# Efficient Parametrization of Complex Molecule-Surface Force Fields

David Z. Gao,[*,†] Filippo Federici Canova,[*,‡] Matthew B. Watkins,[†] and Alexander L. Shluger[†,‡]

[†]*Department of Physics and Astronomy, University College London, Gower Street, London, UK*

[‡]*Advanced Institute for Materials Research, Tohoku University, Katahira, Sendai, Japan*

E-mail: david.gao.10@ucl.ac.uk; felix@wpi-aimr.tohoku.ac.jp

## Abstract

We present an efficient scheme for parametrizing complex molecule-surface force fields from *ab initio* data. The cost of producing a sufficient fitting library is mitigated using a 2D periodic embedded slab model made possible by the quantum mechanics/molecular mechanics (QM/MM) scheme in CP2K. These results were then used in conjunction with genetic algorithm (GA) methods to optimize the large parameter sets needed to describe such systems. The derived potentials are able to well reproduce adsorption geometries and adsorption energies calculated using density functional theory (DFT). Finally, we discuss the challenges in creating a sufficient fitting library, determining whether or not the GA optimization has completed, and the transferability of such force fields to similar molecules.

# Abbreviations

GA, genetic algorithm; QM/MM, quantum mechanics/molecular mechanics; DFT, density functional theory

# Keywords

QM/MM, organic molecules, Genetic algorithm, force field, optimization

# Introduction

The growth and self-assembly of organic molecules on insulating surfaces is of critical importance to many fields including catalysis,[1] molecular electronics,[2–4] and lubrication.[5] While *ab initio* methods can provide a wealth of information on the electronic structure of a system, they are computationally expensive and are usually limited to small systems or ground state calculations. Fortunately, an understanding of the dynamic properties and growth in such systems can be achieved through the use of classical force fields and large scale molecular dynamics (MD) simulations. However, a major challenge that arises is the lack of complete classical interaction models for these kind of systems; the parametrization of such classical models requires a significant investment of resources.

While many of the components that contribute to the full collection of interactions needed to represent molecule-surface systems are readily available, some key contributions are missing. Previous studies have shown that combining existing force fields using newly parametrized contributions can provide a reasonable representation of the complete system. Recently, Wright *et al.* parametrized molecule-surface force fields for small proteins on Au(111) and Au(100) using *ab initio* data[6,7] by incorporating existing protein force fields and generic Au-Au parameters.

Since organic molecules can be described as combinations of elementary building blocks, several transferable force fields are available (CHARMM,[8–10] AMBER,[11] UFF,[12] etc), to describe their intramolecular, bonded interactions. These force fields also incorporate non-bonded interactions, which are represented using Lennard-Jones (LJ) potentials. LJ parameters are obtained using the Lorenz-Berthelot[13] mixing rules. Only minor adjustments are necessary to reproduce experimentally observable properties (such as crystal structure, density and viscosity).

Classical models are also available for a wide range of solid surfaces and are usually optimized to reproduce empirical data (lattice structure, compressibility, and/or vibrational spectra). In most cases LJ potentials are not able to reproduce the strong short-range interactions in solids, and more flexible functional forms are used. Unfortunately, these potentials cannot simply be mixed with LJ models for organics to obtain molecule-surface interactions and are unavailable in the literature for novel systems. They must be reparametrized for every molecule-surface combination. This is a complicated task for several reasons.

In contrast to systems composed entirely of fluids or solids where experimentally measured macroscopic properties can be used as fitting criteria, experimental data is rarely available for systems composed of large organic molecules adsorbed on surfaces. In these situations, atomistic models must be based on higher level quantum mechanical calculations instead. However, quantum chemistry techniques have difficulties treating large organic molecules on surfaces as such systems easily exceed the capabilities of present computers.

One solution that was employed in the past was to separate the molecule into smaller fragments of reduced computational cost.[14] A force field representing each fragment is then separately parametrized using *ab initio* data and the resulting force fields are combined to describe the full molecule. Unfortunately, this procedure is both time consuming to formulate and to implement, requiring a significant amount of chemical intuition.

Furthermore, despite the fact that this approach has been shown to produce reasonable force fields, inaccuracies often appear in the the properties of larger molecules represented in such a way.[15–17] The final force field parameters for the entire molecule should be adjusted to reproduce experimental measurements. However, such measurements are often unavailable. Studying large molecules on surfaces further increases the complexity of the system and as a result increases the error that may arise from an incomplete description of the full molecule. These challenges highlight the need for an efficient systematic procedure for parametrizing molecule-surface interactions in complex systems, which is currently lacking.

In this paper we present an efficient scheme for fitting classical force fields for entire organic molecules on surfaces by matching classical molecule-surface interactions to ab-initio data using genetic algorithm (GA) methods. We employ a periodic QM/MM embedding scheme[18,19] to treat large systems quantum mechanically and avoid the need for fragment analysis. This method reduces the number of atoms within the system that must be treated quantum mechanically, allowing us to calculate the full molecule-surface interaction and produce training data sets. However, the number of interaction parameters that had to be simultaneously optimized was too large for conventional least-square methods due to complexity of the system.

In order to meet this challenge, we employed GA methods which have been shown to produce reasonably good solutions,[20–24] for many small systems. We illustrate this approach on a particular system: a large 1,4-bis(cyanophenyl)-2,5-bis(decyloxy)benzene (CDB) molecule adsorbed onto KCl(100) in vacuum. Our results show how *unlikely* atomic configurations should be included in the training set to avoid over-training the model. Since the CDB molecule contains many organic building blocks, we tested the transferability of our parameters using a variation of the molecule on KCl(100). Our results show that while the parameters produced are qualitatively transferable, they can not quantitatively predict values such as the adsorption energy of other molecules. Fortunately, the methods presented are efficient enough that it is realistic to reparametrize the force field for each unique molecule.

# Density Functional Theory

Large flexible organic molecules on surfaces can be difficult to represent using classical force fields. Such systems contain many degrees of freedom, resulting in complex molecule-surface interactions. Critically, if there are large electronic relaxations present in the system, special treatment such as charge variable potentials may be required. On wide band gap insulators however, many molecules are primarily physisorbed and exhibit little or no charge transfer. One previously studied example is the CDB molecule on KCl(100).[25] This molecule possesses two interchangeable functional groups, two hydrocarbon chains, and a variable length aromatic central body as shown in Figure 1.

