Small Basis Set Density Functional Theory Method for Cost-efficient, Large-scale Condensed Matter Simulations

Elisabeth Keller,[†,‡] Jack Morgenstein,[¶] Karsten Reuter,[†] and Johannes T. Margraf[*,‡]

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

‡*Physikalische Chemie V, Universität Bayreuth, Bayreuth, Germany*

¶*Duke University, North Carolina, Durham, USA*

E-mail: johannes.margraf@uni-bayreuth.de

Abstract

We present an efficient first-principles based method geared towards reliably predicting the structures of solid materials across the periodic table. To this end, we use a density functional theory (DFT) baseline with a compact, near-minimal *min+s* basis set, yielding low computational costs and memory demands. Since the use of such small basis set leads to systematic errors in chemical bond lengths, we develop a linear pairwise correction (LPC), available for elements Z = 1-86 (excluding the lanthanide series), parameterized for use with the PBE exchange-correlation functional. We demonstrate the reliability of this corrected approach for equilibrium volumes across the periodic table and the transferability to differently coordinated environments and multi-elemental crystals. We examine relative energies, forces and stresses in geometry optimizations and MD simulations.

I. Introduction

In materials science, first-principles simulations are the state-of-the-art approach for obtaining detailed atomistic insight into the structure and properties of bulk materials, surface-adsorbate systems, interfaces and nanoparticles. To this end, Kohn-Sham density functional theory (KS-DFT) methods employing generalized gradient approximation (GGA) functionals, such as the Perdew–Burke-Ernzerhof (PBE) functional,[1] are extremely popular. PBE reliably describes equilibrium structures, vibrational spectra, binding and cohesive energies for a broad range of materials.[2] However, to capture the behaviour of structurally complex systems (*e.g.* defects, interfaces, and amorphous phases), large simulation cells are required, with concomitantly large demands of CPU time and memory. Additionally, to describe dynamic properties or finite-temperature effects, molecular dynamics (MD) simulations with millions of simulation steps are required. Such simulations are hindered by the computational cost of typical DFT calculations.

In terms of computational scaling, the bottleneck of KS-DFT calculations lies in the solution of the KS eigenvalue problem. Commonly used direct eigensolvers, such as ELPA,[3] scale cubically ($\mathcal{O}(N^3)$) with system size. This cubic scaling KS solution step therefore dominates the total cost of the self-consistent field (SCF) cycle in the limit of large simulation cells. Considerable research is being directed towards more cost-efficient ways to solve the eigenvalue problem. These efforts include iterative eigensolvers (*e.g.* SLEPc[4]) with better than cubic scaling or density matrix solvers (*e.g.* NTPoly[5]) that bypass the diagonalization step to reach linear scaling. Both iterative eigensolvers and density matrix solvers only reach their full potential when applied to sufficiently large sparse matrices, though (*e.g.* for lower dimensional systems and/or insulators), whereas the computational cost is unfavourable for small to medium-sized systems and dense bulk systems with small band-gaps (metals and semiconductors). Here they are still outperformed by direct eigensolvers. Consequently, despite significant progress in hardware and algorithms, KS-DFT calculations for systems with thousands of atoms are in general far from routine.

Semiempirical electronic structure methods are low-cost alternatives to DFT, which overcome its computational limitations. Popular semiempirical methods in quantum chemistry include the PMn methods (*e.g.* PM6,[6] etc.), the extended tight-binding methods GFNn-xTB (GFN-xTB,[7] GFN2-xTB[8]), and the density functional tight-binding (DFTB) approach.[9–11] These semiempirical methods offer low computational cost but lack the robustness and transferability of first-principles methods, typically relying on system-specific parameterizations.

Improved robustness and accuracy can be obtained when semiempirical methods are built on top of a first-principles baseline. This is for example done in the HF-3c method,[12] which is presently the most cost-efficient method of a set of "3c" methods developed by Grimme and co-workers.[13] Specifically, the HF-3c method uses Hartree-Fock (HF) in combination with a near-minimal basis set.[12] Clearly, the lack of electron correlation and basis-set incompleteness introduces significant errors in energies and geometries. HF-3c corrects these with three atom-pairwise empirical corrections. These include corrections for the dispersion interaction, the basis set superposition (BSSE) error and a short-range basis correction targeting the basis set incompleteness error (BSIE).[12] Notably, the "3c" approach has also been extended to DFT, *e.g.* with the PBEh-3c[14] and r2scan-3c[15] methods. However, all 3c methods are tailored to obtaining geometries and thermodynamic properties of molecular systems. Yet, there is a similar demand for cost-efficient methods for obtaining reliable geometries of inorganic bulk systems at the DFT level, *e.g.* for initial screenings in materials discovery or for generating training data for machine learning (ML) potentials. The current work therefore introduces such an approach for bulk materials within the FHI-aims DFT code.[16]

II. Method

As a basis for the proposed method, we employ a cost-efficient, first-principles model using the semilocal PBE functional with a near-minimal basis set termed *min+s*. Since we observed systematic underbinding at this level of theory (and consequently overestimated lattice constants), we propose a simple empirical correction term. Together, the *min+s* basis set and the proposed linear pairwise correction (LPC) represent a robust and cost efficient method for structural relaxations of materials across the periodic table. The method is described in detail below.

