





RESEARCH ARTICLE | JUNE 12 2023

Black box vs gray box: Comparing GAP and GPrep-DFTB for ruthenium and ruthenium oxide

Special Collection: [Modern Semiempirical Electronic Structure Methods](#)

C. Panosetti ; Y. Lee ; A. Samtsevych ; C. Scheurer 



J. Chem. Phys. 158, 224115 (2023)

<https://doi.org/10.1063/5.0141233>



View
Online



Export
Citation

CrossMark



The Journal of Chemical Physics

Special Topic: Adhesion and Friction

Submit Today!

AIP
Publishing

AIP
Publishing

Black box vs gray box: Comparing GAP and GPrep-DFTB for ruthenium and ruthenium oxide

Cite as: J. Chem. Phys. 158, 224115 (2023); doi: 10.1063/5.0141233

Submitted: 4 January 2023 • Accepted: 23 May 2023 •

Published Online: 12 June 2023



View Online



Export Citation



CrossMark

C. Panosetti,^{a)} Y. Lee, A. Samtsevych, and C. Scheurer

AFFILIATIONS

Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Note: This paper is part of the JCP Special Topic on Modern Semiempirical Electronic Structure Methods.

^{a)} Author to whom correspondence should be addressed: panosetti@fhi.mpg.de

ABSTRACT

The increasing popularity of machine learning (ML) approaches in computational modeling, most prominently ML interatomic potentials, opened possibilities that were unthinkable only a few years ago—structure and dynamics for systems up to many thousands of atoms at an *ab initio* level of accuracy. Strictly referring to ML interatomic potentials, however, a number of modeling applications are out of reach, specifically those that require explicit electronic structure. Hybrid (“gray box”) models based on, e.g., approximate, semi-empirical *ab initio* electronic structure with the aid of some ML components offer a convenient synthesis that allows us to treat all aspects of a certain physical system on the same footing without targeting a separate ML model for each property. Here, we compare one of these [Density Functional Tight Binding with a Gaussian Process Regression repulsive potential (GPrep-DFTB)] with its fully “black box” counterpart, the Gaussian approximation potential, by evaluating performance in terms of accuracy, extrapolation power, and data efficiency for the metallic Ru and oxide RuO₂ systems, given exactly the same training set. The accuracy with respect to the training set or similar chemical motifs turns out to be comparable. GPrep-DFTB is, however, slightly more data efficient. The robustness of GPrep-DFTB in terms of extrapolation power is much less clear-cut for the binary system than for the pristine system, most likely due to imperfections in the electronic parametrization.

© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0141233>

I. INTRODUCTION

The last couple of decades marked the beginning of what we may call the era of machine learning in atomistic computational modeling. The first approaches emerged toward addressing the most natural problem a computational chemist has to tackle—the solution of the Born–Oppenheimer ground state Schrödinger equation. Machine learning interatomic potentials, which address precisely that issue, are by far the most established and advanced class of ML applications in computational molecular and materials science. From the pioneering works of Behler and Parrinello¹ and Bartók *et al.*² with the development of the first neural network potentials (NNP) and Gaussian approximation potentials (GAP), respectively, to the myriad present day approaches (e.g., Refs. 3–8, only to name a few), the toolbox available to this day is likely to suit any need. For a more comprehensive overview, we refer the reader to the excellent review in Ref. 9.

However, there is reason to believe that explicit electronic structure methods are not and will never be completely surpassed. First and foremost, the explicit electronic structure allows us to compute properties as observables within the quantum mechanics formalism in a rigorous and straightforward manner. In addition, properties that are not observable (e.g., partial charges) are much more easily evaluated on the basis of explicit electronic structure. One notorious challenge for ML interatomic potential is, in fact, the correct treatment of charges due to the use of local representations of atomic environments, which neglect long-range electrostatic interactions—work in that direction is an active field of research, with the recent development of approaches such as kernel charge equilibration (kQEq).¹⁰ Furthermore, one simply cannot do, e.g., theoretical spectroscopy with ML potentials alone (except Raman or rotational/vibrational). Clearly, spectra can be learned just like any other data (see, e.g., Ref. 11), but a theoretical spectroscopy ML model for a certain physical system is not necessarily related to

an ML model for structural, thermodynamic, and dynamic properties of the same system unless specifically targeted to combine different models to learn the potential energy surface jointly with electronic properties.¹² Relatedly, in parallel to the development of ML interatomic potentials, a plethora of approaches are emerging to directly target electronic structure, such as learning the electron density¹³ or learning molecular orbitals.¹⁴ These offer the immediate advantage that any property that is represented by a quantum mechanics (QM) observable is directly accessible within the same formalism as QM—a more general approach than directly targeting properties.

A convenient synthesis of the best of two worlds is also met in approaches that combine some form of ML with some form of, usually approximate, QM. Notable examples are based on density functional tight binding¹⁵ (DFTB) in combination with ML to either fit the repulsive potential^{16,17} or to use the DFTB model as an approximate but physically robust baseline to then learn only the difference with some more accurate (typically DFT or beyond) model (see, e.g., Ref. 18). Both of these types of approaches are essentially forms of Δ -learning.

In this work, we intend to explore the complex answer to one simple question: how will two analogous approaches—one fully ML, say “black box,” and one only partially ML, say “gray box”—perform in terms of accuracy, transferability, and data efficiency given the exact same training set?

For this purpose and without any pretense of completeness, we choose two widely established methods for which all the required software is readily and publicly available: (i) the GAP approach as the “black box” contestant, and (ii) self-consistent-charge density functional tight binding¹⁵ (SCC-DFTB, from here on referred to as DFTB) with a Gaussian Process Regression repulsive potential,¹⁶ from now on referred to as GPrep-DFTB, as the semi-empirical, “gray box” contestant. Essentially, the second is an adapted Δ -learning version of the first, where the semi-empirical physical baseline is represented by repulsion-less DFTB (cf. Sec. II B 2) and semilocal DFT is the reference method—the latter being directly machine learned in the full-ML GAP approach. To compare the two models, we chose a monoatomic system (metallic ruthenium) and a binary system (ruthenium oxide) as benchmarks. This choice is first motivated by the fact that a training set for ruthenium oxide was already available in our group. However, the technological relevance of the associated materials in heterogeneous catalysis is a pleasant side aspect—this work will carry the additional benefit of making two cheap models (and, more importantly, the associated datasets) available for public use for the large-scale modeling of these materials. Furthermore, transition metal oxides, in general, are a challenging class of materials for DFTB due to the large internal charge transfer between metal and oxygen and the many oxidation states that appear and coexist. In that regard, this comparison is all but trivial.

Intuitively, one may expect that a “gray box” approach would guarantee more robust transferability and larger stability with respect to extrapolation. Overall accuracy may be expected to be comparable. One could expect GAP to be more accurate within the training set thanks to the many-body descriptor, with GPrep-DFTB being intrinsically more of a compromise due to the less sophisticated pairwise nature of the ML descriptor. However, for the same reason regarding transferability, there is a chance that GPrep-DFTB

is on average more accurate across a larger configurational space. In the following, we will discuss whether these intuitive expectations are met.

II. METHODS

A. Gaussian approximation potential (GAP)

Gaussian Approximation Potential (GAP) is one of the widely used descriptor-based machine-learning interatomic potentials using Gaussian process regression (GPR).^{2,19,20} The GAPs herein are trained for a surrogate model of the potential energy surface (PES) of particularly stable surfaces, and they calculate the total energy E_{GAP} of the system from its atomic coordinates \mathbf{X} as

$$E_{\text{GAP}}(\mathbf{X}) = \underbrace{\sum_{i,j} \delta_{2\text{B}}^2 \sum_{m=1}^{M_{2\text{B}}} c_{m,2\text{B}} \cdot k_{2\text{B}}(r_{ij}, r_m)}_{E_{2\text{B}}} + \underbrace{\sum_i \delta_{\text{MB}}^2 \sum_{m=1}^{M_{\text{MB}}} c_{m,\text{MB}} \cdot k_{\text{MB}}(\chi_i, \chi_m)}_{E_{\text{MB}}} \quad (1)$$

In detail, it consists of a two-body (2B) $E_{2\text{B}}$ and a many-body (MB) E_{MB} energy contribution. $E_{2\text{B}}$ sums over all atomic pairs of atoms i and j , while E_{MB} sums over each atom i . The second sum goes over a set of $M_{2\text{B}/\text{MB}}$ representative data points and includes the regression coefficients $c_{m,2\text{B}/\text{MB}}$ and the kernel basis functions $k_{2\text{B}/\text{MB}}$. The k_{MB} measures the similarity between two local geometric descriptors (representations) χ computed from \mathbf{X} . The 2B contribution simply depends on interatomic distances r_{ij} within a specific cutoff radius r_{cut} . The MB contribution is based on the Smooth Overlap of Atomic Positions (SOAPs), which provides a translationally, rotationally, and permutationally invariant local atomic representation.²¹

The RuO₂ GAP employed in this work is identical to that employed in our previous work.²² To train the Ru GAP, the same active-learning algorithm proposed by Timmermann *et al.* is used.²² As a brief recap, in the case of the RuO₂ GAP, an initial model is trained on a set of atomic information, gas phase O₂ dimer data with varying O–O bond lengths, rutile RuO₂ bulk structures at optimized and constrained lattice constants and those with displaced internal coordinates, and different terminations of all five low-index facets in the bulk-truncated geometry and the DFT local optimized geometry. Similarly, for the Ru GAP, the initial training set includes single-atom information, optimized and constrained bulks with face centered cubic (fcc) and hexagonal close packed (hcp) crystal structures, those with displaced internal coordinates, and their surface structures in bulk-truncated and DFT local optimized geometries. Ru facets are limited to surfaces up to a maximum Miller index of two for non-cubic hcp and three for cubic fcc, generating a total of 12 and 13 surface orientations for hcp and fcc Ru, respectively. Each potential is then refined by generating new surface configurations via GAP-driven global optimization (simulated annealing) for all low-index surfaces. The generated structures are compared to the structures in the training database via SOAP similarity, and some of those that are dissimilar to the database, selected via farthest point sampling (FPS),^{22,23} are added to the training set after DFT geometry

relaxations. In detail, we measure the similarity between two surface structures, A and B , via a kernel distance,

$$\kappa(A, B) = \sqrt{2 - 2 \min_{\substack{a \in A \\ b \in B}} (k_{\text{MB}}(\chi_a, \chi_b))}. \quad (2)$$

New structures having a larger κ than a system-specific parameter κ_{crit} are then considered to be sufficiently dissimilar. The values of κ_{crit} for Ru and RuO₂ were tested and selected as 0.050 and 0.075, respectively. This refinement cycle is repeated until the FPS does not capture further unknown basins from the global optimization. Each converged training set consequently contains collective PES minima for all considered facets. Most technical hyperparameters were optimized by heuristics and four-fold cross-validations on the initial training sets. The cutoff radius of the atomic descriptor was selected via the convergence of the force locality, which is illustrated in Sec. II A. All hyperparameters used for the 2B and MB SOAP descriptors are tabulated in Table I. More details on the hyperparameter selection process are described in Ref. 22.

Converged training databases for RuO₂ and Ru contain 182 structures (of which 148 are surfaces) and 197 (of which 140 are surfaces), respectively.

B. GPrep-DFTB

1. DFTB electronic parametrization

To briefly recall the idea behind DFTB, for the details of which we refer the reader to, e.g., Refs. 24 and 25, the self-consistent charge DFTB (SCC-DFTB) energy is obtained as a second order expansion of the DFT energy around a non-interacting density. Using tight binding approximations, the latter is recast into simple algebraic expressions that only depend on a handful of parameters (the so-called electronic part of the interaction), plus a repulsive energy that, in principle, lumps all the energy contributions together that are missing from the electronic energy. The parameters involved in the electronic interaction are the so-called onsite energies ϵ_l , corresponding to free atom eigenvalues of the valence orbitals of angular momentum l , the Hubbard parameters U_l (in principle, orbitally resolved, but most often a single value U), and the coefficients of a confinement potential used to mimic the compression of atomic densities upon chemical bonding. Onsite energies were

TABLE II. Optimized electronic parameters. The values are given in atomic units (Hartree for energy, Bohr for distances) for consistency with the standard Slater–Koster file format.

