecule. When a hexagonal facet of the molecule becomes parallel with the sloping terrace of anatase (see Fig. 2a.), the overlap between the electronic density of the aromatic ring and the surface terrace atoms is maximised. This effect stabilized the  $\beta$  global minimum structure. Similarly, we discovered that orienting the  $C_h-C_h$  bond directly towards the surface terrace increased molecule-surface interactions and lead to the second-lowest energy minimum in  $\alpha$ . These  $C_{60}$  reactive sites both feature bond conjugation and out-of-plane electronic density associated with  $p_z$  carbon atom orbitals that facilitate molecular physisorption to this surface.

Translation of the molecule across the surface produced slowly-varying AES with fewer minima, illustrated in Fig. 2c. The converged models were smooth, despite the corrugation and complex chemistry of the (101)  $TiO_2$  anatase surface. The 1D energy curves associated with  $x, y$  and  $z$  variables exhibited similar molecule-surface interaction strength, with an overall global minimum of -1.5eV. The  $x$  variable curve featured a double energy minimum. We recognized the two sloping terraces in the anatase unit cell, which give rise to surface corrugation in the [101] crystallographic direction, as the features behind this behavior (see surface model in Fig. 2a). The  $y$  variable revealed the simple energy variation inside the terrace groove (the [010] direction) between the  $Ti_{5c}$  and the  $O_{3c}$  atoms as adsorption sites. In the  $z$  direction, we observed a dispersion-like shape of the AES. Further calculations helped us confirm that the height of the molecule above the surface adjusts the magnitude of molecule-surface interactions, but does not alter the AES profiles of other variables. This lead us to conclude that the  $z$  degree of freedom is uncoupled from the others and may be treated separately.

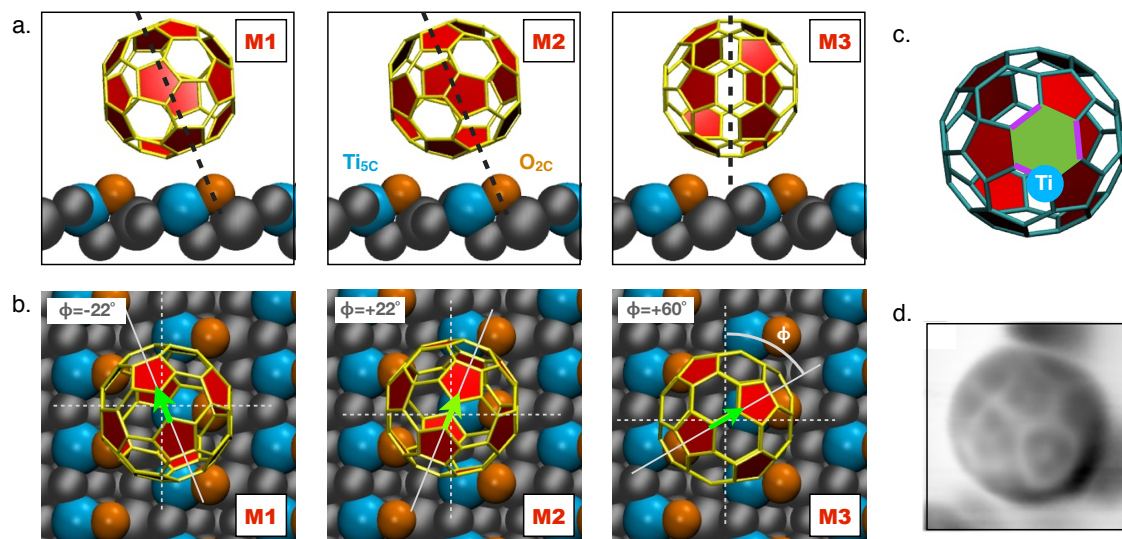
We gained chemical insight into the reactive sites of both the  $C_{60}$  molecule and the  $TiO_2$  surface at a very modest computational effort. All 1D AES were inferred in up to 12 data acquisitions at seemingly random locations: these were

nevertheless selected by the BOSS acquisition function for speedy energy mapping. The complex and fast-varying AES for molecular rotations were learned just as efficiently as the simpler ones for translations, pointing to the transferability of the method. To verify the accuracy of BOSS inference, we carried out stepwise sampling of each 1D AES in 25-point resolution and computed the true energy variations. Figs. 2b and 2c. illustrate the good agreement between true and inferred AES curves. In short, BOSS sampling allowed us to compute excellent high-fidelity energy models at half the computational cost.