A. PBE/*min+s* baseline

1. Basis set specification

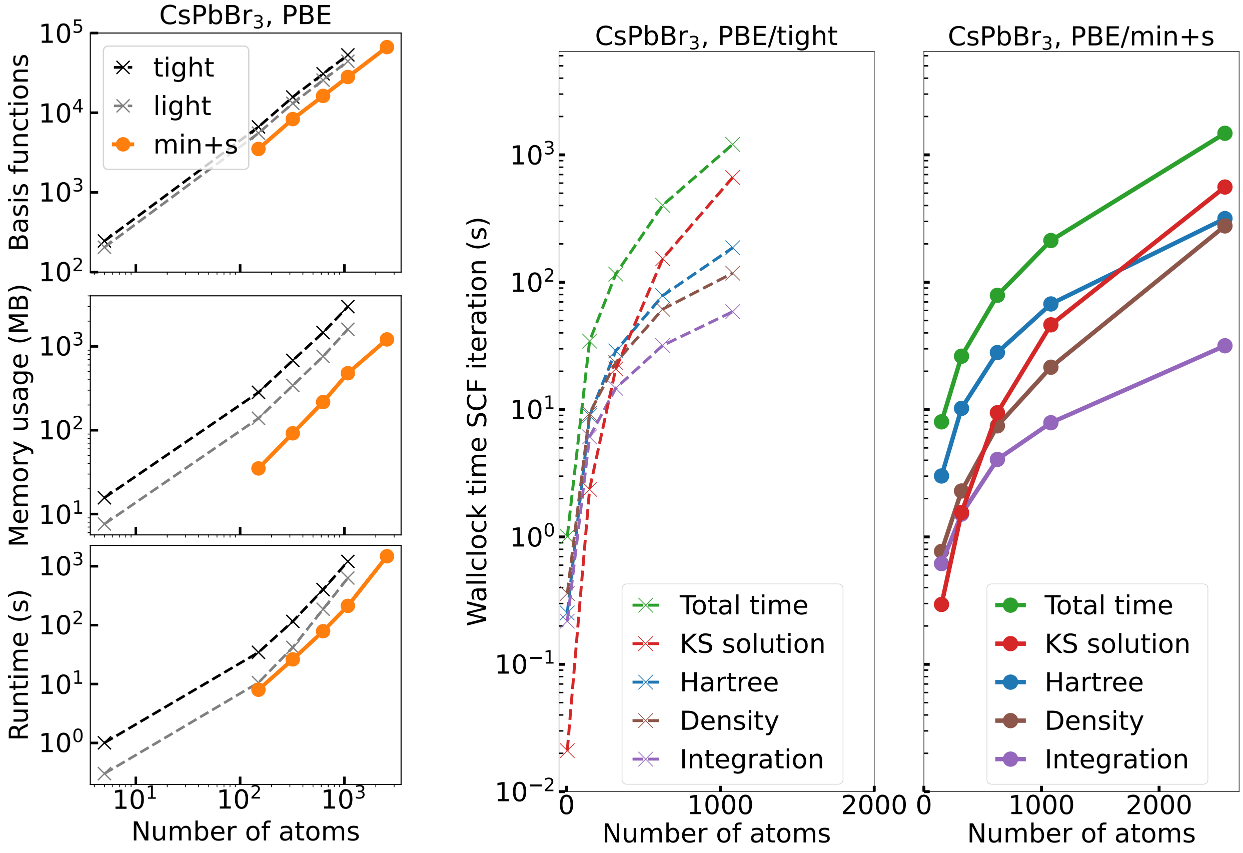
In FHI-aims, KS orbitals are expressed in terms of numeric atom-centered orbital (NAO) basis functions ϕ_i:[16]

$$\phi_i(\boldsymbol{r}) = \frac{u_i(r)}{r} Y_{lm}(\Omega) \tag{1}$$

with localized numerical tabulated radial functions $u_i(r)$ and spherical harmonics $Y_{lm}(\Omega)$. For each element, the NAO basis set is hierarchically constructed from a minimal free-atom basis set by iteratively adding additional radial functions until a required level of energy convergence is reached. FHI-aims provides predefined numerical settings for different levels of convergence. These settings define a set of NAO basis functions and correspondingly adjusted integration grids, multipole expansions for the Hartree potential, etc. The most commonly used settings in FHI-aims are termed *light* and *tight*. For GGA functionals, such as PBE, the *tight* settings are essentially fully converged and recommended for highly precise energy calculations, whereas the *light* settings are much more cost-efficient and often used for structural relaxations and ab initio molecular dynamics.[16] For large systems, where the solution of the KS equations dominates the computational cost, even *light* settings can become computationally prohibitive in terms of CPU time and memory consumption, however.

In this work, we employ a cost-efficient near-minimal NAO basis set, which we denote as *min+s*. The *min+s* basis consists of a minimal set of basis functions (which includes the full valence shell of the corresponding element) and one additional s-type function, which grants some amount of radial flexibility to the basis set at negligible computational cost. All basis functions are localized within a basis set cutoff inherited from the *light* settings, where the cutoff is chosen to be as small as possible without significantly affecting the accuracy of computations. For completeness, the *min+s* basis set cutoffs for each element are provided in the Supplementary Information. To further reduce the computational cost, only the chemically relevant valence and shallow-core electrons are considered by using the frozen core approximation in the solution of the KS equations. Here, we are using the implementation by Yu et al. (FC99+C+V) with an energy cutoff of -100 eV (if not otherwise stated).[17,18]

2. Computational cost



(a) (b)

Figure 1: Scaling behaviour of the computational cost for self-consistent field (SCF) iterations employing the *min+s*, *light*, and *tight* settings as a function of system size for $CsPbBr_3$ supercells with 5, 150, 320, 625, 1080, 2560 atoms. a) Comparison of the number of basis functions, highest tracked memory usage and total runtime. b) Subtimings for the first SCF iteration with the *min+s* and *tight* basis sets. The 5, 150, 320, 625, 1080 and 2560-atom supercells have been computed on 2 nodes with 40 CPUs each with large memory nodes 192 GB on the HPC cluster Cobra (processor type: Intel Skylake 6148, processor clock: 2.4 GHz). For *min+s* settings 5-atom computations were excluded due to ELPA errors from too small matrices for the chosen number of CPUs. For *light* and *tight* settings the 2560-atom supercell could not be computed due to memory limitations.