Parameter	Ru (metal)	Ru (RuO ₂)	O (RuO ₂)
ϵ_d (hartree)	−0.229 194	−0.229 194	...
ϵ_p (Ha)	−0.031 323	−0.031 323	−0.338 655
ϵ_s (Ha)	−0.171 091	−0.171 091	−0.928 529
U (Ha)	0.329 726	0.329 726	0.495 405
r_0 (bohr)	10.000	2.863	5.970
r_{cut} (bohr)	46.629	4.362	7.422

kept fixed at the computed free atom eigenvalues (see Sec. IV), while U values were kept fixed at those tabulated in Ref. 26. A Woods–Saxon potential was employed to confine the atomic densities, with two parameters per atomic species that were optimized with a particle swarm optimization (PSO) approach as proposed by Chou *et al.*²⁷ Electronic parameters for Ru and for the Ru–O system were optimized separately. The cost function consists of a measure of deviation (using distance matrix as metrics, as performed in Ref. 28) between the DFTB and DFT band structures at (i) equilibrium hcp Ru geometry as well as compressed with a factor 0.9 for the monoatomic Ru system, and (ii) equilibrium rutile RuO₂ geometry as well as compressed with a factor 0.9 for the composite Ru–O system. The resulting optimal confinement parameters are reported in Table II, and the corresponding band structures are shown in Fig. 1.

As one may immediately notice, the “optimal” DFTB band structure does not necessarily correspond to a perfect overlap with the valence DFT band structure. There are many reasons to prefer some compromise: first and foremost, it is commonly known in the DFTB parametrization community that “too perfect” band structures often make repulsion hard to fit. This is mostly due to the fact that a good band structure at equilibrium geometry does not guarantee a good band structure at out-of-equilibrium geometry, where artifacts may appear due to the limited DFTB basis set. This is especially important at compressed geometries, where the DFT–DFTB force residues, which are ultimately fitted in the repulsion parametrization, dominate. More specifically in this case, for

TABLE I. Hyperparameters used for the GAP models in this work.

Description	Symbol	RuO ₂		Ru	
		2B	SOAP	2B	SOAP
Cutoff (Å)	r_{cut}	5.0		5.5	
Kernel width (Å)	σ	1.0	0.600	1.0	0.688
Scaling factor (eV)	δ	0.326	0.086	0.414	0.174
SOAP basis	$n_{\text{max}}/l_{\text{max}}$...	8/4	...	8/4
Number of sparse points	M	25	2000	25	2000
Kernel exponent	ζ	...	2	...	2
Regularization factors	σ_e (eV)/ σ_f (eV/Å)	0.001/0.01		0.001/0.01	

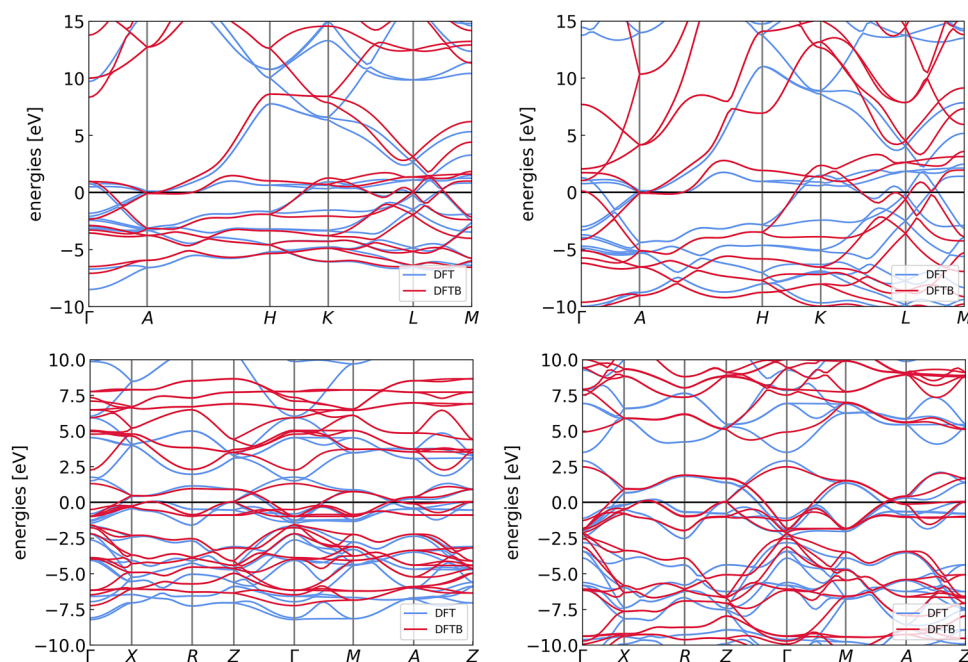


FIG. 1. PSO-optimized band structures for hcp Ru (top) and rutile RuO_2 (bottom). The DFTB band structure is shown in magenta, while the DFT reference is shown in blue. Equilibrium geometry band structures are shown on the left side. The right side shows band structures for compressed geometries with a compression factor of 0.9. All the band structures are aligned with respect to the DFT Fermi level. The spurious conduction bands visible between the Gamma and A special points for hcp Ru are problematic because they can intrude into the valence region upon compression. The chosen confinement potential avoids that; however, a truly optimal confinement potential that would yield a perfect valence band structure for equilibrium hcp Ru would suffer from this issue.

pure Ru, spurious conduction bands appear around the Gamma point, which does not matter much at equilibrium and elongated geometries (because they remain unoccupied), but intrudes the valence region at compressed geometries. This caused the electronic interaction to dip too much at shorter distances, making it impossible to fit a good repulsion (see Appendix B). For this reason, we also included compressed band structures in the cost function, and the final confinement represented a compromise allowing a decent band structure at equilibrium geometry (especially around the Fermi level) while preventing spurious bands from intruding too much at compressed geometries. We did not observe a similar problem for the Ru–O system; nonetheless, for consistency, we used an analogous cost function.

One thing that is worthwhile pointing out is that the optimal confinement for the individual atoms in a pure system does not necessarily coincide with the optimal confinement for a composite system. In particular, for Ru, the optimal confinement for the pure system is much longer and softer than that for the Ru–O system. The wavefunction range, which results from both r_0 and r_{cut} , is a useful, single interpretable indicator of the effect of confinement. For Ru in the pure system, the resulting wavefunction range is circa 5.0 Å compared to circa 1.8 Å for the composite system. In other words, in the oxide, optimal band structures are found because the atomic size of Ru is almost three times smaller in the oxide than in the pristine metal. Nicely, this reflects the fact that the atomic sizes change significantly upon oxide formation [Ru(IV) in RuO_2 is much smaller than Ru(0)]. The compression ratio naturally found by PSO is comparable with the ratio of neutral and ionic radii of Ru. One has to keep in mind, however, that the resulting electronic parameters for Ru in Ru–O are most likely not at all optimal to describe crystalline Ru, at least not without a significant effort in the repulsion parametrization and certainly with a cost in terms of accuracy-transferability trade-

off. We did not explore this further because it would be out of scope for the present work.

2. DFTB repulsion parametrization: GPrep

In the following, we briefly recap the derivation of GPrep, following the same notation as Ref. 16. In standard DFTB implementations, the repulsive potential is empirically approximated by a sum of pairwise potentials,

$$E_{\text{rep}} = E_{\text{DFT}} - E_{\text{DFTB}} \simeq \frac{1}{2} \sum_I \sum_{J(\neq I)} V_{\text{rep}}(R_{IJ}), \quad (3)$$

where the sum runs over atom pairs IJ with interatomic distance $R_{IJ} = |\mathbf{R}_I - \mathbf{R}_J|$. The GPrep approach¹⁶ can be considered a simplified version of the GAP approach. The main simplification lies in the use of a pairwise descriptor rather than the many-body SOAP descriptor. In principle, nothing precludes the usage of many-body descriptors or, more generally, of many-body formulations of the DFTB repulsion itself, regardless of whether the repulsion is fitted using machine learning or not (see, e.g., Refs. 17, 29, and 30). However, a pairwise repulsion directly complies with Eq. (3) and allows for the generation of parameter files in the standard .skf format compatible with all the available software implementations of DFTB.

In a Gaussian Process Regression (GPR) formulation, V_{rep} is modeled as a linear combination of kernel functions,

$$V_{\text{rep}}(R) = \sum_{I, J \in \{\mathbf{X}\}} \alpha_{IJ} \cdot k(R, R_{IJ}), \quad (4)$$

where the sum is over all pairs in the set of reference structures $\{\mathbf{X}\}$, with regression coefficients α_{IJ} and the kernel function $k(R, R_{IJ})$.

In line with the GAP approach, we use a sparse formulation of GPR by defining a set of sparse training points and constructing the covariance matrix by projecting the full dataset onto those. We use the Gaussian (squared exponential, SE) kernel,

$$k^{\text{SE}}(R, R') = \exp\left(\frac{-|R - R'|^2}{\theta^2}\right), \quad (5)$$

where θ is a length-scale parameter.

To ensure that $V_{\text{rep}}(R)$ smoothly vanishes at R_{cut} , we multiply the SE kernel with a damping function so that

$$k^{\text{damp}}(R, R') = e^{-\beta R} f_{\text{cut}}(R) k^{\text{SE}}(R, R'), \quad (6)$$

where the parameter β enforces a smooth decay over a transition length d . Among the GPrep hyperparameters, only the cutoff radius has a direct physical meaning. A common practice within the DFTB community is to choose between the first and second nearest neighbor distances of the physical system one is interested in (see, e.g., the discussion in Refs. 24 and 29). It is immediately obvious that such a choice is system-dependent. However, due to the two-body approximation, certain parametrizations require larger cutoff radii to attain the desired accuracy (see, e.g., Ref. 28) or a many-body correction to the repulsive potential (see, e.g., Refs. 29–31). Here, for the sake of simplicity, we limit ourselves to choosing the shortest possible cutoff radii based on the bond length distributions of our training data, as shown in Appendix C. The employed hyperparameters are reported in Table III. Since forces can be computed analytically as derivatives of $k^{\text{damp}}(R, R')$, we directly train on forces. The training targets are defined as the force components of the total force acting on every atom belonging to the set of training structures, associated with a list of all the bonds between said atom and atoms within a chosen cutoff (in principle independent from the repulsive cutoff, which is a hyperparameter of the GPR process) and the corresponding bond types and bond lengths.

As mentioned earlier, GPrep can be considered a form of Δ -learning, where the baseline model is repulsion-less DFTB (i.e., the electronic part of the interaction). As such, one may argue that the repulsive potential is a simpler object to fit than the total energy as in the GAP approach, as a large part of the underlying physics is already captured in the much less empirical electronic parametrization—to recall, for the latter, the only truly empirical parameters are those of the confinement potential, which is entirely “artificial.”

For Ru and RuO₂, we perform distinct parametrizations for both the electronic and the repulsive parts. We will thus optimize the Ru confinement potential for the metallic system and the Ru and O confinement potentials for the oxide system separately. Consistently, we train two separate sets of repulsive potentials: one for pure Ru, consisting of only the Ru–Ru pairwise repulsive potential, and one for RuO₂, comprising Ru–Ru, Ru–O (=O–Ru), and O–O repulsive potentials. Due to the difference in electronic parts, the Ru–Ru potentials of the two sets will significantly differ. We use the same training sets previously converged for the GAP training; however, we would not expect significantly different trends and conclusions had we performed the other way around—converging a training set for GPrep and then using it to train a GAP. In principle, and as a further refinement, one may join the two training sets to generate both GAP and GPrep-DFTB models to describe both the metallic and the oxide systems on the same footing—perhaps at the expense of some accuracy. The resulting repulsive potentials are plotted in Appendix C.

III. COMPARISON CRITERIA

In an attempt to cover a comprehensive enough picture of performance, we choose the following evaluation criteria.

A. Accuracy

As a property-based criterion, we compare lattice parameters, bond lengths, mechanical properties (Birch–Murnaghan equation of state), and relative energy spacing for bulk crystals represented in the training set, i.e., hcp and fcc Ru and rutile RuO₂. For RuO₂, we also evaluated the surface phase diagrams for 100 and 110 surfaces within the *ab initio* thermodynamic framework.³² This is a particularly challenging test for force-trained GPrep, as a good fit for forces does not guarantee an equally good fit for relative energetics across different compositions. This is mainly due to the repulsive cutoff, a hyperparameter of the ML procedure. The truncation of the repulsive potential at a finite distance, where the effective interaction is not necessarily vanishing, causes different energy offsets for each atomic pair. Even if these are small in the pairwise interaction, the resulting errors can quickly accumulate (and couple) in extended calculations with many atoms.

TABLE III. Hyperparameters used for the GPrep models in this work.

Description	Symbol	RuO ₂			Ru
		Ru–Ru	Ru–O	O–O	Ru–Ru
Number of sparse points	M	21	17	17	21
Start (Å)	r_{start}	2.0	1.0	1.0	2.0
Cutoff (Å)	r_{cut}	4.0	2.6	3.6	4.0
Kernel width	θ	1.0	1.0	1.0	0.6
Damping exponent	β	1.0	1.0	2.0	1.0
Regularization factor	σ_n		0.05		0.05

B. Extrapolation power

As a standard quantitative criterion, we evaluate the RMSE of forces vs some validation set(s). As an additional, property-based criterion, we compare lattice parameters, bond lengths, mechanical properties (Birch–Murnaghan equation of state), and (where applicable) relative energy ordering and spacing for bulk crystals *not* represented in the training set, i.e., bcc and sc Ru and anatase RuO₂.