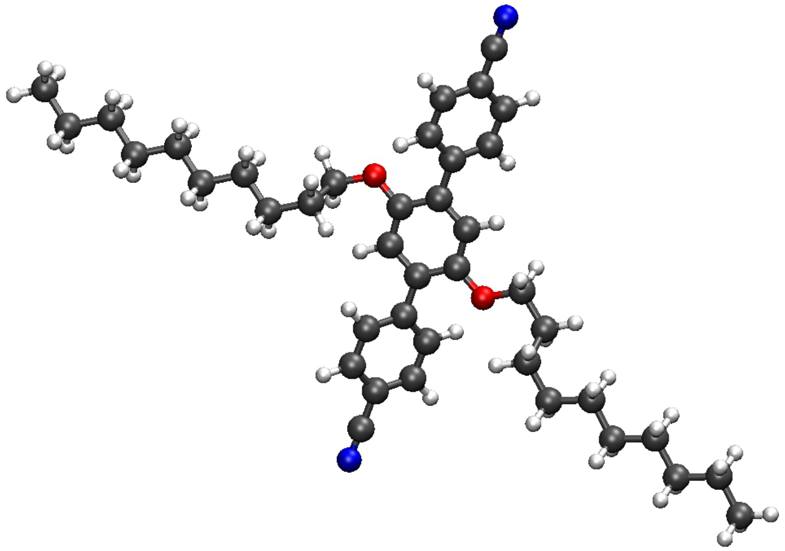

Figure 1: The structure of a CDB molecule with N atoms shown in blue and O atoms in red. The molecule consists of a central aromatic core with a CN or methyl functional group on either end and two hydrocarbon arms joined to the main body by an O atom. Experimentally, the length of the body and the functional groups was varied in order to produce a variety of structures.[25]

These components all contribute to the various competing molecule-molecule (MM) and molecule-surface (MS) interactions that dominate structure formation and growth processes.[25] We previously performed DFT simulations using the CP2K code and a mixed Gaussian and plane waves (GPW) approach[26,27] to study the properties of CDB on KCl(100).

The PBE/GGA[28,29] functional was used to represent the surface and the molecule and the MOLOPT[30] basis set was used to minimize basis set superposition error (BSSE). Finally, we employed long range dispersion corrections[31] in order to represent the vdW interactions in the system. The values for surface K atoms were approximated using the parameters for Ar while the rest of the atom types were assigned default values.

On the KCl(100) surface, the CDB molecule was found to adsorb with a total energy of 3.1 eV with a 2.4 eV contribution from dispersion corrections.[25] Since these dispersion corrections account for such a significant percentage of the total adsorption energy, it would be ideal to validate these results using experimental data or more accurate *ab initio* methods. Unfortunately such results are not available for these types of systems at this time. The molecule sits on the surface with CN groups positioned near cations, and shows negligible charge transfer using Mulliken population analysis. Additionally, charge density difference plots show that there is very little polarization or charge redistribution in the adsorbed molecule, as shown in Figure 2.

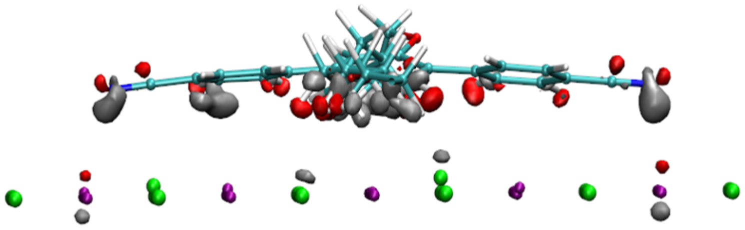

Figure 2: The charge density difference plot of a CDB molecule adsorbed onto KCl(100). The molecule sits with the two CN groups and one of the two O atoms interacting with cations on the surface. The positive (0.001 electrons/Å$^3$) isosurface is shown in silver while the negative (-0.001 electrons/Å$^3$) isosurface is shown in red. In this system, charge redistribution between CDB and the surface is negligible.

The interaction between the molecule and the surface can be further characterized by examining the electronic structure of the system. PBE/GGA is known to often underestimate the band gap and in this case produces a HOMO-LUMO gap of 5.6 eV compared to an experimental value of 7.6 eV.[32] Our calculations show that the HOMO and LUMO of the CDB molecule sit in the gap of KCl, as shown in Figure  3. This provides further evidence that charge transfer is unlikely in this system and indicates that it may be an excellent candidate for representation using a classical force field.

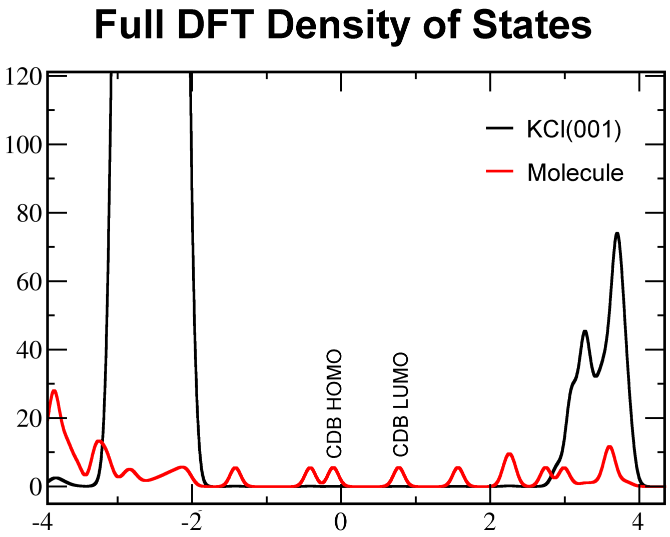

Figure 3: Electronic density of states for a single CDB molecule on the KCl(100) surface. The HOMO and LUMO of the molecule lie in the band gap of KCl, indicating a primarily physisorbed system.

Since the CDB molecule is large, flexible, and has many degrees of freedom, generating classical potentials directly from *ab initio* calculations of the entire molecule would be prohibitively expensive; we employed an embedded slab or QM/MM method to greatly reduce the computational cost of generating a fitting library.