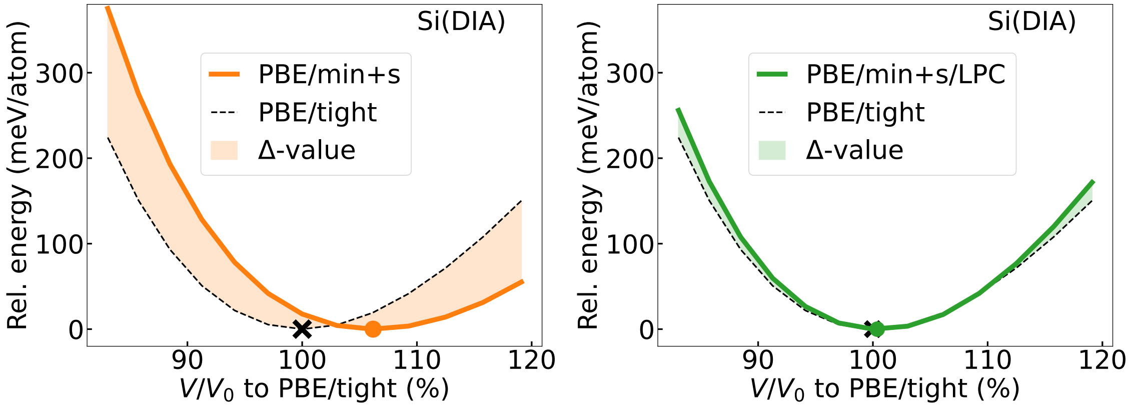
In Fig. 1a, the computational cost of the *min+s* basis set is compared to the *light* and *tight* settings for a single SCF iteration on $CsPbBr_3$ supercells with up to 2560 atoms. PBE/*min+s* reduces the total wallclock time on average to 52% and 21% compared to PBE/*light* and PBE/*tight*. Beyond the runtime, large-scale simulations are also often limited by their memory demand. Here, PBE/*min+s* reduces the required memory usage on average to 15%, 28% compared to PBE/*light* and PBE/*tight*, respectively. The computational time and memory savings due to *min+s* are thereby mainly due to the reduced number of basis functions and lighter integration grids (compared to *tight*). However, the overall reduction varies according to the contained elements, as well as the size and density of the examined system. Generally, the savings are largest for elements with low atomic numbers and decrease for heavy elements with many core orbitals. In this respect $CsPbBr_3$ is far from the best case scenario. Nevertheless, the *min+s* settings clearly lead to substantial savings compared to the *light* and especially *tight* settings.

To obtain better insight into the time savings for different system sizes, we examine the major contributors to the total wallclock time of the first SCF iteration in Fig. 1b. For small system sizes the major contributors to the total time are linear scaling grid-based computational steps (Hartree potential, density update, integration step). For large system sizes the cubic scaling KS solution dominates. By using the *min+s* basis set (Fig. 1b, right) instead of the *tight* basis set (Fig. 1b left), the computational cost for both grid-based computational steps and the KS solution is decreased, yielding lower computational cost for small, medium and large-scale systems. Even better, the relative savings increase with increasing system size (see Supplementary Information Tab.1) due to the increased sparsity of matrices obtained with the *min+s* basis set. Crucial for enabling large-scale computations is the crossover point, for which the cost of the cubic scaling KS solution exceeds the linear scaling grid-based computational steps. PBE/*min+s* shifts this crossover point to significantly larger system sizes (around 1000 atoms). In turn, this pushes the KS bottleneck to larger system sizes and enables cost-efficient large-scale simulations.

3. Basis set incompleteness errors for crystals

Table 1: Test set of 128 materials including noble gases,[19] ionic binary compounds,[20] covalent semiconductor binary compounds,[20] metalloids,[19] metals[19] and molecular elemental crystals.[19]

Bonding type	Number of materials	Materials
Noble gases	6	He, Ne, Ar, Kr, Xe, Rn[19]
Ionic	21	alkali halides AB with A = Li, Na, K, Rb, Cs and B = F, Cl, Br, I[20]
Covalent	37	AlAs(ZB), AlN(WUR), AlN(ZB), AlP(ZB),AlSb(ZB), BAs(ZB), BP(ZB), CdS(WUR), CdS(ZB), CdSe(WUR), CdSe(ZB), CdTe(ZB), GaAs(ZB), GaN(WUR), GaN(ZB), GaP(ZB), GaSb(ZB), HgS(ZB), HgSe(ZB), HgTe(ZB), InAs(ZB), InN(WUR), InP(ZB), InSb(ZB), MgO(RS), MgS(RS), MgSe(RS), PbS(RS), PbSe(RS), PbTe(RS), SiC(ZB), ZnO(WUR), ZnS(WUR), ZnS(ZB), ZnSe(ZB), ZnTe(ZB), C(DIA)[20]
Metalloid	8	B, Si, Ge, As, Se, Sb, Po, Te[19]
Metallic	47	Li, Na, K, Rb, Cs, Be, Mg, Ca, Sr, Ba, Sc, Y, Lu, Ti, Zr, Hf, V, Nb, Ta, Cr, Mo, W, Mn, Tc, Re, Fe, Ru, Os, Co, Rh, Ir, Ni, Pd, Pt, Cu, Ag, Au, Zn, Cd, Hg, Al, Ga, In, Sn, Tl, Pb, Bi[19]
Molecular elemental crystals	9	H_2, O_2, N_2, F_2, P, S, Cl_2, Br_2, I_2[19]



(a) (b)

Figure 2: Energy-volume curves for bulk silicon obtained with a) PBE/*min+s* and b) PBE/*min+s*/LPC, both relative to PBE/*tight*. The Δ-values show the dissimilarity of the E(V) curves obtained from PBE/*min+s* and PBE/*min+s*/LPC compared to PBE/*tight*. The equilibrium volumes obtained with PBE/*min+s*, PBE/*min+s*/LPC and PBE/*tight* are marked by dots and crosses.