C. Data efficiency

We compare the learning curves for the two models. We evaluate two different forms of learning curves. One is the standard RMSE vs number of training points. Additionally, we evaluate the RMSE vs the training set generations employed in the iterative training workflow for GAP, as described in detail in Ref. 22. Here, the validation set is different for each generation, comprising all the farthest point samples (e.g., for generation 4, it includes all farthest point samples from generation 4, 5, 6, etc.) and randomly selected structures from DFT geometry optimizations. Since the total number of farthest point samples is getting smaller as generations go by, the total number of validation structures decreases. As GAP is trained on energies as well as forces, the RMSE of energies was also used as a loss measure in Ref. 22 as well as the RMSE of forces. Since for GPrep, we did not follow an iterative procedure (which is, however, in principle possible and good practice—see, e.g., Ref. 28) but exclusively used the final GAP training sets, we rather constructed a learning curve by evaluating the RMSE of force residues (as GPrep is trained on those) for random subsets of the final training set with an increasing number of samples. Albeit substantially different, both forms of learning curves give an idea of how many DFT calculations are necessary to obtain a good fit. For completeness, we construct an additional learning curve for GPrep analogous to that of GAP, using the same training and validation sets.

IV. COMPUTATIONAL DETAILS

Structural relaxations, pre- and post-processing, visualization, and analysis were performed with the Atomic Simulation Environment (ASE).³³

A. Density functional theory calculations

All DFT calculations are performed using a plane-wave basis set within SG15 optimized norm-conserving Vanderbilt pseudopotentials³⁴ as implemented in the QuantumEspresso software package.³⁵ The semi-local generalized gradient approximation (GGA) in the revised Perdew–Burke–Ernzerhof (rPBE)³⁶ form is used as the electronic exchange and correlation (*xc*) functional. GGA functionals are routinely used in computational studies of metals and metallic oxides and have been generally shown to be adequate for the description of RuO₂.^{22,32,37–40} The kinetic cutoff energy for the expansion of the wave function is set to 80 Ry. Brillouin-zone integrations are carried out on a grid of *k*-points with reciprocal distances of 0.02 Å⁻¹, producing, e.g., (11 × 11 × 16) *k*-point grids for bulk rutile RuO₂. Optimized lattice parameters for the bulk RuO₂ are obtained by minimizing the stress tensor and all internal degrees of freedom until the external pressure falls below 0.5 kbar. Geometry optimization for

all slab calculations employs Broyden–Fletcher–Goldfarb–Shanno (BFGS) minimization^{41–43} until residual changes in total energy and all force components fall below 1.4×10^{-2} meV and 0.3 meV/Å, respectively.

B. Density functional tight binding calculations

All the DFTB calculations were performed using the implementation in dftb+²⁵ version 21.1 with the same *k*-point density as the reference DFT calculations. A Fermi filling with an electronic temperature of 0.001 hartree (corresponding to ~316 K) was employed. The SCC convergence criterion was set to 10⁻⁵. Structural relaxations were performed using the BFGS algorithm as implemented in ASE until the maximum force acting on each atom is less than 10 meV/Å. The cell optimization was turned on freely for all the cell degrees of freedom as well as the atomic coordinates using the UnitCellFilter module in ASE. All the DFTB calculations are not shell-resolved, consistent with the usage of a single Hubbard *U* value per atomic species. Slater–Koster tables for the electronic parametrization were generated with a developer version of hotbit²⁴ (available from the authors upon request) interfaced with the particle swarm optimization as implemented in the Python package pyswarm.⁴⁴ The all-electron DFT calculations of confined atomic wavefunctions for the Slater–Koster table were performed at the LDA level. The free-atom eigenvalues (on-site energies) were calculated with the all-electron electronic structure code FHI-aims with the rPBE functional, ZORA relativistic corrections, colinear spin polarization, and a custom basis set based on the so-called “tight” defaults, employing all the available tiers of basis functions and extending the damping cutoff to 8 Å.

C. Ru and RuO₂ extended validation set

Extended validation sets of bulk Ru and bulk RuO₂ structures have been generated using the evolutionary algorithm USPEX.^{45–47} We performed USPEX calculations for 10 generations, where the first generation of structures (200 items) was created randomly using the Python library PyXtal⁴⁸ and the topology-based crystal structure generator.⁴⁹ After each new generation consisted of 20% randomly generated structures, the remaining 80% were created using heredity, soft-mutation, and transmutation operators. For a better representation of short-range interactions, we included structures with interatomic distances as short as 1.2 Å. Those were generated by lowering the threshold IonDistance parameter. The number of atoms in the generated structures varied from 8 to 32. Partial structural relaxations were performed on each structure using DFT with the same functional and computational parameters used in the generation of the training set, and snapshots extracted from the relaxation trajectories were added to the validation set. At this point, we have assembled 21 031 structures of Ru configurations and 24 385 of RuO₂ configurations. Out of these, we selected the 300 most diverse structures per set using FPS,^{22,23} which were then used for the final force validation. The full dataset remains available for further training or validation (see Data Availability statement).

V. RESULTS AND DISCUSSION

A. Ru

1. Ru crystals

Tables IV–VII show the resulting cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for hcp Ru, fcc Ru (these two were included in the training set), bcc Ru, and sc Ru (not in the training set), respectively. All the values are calculated with respect to orthorhombic supercells.

For hcp and fcc crystals, which were both represented in the training set, the performance of GAP and GPrep-DFTB is comparable. All the structural properties are reproduced correctly. The supercell angles, not reported in the table, are all consistently 90° .

TABLE IV. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for hcp Ru.

Method	a (Å)	b (Å)	c (Å)	c/a	R_{NN} (Å)	dE (eV/f.u.)	B (GPa)
DFT	2.723	4.716	4.282	1.573	2.656	0	290.617
GAP	2.737	4.741	4.359	1.728	2.692	0	288.133
GPrep-DFTB	2.754	4.771	4.305	1.563	2.676	0	374.021

TABLE V. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for fcc Ru.

Method	a (Å)	b (Å)	c (Å)	R_{NN} (Å)	dE (eV/f.u.)	B (GPa)
DFT	2.692	2.692	3.807	2.692	0.114	285.386
GAP	2.725	2.725	3.853	2.725	0.105	284.273
GPrep-DFTB	2.720	2.720	3.848	2.721	0.124	369.200

TABLE VI. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for bcc Ru.

Method	a (Å)	b (Å)	c (Å)	R_{NN} (Å)	dE (eV/f.u.)	B (GPa)
DFT	2.642	2.642	2.642	2.642	0.609	259.199
GAP	2.989	2.989	2.989	2.588	0.597	132.048
GPrep-DFTB	3.076	3.076	3.076	2.664	0.859	387.447

TABLE VII. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for sc Ru.

Method	a (Å)	b (Å)	c (Å)	R_{NN} (Å)	dE (eV/f.u.)	B (GPa)
DFT	2.511	2.511	2.511	2.511	1.102	204.922
GAP	2.611	2.611	2.611	2.611	0.892	346.780
GPrep-DFTB	2.550	2.550	2.550	2.550	1.281	287.200

Both GAP and GPrep-DFTB show some tendency to expand the hcp cell parameters, with the expansion of a and b slightly more pronounced for GPrep-DFTB. However, GPrep-DFTB expands proportionally in the c direction, resulting in a more accurate c/a ratio. For fcc, both models show a slight, consistent expansion in all three directions. The energy spacing is reproduced well, being slightly underestimated for GAP and slightly overestimated for GPrep-DFTB. Here, GAP reproduces perfectly the mechanical properties, with the bulk modulus being almost exactly the same as in DFT. GPrep-DFTB, conversely, shows a consistent overestimation, indicating that the resulting interaction is somewhat stiffer than the DFT reference.

For bcc and sc crystals, which were both not included in the training set, the performance of GAP and GPrep-DFTB differs significantly. GAP reproduces the cell parameters for bcc better than GPrep-DFTB, as well as the energy difference, but the compression–expansion curve presents a local minimum and a local maximum, indicating the presence of a spurious additional local minimum at a larger volume. Therefore, the number in Table VI is not to be interpreted as a true bulk modulus, as it is merely the result of an ill-defined Birch–Murnaghan fit. GPrep-DFTB produces (consistently with all the other crystals) slightly expanded cell parameters and an overestimated bulk modulus, as well as a slightly overestimated energy difference, but the Birch–Murnaghan equation of state is well defined.

For the sc crystal, GPrep-DFTB consistently outperforms GAP. The bulk modulus is still slightly overestimated, but less than the other crystals and closer to DFT than GAP. The structural and energetic parameters are much closer to the reference than GAP, as is the bulk modulus.

These results are visually summarized in Fig. 2. It has to be noted that the value of the bulk modulus is extremely sensitive to rather small variations in the computed energy of compressed and extended geometry. From a simple visual inspection of Fig. 2, it is quite clear that even a rather large numerical discrepancy in the bulk modulus values, as observed for GPrep-DFTB with respect to the DFT reference, is not really a major failure. More importantly, the ordering of bulk moduli is correct, and the values are proportional. Therefore, despite a systematic overestimation, an evaluation of relative values appears to be robust.

Not surprisingly, the compressive branch of the EOS curve seems to be generally more affected than the expansive branch, indicating a certain tendency of GPrep to produce systematically more repulsive potentials at short distances than they should be. GAP clearly does not suffer from that, as it is trained on energies as well as forces. However, the quite clear drop in performance on the EOS curves for unseen types of structures shows how critical the diversity of the training set is.

Overall, if we consider the correct reproduction of properties that are in some form represented in the training set as a measure of accuracy, we may conclude that both models perform equally well. However, GPrep-DFTB seems to do a better job at capturing unseen features, such as the structural and energetic properties of crystals, that were in no way present in the training set. This hints at better transferability or extrapolation power, as it is reasonable to expect from a semi-empirical model where part of the interaction is fundamentally approximated *ab initio*. That is to say, the electronic part of the interaction is only dependent on directly

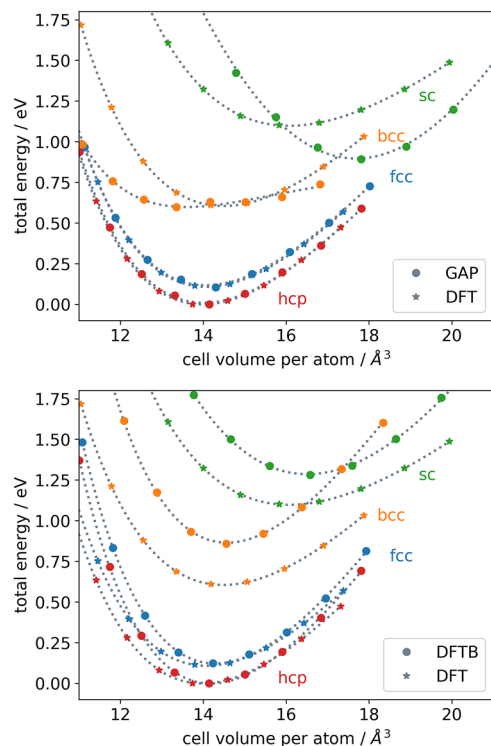


FIG. 2. Birch–Murnaghan fit for the volume–energy equation of state for hcp Ru (red dots), fcc Ru (blue dots), bcc Ru (orange dots), and sc Ru (green dots). Top and bottom show the plots for GAP and GPrep–DFTB, respectively, together with the DFT curves. The curves obtained with the GAP model are in excellent agreement with DFT for hcp Ru and fcc Ru, while for bcc Ru and sc Ru, the predictions are unreliable. The curves obtained with the GPrep–DFTB model are in good agreement with DFT for hcp Ru and fcc Ru, with some overestimation of the bulk modulus, and for bcc Ru and sc Ru, the predictions are equally robust.

computable properties (the on-site energies and Hubbard values) and the confinement potential, which is represented by two empirical parameters (or even one if a quadratic confinement is used) per atomic species fitted to simple properties as band structures of one

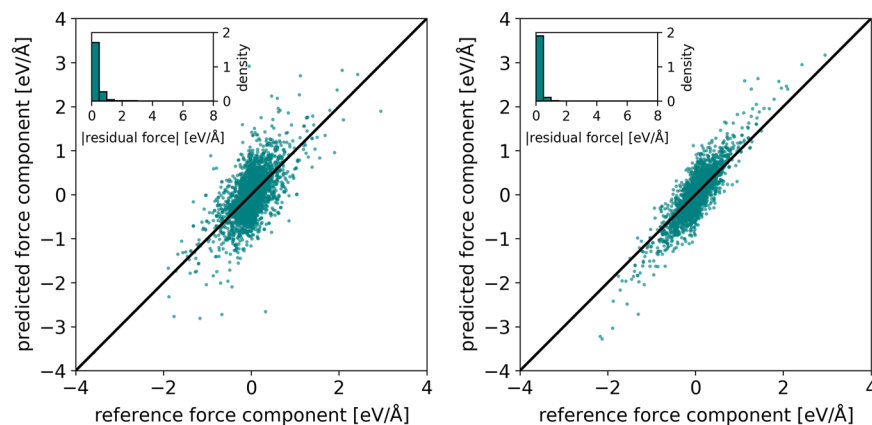


FIG. 3. Correlation plot between reference (DFT) forces and surrogate model forces (left: GAP; right: GPrep–DFTB) for validation set 1. Validation set 1 includes structures not included, but relatively similar to those included in the training set. The RMSE is 0.364 eV/Å for GAP and 0.231 eV/Å for GPrep–DFTB. Insets show the error density histograms. Both distributions are narrowly accumulated toward 0.

or a few selected structures. In this sense, the electronic part does not depend on any extended training set, while the repulsive part of the interaction, which does depend on an extended training set, is only a correction and also a simpler object (pairwise potential). The latter may be less accurate, being a less sophisticated object, but should ensure a more robust “one fits all” representation—perhaps not perfect for every chemical motif, but at least adequate for any. To be fair, the accuracy of GPrep repulsive potentials can be tuned to a great extent by tuning hyperparameters (as discussed, e.g., in Ref. 16 or shown practically in Ref. 50, where the hyperparameters of a pre-existing parametrization²⁸ were re-tuned to reproduce correct energetics) even without extending the training set, but surely at the expense of at least some transferability.