# Periodic QM/MM

One of the major limitations to performing ab initio calculations on this system is that the surface requires several atomic layers of atoms in order to produce a proper band gap. A calculation of bulk KCl in this case gives a band gap of 5.6 eV while a 4 layer slab represented periodically in X and Y produces a band gap of 5.4 eV due to surface states. One strategy to reduce the computational cost of such calculations is to simply reduce the number of atoms in the simulation. However, it is critically important to ensure that the electronic structure of the surface is preserved so that the molecule-surface interaction is accurately represented. In order to accomplish this, many of the KCl atoms can be represented classically using the 2-dimensional embedding scheme implemented by Laino et al,[18,19] as shown in Figure 4. Atoms within the CDB molecule and the first layer of surface atoms that they interact with are treated quantum mechanically, while the remaining surface atoms are represented using a set of classical potentials.

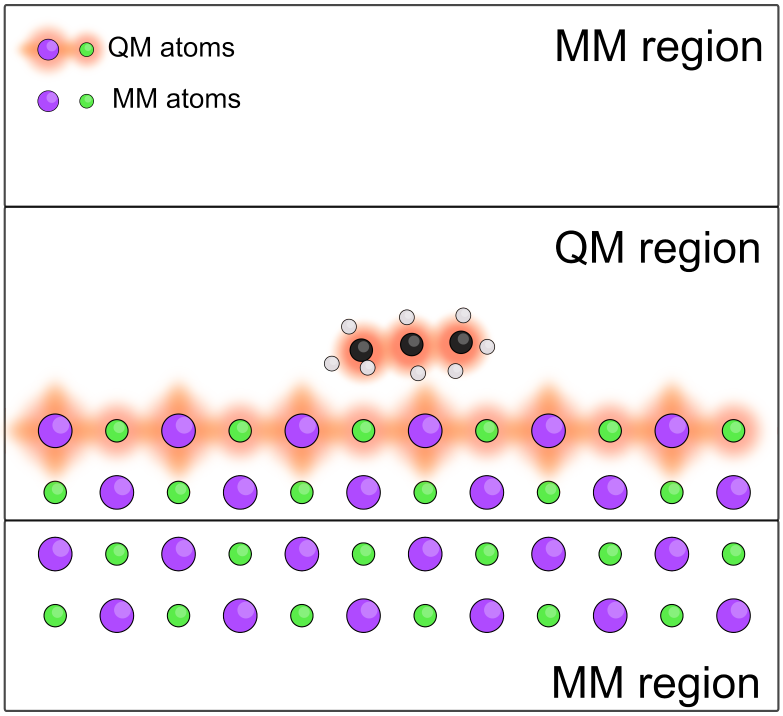

Figure 4: A schematic representation of the QM/MM model system. Quantum mechanically represented atoms are highlighted in red and the boundaries of the QM and MM regions are highlighted. MM atoms closer to the QM region have their charge density projected onto finer grids than those at the edge of the simulation box.

We have chosen a set of classical potentials derived by Catlow et al.[33] to represent the MM region. Since the shell model is not currently implemented within the QM/MM scheme in CP2K, we have fixed the positions of the shells to that of the cores in this model. The

QM region was represented using the methods described in the previous section: the GPW approach,[26,27] PBE/GGA[28,29] functional, MOLOPT[30] basis sets, and long range dispersion corrections.[31]

The first step in using such a representation is to verify that the properties of the surface do not change significantly in comparison to a full DFT calculation. Examining the electronic density of states of KCl represented using the QM/MM approach shows that the HOMO-LUMO gap of the material is only slightly reduced when fewer atoms are treated quantum mechanically, as shown in Figure 5. At the extreme limit, where there is only one QM layer and three MM layers, the band gap of KCl(100) is calculated to be 4.4 eV. Since the HOMO of the molecule lies nearly 2 eV above that of the surface, and the LUMO of the molecule lies nearly 2 eV below that of the surface as shown in Figure 3, a 1 eV reduction in the HOMO-LUMO gap of KCl(100) did not have any effect on the electronic structure of the adsorbed system. Similarly, increasing the band gap by using a hybrid functional also did not qualitatively change the interaction between CDB and KCl.

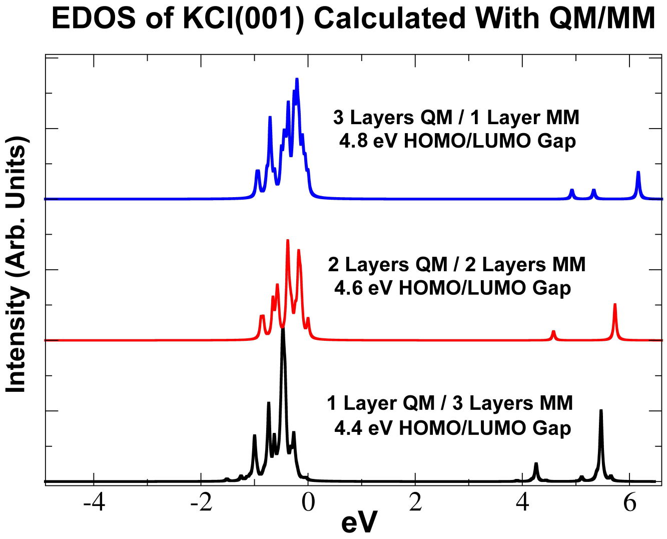

Figure 5: The electronic density of states for KCl(100) represented using QM/MM.

One common problem in previously employed embedded cluster methods is that electron density tends to redistribute from quantum mechanically treated atoms to the nearest classical atoms.[34,35] In previous studies, pseudopotentials are often applied to the cations nearest the QM region.[36] However, in CP2K the classical atoms are not represented as point charges, but are treated as a Gaussian charge distribution, greatly reducing this effect in our system. The size of the function used was chosen based on the ionic radius of the atom in question. In this work we used a value of 152 pm for K and 167 pm for Cl and charges of +1 and -1 respectively. An isosurface of the electron density of the system (0.001 electrons) is shown in Figure 6, showing that there is effectively no charge density present on classically treated atoms.