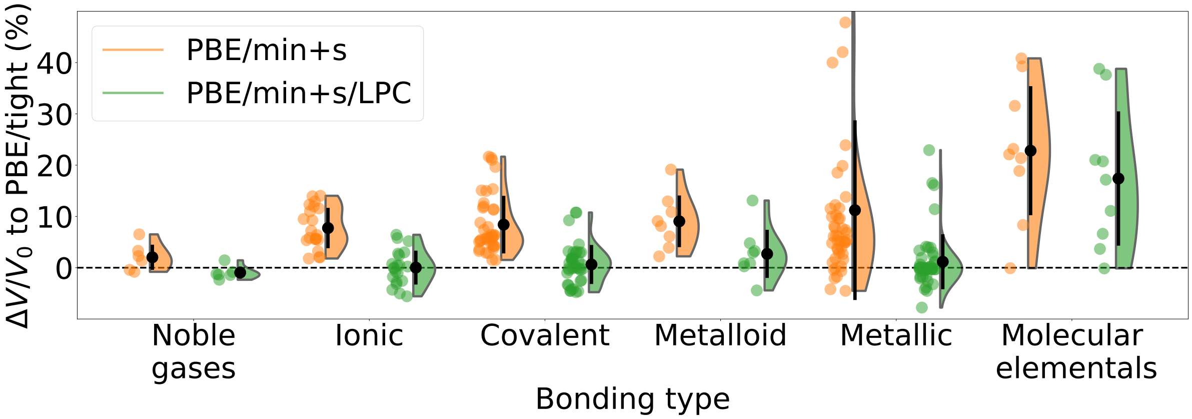


Figure 3: Performance of PBE/*min+s* and PBE/*min+s*/LPC for equilibrium volumes of bulk systems with different bonding types, relative to PBE/*tight* (dataset see Tab.1). The mean errors and standard deviations are indicated in black. The original data is plotted on the left side of each distribution.

Using the compact *min+s* basis clearly leads to computational advantages, but this inevitably has negative effects on the predictive accuracy of the calculations, due to basis set incompleteness errors (BSIE). For example, it has been observed that small basis sets lead to systematically overestimated bond lengths in organic molecules.[12,21] To examine whether similar systematic trends can be observed in solids, we computed energy-volume curves and derived equilibrium properties (such as equilibrium volumes V_0) of bulk crystals across the periodic table. Note that for simplicity we will use the term BSIE for the discrepancy between *min+s* and *tight* settings in the following, although there are also (smaller) contributions from grid densities and other factors to this difference.

For illustration, the PBE/*min+s* and PBE/*tight* energy-volume curves of bulk silicon in the diamond crystal structure are shown in Fig. 2a. The corresponding equilibrium volumes are obtained from a Birch–Murnaghan equation-of-state fit,[22,23] performed using the atomic simulation environment (ASE).[24] This reveals that PBE/*min+s* overestimates the equilibrium volume V_0 of silicon by 6%, compared to PBE/*tight*. Beyond this, the BSIE also impacts the shape of the energy volume curve. This discrepancy in shape and equilibrium volume between two energy-volume curves can be assessed with the Δ-value introduced by Lejaeghere et al.:[19]

$$\Delta = \sqrt{\frac{\int \Delta E^2(V)\mathrm{d}V}{\Delta V}}. \tag{2}$$

These Δ-values will be used as an optimization target further below.

Importantly, the overestimation of equilibrium volumes due to BSIE is not just observed for silicon, but for a wide range of elements and bonding types including ionic, covalent, metallic, and molecular systems. This is shown for a dataset of 128 crystals in Fig. 3, (see Tab. 1). The volumes are commonly overestimated by 5-10%, and in some cases by more than 40%. Analyzing the BSIE for different bonding types, we find that the error is smallest for systems which are reasonably similar to free atoms, such as noble gases or ionic materials. In contrast, the overestimation increases for more complex bonding situations, *e.g.* in covalent, metallic and molecular systems. This can be rationalized by considering that the minimal basis set is obtained from isolated atom calculations in the NAO scheme. A near minimal basis is thus well suited for free atoms (the limit of infinite volume in an energy volume curve), and the BSIE will be more pronounced for smaller volumes. Overall, this results in underbinding and an overestimation of equilibrium volumes.