2. Force correlation plots

Figure 3 shows correlation plots of predicted total force components vs reference force components for validation set 1, corresponding to the last validation set in the iterative GAP training workflow. This validation set consists of bulk and surface structures not included in the training set but still based on hcp and fcc bulk motifs. Technically, this would be more of a test set. The accuracy of GPrep–DFTB is higher than that of GAP, as a further indication of superior extrapolation power. However, the distributions are overall comparable, as further evidenced by the error density histograms, which show a narrow distribution for both models. The RMSE is 0.231 eV/Å for GPrep–DFTB and 0.364 eV/Å for GAP.

Figure 4 shows correlation plots of predicted total force components vs reference force components for validation set 2, consisting of completely different bulk structures generated as described in Sec. IV. This validation set includes much higher forces than validation set 1. The RMSE is 1.842 eV/Å for GPrep–DFTB and 3.132 eV/Å for GAP. Here, the overall distribution is significantly narrower and straighter for GPrep–DFTB (as indicated also by the RMSE) than for GAP, albeit with a positive deviation in the slope (which is also present in both models for validation set 1), suggesting a systematic overestimation of forces. Interestingly, though, the GPrep–DFTB force correlation for validation set 2 exhibits a subset of forces with extremely large scatter and a tendency toward underestimation of forces for those, indicating that, despite a generally better extrapolation power than GAP, some structural motifs appear to be

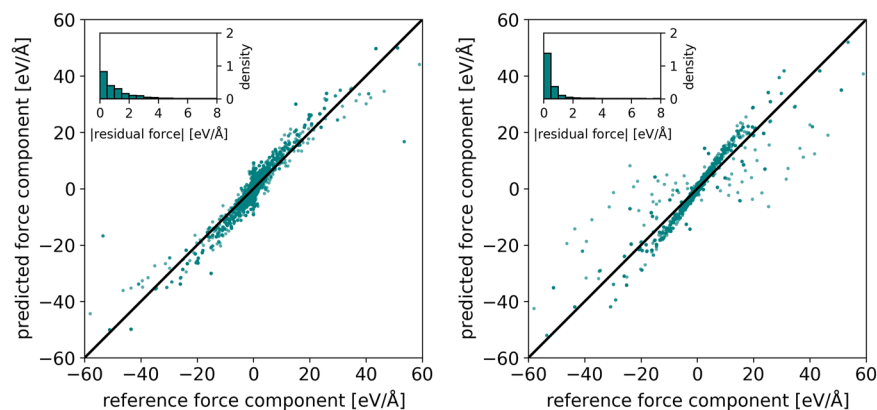


FIG. 4. Correlation plot between reference (DFT) forces and surrogate model forces (left: GAP; right: GPrep-DFTB) for validation set 2. Validation set 2 includes structures generated by USPEX structure prediction that were completely unseen in the training procedure. The RMSE is 3.132 eV/Å for GAP and 1.482 eV/Å for GPrep-DFTB. Insets show the error density histograms. Both distributions are narrowly accumulated toward 0, with the GPrep-DFTB distribution narrower than that of GAP.

problematic. However, the error density histograms show narrow distributions for both models, with the GPrep-DFTB distribution narrower than that of GAP despite the higher RMSE.

3. Learning curves

Figure 5, left, shows learning curves for GAP (top) and GPrep-DFTB (bottom), constructed using increasing numbers of data

points randomly drawn from the full training set. The validation error is calculated with respect to the full training set; hence, it converges to the training error. The RMSE values for GAP are calculated with respect to force residues (repulsion-less DFTB vs DFT) rather than total force components, as those are what GPrep learns directly. The training error is plotted in blue, and the validation error is plotted in orange. GPrep shows larger RMSE and oscillations for smaller training sets but converges more quickly than GAP.

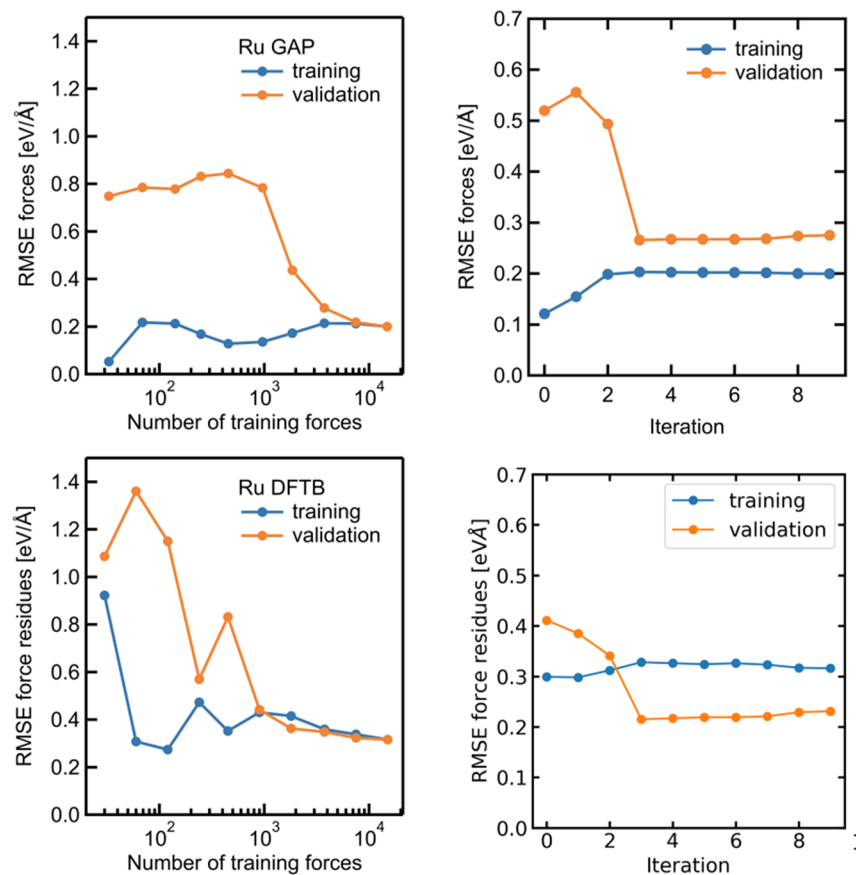


FIG. 5. Left: learning curves for GAP (top) and GPrep-DFTB (bottom), constructed using increasing numbers of data points randomly drawn from the full training set. The validation error is calculated with respect to the full training set; hence, it converges to the training error. GPrep shows larger RMSE and oscillations for smaller training sets but converges more quickly than GAP. Right: learning curves for GAP (top) and GPrep-DFTB (bottom), based on the iterative GAP training procedure. The GAP curve corresponds to the iterative procedure used in Ref. 22. Here, the validation sets used in each point are constructed as explained in Sec. III C and roughly correspond to validation set 1. The RMSE values for both energy and forces fall below 0.3 meV/Å for both models already at the third iteration (3 generations) and exhibit no considerable change afterward. However, for GAP, it takes substantially longer (9 generations) to attain convergence based on the active learning algorithm and the FPS.²² For GPrep, the training RMSE does not change significantly across generations, and the validation RMSE trend is slightly flatter.

Figure 5, right, shows learning curves for GAP (top) and GPrep-DFTB (bottom) with respect to the generations used in the iterative GAP training procedure. The GAP curve corresponds to the iterative procedure used in Ref. 22. Here, the validation sets used at each point are constructed as explained in Sec. III C and roughly correspond to validation set 1. The RMSE values for both energy and forces fall below 0.3 meV/\AA for both models already at the third iteration (3 generations) and exhibit no considerable change afterward. However, for GAP, it takes substantially longer (9 generations) to attain convergence based on the active learning algorithm and the FPS.²² For GPrep, the curve was constructed strictly following that for the GAP model (but only for force residues), obtained using the same training and validation sets as for GAP (i.e., without repeating the procedure from scratch). As can be seen, the loss is somewhat flatter with respect to the GAP training set generation. Similarly to what was observed for GAP, the validation RMSE converges at the third generation, with a slightly smaller variation. The training RMSE does not change significantly across generations. The training behavior as a function of the number of force residues (not shown here) does not change across generations, consistently converging before 5000.

GPrep appears to be less affected by the iterative refinement of the training set. In our own experience, indeed, the repulsive potentials change dramatically only when entirely different structural motifs that would not be captured by sampling and FPS evaluation are added to the training set, as observed for the parametrization of graphite intercalation compounds.^{28,50}

In terms of data greediness, this is an indication that GPrep training is slightly more data efficient than GAP training. This is not surprising since GPrep has to learn a much simpler object (a pairwise potential) than GAP, and this simpler object, in turn, only represents a correction to the total energy.

B. RuO

1. RuO crystals

Analogously to how we performed for Ru crystals, we compare cell parameters, selected bond lengths, and mechanical properties for ideal RuO₂ crystals. To recall, for the RuO₂ system, the training contained exclusively rutile-based structures. For an estimation of the extrapolation power, we also compare anatase, which was not included in the training set.

Table VIII reports cell parameters, Ru–O first nearest neighbor distance, and bulk modulus for rutile RuO₂. All the values are calculated with respect to orthorhombic supercells. The supercell angles, not reported in the table, are all consistently 90° . Here, both GAP and GPrep-DFTB are extremely accurate, showing excellent agreement with the DFT references, with only a slight overestimation of the bulk modulus for GPrep-DFTB in line with what was observed for metallic Ru as reported in Sec. V A 1.

For anatase RuO₂ (Table IX), the structural details predicted by the GAP model are closer to the DFT reference, while GPrep-DFTB slightly expands in the *a* and *c* directions of the unit cell and slightly compresses in the *c* direction. The bulk modulus is again overestimated with GPrep-DFTB and marginally smaller with GAP. Consistent with metallic Ru, the GPrep-DFTB overestimation of the bulk modulus is systematic with correct proportions. However, the

TABLE VIII. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for rutile RuO₂. R_{NN} refers to Ru–O distances.

Method	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	R_{NN} (Å)	<i>dE</i> (eV/f.u.)	<i>B</i> (GPa)
DFT	4.543	4.543	3.140	1.964	0	239.910
GAP	4.556	4.556	3.143	1.970	0	235.812
GPrep-DFTB	4.544	4.544	3.146	1.974	0	281.962

TABLE IX. Cell parameters, selected bond lengths, and Birch–Murnaghan bulk modulus for anatase RuO₂. R_{NN} refers to Ru–O distances.

Method	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	R_{NN} (Å)	<i>dE</i> (eV/f.u.)	<i>B</i> (GPa)
DFT	3.866	3.866	9.948	1.993	0.457	177.224
GAP	3.846	3.846	9.939	1.980	0.648	180.243
GPrep-DFTB	3.938	3.938	9.808	1.968	0.413	248.934

energy difference is much closer to the DFT reference for GPrep-DFTB than for GAP.

Overall, both the accuracy (rutile structure) and extrapolation power (anatase structure) appear comparable, with a slightly better performance of GPrep-DFTB with respect to relative energetics and of GAP with respect to mechanical properties and structural details of anatase. These results are visually summarized in Fig. 6.

2. Force correlation plots

Figure 7 shows the force correlation plots for GAP (left) and GPrep-DFTB (right) for validation set 1, corresponding to the last validation set used in the iterative GAP training procedure. This validation set consists of surface structures for all low-index surfaces of rutile RuO₂, as described in Sec. II A. The RMSE is 0.245 eV/\AA for GAP and 0.603 eV/\AA for GPrep-DFTB. Despite the not-so-large RMSE values, the correlation appears surprisingly bad for both models, especially considering the good agreement with the DFT reference in the structural details at equilibrium crystal geometries as well as mechanical properties.