**Fig. 3.** Energy maps inferred in 2D BOSS simulations, with white circles indicating the locations of data acquisitions: a. map of simultaneous  $C_{60}$   $\alpha - \beta$  molecular rotations; b. map of  $x-y$  molecular translations, with the  $C_{60}$  cage in its inferred optimal orientation configuration. c. Convergence of the adsorption energy corresponding to the global minimum configuration actively inferred in a 5D BOSS search (black line). The accuracy of the inferred result improved with configurational sampling (red data points) as the 5D AES model was refined.

Next, we conducted 2D BOSS simulations while constraining the remaining degrees of freedom to reference values. The most important resulting AES landscapes are presented in Figs. 3a. and 3b. for molecular rotation and translation respectively. As expected, the 2D  $\alpha - \beta$  molecular rotation



**Fig. 4.** a. Side view of the three lowest energy adsorption configurations M1, M2 and M3 obtained by full structural relaxation from BOSS-predicted minima. Pentagonal facets of C<sub>60</sub> are colored in red for visual distinction and the symmetry axis for molecular rotation is indicated by the black dashed line; b. Top view of the three lowest energy adsorption configurations M1, M2 and M3. Reactive under-coordinated atoms on the surface are shown in blue (Ti<sub>5c</sub>) and orange (O<sub>2c</sub>) to highlight molecular registry on the surface. The green arrow illustrates the direction of the typical oval feature observed in all three structures, along the bond between two hexagons. Angle  $\phi$  denotes the orientation of the bond with respect to the [010] crystallographic direction; c. Underside of the C<sub>60</sub> cage directly above the Ti<sub>5c</sub> surface binding site. Molecular binding is facilitated by the hexagonal facet (green) and the nearby C<sub>1h</sub>-C<sub>1h</sub> bond (purple); d. Frequency shift response sub-molecular AFM image of C<sub>60</sub> on the (101) surface of TiO<sub>2</sub> anatase. Adapted with permission from Moreno, et. al., Nano Letters, 12, 2257 (2015). Copyright (2015) American Chemical Society.

produced multiple energy minima that can be observed in the blue regions of Fig. 3a. We identified them by conducting inexpensive local minima searches along the AES in a post-processing step. All 16 local minima energies were within 0.2eV of the global minimum at -1.94eV. So many comparable minima point to competing adsorption configurations of interest. Some of these local minima could be relevant in comparison with surface experiments, where thermal effects may stabilise adsorption configurations that are not lowest in energy. The alternative minima thus constitute a valuable finding, but also considerably increase the complexity of the high-dimensional BOSS models.

Conversely, we observed a single minimum in the 2D  $x$ - $y$  molecular translation AES map in Fig. 3b. The location of the minimum points to the optimal adsorption site on this surface. This depends strongly on the molecular sites exposed to the surface. When switching the molecular orientation from a sub-optimal reference structure to its global minimum setting, we observed the  $x$ - $y$  global minimum moving from the O<sub>3c</sub> atom site to the Ti<sub>5c</sub> atom site. The 2D map correctly reflected the symmetry of the underlying two sloping grooves on anatase, one of which is 0.5 unit cell shifted in the [010] direction with respect to the other. The single minimum was extended in the [010] direction, suggesting a degree of molecular translation along the grooves of the anatase surface. Nevertheless, the diffusion barrier value of 0.2eV in this case tells us that C<sub>60</sub> molecules are not mobile along the grooves at room temperature, but remain in the vicinity of the optimal surface binding site.