B. Linear pairwise correction (LPC) for BSIE

1. Method definition

Having observed the systematic BSIE effect on energy volume curves, we now aim to correct the potential energy surface (PES) in such a way that overestimated bond lengths are shortened without adversely affecting the PES and the related thermodynamic ensemble. To this end we draw on the literature of minimally-invasive biases that have been developed for large-scale biomolecular simulations. In particular, Pitera and Chodera introduced a linear form of bias based on a maximum entropy argument, which distorts the unbiased statistical ensemble the least.[25,26] With this goal in mind, we define a simple linear pairwise correction (LPC) $e_{\mathrm{LPC},AB}$, which is fast to evaluate and easy to parameterize:

$$e_{\mathrm{LPC},AB} = c_{Z_A Z_B} \cdot (r_{AB} - r_{\mathrm{cut},Z_A Z_B}) \tag{3}$$

with $r_{AB} = |\boldsymbol{r}_{AB}|$ being the absolute distance between atoms A and B, $c_{Z_A Z_B}$ denoting the element-pair dependent correction strength (with $c_{Z_A Z_B} \geq 0$), and $r_{\mathrm{cut},Z_A Z_B}$ denoting an element-pairwise cutoff radius. The latter provides a measure for the onset of the correction, which should be short-ranged and act mainly on directly bonded atoms, while interactions between next-nearest neighbors and beyond are removed by a switching function (see below). The effect of the LPC on the silicon energy-volume curve is shown in Fig. 2b. This confirms that the underbinding is corrected, without otherwise distorting the potential energy surface.

To avoid the need for parameterizing all element-pairs in the periodic table, the pairwise parameters $c_{Z_A Z_B}$ and $r_{\mathrm{cut},Z_A Z_B}$ are determined by via arithmetic means of the contributing species:

$$c_{Z_\mathrm{A}Z_\mathrm{B}} = \frac{c_{Z_\mathrm{A}} + c_{Z_\mathrm{B}}}{2}, \tag{4}$$

$$r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} = \frac{r_{\mathrm{cut},Z_\mathrm{A}} + r_{\mathrm{cut},Z_\mathrm{B}}}{2}. \tag{5}$$

The use of the arithmetic mean is a common choice for cutoff radii. It is, *e.g.*, also used for Lennard-Jones potentials. For the correction strength $c_{Z_\mathrm{A}Z_\mathrm{B}}$, use of the geometric mean was also explored. However, this proved problematic in cases where the parameterization yielded values of c_{Z_A} close to zero (see below). With the geometric mean, all pairwise corrections involving these elements would be zero, while the arithmetic mean is more well behaved in this case.

The full correction term E_LPC is obtained by summing up the pairwise corrections $e_\mathrm{LPC,AB}$ for each atom pair AB, multiplied by the aforementioned switching function:

$$E_\mathrm{LPC} = \frac{1}{2}\sum_{A}^{N_\mathrm{unit}}\sum_{B\neq A}^{N_\mathrm{super}} e_{\mathrm{LPC},AB} \cdot f_\mathrm{switch}(r_{AB}, r_{\mathrm{cut},Z_AZ_B}). \tag{6}$$

The switching function f_switch of width $\mathrm{w} = 0.5$ Å is given by:[27]

$$f_\mathrm{switch}(r_{AB}, r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}}) = \begin{cases} 1 & \text{if } \mathrm{r}_{AB} < r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} \text{ - } w \\ \frac{1}{2}\ (\ \cos(\frac{\pi}{w} \cdot (r_{AB} - r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} + w)) + 1\ ) & \text{if } r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} \text{ - } w \leq r_{AB} \leq r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} \cdot \\ 0 & \text{if } r_{AB} > r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}} \end{cases} \tag{7}$$

f_switch ensures a smooth transition to zero as $r_{\mathrm{cut},Z_\mathrm{A}Z_\mathrm{B}}$ is approached. This in turn leads to continuous derivatives, which enables force and stress evaluations. Finally, the total corrected energy $E_\mathrm{min+s/LPC}$ consists of the first-principles energy $E_\mathrm{PBE/min+s}$ obtained with the PBE/*min+s* baseline and the LPC correction term E_LPC:

$$E_\mathrm{min+s/LPC} = E_\mathrm{PBE/min+s} + E_\mathrm{LPC}. \tag{8}$$

Expressions for LPC forces and stresses are given in the Supplementary Information.

2. Parameterization

Clearly, the accuracy of the LPC ultimately depends on an appropriate parameterization of the correction strength c_{Z_A} and the cutoff $r_{\text{cut},Z_\text{A}}$. The latter mainly serves to ensure that the correction is applied to all relevant short-range interactions, while leaving long-range interactions unaltered. For most elements, this can be achieved by setting r_{cut,Z_A} to 2.5 times the corresponding elemental single-bond covalent radius ($r_{\text{cut},Z_A} = 2.5r_{\text{cov},Z_A}$), with radii taken from Refs.[28,29] Exceptions are made for some elements which require larger cutoffs to cover the most common bonding situations. These exceptions include noble gases for which van-der-Waals radii are used ($r_{\text{cut},Z_A} = 2.5r_{\text{vdW},Z_A}$ with r_{vdW,Z_A} for He, Ne, Ar, Kr, Xe from Refs.[30] and[31]). Furthermore, slightly larger cutoffs are used for S, Hg, Pb, Se, Be ($r_{\text{cut},Z_A} = 3r_{\text{cov},Z_A}$), all of which display diverse bonding patterns in different crystal polymorphs.

In contrast to the cutoffs, the correction strengths c_{Z_A} need to be more carefully tuned for each element, in order to minimize the systematic underbinding caused by the BSIE. To this end, a training set including a range of common homoelemental bonding situations for each element (namely the dimer, graphite, diamond, ß-tin, bcc and fcc prototypes) was used, as first reported in reference.[32] This consistent set of structures covers coordination numbers from 1-12 and is thus representative of the diverse bonding situations encountered in solids. However, this diversity also has a downside, in that less important high energy configurations (*e.g.* fcc oxygen) can dominate the error when optimizing the parameters, leading to an unbalanced paramterizations.