Both models show a tendency to predict almost arbitrarily large values for very small forces, mostly acting on ruthenium atoms, as evidenced by the vertical feature in both correlation plots. Such a specific feature cannot stem from a simple overestimation, as it occurs for both Ru models but clearly has some deeper flaws. This effect is much more pronounced for GPrep-DFTB and appears to be symmetric, as confirmed upon closer inspection of the predicted force vectors (cf. the dataset given in section Data Availability Statement), which also shows that it only tends to happen for some, mostly undercoordinated surface atoms. The affected atoms are largely the same across the validation set, i.e., undercoordinated Ru atoms at the surface. While for GPrep-DFTB, the inability to correctly resolve forces acting on undercoordinated Ru atoms could be ascribable to imperfections in the electronic part of the parametrization, which may be unable to capture subtleties in different oxidation states for Ru. It is unclear why it tends to happen (albeit to a much lesser extent) for GAP as well. One possible reason could be the narrow distribution of forces in the training data. Retraining GPrep by including validation set 1 in the training set does not significantly

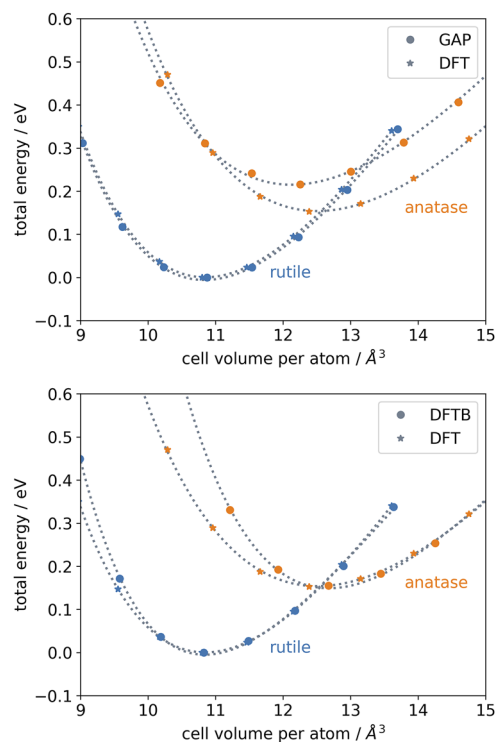


FIG. 6. Birch–Murnaghan fit for the volume–energy equation of state for rutile RuO_2 (red dots) and anatase RuO_2 (blue dots). The top and bottom show the plots for GAP and GPrep-DFTB, respectively, together with the DFT curves. The curves obtained with the GAP model are in excellent agreement with DFT, except for an overestimation of the relative energy. The curves obtained with the GPrep-DFTB model are in excellent agreement with DFT, except for an overestimation of the bulk modulus.

improve the force correlation, as shown in Appendix D, which further points toward imperfections in the electronic part. It is less clear why this happens for GAP, albeit to a smaller extent. To verify whether the vertical feature in the correlation of forces acting on Ru atoms is a result of overfitting, we performed scans of both

the energy and force regularization parameters σ_ε and forces σ_f . Results are reported in Appendix E. A more regular model performs worse than our choice in terms of both RMSE and force correlation; therefore, we can exclude that overfitting is the issue. A less regular model mitigates the effect and exhibits better RMSE, but not dramatically so.

Most likely, the structures affected by this would converge to roughly correct equilibrium geometries thanks to error cancellation. However, such a drawback is not to be overlooked, as it may affect dynamics in, e.g., MD simulations or MD-based structural searches, such as minima hopping or the dimer methods for a transition state search. To verify if both models are capable of driving physical dynamics, we performed two test MD runs in the NVT ensemble for 1 ps at 300 K, starting from an exemplary surface structure. Both simulations run smoothly, and neither ends up in unphysical configurational spaces, so one may conclude that the forces are at least reasonable for both models. Results are reported in Appendix F.

Figure 8 shows the force correlation plots for GAP (left) and GPrep-DFTB (right) for validation set 2, generated analogously to how they were performed for metallic Ru as described in Sec. IV. The RMSE is 3.236 eV/Å for GAP and 702.600 eV/Å for GPrep-DFTB. Despite the astronomically high value of the RMSE for GPrep, the correlation here is clearer for both models, with GAP performing better than GPrep-DFTB. The error density histograms are reasonably narrow for both models, but significantly less than those for metallic Ru.

Again, one may conclude that GAP and GPrep-DFTB show comparable accuracy, in line with what was observed for pure Ru as reported in Sec. V A. Not too surprisingly, GPrep-DFTB struggles a little more with the multi-component system than it does with the pure metallic system. The training procedure itself is not more expensive, data-greedy, or technically complicated for multi-component systems than it is for monoatomic systems. However, there is an additional complication given by the fact that a different set of hyperparameters (with the exception of the data noise factor σ_n , which is a global regularization hyperparameter) is needed for each atomic species pair. As discussed in Ref. 16, there are multiple choices of hyperparameters giving similar RMSE, which can, however, differ significantly in the fine details of the repulsive potentials.

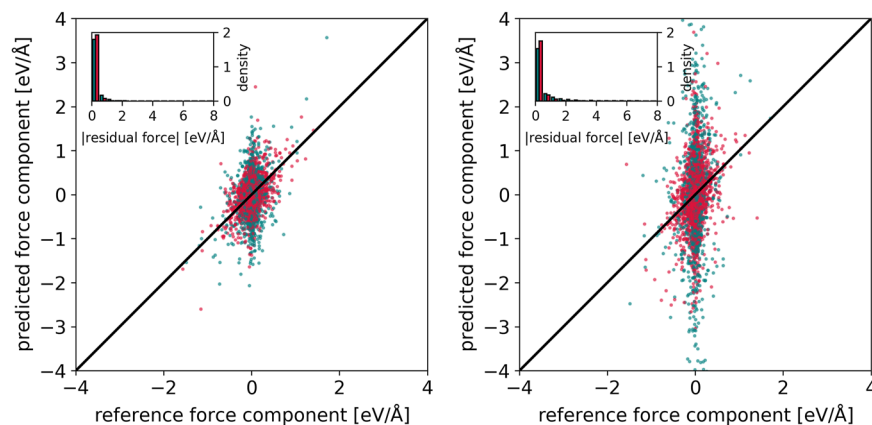


FIG. 7. Correlation plot between reference (DFT) forces and surrogate model forces (left: GAP; right: GPrep-DFTB) for validation set 1. Magenta dots represent forces acting on oxygen atoms, while teal dots represent forces acting on ruthenium atoms. The validation set contains structures not included in the training set that are relatively similar to those in the training set. The RMSE is 0.254 eV/Å for GAP and 0.603 eV/Å for GPrep-DFTB. Insets show the error density histograms. The distributions are similar to each other and both are much broader than observed for metallic Ru.

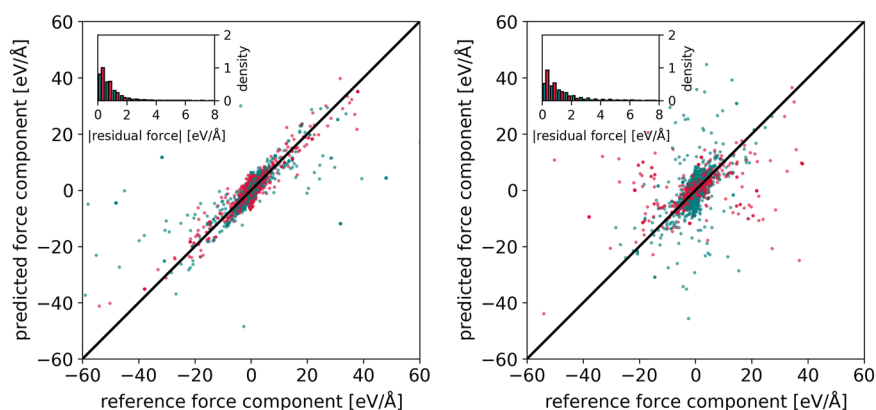


FIG. 8. Correlation plot between reference (DFT) forces and surrogate model forces (left: GAP; right: GPrep-DFTB) for validation set 2. Magenta dots represent forces acting on oxygen atoms, while teal dots represent forces acting on ruthenium atoms. Validation set 2 includes structures generated by USPEX structure prediction that were completely unseen in the training procedure. The RMSE is 3.236 eV/Å for GAP and 702.600 eV/Å for GPrep-DFTB. While the RMSE for GPrep-DFTB is astronomically high, the correlation is much clearer here than for validation set 1. Similarly to what was observed for Ru, the large RMSE is caused by a relatively small number of outliers with a large scatter. However, the main correlation trend is not as sharp as for Ru. Insets show the error density histograms. The distribution for GAP here is narrower than that of GPrep-DFTB.

Throughout this work, the hyperparameters for GPrep were not systematically optimized, choosing safe standard values for Ru and only minor tweaking for RuO₂.

For RuO₂, we limit ourselves to evaluating the force correlation only for one validation set. We do not expect trends to be substantially different for an extended validation set.

3. *Ab initio* thermodynamics

As a final measure of accuracy, we evaluate the surface phase diagram for the 110 and 100 low-index surfaces of rutile RuO₂. The training set is mostly representative of 100 surface terminations; however, 110 are also present. We calculate surface free energies within the *ab initio* thermodynamic framework^{32,51} as

$$\gamma(T, p) = \frac{1}{2A} \left[G^{\text{slab}}(T, p, N_{\text{Ru}}, N_{\text{O}}) - N_{\text{Ru}} \cdot g_{\text{RuO}_2}^{\text{bulk}}(T, p) + (2N_{\text{Ru}} - N_{\text{O}})\mu_{\text{O}} \right],$$

where G^{slab} is the Gibbs free energy of the double surface slab, g^{bulk} is the bulk Gibbs free energy per formula unit, $\mu_{\text{O}} = g_{\text{O}}$ is the oxygen chemical potential, i.e., the Gibbs free energy per molecule of the oxygen gas reservoir, and N_{O} is the number of Ru and O atoms, respectively. Different surface terminations have different oxygen contents. G^{slab} is directly approximated as $E_{\text{tot, DFT}}^{\text{slab}}$.

We thus calculate $E_{\text{tot, DFT}}^{\text{slab}}$ for five different surface terminations: 110 O-bridge, 110 O-cus, 110 O-poor (or Ru-terminated), 100 O-bridge, and 100 O-cus. Both O-bridge terminations are stoichiometric with respect to the bulk. O-cus terminations have two additional O atoms per formula unit, and O-poor terminations have two fewer O atoms per formula unit. Top views of the surfaces are shown in Fig. 9; more details about the individual structures can be found, e.g., in Refs. 32 and 51.

Figure 9 shows the surface phase diagram for 100 and 110 terminations of rutile RuO₂ as obtained by DFT (left), GAP (center),

and GPrep-DFTB (right). Surface free energies as a function of the oxygen chemical potential μ_{O} are plotted in purple for 110 terminations and magenta for 100 terminations. Shaded lines show the free energy of each termination as a function of the oxygen chemical potential, while solid lines mark the most stable terminations at each point. The GAP surface diagram shows overall good agreement, only with a small overestimation of the energy spacing between the 100 O-bridge and 110 O-bridge terminations. The transition points between O-bridge and O-cus terminations are captured qualitatively correctly, and the O-poor termination is correctly predicted to never be stable in the range of allowed chemical potentials (black vertical lines in Fig. 9), although its surface free energy is predicted to be lower than the reference.