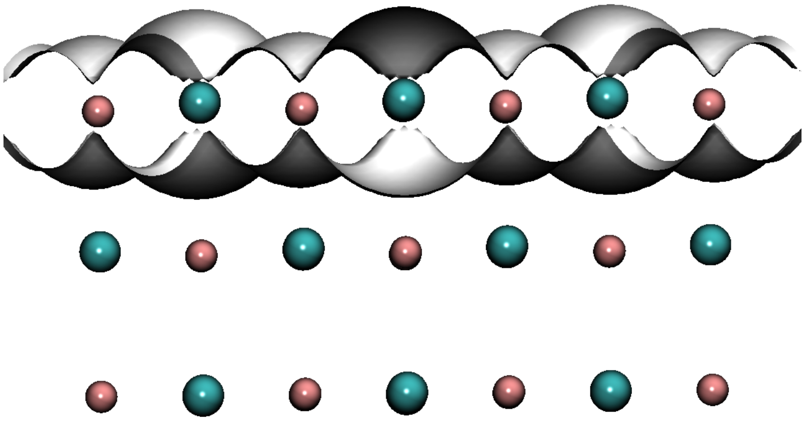

Figure 6: An isosurface (0.001 electrons) of the electron density of the QM/MM system. One layer of atoms has been represented quantum mechanically while the rest are classically treated. The system is periodic in 2 dimensions along X and Y. There is effectively no charge density on any of the surface atoms beyond the first layer.

A summary of the properties of the KCl(001) surface obtained from experiment,[32,37] standard DFT (GGA/PBE) methods, and QM/MM methods is shown in Table 1. These results indicated that the QM/MM representation provides a reasonable approximation for the insulating properties and accurately reproduces the structure of the KCl(100) surface.

This treatment greatly reduced the number of atoms treated quantum mechanically within our system, allowing us to generate a large database of data aimed at describing the full range of possible interactions between a CDB molecule and the KCl(100) surface. This database was then used to parametrize classical potentials.

Table 1: The properties of the KCl(001) surface from experimental data,[32,37] a full DFT (GGA/PBE) representation, and the QM/MM scheme using 1 QM layer and 3 MM atomic layers.

| | Lattice Constant | Surface Rumpling | Band Gap |
|---|---|---|---|
| Experiment | 6.3 Å | 0.03 Å | 7.6 eV |
| DFT (PBE-D2) | 6.3 Å | 0.03 Å | 5.4 eV |
| 1QM/3MM Layers | 6.3 Å | 0.04 Å | 4.4 eV |

# Classical Force Fields

While complete force fields representing CDB molecules on KCl(100) are not readily available, many of the components we need have been previously derived. We can separate the interactions into various categories, as shown in Figure 7, and use previously derived potentials when possible.

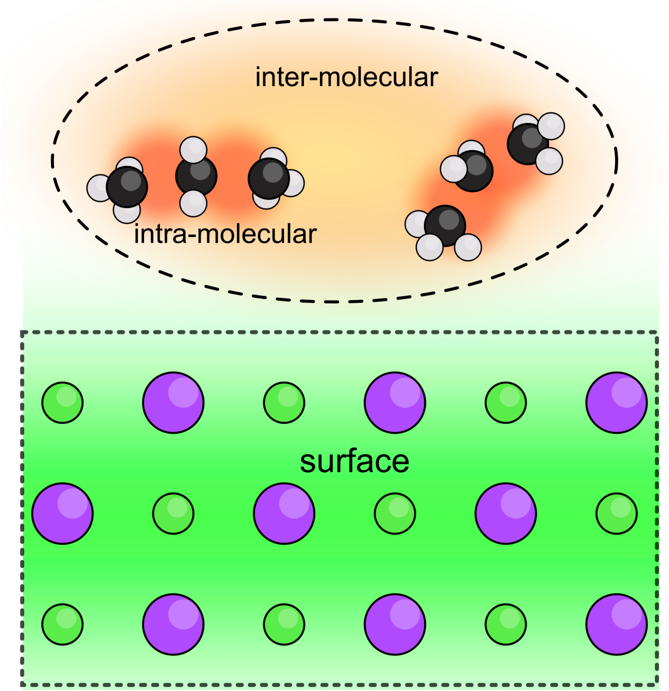

Figure 7: A schematic illustration of the interactions in our system. The area within the dashed line (highlighted in red and orange) represents intramolecular and intermolecular interactions. These contributions were obtained from CHARMM. The area within the dotted line (highlighted in green) represents interactions within the KCl surface. These contributions were obtained from a classical model derived by Catlow et al.[33] Interactions between the CDB molecule and the KCl(100) surface are not available in the literature.

Organic molecules are well studied and a number of popular forcefields are available (CHARMM,[8–10] AMBER,[11] UFF,[12] etc). We have selected the CHARMM force field[8–10] to describe the intramolecular interactions within CDB molecules, due to availability in most MD codes. This provides all of the bonded interactions within CDB, and a set of Lennard-Jones atomic parameters for non-bonded short-range interactions. Since CHARMM does not provide partial charges for the functional groups contained in the CDB molecule, we obtained them using Mulliken population analysis on DFT calculations of an isolated CDB molecule. The KCl surface has also been studied extensively in the past. For the sake of continuity we used the same classical potentials Catlow et al.[33] as in the previously described QM/MM scheme.

Interactions between the molecule and the surface, however, are not available in the literature. Furthermore, experimental fitting data is not available and the analytical forms of the two models selected are incompatible, making it impossible to apply any mixing rules. Therefore all mixed interactions must be parametrized using ab-initio data. It is important to note that all of the molecule-surface interactions within the system are defined explicitly and modelled with new parameters. The interactions included between the molecule and the surface do not change the original force fields chosen to represent interactions within CDB molecule or the KCl surface.

Since the functional forms we chose to represent our organic molecules and surfaces are not the same, it is unclear which analytical expression would best reproduce the molecule-surface interactions in the system. We considered several different analytical forms for fitting the short-range potentials between CDB and KCl atoms. The first two fitting models were of the Lennard-Jones type:

$$\begin{aligned} V_{ij}(r;\mathbf{p}) &= 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \\ \mathbf{p} &= \{\epsilon_{ij}, \sigma_{ij}\}, \end{aligned}$$

where $\mathbf{p}$ is the parameter vector, $i$ is an atom type within the CDB molecule and $j$ represents an atom type within the KCl surface. In the first model, the atomic parameters $\{\epsilon_K, \sigma_K\}$ and $\{\epsilon_{Cl}, \sigma_{Cl}\}$ for K and Cl were parametrized; all mixed interactions were then obtained by mixing these parameters with the ones for CDB atomic types, using Lorenz-Berthelot mixin rules. This Lennard-Jones-Mixing (LJM) model has a total of four fitting parameters. All mixed interactions in the second model were optimized independently. Since the CDB molecule is composed of 13 different atomic types and the surface contains 2 atomic types, this model has a total of 52 fitting parameters. This model is labelled Lennard-Jones-All (LJA).