To ensure that the LPC is robust for the important low energy configurations of each element, the structures are therefore weighted according to a Boltzmann distribution centered on the energetically most stable structure (at the PBE/*tight* level), with the Boltzmann factor p_i of structure $i \in I = \{\text{dimer, graphite, diamond, ß-tin, bcc and fcc}\}$ given as:

$$p_i = \frac{1}{Z}\exp\left(\frac{-\Delta E_i}{kT}\right), \tag{9}$$

with

$$\Delta E_i = E_{0,i} - \min(\{E_{0,j}|j \in I\}). \tag{10}$$

and the normalization factor $Z = \sum_{j\in I}\exp\left(\frac{-\Delta E_j}{kT}\right)$. Here, kT is the product of the Boltzmann constant k and the temperature T. In the fitting procedure kT is fixed to 0.25 eV, which provides a good balance between the correct description of low energy configurations and a qualitative description of higher energy configurations. To further reduce the influence of outliers (*e.g.* in many cases the dimers), structures with Boltzmann factors smaller than ten percent are excluded from the fitting procedure.

Based on these structures, the c_{Z_A} parameters for all elements were optimized with the Nelder-Mead method,[33 34] for fixed $r_{\mathrm{cut},A}$, minimizing the loss function $L(c_{Z_A})$ by summing over the Boltzmann-factor p_i (Eq. 9) weighted Δ_i-values (Eq. 2) for each structure i in the training set :

$$L(c_{Z_A}) = \sum_{i\in I}\Delta_i(E_{\mathrm{min+LPC}}, E_{\mathrm{PBE}/tight})\cdot p_i \tag{11}$$

While this approach for reference data generation worked well for most elements, exceptions were made for H, O, N, F, Cl, Br, and I. This was necessary due to the fact that these elements form molecular dimers, which dominate the loss function when following the procedure described above. As the chemistry of molecular dimers is very different from important classes of solids (*i.e.* hydrides, oxides, nitrides and halides), the corresponding parameters were reoptimized for representative binary compounds, keeping the parameters of the other elements fixed. For further details regarding the dataset, please refer to the Supplementary Information.

Overall, we thus obtained LPC parameters for elements with Z = 1-86 (excluding the lanthanide series) for use with the *min+s* basis set and the PBE functional. The *min+s* basis set and the LPC correction are accessible through the corresponding species defaults in FHI-aims starting from version 231212 in $\mathrm{species\_defaults/defaults\_2020/minimal{+}s}$, which automatically include the keyword to enable the basis set error corrections (currently parameterized for use with the *min+s* basis set and the PBE functional) for energies, forces and stresses.

III. Results and Discussion

Having defined the LPC, we first benchmark PBE/*min+s*/LPC for equilibrium volumes of crystals against a PBE/*tight* reference. Subsequently, we examine the performance of PBE/*min+s*/LPC for geometry optimizations and MD simulations.

Equilibrium volumes of crystalline solids

The performance of PBE/*min+s*/LPC is examined for a test set of monoelemental and binary solid state systems categorized into predominantly ionic, covalent, metalloid and metallic materials and molecular elemental solids (see Tab.1). For this test set, the equilibrium volumes obtained with the PBE/*min+s*/LPC and PBE/*min+s* are compared to the PBE/*tight* reference. As mentioned above, PBE/*min+s* overestimates most equilibrium volumes, whereas PBE/*min+s*/LPC reduces the BSIE significantly and obtains reliable equilibrium geometries for most materials, see Fig. 3.

Even though the LPC was mostly fitted on monoelemental reference structures, the method is transferable to multi-elemental materials. Indeed, basis set errors are generally larger for monoelemental systems (mainly represented in the metallic, covalent and molecular elementals classes) compared to poly-elemental (e.g. ionic) materials. Overall, PBE/*min+s*/LPC shows the largest residual errors for molecular elemental dimer structures such as O_2 or N_2, where the performance is only slightly better than PBE/*min+s*. This can be attributed to the fact that the LPC was fitted to binary compounds (such as oxides and nitrides) for these elements, and is thus not well suited for the corresponding elemental molecular systems. Indeed, the results indicate that fitting on elemental dimers would likely lead to an over-correction for solid binaries, vindicating the selected fitting strategy.

It should be noted that the energy-volume curves are computed by applying uniform strain to the cell. Under these conditions, the curves for molecular dimers mainly reflect a stretching or compression of the covalent bonds in the dimers. In principle, one could also compute energy-volume curves under the condition of fixed dimer geometries. In this case, the curve would likely be dominated by the basis set superposition error (BSSE), which leads to overbinding in non-covalent interactions. BSSE effects are not addressed by the LPC correction, since they occur on a different lengthscale (i.e. on the order of van-der-Waals radii) and have the opposite sign. When using a dispersion correction, BSSE effects can be compensated to a certain degree by adjusting the damping factor and onset of damping of the switching function. Furthermore a method analogous to the semiempirical geometrical counterpoise correction (gCP) of Kruse and Grimme[35] could be developed for solids. These developments are beyond the scope of the current paper, however. At this stage, the PBE/*min*+*s*/LPC method is not recommended for systems that are dominated by van-der-Waals interactions.