For GPrep-DFTB, the quantitative agreement with DFT is even better than GAP for stoichiometric and oxygen-poor terminations (100 O-bridge, 110 O-bridge, and 110 O-poor). Both O-cus terminations, however, are overstabilized, only moderately for the 110 termination but greatly for the 100 termination. This is somewhat puzzling, given that 100 terminations were more represented in the training set than 110 terminations. Correctly capturing relative energetics across different compositions is a particularly challenging test, as the truncation of repulsive potentials at a chosen cutoff introduces different energy offsets for each species pair. The latter issue is not specific to Gprep but rather to DFTB as well as any truncated interatomic potential. For GAP, the two-body contribution of which is also truncated, the issue is mitigated by the many-body contribution. Similarly, for DFTB, many-body formulations of the repulsive potential are equally expected to provide more accurate energetics.²⁹ For DFTB, however, it is possible that this behavior is not solely ascribable to the repulsive potential but also to imperfections in the electronic part of the parametrization. All terminations have different oxidation states for Ru atoms, with the O-bridge ones stoichiometric with respect to the bulk (i.e., Ru is always formally in the IV oxidation state). It is possible that the electronic parametrization is

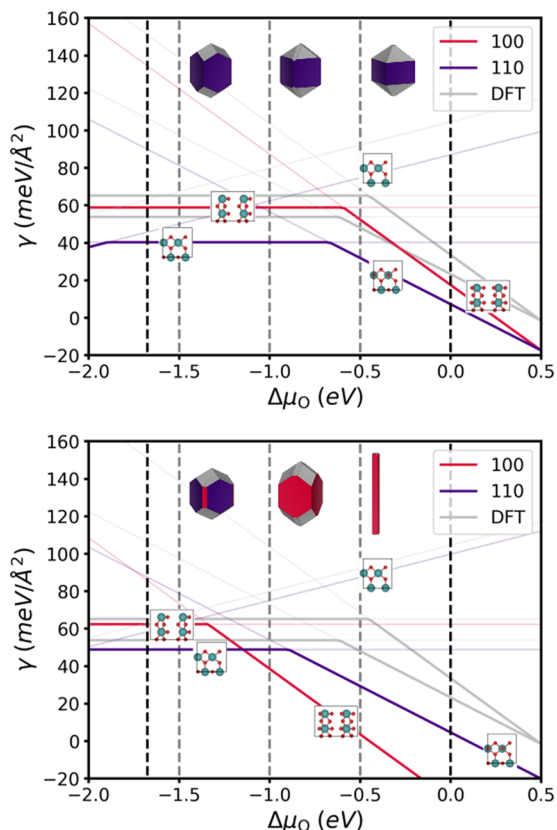


FIG. 9. Surface phase diagram for 100 and 110 terminations of RuO_2 as obtained by DFT (top) and GPrep-DFTB (bottom). Gray lines show the corresponding DFT energies. Black vertical lines mark the allowed range of oxygen chemical potentials, as discussed in Ref. 32. Surface free energies (shaded lines) as a function of the oxygen chemical potential μ_{O} are plotted in purple for 110 terminations and magenta for 100 terminations. Solid lines mark the most stable terminations at each point. Top views of the surfaces are also shown here by the corresponding free energy lines; more details about the individual structures can be found, e.g., in Refs. 32 and 51. The Wulff reconstructions at $\Delta\mu_{\text{O}} = -1.5$, -1.0 , and -0.5 eV are also shown. To construct the latter, the (DFT) surface-free energies of the remaining facets are taken from Ref. 22. The GAP surface diagram shows overall better agreement with the reference DFT. The GPrep-DFTB surface diagram shows better quantitative agreement with the reference DFT for the stoichiometric (110 O-bridge and 100 O-bridge) and oxygen poor (110 O-poor) terminations but severely over-stabilizes the oxygen rich terminations (110 O-cus and 100 O-cus).

not able to fully capture subtle changes in the charge transfer for the (relatively complex) non-stoichiometric surface terminations. A strong indication that this is indeed the case here emerges from analyzing the Mulliken charges resulting from the self-consistent charge procedure in DFTB. These directly enter the DFTB energy expression as part of the approximated electronic interaction. As the Mulliken charges are basis-set dependent, they do not necessarily correspond to those obtained by DFT; however, they do correlate. As shown in Appendix G, for all the surface terminations, the oxygen atoms in bridge positions show consistently underestimated negative charges. While this does not significantly affect the stoichiometric and O-poor terminations, for the O-rich (cus) terminations,

the additional oxygen atoms at the surface feel a lower Coulombic repulsion than they should, hence the over-stabilization. The effect is much more evident for the 100-cus termination, which presents twice as many extra oxygen atoms as the 110-cus termination. This drawback is not trivial to fix and depends on the intricate interplay between on-site energies, Hubbard U values (not optimized in this work), and confinement coefficients. Solving it, or even just mitigating it, requires a complete reparametrization of the electronic part, possibly including the DFT-DFTB correlation of Mulliken charges into the PSO cost function. Therefore, we do not attempt to solve it here but rather leave it as a reminder of how subtle details can go wrong in the electronic parametrization, which is commonly regarded in the DFTB community as significantly easier than the parametrization of the repulsive potential.

VI. CONCLUSION

We have compared GAP (fully ML interatomic potential, “black box”) and GPrep-DFTB (semi-empirical density functional tight binding with machine-learned repulsive potential, “gray box”) models for metallic Ru and oxide RuO_2 , using exactly the same training sets, in terms of accuracy, extrapolation power, and data efficiency. To do so, we evaluated the structural and mechanical properties of known crystals (hcp and fcc Ru, rutile RuO_2) represented in the training sets as a measure of accuracy, as well as those of other ideal crystals (body centered cubic, bcc, simple cubic, sc Ru, and anatase RuO_2) not represented in the training sets as a measure of extrapolation power. Furthermore, we evaluated the force correlation vs validation sets, both containing structural motifs relatively similar to those present in the training sets and (for Ru) completely different structural motifs. For RuO_2 , we also evaluated the surface free energy diagram for 100 facets and 110 facets. Finally, we evaluated data efficiency by discussing the learning curves of the two models.

Not surprisingly, the accuracy with respect to structural motifs, forces, and energetics within the coverage of the training set is fully comparable. GPrep-DFTB shows generally better extrapolation power, albeit with some exceptions, such as the massive over-stabilization of the $\text{RuO}_2(100)$ O-cus terminated surface. This serves as a warning that no method is bulletproof, and any approximate model should be subject to continuous scrutiny. To be fair, the correct treatment of multicomponent systems, especially in terms of relative energetics across compositional ranges, is a challenging task for any approximate interatomic model, from classical force fields to DFTB. While GPrep-DFTB solves the problem of multidimensional fitting of the repulsive potential from a technical point of view, it does not necessarily solve the existence of multiple offsets in the resulting potentials caused by the truncation of the latter at a chosen cutoff (unless employing extremely long cutoff radii, which is problematic for a number of different reasons). However, it is unclear whether, in this context, the shortcomings can be addressed with a more careful optimization of hyperparameters or if the underlying electronic parametrization has to be revisited.

The intuitive notion that a “gray box” approach is more transferable than a “black box” approach is hereby quite clearly supported for single-species metallic Ru but not at all clear-cut for the binary oxide system. The “black box” approach employed here, anyway, does not perform badly at all for Ru, either, and outperforms the

“gray box” in some aspects for RuO₂. In terms of data efficiency, we observe a moderately better performance of GPrep, which seems to be less sensitive than GAP to small changes in the diversity of the training set, as shown by the flatter learning behavior with respect to the number of generations, as well as being slightly faster in learning vs the number of data points. Converging a GAP requires several iterations of one to two hundred calculations of considerable size. GPrep converges with only a few thousand force residue data points, corresponding to only a handful of expensive DFT calculations. The increased extrapolation power ensures that one should not worry too much about showing all the possible motifs. Clearly, iterative training approaches can be applied as well (as performed, albeit not in an automated manner but rather driven by chemical intuition, in Ref. 16), but the fact that they are not strictly necessary is good news. As a possible outlook, we propose that one may, in principle, bootstrap a GPrep-DFTB model with only a handful of DFT calculations, then generate large training sets with GPrep-DFTB and train a GAP on those. Once the GAP is converged, one may, in principle, retrain it on DFT, but only on the final training set, thus avoiding expensive DFT calculations throughout the entire iterative training workflow. Relatedly, since the two models do show overall similar performance, the GAP can also be used to generate a large number of extremely diverse candidate training structures, out of which a subset of the most dissimilar ones can be chosen using FPS and used to tune the DFTB repulsion for better transferability. All in all, we encourage continuous scrutiny across the two approaches, as the complete interchangeability of training sets enables parametrization scenarios where GPrep-DFTB can benefit from GAP and vice versa.

ACKNOWLEDGMENTS

All the authors jointly and gratefully acknowledge funding through the Kopernikus/P2X-2 program (Grant No. 03SFK2V0-2) of the German Federal Ministry of Education and Research (BMBF) and the computational and data resources provided by the Max Planck Computing and Data Facility (MPCDF). C.P. wishes to dedicate this paper to the memory of Prof. Michelangelo De Maria.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

C. Panosetti: Conceptualization (lead); Data curation (equal); Formal analysis (equal); Investigation (lead); Methodology (equal); Validation (lead); Writing – original draft (lead); Writing – review & editing (equal). **Y. Lee:** Data curation (equal); Formal analysis (equal); Investigation (supporting); Methodology (equal); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (equal). **A. Samtsevych:** Data curation (equal); Formal analysis (equal); Investigation (supporting); Methodology (equal); Validation (equal); Writing – original draft (supporting); Writing – review & editing (equal). **C. Scheurer:** Funding acquisition (lead); Investigation (supporting); Methodology (equal); Project

administration (lead); Supervision (lead); Writing – original draft (supporting); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in The Open Research Data Repository of the Max Planck Society Edmond, <https://edmond.mpg.de/>, at <https://doi.org/10.17617/3.CRSJQV>.

APPENDIX A: GAP LOCALITY TEST

The GAP estimates the total energy of the system by the sum of local contributions, which are divided by an applied cutoff radius r_{cut} , and neglects any long-range interactions originated from electrostatics or dispersion outside of the cutoff. To minimize such errors from long-range interactions, we have tested a force locality for symmetry inequivalent atoms i in Ru and RuO₂ surface models. Based on bulk-truncated surface supercells \mathbf{X} , we generated a set of perturbed configurations $\{\mathbf{X}'\}$ in which all atom positions outside of r_{cut} from atom i are randomly displaced. The induced force at center atom i is then measured as the force difference of the ground-state and the perturbed configurations $\Delta F_i = |F_i^{\mathbf{X}'} - F_i^{\mathbf{X}}|$ as a function of r_{cut} , as illustrated in Fig. 10 for Ru (top) and RuO₂ (bottom). In detail, the random perturbations are applied uniformly with a standard deviation of 0.05 Å, so the maximum displacement of atoms is not larger than 0.2 Å. As a result, ΔF_i is converged for all atoms i at r_{cut} of 5.5 and 5.0 Å for Ru and RuO₂, respectively.

APPENDIX B: SPURIOUS BANDS IN ELECTRONIC PARAMETRIZATION

Figure 11 shows band structures for hcp Ru at equilibrium geometry and compressed geometry with a compression factor of 0.9. The equilibrium band structure is in perfect agreement with DFT in the valence region. However, as mentioned in Sec. II B 1, spurious conduction bands are present between Γ , A and H special points for hcp Ru. A lack of agreement in the conduction bands is common for DFTB due to the minimal basis set approximations. The spurious bands appearing here are not problematic at equilibrium, as they are not occupied and thus do not affect total energies. Upon compression, though, these bands spill into the valence region, where they end up being occupied. As a result, the electronic part of the DFTB interaction becomes artificially overattractive at short distances. In all the attempted repulsion parameterizations with these electronic parameters, GPrep could not properly compensate for this effect and, as a result, the total DFTB interaction presented spurious minima at distances around 2 Å. This caused, e.g., the simple cubic crystal to systematically relax to a much more compressed geometry.

APPENDIX C: REPULSIVE POTENTIALS

Figure 12 shows bond distributions in the Ru training set (Ru–Ru bonds, top left) and in the RuO₂ training set (Ru–Ru bonds, top right; Ru–O bonds, bottom left; O–O bonds, bottom right). Peaks are centered around the nearest neighbor's distances. The

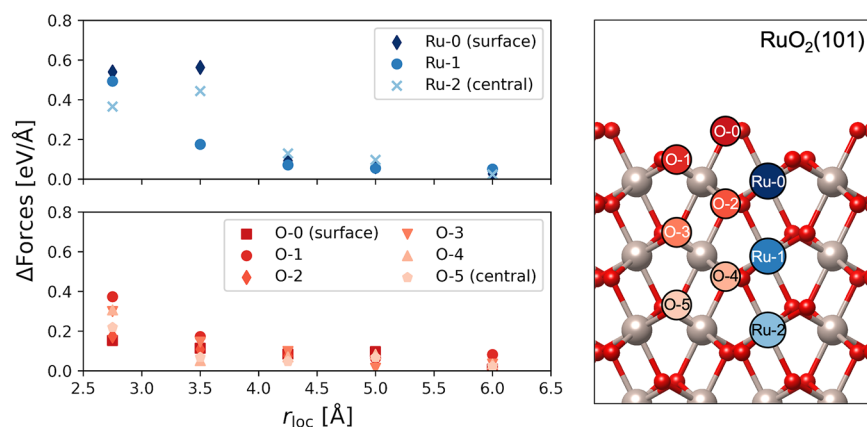
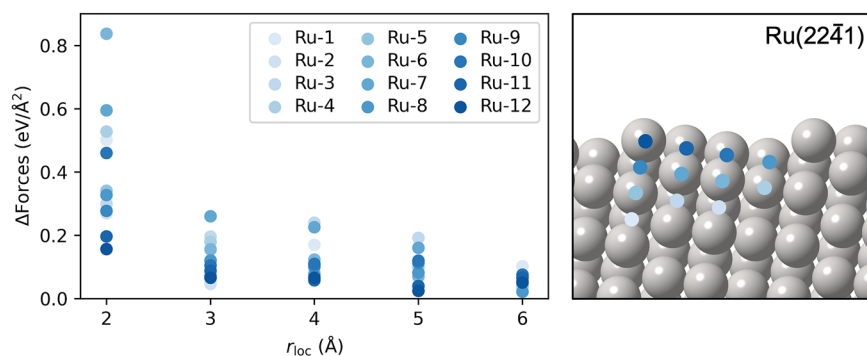


FIG. 10. Force locality in Ru(2241) (top) and RuO₂(101) (bottom).