We then considered the Morse potential form[38] (MRS):

$$\begin{aligned} V_{ij}(r;\mathbf{p}) &= D_{ij}\left[e^{-2\alpha_{ij}(r-\rho_{ij})} - 2e^{-\alpha_{ij}(r-\rho_{ij})}\right], \\ \mathbf{p} &= \{D_{ij}, \alpha_{ij}, \rho_{ij}\}, \end{aligned}$$

with a total of 78 parameters, and the Fumi-Tosi potential[39] (FT) with a total of 130 parameters:

$$\begin{aligned} V_{ij}(r;\mathbf{p}) &= A_{ij}\exp\left(\frac{\sigma_{ij}-r}{\rho_{ij}}\right) - \frac{C_{ij}}{r^6} + \frac{D_{ij}}{r^8}, \\ \mathbf{p} &= \{A_{ij}, \sigma_{ij}, \rho_{ij}, C_{ij}, D_{ij}\}. \end{aligned}$$

We then evaluated each of these potential forms to determine which one was best for describing the interactions between CDB molecules and the KCl(001) surface. Since these forms were originally designed to represent a variety of materials, they will not all be able to accurately represent the surface.

# Genetic Algorithm

Since each of the classical models described so far contains a large number of interdependent force field parameters, we could not use simple systematic optimization algorithms. In order to address this challenge, we used a home-built GA code with the algorithm shown in Figure 8a.

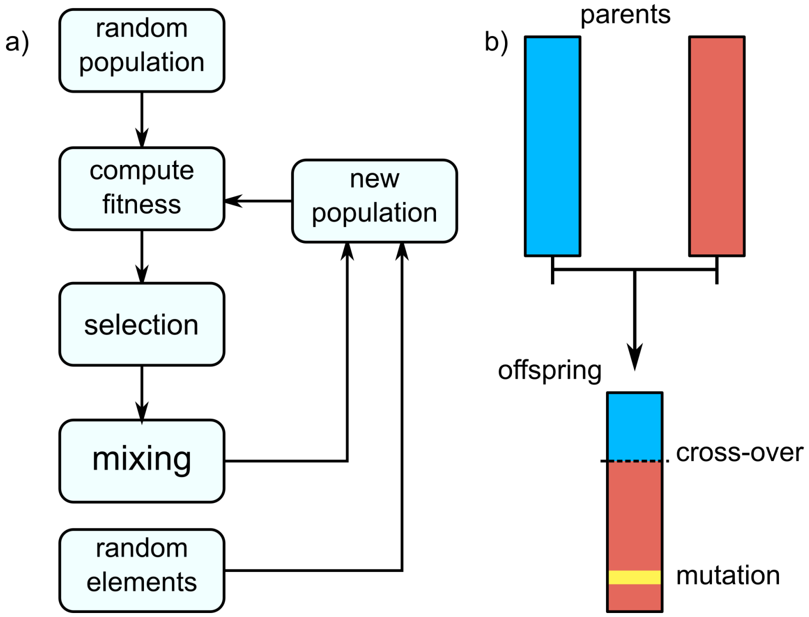

Figure 8: (a) Schematic description of the GA algorithm and (b) the mixing procedure with cross-over and mutation operations.

The main idea of the algorithm is to create a population of $N$ elements $e_0, e_1, ..., e_{N-1}$ with random genomes, evaluate their *fitness*, and intermix the best ones to generate an offspring population. In this study, we defined fitness as a measure of how well the classical force field reproduces *ab initio* data. Elements with higher fitness are more likely to be selected for mixing; their characteristics will be passed on to the next generation. After many generations, the genome eventually adapts to produce an accurate classical force field.

In our implementation, the genome of $e_k$ is the parameter vector $\mathbf{p}^{(k)}$ associated with the chosen analytical model.

$$\mathbf{p}^{(k)} = \left\{ p_0^{(k)}, p_1^{(k)}, ... \right\} . \tag{1}$$

Each parameter $p_i^{(k)}$ is constrained within the interval $[p_{min}, p_{max}]$, which prevents the algorithm from finding unphysical solutions, such as negative values for lengths. Furthermore, we rounded all parameter values to the second decimal place in order to facilitate identifying similar elements. When duplicate elements appear in the population, only one of them is preserved while the remainder are randomized. Finally, 5% of the total population is randomized every generation, introducing new genes into the system and reducing the rate of stagnation.

## Defining Fitness

Experimental values for the physical properties of CDB molecules adsorbed onto KCl(001) were not available, so our aim with this example was to show that the method can rapidly generate fitting data and parametrize a force field to represent the potential energy surface of the system obtained from DFT. The data available included atomic forces and energies at various adsorption geometries on the surface. We chose to parametrize our force field using forces and to evaluate them using previously published values for adsorption energy and geometry.

In order to select the best members of each population, we have defined the fitness of an element, $f\left(\mathbf{p}^{(k)}\right)$, as the difference between the forces on the molecular atoms calculated using DFT methods and the forces on the same atoms produced using the classical model defined by the parameters of that element:

$$f\left(\mathbf{p}^{(k)}\right) = -\left\langle \left| \vec{F}^{\mathrm{MD}} - \vec{F}^{\mathrm{DFT}} \right|^2 \right\rangle_{\mathrm{frames}}, \tag{2}$$

which is the squared mismatch between forces from DFT and the classical model, averaged over the configurations (or frames) in the training set. As the classical potential is trained to fit the DFT data, this discrepancy $f\left(\mathbf{p}\right) \leq 0$ approaches zero. The mismatch is calculated between the total force acting on the CDB molecule:

$$\vec{F}^{\mathrm{MD}} = \sum_{i}^{CDB} \vec{F}_i^{MD} \tag{3}$$

$$\vec{F}^{\mathrm{DFT}} = \sum_{i}^{CDB} \vec{F}_i^{DFT} \tag{4}$$

Atomic forces should not be used as fitness criterion as they cannot be partitioned into surface and intra-molecular components in the DFT data. Moreover, CHARMM intra-molecular interactions are different from DFT intra-molecular interactions, leading to an intrinsic mismatch between atomic forces that the GA will attempt to minimize in any way possible. The resulting molecule-surface interactions will then be unphysically optimized. We chose to consider the total force on CDB, since Newton's third law ensures that the contributions from intra-molecular forces cancel out numerically. This way we avoided incorrectly compensating for the mismatch between CHARMM and DFT interactions within the CDB molecule. Similarly, the forces on the surface that can be attributed to molecule-surface interactions are equal and opposite of those on the molecule, and were ignored in our analysis.