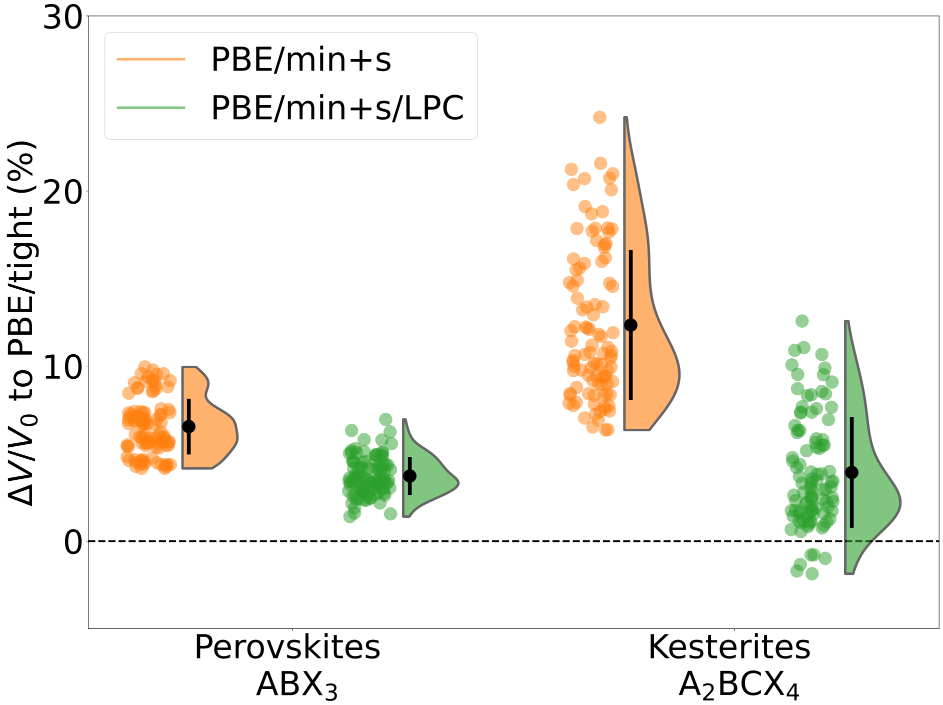


Figure 4: Performance of PBE/*min+s* and PBE/*min+s*/LPC for equilibrium volumes of 100 organo-metal halide perovskites ABX_3 (structures from Ref.[36]) and 100 kesterites A_2BCX_4 (structures from Ref.[37]), relative to PBE/*tight*. The mean errors and standard deviations are indicated in black. The raw data is plotted on the left side of each distribution.

To test the transferability of the PBE/*min+s*/LPC method further, we examined its performance for a test set of 100 organo-metal halide perovskite materials ABX_3 (structures from Ref.[36]) and 100 quarternary kesterite materials A_2BCX_4 (the first 100 structures from Ref.[37]), see Fig. 4. Because these materials consist of three to four different elements and may include small organic molecules, this represents a challenging test for the PBE/*min+s*/LPC method.

We find that PBE/*min+s*/LPC remains a consistent improvement over PBE/*min+s* here, although the volumes are still overestimated by ca. 5% on average. The improvement is particularly significant for the kesterites, where volumes are overestimated by up to 25% at the PBE/*min+s* level. Notably, the residual error of PBE/*min+s*/LPC is fairly systematic, so that most systems are still underbound. This points to limitations of the simple functional form of the LPC, which cannot distinguish different crystal environments. Nonetheless, the current approach represents an improvement over the baseline, even in this extrapolative setting.

Applications

One of the main use-cases for methods like PBE/*min*+*s*/LPC is the (pre-)relaxation of medium to large simulation cells, *e.g.* in the context of *ab initio* thermodynamics studies or materials screening.[38] While we have established that the proposed method will yield improved equilibrium volumes, the task of geometry optimization can potentially start from structures that are far from equilibrium.

To demonstrate this, the relative volume deviations for full unit-cell relaxations of a compressed and rattled 512-atom silicon supercell are shown in Fig. 5. The initial lattice constant was 4.95 Å (corresponding to a volume compression of 26% in the 512-atom supercell) and the atomic positions were randomly displaced from the diamond structure with a standard deviation of 0.21 Å. This cell was subsequently relaxed using the rust-radius enhanced Broyden-Fletcher-Goldfarb-Shann (BFGS) algorithm,[16,39,40] with a convergence criterium of $|F_{\text{max}}| \leq 0.01$ eV/Å. Note that this medium-sized system was deliberately chosen to allow comparison with fully converged PBE/*tight* calculations. Even here, the PBE/*min*+*s* based models show significant computational benefits, with 3-4 times lower memory demands and calculation times. Full timings and memory usage statistics of each method are provided in the Supplementary Information.