Consequent 3D and 4D BOSS searches were completed successfully and produced consistent AES global minima findings, so we proceeded to full dimensionality of the problem. Since we identified that the  $z$  variable produces only a vertical shift in the adsorption energy, we eliminated this trivial degree of

freedom by constraining the  $z$  to the location of its global minimum value (clearly seen in Fig. 2c.). Thus we carried out the full adsorption site BOSS search in 5D and obtained the converged AES model after 500 data acquisitions, as shown in Fig. 3c. The maximum adsorption energy was identified as -1.88eV, in good agreement with low-dimensional global minimum observations. The global minimum orientation of the C<sub>60</sub> cage featured the hexagonal facet parallel to the anatase terrace, as observed earlier. We mined the 5D model to extract the multiple local minima structures and found them largely associated with molecular rotation. The optimal surface adsorption site was located above the under-coordinated Ti<sub>5c</sub> surface atom; the same site has been identified as the most reactive site on this surface by earlier studies of small adsorbates (43, 44). In the [101] direction, the molecule adsorbs -1 Å from the Ti<sub>5c</sub> site, and towards the deeper region of the sloping terrace.

Verifying the accuracy of the BOSS global minimum predictions was not an easy task since the optimal result was not known. Ideally, we would check the quality of BOSS AES models against true energy surfaces obtained by grid-based sampling. This was only possible in 1D and the agreement was excellent; in 2D simulations even coarse sampling would require a minimum of 500 data acquisitions. Here we appreciated the efficiency of the Bayesian sampling scheme. Only 90 data evaluations were needed to converge the complex 2D  $\alpha - \beta$  energy map in Fig. 3a., supplying information on all energy minima and barriers at the same time. The  $x - y$  map in Fig. 3b. required only 45 data points to converge. Grid-based sampling in more than 2D quickly becomes computationally prohibitive, but BOSS allowed us to build up and converge good quality models in 3D-5D through multi-dimensional sampling.

In another accuracy check, we compared and tracked the global minimum solutions by both value and location as the

model dimensionality grew. Apart from the  $z$  variable, other degrees of freedom were not independent and some results were interpreted as local minima in the light of the constrained search dimensionality. Nevertheless, increasing model dimensionality produced consistent global minimum solutions. We observed that the multiple energy minima associated with three molecular rotation variables presented the biggest challenge in a 5D search, which contributed to the relatively large sampling needed to obtain a converged result.

**Full C<sub>60</sub>/TiO<sub>2</sub> interface structure search.** The full BOSS AES search converged with a global adsorption energy minimum of  $E_{\text{BOSS}} = -1.9\text{eV}$  within the constraint of the structural 'building blocks'. To verify the quality of the prediction, we removed this approximation and allowed all degrees of freedom to relax. We selected for full structural relaxations the global minimum and six unique local minima located by BOSS within a 0.1eV energy window of the global minimum. Such an approach would allow us to compare a range of low-energy adsorption configurations with experimental data, where C<sub>60</sub> molecules evaporated on a hot surface may have acquired similar thermal energy.

After the seven full structural optimizations, all BOSS-predicted minima structures were reduced to one of three different configurations shown in Fig. 4a. We identified the global minimum of adsorption energy as  $E_{\text{GL}} = -2.0\text{eV}$ , only 0.1eV below the value predicted by BOSS of  $E_{\text{BOSS}} = -1.9\text{eV}$ . The close correspondence between the BOSS-inferred minimum and the true energy minimum was explained by analysis of optimal structures: we noted minimal distortions of the molecule and surface bond lengths, indicating that the 'building block' approximation was particularly appropriate in this case study.