Each element in the population is then ranked according to their fitness and $N$ pairs were selected at random for mixing. The probability of choosing a particular $e_k$ is $\exp{(k/\lambda)}$ with:

$$\lambda = \frac{\log{(N)}}{N-1}. \tag{5}$$

This way, the probability $P(e_{N-1})$ of choosing the worst element is low:$P(e_0)/N$. The selection operation is then performed made based on the position of $e_k$ in the sorted population, rather than the fitness value; this ensures that bad elements are occasionally selected to enhance diversity.

The mixed genome $\mathbf{p}^{(\mathrm{new})}$ of $e_i, e_j$ is generated using a standard cross-over operation, where a random portion of the genome $\mathbf{p}^{(i)}$ is complemented by a portion from the other member $\mathbf{p}^{(j)}$, as shown in Figure 8b. The offspring genome can then be expressed as:

$$\mathbf{p}^{(\text{new})} = \left\{ p_0^{(i)}, ..., p_m^{(i)}, p_{m+1}^{(j)}, ..., p_n^{(j)} \right\}, \tag{6}$$

where $m$ is an index chosen at random. Furthermore, there is a 0.2% chance of randomly mutating one of the values $\mathbf{p}_k^{\text{new}}$ during each operation.

## Acceptance Criteria

When using GA optimization algorithms, it is important to remember that any perceived convergence can be misleading. The technique has not been mathematically proven to eventually find the best solution to the problem, and indeed it is possible that a sudden random mutation can greatly improve the fitness of an element within the population. This results in great difficulties when determining how many generations are needed before an acceptable set of results has been produced.

In the context of the methods discussed here, however, we were able to define an acceptance criterion as a replacement for the more commonly used convergence criteria. Since the model uses the CHARMM force field to represent interactions within the CDB molecules, a certain amount of error between DFT and classical forces already exists within the system. For the isolated CDB molecule in the minimized lowest energy configuration, we calculated the difference in forces produced by DFT and CHARMM, to be on the order of 5%. Since this comparison was made at the lowest energy configuration of the molecule, these forces are very close to zero already.

If the average difference in the force on the CDB molecule when it is adsorbed onto KCl(001) between our QM/MM simulations and the optimized force field is comparable, then it is reasonable to consider it to be fairly high quality. Since this average value would include comparisons performed at geometries other than the minimum energy configuration, it would indicate that the force field is able to provide a good description of the molecule-surface interactions between CDB and KCl for a broad range of geometries and positions of the molecule. For this study, we then set an acceptance criterion of 5% average error in forces when compared to QM/MM data. It is important to note that the acceptance criteria simply represents a numerical comparison between the mismatch between the optimized interactions and the original CHARMM force field and cannot be used as a physical evaluation of the force field alone.

# Results

Since experimental data describing the properties of CDB molecules on KCl(100) were not available, we used the previously described method to generate classical force fields for CDB on the KCl(100) surface.

The training set created for CDB on KCl(100) consisted of 210 atomic configurations for the system. The first 80 frames (or configurations) were obtained from QM/MM MD calculations starting from the ground state of the adsorbed molecule. The starting position of the molecule was then rotated in 15° increments and a new trajectory was calculated at each starting position, resulting in 90 additional frames. The final 40 frames were created by artificially positioning the molecule at varying heights above the surface. These frames represent situations that would not normally be probed via room temperature MD and provide information about the system when the molecule is farther or closer to the surface plane.

For each previously described functional form, a population of $N = 1024$ elements was evolved over 1000 generations. Since GA relies heavily on randomness, the calculation was repeated 5 times for each model. In our preliminary tests, we found that populations as small as $N = 256$ and $N = 512$ could not generate reproducible results, due to a rapid convergence of the genes. Large populations of $N = 1024$ or more members tend to remain diverse for enough generations to produce similar results.

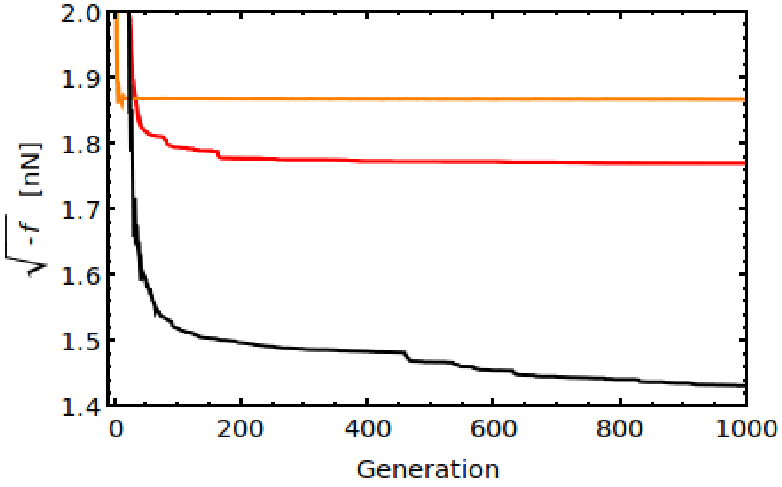

Figure 9: Fitness of the best element for LJM (dashed, orange), LJA (dotted, red), and MRS (solid black) models over 1000 GA generations. Solid and dashed lines represent different runs with random initial conditions.

Figure 9 shows the fitness (square root of $f(\mathbf{p})$) of the best element over 1000 generations for the LJM, LJA and MRS models. The absolute best fit was achieved by the MRS analytical form, with a mean discrepancy of 1.40 nN. LJA and LJM also produced reasonably low mean discrepancies of 1.78 nN and 1.86 nN, respectively. The LJM model converges rapidly since it contains fewer parameters than the others, but results in the worst fitness out of these functional forms. Finally, the FT produced a fitness value three orders of magnitude worse than these models and will not be discussed further.

When the fitness of LJ and MRS models is translated into forces, the average error on the total CDB force is always within 5% of the DFT reference values; all of these models satisfy the acceptance criteria. However, there are important differences between the LJ and MRS models that are not immediately apparent.