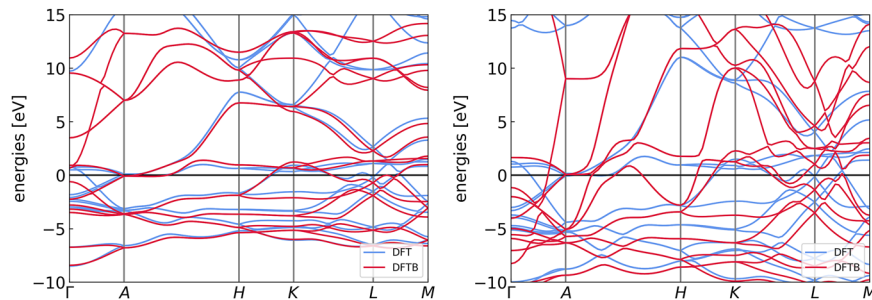


FIG. 11. Band structures for hcp Ru obtained by including only the equilibrium band structure in the PSO cost function. The DFTB band structure is shown in magenta, while the DFT reference is shown in blue. The optimal equilibrium geometry band structure is shown on the left side. The right side shows the band structure for a compressed geometry with a compression factor of 0.9. All the band structures are aligned with respect to the DFT Fermi level. The equilibrium band structure is in perfect agreement with DFT in the valence region. However, the spurious conduction bands visible between Γ , A and H special points for hcp Ru here clearly bleed into the valence region upon compression.

cutoff radii were chosen to only include the first peak for each distribution. This ensures that the shortest possible repulsive potentials are generated, thus avoiding spurious contributions from higher order nearest neighbors.

Figure 13 shows repulsive potentials for Ru–Ru (top left, Ru parametrization, top right, RuO₂ parametrization), Ru–O (bottom left), and O–O (bottom right) as obtained by GPrep. Green-shaded areas mark bond ranges within which the first nearest neighbor distances of all evaluated structures fall. These serve merely as a visual aid to the regions of the potentials that are most important for

equilibrium geometries. As a normal consequence of the different electronic parametrization, the repulsive potentials for the Ru–Ru pair for pure Ru and RuO₂ differ greatly.

APPENDIX D: FORCE CORRELATION WITH EXTENDED TRAINING SET

Figure 14 shows the force correlation plot between DFT forces and GPrep-DFTB forces for RuO₂ obtained after refitting the GPrep

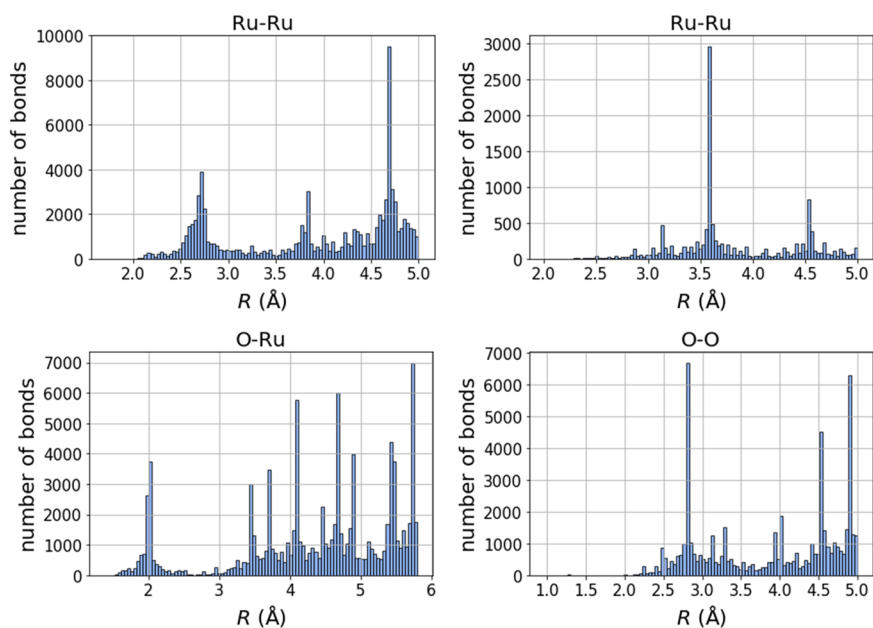


FIG. 12. Bond distributions in the Ru training set (Ru–Ru bonds, top left) and in the RuO₂ training set (Ru–Ru bonds, top right; Ru–O bonds, bottom left; O–O bonds, bottom right). Peaks are centered around the nearest neighbor's distances.

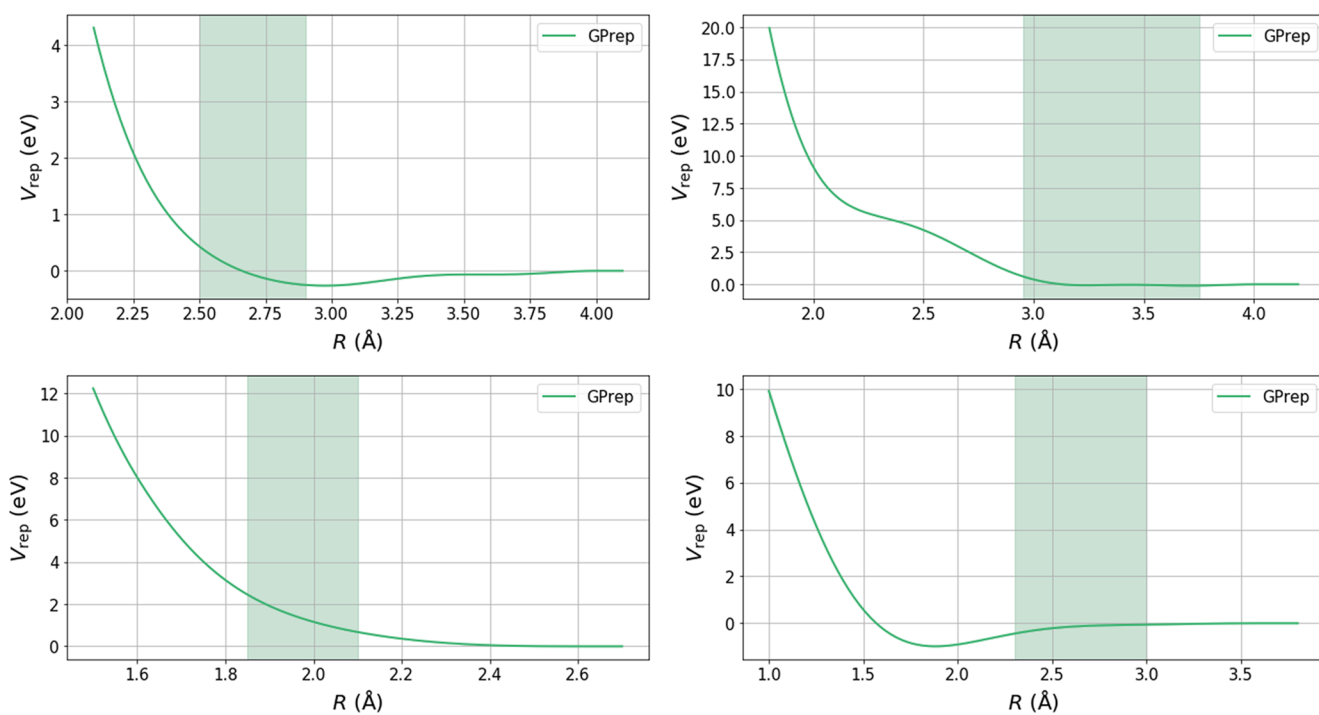


FIG. 13. Repulsive potentials for Ru–Ru (top left, Ru parametrization; top right, RuO₂ parametrization), Ru–O (bottom left), and O–O (bottom right). Green-shaded areas mark bond ranges within which the first nearest neighbor distances of all evaluated structures fall.

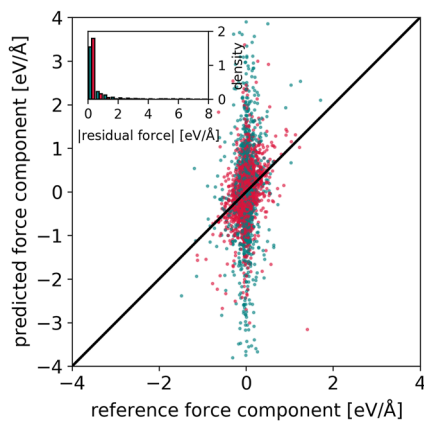


FIG. 14. Correlation plot between reference (DFT) forces and surrogate model forces (left: GAP; right: GPrep-DFTB) for validation set 1, obtained after refitting the GPrep repulsion and including validation set 1 in the training set. Magenta dots represent forces acting on oxygen atoms, while teal dots represent forces acting on ruthenium atoms. The RMSE is 0.560 eV/Å.

repulsion, including validation set 1 in the training set. Magenta dots represent forces acting on oxygen atoms, while teal dots represent forces acting on ruthenium atoms. The RMSE is 0.560 eV/Å. The inclusion of the validation set into the training set does not significantly improve the force correlation, neither quantitatively nor qualitatively, suggesting that the issue is not in the training set but rather in the electronic parametrization.

APPENDIX E: GAP REGULARIZATION SCAN

To clarify whether the vertical feature in the correlation of forces acting on Ru (see Fig. 7) is the result of overfitting, we systematically varied the regularization factor for both energy σ_ϵ and forces σ_f and evaluated the RMSE of forces for different values of σ_ϵ and σ_ϵ/σ_f ratios. The results, reported in Fig. 15, show that the best choices for regularization parameters could be [0.0005, 0.0025] or [0.001, 0.005], but our choice (orange triangle at 10^{-3}) of [0.001, 0.01] is also similarly good. Of note, changing the regularization for GPrep does not have a significant effect, confirming that, indeed, for DFTB, the imperfections of the electronic parametrization are predominant.

Figure 16 shows force correlation plots for a less regular model (RMSE = 0.190 eV/Å) and a more regular model (RMSE = 0.361 eV/Å). It is evident that larger regularization accentuates the feature, while lower regularization mitigates it. Therefore, this is not the result of overfitting but rather of underfitting. As such, the RuO₂ GAP (which is already published) may certainly be revised, but the force correlation performance shown in Fig. 8 reassures us that the model is not bad at all.

APPENDIX F: EXEMPLARY MD SIMULATIONS

Figure 17 shows the temperature, potential energy, and total energy for two test MD runs in the NVT ensemble for 1 ps at 300 K, starting from an exemplary RuO₂ surface structure. Both simulations run smoothly, equilibrate quickly, show no energy drifts after

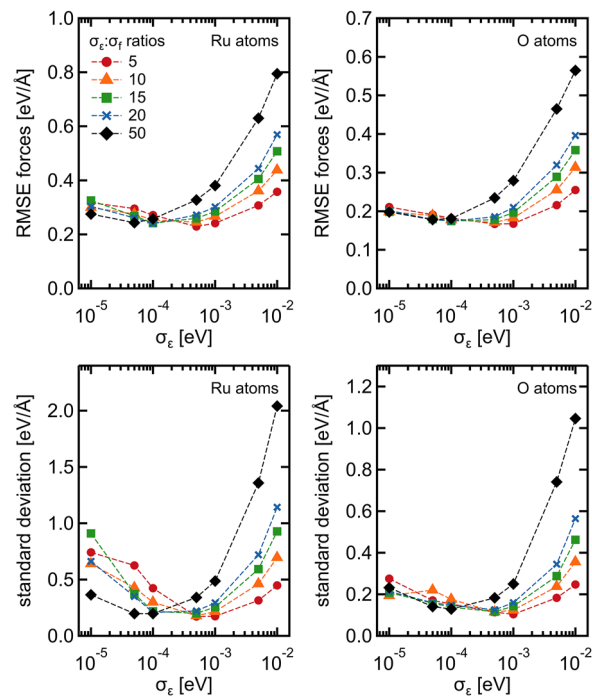


FIG. 15. RMSE of energy and forces for different choices of the GAP regularization parameters σ_ϵ and forces σ_f .

equilibration, and neither ends up in unphysical configurational spaces, so one may conclude that the forces are at least reasonable for both models. However, some qualitative aspects differ, indicating that the resulting potential energy surfaces are substantially different. The GAP run exhibits much larger atomic displacements at the beginning of the run (before equilibration) and quickly transitions to a different structure reconstruction even before reaching thermalization. The DFTB run equilibrates smoothly and stays in the same local basin. The new GAP local minimum turns out to be significantly more stable on DFTB energetics as well as on DFT energetics. The initial and final structures, locally optimized on GAP energetics, are shown in Fig. 17, overlaid on the GAP plot.