We discovered the global minimum of the adsorption energy to be two-fold degenerate, with both structures M1 and M2 in Fig. 4a. corresponding to the same energy. If one defined an axis of C<sub>60</sub> rotation perpendicular to the anatase terrace for both low energy configurations, as illustrated, the M2 molecule could be mapped into the M1 orientation by a 180deg. rotation around this axis to produce identical interactions with the surface. The two structures were obtained by optimizing the structure of different local minima, but matched in energy to within a 1meV tolerance. The M3 configuration in Fig. 4a. was the only local energy minimum we found within a 0.1eV energy window from the global minimum. With an energy of  $E_{\text{loc}} = -1.93\text{eV}$ , it featured a highly symmetric C<sub>60</sub> orientation less compatible with the corrugation of the underlying substrate.

As predicted by BOSS, the parallel alignment of a hexagonal C<sub>60</sub> facet to the sloping terrace of anatase emerged as the key feature associated with surface adsorption. The global minimum geometries M1 and M2 were slightly tilted away from this configuration. Analysis of the adsorbed C<sub>60</sub> underside, presented in Fig. 4c., revealed that tilt allows one of the nearest C<sub>h</sub>-C<sub>h</sub> bond to approach the surface as well. This was another energy-lowering microscopic feature identified by early BOSS simulations. The most stable adsorbed molecule configurations thus feature a lack of symmetry with respect to the substrate that is difficult to predict by chemical intuition; any symmetric initial guess structure would likely fail to reach the deeper energy minimum during structure optimization.

We present the top-down view of the three relevant absorption configurations in Fig. 4b. and note a repeating structural

motif. An oval feature containing two hexagonal and two pentagonal facets is exposed at the top of the molecule in all three cases. For comparison, we show a detailed AFM image of this molecular adsorbate in Fig. 4d. A similar oval feature appears, although the hexagonal and pentagonal facets are difficult to identify. The image suggests that computed molecular adsorption structures resemble the one in the experiment, even if it is not possible to conclusively identify any one of the three. We defined the direction of the feature along the bond separating the two hexagons (also along the long axis of the oval) and computed its orientation with respect to the [010] crystallographic direction. As illustrated in Fig. 4b, this allowed us to determine the angle of the feature with respect to the lines of O<sub>2c</sub> atoms that are typically observed in AFM surface experiments on anatase. Should experimental data including substrate information become available, our analysis would help us explain experimental configurations and also verify our computed molecular adsorption structures.

Our final consistency check was to increase the accuracy of the computational settings (see Methods section). The three configurations were reoptimized with higher accuracy simulation settings and we obtained a better estimate for the maximum adsorption energy at  $E_{\text{GL}} = -1.6\text{eV}$ . The adsorption configurations remained the same, confirming that computationally cheaper lighter settings yield good quality geometries and present a useful tool in structural studies.

## Conclusions

We proposed a novel structure search scheme that combines a smart ML sampling strategy and a natural "building block" representation with accurate quantum mechanical calculations. As first step in targeting the structure of large-scale molecular films and organic/inorganic interfaces, we employed it to learn the adsorption structure of a single molecule: C<sub>60</sub> on the (101) surface of TiO<sub>2</sub> anatase. The BOSS method produced a computationally tractable study of molecular adsorption as function of key degrees of freedom, molecular registry and orientation, with readily available chemical insight. The benefits of smart multi-dimensional sampling were particularly evident in low dimensional searches, where 10-100 data acquisitions sufficed to converge the model predictions. In higher dimensions, model convergence was slowed down by the complexity of multiple minima in correlated degrees of freedom and 500 data acquisitions were required. Model refinement could be made more robust by employing prior belief functions or different GP kernels. In an intuition-led force minimization adsorption study, 500 single point calculations would supply optimized structures from only five different initial guess configurations. With BOSS, they deliver optimal configurations amongst a much wider configurational pool, alongside with the local minima and the barriers between them.