Figure 10 shows a comparison between DFT reference forces and classical forces obtained using the LJA model over the entire fitting data set. Large differences were observed throughout the entire set. It is quite evident from the last 40 frames, where the molecule is lifted above the equilibrium adsorption height, that the LJ model underestimates the the attractive interactions between the molecule and the surface. These results can be compared

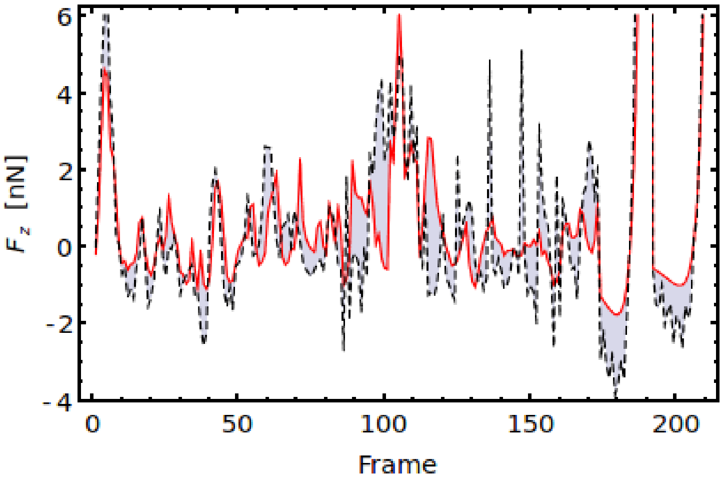

Figure 10: Comparison between DFT reference forces (black dashed) and ones obtained from the best fitted LJA model (solid red). Only the component normal to the surface plane is shown.

to the MRS model, which is shown in Figure 11.

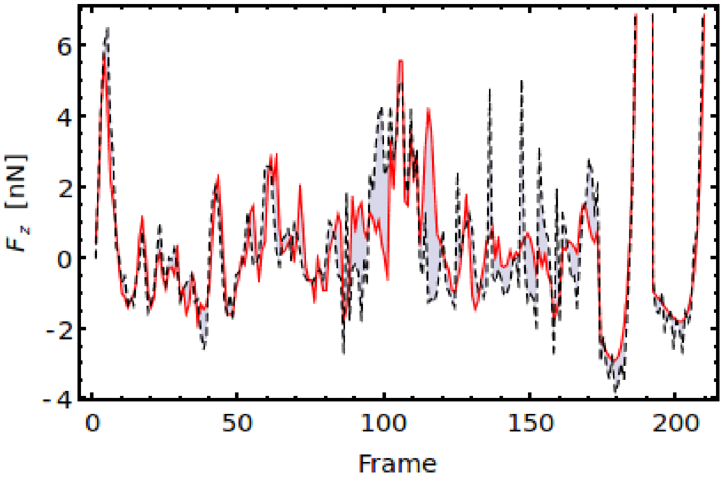

Figure 11: Comparison between DFT reference forces (black dashed) and ones obtained from the best fitted Morse model (solid red). Only the component normal to the surface plane is shown.

This improved model is able to reproduce the first 80 training frames, from near-ground state DFT-MD, as well as the last 40, indicating a better representation of the distance dependence of the molecule-surface interaction.

Finally, we evaluated these force fields by examining physical properties of the system that were not directly used as fitting criteria. The LJ and MRS models both predict that the lowest energy configuration of an isolated CDB molecule on KCl(100) is where the cyano groups are anchored to surface cation sites and the body of the molecule is aligned in the [110] direction. This is consistent with previously published DFT results.[25] These models can also be evaluated by comparing the adsorption energies they produce to DFT data. While the exact value of this adsorption energy depends greatly on the dispersion correction used in the training set, the goal of our GA scheme was simply to reproduce DFT results. The adsorption energy of CDB was estimated using DFT to be 3.1 eV[25] while LJA gives an adsorption energy of 2.6 eV and MRS gives better agreement with an adsorption energy of 2.8 eV.

It is clear that the final force fields must be evaluated using properties that were not explicitly incorporated into the fitting data. Furthermore, it is important to remember that it is difficult for a single classical force field to reproduce all of the physical properties of a material. If a particular property of a system is critical, then it should be directly included in the fitting data.

## Discussion and Conclusions

One of the main challenges with GA based optimization schemes is that convergence is not guaranteed, even to a local minimum. A convergence criteria, i.e. RMS deviation between successive generations, can not be defined since the step size between generations can not be defined. Furthermore, the variational principle is not applicable in these systems. The fitness of the best element can remain constant for 100 generations before a random mutation appears to improve the population, as shown in Figure 9. However, within the framework of this scheme, an acceptance criterion was defined based on the average error between reference forces from DFT and the CHARMM force field. For the isolated CDB molecule, CHARMM interactions already result in a 5% error in forces when compared to DFT. Each parametrized model that then gives a comparable error is of acceptable quality.

When reasonable interaction models are considered, the algorithm was able to find a solution giving less than 5% error within 1000 generations.

Another important challenge that we encountered was ensuring the completeness of the training set. Simply taking atomic configurations from MD trajectories was not sufficient to represent the full range of interatomic interactions, since the CDB molecule is always found near its equilibrium position,. The model was *overtrained* to reproduce that one particular state. In our initial attempts, LJ models were driven towards unphysically strong interactions between certain atom types in order to reproduce the DFT-MD training set. The algorithm did not know that such strong interactions produced catastrophic instabilities when CDB atoms came as little as 0.1 nm closer to KCl. The inclusion of additional configurations in the repulsive regime balanced the fitness criteria, and the models produced were found to be stable throughout up to 10 ns of MD.

It is important to note that not all functional forms are appropriate for all systems. CDB-KCl interactions are strong, as indicated by the large adsorption energy calculated using DFT. Our fit shows that LJ models fail to reproduce such interactions because they are too soft. Therefore, Morse potentials were better suited for the task.

Moreover, the LJ parameter $\sigma$ controls the width (and slope) and position of the energy well simultaneously, while these are independently tuned by $\alpha$ and $r_0$ in the Morse model. This increased freedom in shaping the energy profiles mathematically guarantees a better fit. On the other hand, the Fumi-Tosi potential relies on a delicate balance between the exponential and the $1/r^n$ terms. Minor changes in one parameter may cause the shape of the energy profile to change dramatically, so that the minimum disappears. Since we were not able to restrict the parameter search space to a sensible, narrow range beforehand, it would take several more GA iterations to find appropriate values, and this method becomes inefficient.