APPENDIX G: MULLIKEN CHARGES

Figure 18 shows Mulliken charges as obtained by DFT and DFTB with single-point calculations at the DFT-relaxed geometries for the 100 O-cus, 110 O-cus, 100 O-bridge, and 110 O-bridge surfaces. As the Mulliken charges are basis-set dependent, the values are not expected to be the same. However, we find an excellent correlation between DFT and DFTB charges with a proportionality factor of ~ 3.15 . The correlation plot shows two significant clusters of outliers (circled in Fig. 18), which correspond to bridge atoms in both 100 O-cus and 100 O-bridge. These atoms have a less negative partial charge in DFTB than they should. This causes a diminished electrostatic repulsion between these and the other oxygen atoms in the structure. For the cus termination, there are additional oxygen atoms on the cus positions, which should have a corresponding energy

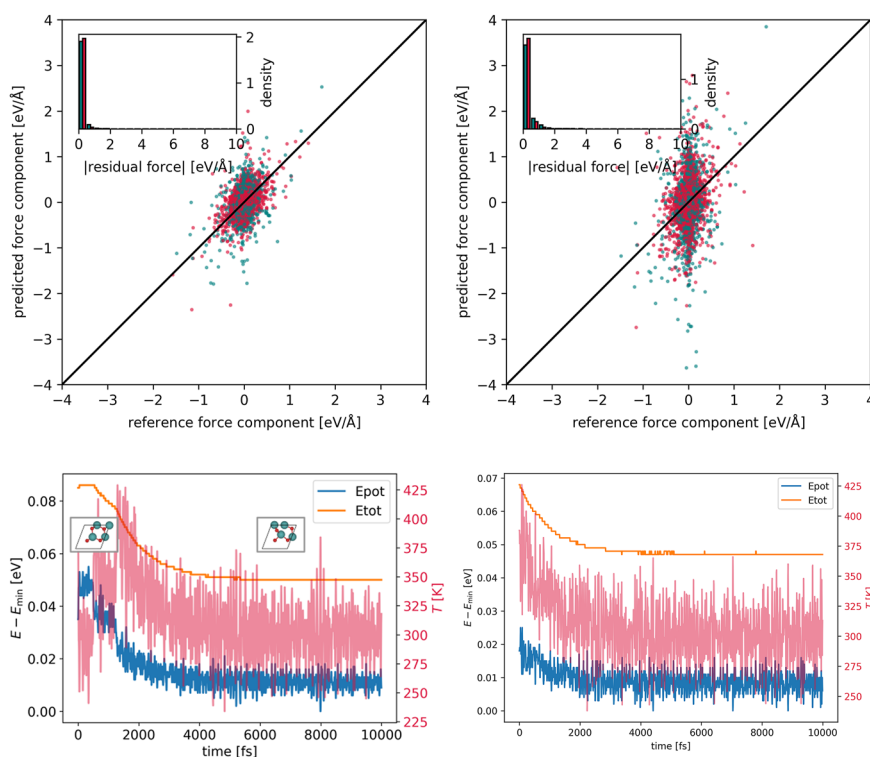


FIG. 16. Force correlation plots for a less regular GAP model (left) and a more regular GAP model (right), with $[\sigma_e, \sigma_e/\sigma_f] = [0.0005, 0.0025]$ and $[0.01, 0.1]$, respectively. The RMSE values are 0.190 and 0.361 eV/Å, respectively.

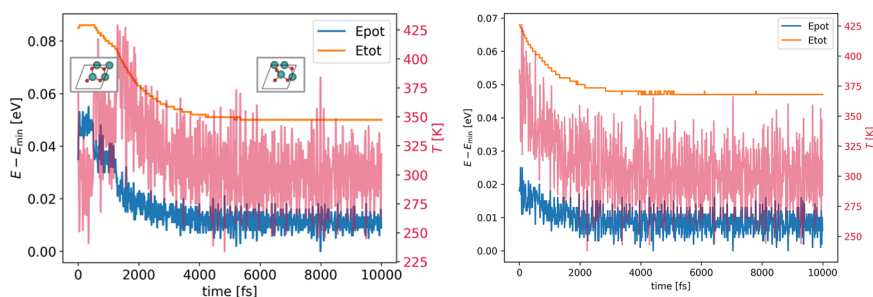


FIG. 17. Potential energy, total energy, and temperature for the test MD runs for 10 ps in the NVT ensemble for an exemplary RuO₂ surface structure for GAP (left) and GPrep-DFTB (right).

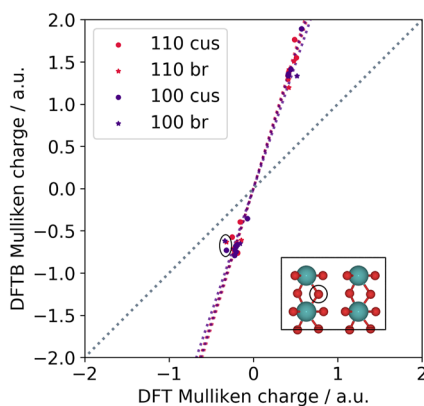


FIG. 18. Mulliken charges as obtained by DFT and DFTB with single-point calculations at the DFT-relaxed geometries for the 100 O-cus, 110 O-cus, 100 O-bridge, and 110 O-bridge surfaces.

penalty, the lack of which causes the overstabilization of this termination. As shown in the correlation plot, for the 110 surface, there are similar outliers (also corresponding to bridge oxygen atoms), but with a smaller effect due to both the fact that the deviation is smaller and the fact that there are fewer additional oxygen atoms in the 110 O-cus termination. Correspondingly, a smaller overstabilization is observed for the 110 O-cus termination.

REFERENCES

- J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- A. V. Shapeev, *Multiscale Model. Simul.* **14**, 1153 (2016).
- L. Zhang, J. Han, H. Wang, R. Car, and W. E, *Phys. Rev. Lett.* **120**, 143001 (2018).
- R. Drautz, *Phys. Rev. B* **99**, 014104 (2019).
- J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, *npj Comput. Mater.* **6**, 20 (2020).
- D. P. Kovács, C. van der Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner, and G. Csányi, *J. Chem. Theory Comput.* **17**, 7696 (2021).
- I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, and G. Csányi, “Mace: Higher order equivariant message passing neural networks for fast and accurate force fields,” *arXiv:2206.07697* (2022).
- O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, *Chem. Rev.* **121**, 10142 (2021).
- C. G. Staacke, S. Wengert, C. Kunkel, G. Csányi, K. Reuter, and J. T. Margraf, *Mach. Learn.: Sci. Technol.* **3**, 015032 (2022).
- K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, *Adv. Sci.* **6**, 1801367 (2019).
- P. O. Dral and M. Barbatti, *Nat. Rev. Chem.* **5**, 388 (2021).
- A. M. Lewis, A. Grisafi, M. Ceriotti, and M. Rossi, *J. Chem. Theory Comput.* **17**, 7203 (2021).
- K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, *Nat. Commun.* **10**, 5024 (2019).
- M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, *Phys. Rev. B* **58**, 7260 (1998).
- C. Panosetti, A. Engelmann, L. Nemeč, K. Reuter, and J. T. Margraf, *J. Chem. Theory Comput.* **16**, 2181 (2020).

- ¹⁷M. Stöhr, L. M. Sandonas, and A. Tkatchenko, *J. Phys. Chem. Lett.* **11**, 6835 (2020).
- ¹⁸S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, *Chem. Sci.* **12**, 4536 (2021).
- ¹⁹V. L. Deringer, M. A. Caro, and G. Csányi, *Adv. Mater.* **31**, 1902765 (2019).
- ²⁰V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, *Chem. Rev.* **121**, 10073 (2021).
- ²¹A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- ²²J. Timmermann, Y. Lee, C. G. Staacke, J. T. Margraf, C. Scheurer, and K. Reuter, *J. Chem. Phys.* **155**, 244107 (2021).
- ²³A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, *Sci. Adv.* **3**, e1701816 (2017).
- ²⁴P. Koskinen and V. Mäkinen, *Comput. Mater. Sci.* **47**, 237 (2009).
- ²⁵B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshayé, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W. z. Yu, and T. Frauenheim, *J. Chem. Phys.* **152**, 124101 (2020).
- ²⁶M. Wahiduzzaman, A. F. Oliveira, P. Philippsen, L. Zhechkov, E. van Lenthe, H. A. Witek, and T. Heine, *J. Chem. Theory Comput.* **9**, 4006 (2013).
- ²⁷C.-P. Chou, Y. Nishimura, C.-C. Fan, G. Mazur, S. Irle, and H. A. Witek, *J. Chem. Theory Comput.* **12**, 53 (2015).
- ²⁸C. Panosetti, S. B. Anniés, C. Grosu, S. Seidlmayer, and C. Scheurer, *J. Phys. Chem. A* **125**, 691 (2021).
- ²⁹A. K. A. Kandy, E. Wadbro, B. Aradi, P. Broqvist, and J. Kullgren, *J. Chem. Theory Comput.* **17**, 1771 (2021).
- ³⁰N. Goldman, K. E. Kweon, B. Sadigh, T. W. Heo, R. K. Lindsey, C. H. Pham, L. E. Fried, B. Aradi, K. Holliday, J. R. Jeffries, and B. C. Wood, *J. Chem. Theory Comput.* **17**, 4435 (2021).
- ³¹C. H. Pham, R. K. Lindsey, L. E. Fried, and N. Goldman, *J. Phys. Chem. Lett.* **13**, 2934 (2022).
- ³²K. Reuter and M. Scheffler, *Phys. Rev. B* **65**, 035406 (2001).
- ³³S. Bahn and K. Jacobsen, *Comput. Sci. Eng.* **4**, 56 (2002).
- ³⁴D. R. Hamann, *Phys. Rev. B* **88**, 085117 (2013).
- ³⁵P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Del Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- ³⁶B. Hammer, L. B. Hansen, and J. K. Nørskov, *Phys. Rev. B* **59**, 7413 (1999).
- ³⁷Y. D. Kim, A. P. Seitsonen, S. Wendt, J. Wang, C. Fan, K. Jacobi, H. Over, and G. Ertl, *J. Phys. Chem. B* **105**, 3752 (2001).
- ³⁸J. Rossmeisl, Z.-W. Qu, H. Zhu, G.-J. Kroes, and J. Nørskov, *J. Electroanal. Chem.* **607**, 83 (2007).
- ³⁹T. Wang, J. Jelic, D. Rosenthal, and K. Reuter, *ChemCatChem* **5**, 3398 (2013).
- ⁴⁰R. R. Rao, M. J. Kolb, N. B. Halck, A. F. Pedersen, A. Mehta, H. You, K. A. Stoerzinger, Z. Feng, H. A. Hansen, H. Zhou, L. Giordano, J. Rossmeisl, T. Vegge, I. Chorkendorff, I. E. L. Stephens, and Y. Shao-Horn, *Energy Environ. Sci.* **10**, 2626 (2017).
- ⁴¹C. G. Broyden, *IMA J. Appl. Math.* **6**, 222 (1970).
- ⁴²D. Goldfarb, *Math. Comput.* **24**, 23 (1970).
- ⁴³D. F. Shanno, *Math. Comput.* **24**, 647 (1970).
- ⁴⁴A. Lee and S. Castillo-Hair, see <https://pythonhosted.org/pyswarm/> for “Particle swarm optimization (pso) with constraint support” (last accessed December 2022).
- ⁴⁵A. R. Oganov and C. W. Glass, *J. Chem. Phys.* **124**, 244704 (2006).
- ⁴⁶C. W. Glass, A. R. Oganov, and N. Hansen, *Comput. Phys. Commun.* **175**, 713 (2006).
- ⁴⁷A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, *Comput. Phys. Commun.* **184**, 1172 (2013).
- ⁴⁸S. Fredericks, K. Parrish, D. Sayre, and Q. Zhu, *Comput. Phys. Commun.* **261**, 107810 (2021).
- ⁴⁹P. V. Bushlanov, V. A. Blatov, and A. R. Oganov, *Comput. Phys. Commun.* **236**, 1 (2019).
- ⁵⁰S. Anniés, C. Panosetti, M. Voronenko, D. Mauth, C. Rahe, and C. Scheurer, *Materials* **14**, 6633 (2021).
- ⁵¹Y. Lee, C. Scheurer, and K. Reuter, *ChemSusChem* **15**, e202200015 (2022).