Structural optimizations based on BOSS-inferred AES models and optimal structures produced the definitive molecular adsorption geometries despite the approximations employed, here BOSS was both efficient and accurate. Refined adsorbate structure models compared well with high-resolution experimental images of these materials, additionally confirming the accuracy of BOSS predictions. Comprehensive adsorption energy surface (AES) information obtained allowed us to consider also local minima, and could be mined to extract important energy barriers. The 'building block' approach performed very



well with C<sub>60</sub> adsorbed on TiO<sub>2</sub> anatase, and would allow us to easily extend our approach to molecular aggregates and make it truly multi-scale. In short, our BOSS scheme delivers on many fronts in a successful study of molecular surface adsorption and further work will see it applied to more complex configurational studies of surface-supported molecular aggregates and films.

**ACKNOWLEDGMENTS.** This work was supported by the Academy of Finland through its Centres of Excellence Programme under Project Nos. 251748 and 284621, and also through the European Union's Horizon 2020 research and innovation program under Grant agreement No. 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. J.C. was funded by the ERC grant no. 742158. Computer time was provided by the Centre for Scientific Computing (CSC, Finland) at the Taito supercomputer.

- Theobald JA, Oxtoby NS, Phillips MA, Champness NR, Beton PH (2003) Controlling molecular deposition and layer structure with supramolecular surface assemblies. *Nature* 424(6952):1029–1031.
- Barth JV, Costantini G, Kern K (2005) Engineering atomic and molecular nanostructures at surfaces. *Nature* 437(7059):671–679.
- Grill L, et al. (2007) Nano-architectures by covalent assembly of molecular building blocks. *Nat. Nano* 2(11):687–691.
- Schlesinger R, et al. (2015) Efficient light emission from inorganic and organic semiconductor hybrid structures by energy-level tuning. *Nat. Comm.* 6:6754.
- Denny Jr. MS, Moreton JC, Benz L, Cohen SM (2016) Metal–organic frameworks for membrane-based separations. *Nat. Rev. Mater.* 1:16078.
- Huang N, Wang P, Jiang D (2016) Covalent organic frameworks: a materials platform for structural and functional designs. *Nat. Rev. Mater.* 1:16068.
- fang Geng Y, et al. (2017) Stm probing the supramolecular coordination chemistry on solid surface: Structure, dynamic, and reactivity. *Coordination Chemistry Reviews* 337:145 – 177.
- Song Y, et al. (2017) Self-assembly and local manipulation of au-pyridyl coordination networks on metal surfaces. *ChemPhysChem* 18(15):2088–2093.
- Howarth AJ, et al. (2016) Chemical, thermal and mechanical stabilities of metal–organic frameworks. *Nat. Rev. Mater.* 1:15018.
- von Lilienfeld OA (2014) *Towards the Computational Design of Compounds from First Principles*. (Springer International Publishing).
- Curtarolo S, et al. (2013) The high-throughput highway to computational materials design. *Nature Materials* 12(3):191–201.
- Jones RO (2015) Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* 87.
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by Simulated Annealing. *Science* 220(4):671–680.
- Goedecker S, Hellmann W, Lenosky T (2005) Global Minimum Determination of the Born-Oppenheimer Surface within Density Functional Theory. *Physical Review Letters* 95(5):055501.
- Behler J (2014) Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics: Condensed Matter* 26(18):183001.
- Bartok AP, Payne MC, Kondor R, Csanyi G (2010) Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* 104(13).
- Li Z, Kermod JR, De Vita A (2015) Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Physical Review Letters* 114(9):096405.
- D'Avezac M, Zunger A (2008) Identifying the minimum-energy atomic configuration on a lattice: Lamarckian twist on Darwinian evolution. *Physical Review B* 78(6):064102.
- Wang Y, Lv J, Zhu L, Ma Y (2012) CALYPSO: A method for crystal structure prediction. *Computer Physics Communications* 183(10):2063–2070.