Finally, we examined the transferability of this particular force field by considering a variation of the CDB molecule. This variant is called 1,3,5-tri(4”-cyano-4,4’-biphenyl) benzene

(TCB) and is composed of three cyano-benzene functional groups attached to a central ring. A direct mapping of the atom types within the CDB molecule resulted in a charged TCB molecule, which is unphysical. In order to solve this problem, we reassigned the charges on TCB atoms from Mulliken population analysis using DFT results. However, the MRS potentials were parametrized to compliment existing Coulomb interactions within the system; the charges within the molecule should not be adjusted without simultaneously reparametrizing all the components. Such a treatment resulted in a force field that could qualitatively predict the lowest energy adsorption geometry of the molecule, but severely underestimated the value of the total adsorption energy by 50%. It is clear that while it may be possible to manually adjust the charges within the molecule in order to produce a higher quality force field using the same MRS components, this would require significant chemical intuition and can not be done systematically. Fortunately, the scheme presented here is efficient enough that it is possible to reparametrize the force field for each molecular variant directly from *ab initio* data.

To summarize, in this study we have developed an efficient scheme for fitting molecule-surface force fields by combining QM/MM embedding techniques with GA. This scheme avoids the need to evaluate large molecules as a combination of smaller fragments and allows us to optimize all the parameters within our system simultaneously. We showed how GA methods can produce a reasonably good fit at a moderate computational cost, suggesting that it can be employed as a routine method.

## Acknowledgement

The authors would like to thank Prof. K. Akagi for computational resources and helpful discussion. David Gao is grateful to Chevron Oronite Company for financial support. Matthew Watkins is grateful to the Leverhulme Trust (grant F/07134/CK) for financial support. The authors acknowledge the use of the HECToR High Performance Computing Facility via our membership to the UK‘s HPC Materials Chemistry Consortium, which is funded by EPSRC (EP/F067496). David Gao acknowledges the use of the Serenity computing cluster, and support services provided by Nanolayers.com for the completion of this work.

## References

1. Besenbacher, F.; Lauritsen, J. V.; Wendt, S. *Nano Today* **2007**, *2*, 30–39.

2. Joachim, C.; Gimzewski, J. K.; Aviram, A. *Nature* **2000**, *408*, 541–548.

3. Heath, J. R. *Rev. Mater. Res.* **2009**, *39*, 1–23.

4. Song, H.; Reed, M. A.; Lee, T. *Adv. Mater.* **2011**, *23*, 1583–1608.

5. Barnes, A. M.; Bartle, K. D.; Thibon, V. R. A. *Tribol. Int.* **2001**, *34*, 389–395.

6. Wright, L.; Rodger, M.; Corni, S.; Walsh, T. *J. Chem. Theory Comput.* **2013**, *9*, 1616–1630.

7. Wright, L.; Rodger, M.; Walsh, T.; Corni, S. *J. Phys. Chem. C* **2013**, *117*, 24292–24306.

8. Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. *et al. J. Comput. Chem.* **2009**, *30*, 1545–1614.

9. Brooks, B.; Bruccoleri, R.; Olafson, D.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

10. MacKerel Jr., A.; Brooks III, C.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In ; v. R. Schleyer et al., P., Ed.; The Encyclopedia of Computational Chemistry; John Wiley & Sons: Chichester, 1998; Vol. 1; pp 271–277.

11. Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K. J.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

12. Rappe, A.; Casewit, C.; Colwell, K.; III, W. G.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

13. Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford University Press: Oxford, UK, 1987.

14. Walker, P.; Mezey, P. *J. Am. Chem. Soc.* **1993**, *115*, 12423.

15. Sato, F.; Hojo, S.; Sun, H. **2003**, *107*, 248–257.

16. Vellore, N. A.; Yancey, J. A.; Collier, G.; Latour, R. A.; Stuart, S. J. *Langmuir* **2010**, *26*, 7396–7404.

17. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory and Comput.* **2012**, *8*, 3257–3273.

18. Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1*, 1176–1184.

19. Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2006**, *2*, 2178–2189.

20. Hunger, J.; Beyreuther, S.; Huttner, G.; Allinger, K.; Radelof, U.; Zsolnai, L. *Eur. J. Inorg. Chem.* **1998**, *1*, 693–702.

21. Hunger, J.; Huttner, G. *J. Comput. Chem.* **1999**, *20*, 455–471.

22. Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2001**, *22*, 1219–1228.

23. Strassner, T.; Busold, M.; Herrmann, W. A. *J. Comput. Chem.* **2002**, *23*, 282–290.

24. Larsson, H. R.; van Duin, A. C. T.; Hartke, B. *J. Comput. Chem.* **2013**, *34*, 2178–2189.

25. Amrous, A.; Bocquet, F.; Nony, L.; Para, F.; Loppacher, C.; Lamare, S.; Palmino, F.; Cherioux, F.; Gao, D. Z.; Federici-Canova, F. *et al. Adv. Mater. Interfaces* **2014**, *1*, 1400414.

26. Lippert, G.; Hutter, J.; Parrinello, M. *M. Mol. Phys.* **1997**, *92*, 477–487.

27. VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

28. Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

29. Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

30. VandeVondele, J.; Hutter, J. *J. Chem. Phys.* **2007**, *127*, 114105–1–9.

31. Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

32. Roessler, D. M.; Walker, W. C. *Phys. Rev.* **1968**, *166*, 599.

33. Catlow, C. R. A.; Diller, K. M.; Norgett, M. *J. Phys. C: Solid State Phys.* **1977**, *10*, 1395–1412.

34. Sushko, M.; P. Sushko, U. A.; Shluger, A. *J. Comput. Chem.* **2010**, *31*, 2955–2966.

35. Das, D.; Eurenius, P.; Billings, E.; Sherwood, P.; Chatfield, C.; Hodoscek, M.; Brooks, B. *J. Chem. Phys.* **2002**, *117*, 10534–10547.

36. Berger, D.; Logsdail, A.; Oberhofer, H.; Farrow, M.; Catlow, C.; Sherwood, P.; Sokol, A.; Blum, V.; Reuter, K. *J. Chem. Phys.* **2014**, *141*, 024105.

37. Vogt, J.; Weiss, H. *Surf. Sci.* **2001**, *491*, 155–168.

38. Morse, P. M. *Phys. Rev.* **1929**, *34*, 57–64.

39. Fumi, F. G.; Tosi, M. P. *J. Phys. Chem. Solids* **1964**, *25*, 31.