- Nelson LJ, Hart GLW, Zhou F, Ozolins V (2013) Compressive sensing as a paradigm for building physics models. *Physical Review B* 87(3):035125.
- Bhattacharya S, Levchenko SV, Ghiringhelli LM, Scheffler M (2013) Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of MgMOx. *Physical Review Letters* 111(1):135501.
- Kiyohara S, Oda H, Tsuda K, Mizoguchi T (2016) Acceleration of stable interface structure searching using a kriging approach. *Japanese Journal of Applied Physics* 55(4):045502.
- Xue D, et al. (2016) Accelerated search for materials with targeted properties by adaptive design. *Nature Communications* 7:11241.
- Bhattacharya S, Levchenko SV, Ghiringhelli LM, Scheffler M (2014) Efficient ab initio schemes for finding thermodynamically stable and metastable atomic structures: benchmark of cascade genetic algorithms. *New Journal of Physics* 16(1):123016.
- Ueno T, Rhone TD, Hou Z, Mizoguchi T, Tsuda K (2016) COMBO: An efficient Bayesian optimization library for materials science. *Materials Discovery* 4:18–21.
- Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J (2017) Fundamentals and recent developments in approximate bayesian computation. *Systematic Biology* 66(1):e66–e82.
- Brochu E, Cora VM, de Freitas N (2010) A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv.org*.
- Gutmann MU, Corander J (2016) Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *Journal of Machine Learning Research* 17(125):1–47.
- Blum V, et al. (2009) Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* 180(11):2175–2196.
- Behler J, Lorenz S, Reuter K (2007) Representing molecule-surface interactions with symmetry-adapted neural networks. *Journal of Chemical Physics* 127(1):014705–014705.
- Oganov AR, Glass CW (2006) Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of Chemical Physics* 124:244704.
- Carr SF, Garnett R, Lo CS (2016) Accelerating the search for global minima on potential energy surfaces using machine learning. *The Journal of Chemical Physics*.
- Obersteiner V, Scherbela M, Hörmann L, Wegner D, Hofmann OT (2017) Structure Prediction for Surface-Induced Phases of Organic Monolayers: Overcoming the Combinatorial Bottleneck. *Nano Letters* 17:4453–4460.
- Packwood DM, Han P, Hitosugi T (2017) Chemical and entropic control on the molecular self-assembly process. *Nature Communications* 8:14463.
- Grätzel M, et al. (1998) Solid-state dye-sensitized mesoporous TiO<sub>2</sub> solar cells with high photon-to-electron conversion efficiencies. *Nature* 395(6702):583–585.
- Yoo S, et al. (2007) Analysis of improved photovoltaic properties of pentacene/C<sub>60</sub> organic solar cells: Effects of exciton blocking layer thickness and thermal annealing. *Solid-State Electronics* 51(10):1367–1375.
- Cheyens D, Gommans H, Odijk M, Poortmans J, Heremans P (2007) Stacked organic solar cells based on pentacene and C<sub>60</sub>. *Solar Energy Materials and Solar Cells* 91(5):399–404.
- Moreno C, Stetsovych O, Shimizu TK, Custance Ó (2015) Imaging Three-Dimensional Surface Objects with Submolecular Resolution by Atomic Force Microscopy. *Nano Letters* 15:2257–2262.
- Rasmussen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning*. (MIT Press), 2nd edition edition.
- Perdew JP, Burke K, Ernzerhof M (1996) Generalized Gradient Approximation Made Simple. *Physical Review Letters* 77(18):3865–3868.
- Tkatchenko A, Scheffler M (2009) Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters* 102(7):73005.
- Marek A, Blum V, Johanni R, Havu V (2014) The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *Journal of Physics: Condensed Matter* 26:213201.
- Tilocca A, Selloni A (2004) Methanol adsorption and reactivity on clean and hydroxylated anatase (101) surfaces. *The Journal of Physical Chemistry B* 108(50):19314–19319.
- He Y, Tilocca A, Dulub O, Selloni A, Diebold U (2009) Local ordering and electronic signatures of submonolayer water on anatase TiO<sub>2</sub>(101). *Nature Materials* 8(7):585–589.