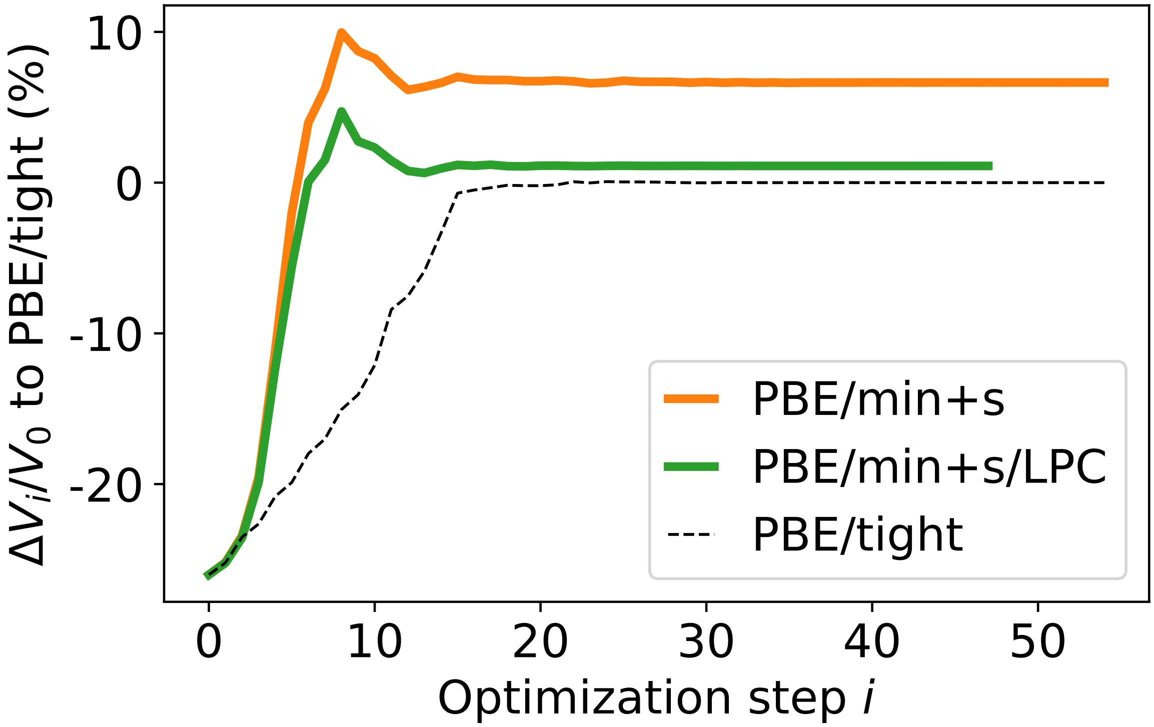
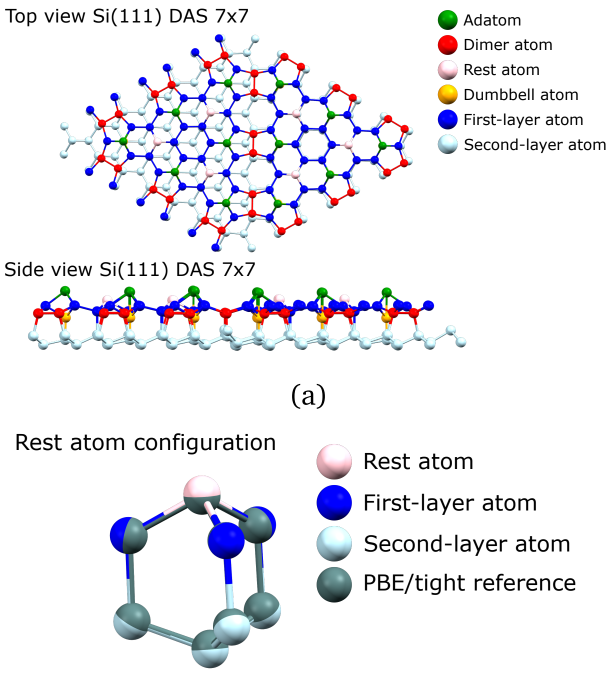


Figure 5: Geometry optimization path obtained with LPC corrected PBE/"minimal+s" (energy cutoff -200 eV), PBE/*min+s* and PBE/*tight* starting from a stretched and rattled Si(DIA) 512-atom supercell. For each optimization step i the relative volume deviation compared to the relaxed reference geometry (PBE/*tight* $|\mathrm{F}_{\mathrm{max}}| \leq 0.01$ eV/Å) is plotted.

To gauge how the LPC affects the shape of the PES far from equilibrium, we can examine the optimization paths PBE/*min+s*/LPC compared to PBE/*min+s* and PBE/*tight*, see Fig. 5. We observe that the initial optimization steps are nearly identical for PBE/*min+s*/LPC and PBE/*min+s*, displaying a fast expansion of the cell volume. In contrast, the PBE/*tight* benchmark approaches the equilibrium volume at a gradual pace. After the first five steps, the LPC correction steers the optimization towards the PBE/*tight* equilibrium volume, diverging from the PBE/*min+s* path, which overestimates the final relaxed volume by 7 %, compared to a 1 % overestimation with PBE/*min+s*/LPC. This behaviour illustrates the conservative nature of the LPC, in that it yields the correct results around equilibrium and otherwise leaves the baseline method mostly unaffected. This is an important property, because it avoids the generation of spurious minima on the PES, which can result from non-linear corrections.[41,42]



(a)

(b)

Figure 6: a) The relaxed Si(111)-DAS 7x7 surface reconstruction relaxed at the PBE/*min+s*/LPC level, in top and side view. The lower eight sub-surface layers are not shown for clarity. The Si atoms are color-coded analogously to Ref.:[43] dumbbell atoms (orange), rest atoms (pink), dimer atoms (red), adatoms (green), bulk-like atoms in top layer (dark blue) and second-layer atoms (*light* blue). The initial geometry was taken from Ref.[43] b) Relaxed rest atom configurations from the Si(111)-DAS 7x7 surface overlayed with the PBE/*tight* reference (dark green).

Moving beyond bulk systems, it is also important to establish the transferability of the LPC to surface systems. Surface slab calculations often require large supercells, in order to accommodate complex reconstructions and obtain sufficiently bulk-like properties for central atoms. An example for such a system is the Si(111)-dimer-adatom stacking fault (DAS) 7×7 reconstruction.[44] We use an initial structure with 1033 atoms obtained from Ref.[43] Therein, Shen et al. employed an ML force field to relax the structure, due to the high computational cost of first-principles methods for such systems.

We performed local relaxations of this surface structure, both at the PBE/*min*+*s*/LPC and PBE/*tight* levels, with a convergence criterium of $|F_{\max}| \leq 0.01$ eV/Å. Importantly, the PBE/*min*+*s*/LPC structure displays all the characteristic features of the DAS 7×7 reconstruction: two triangular faulted and unfaulted half unit cells with adatoms, rest atoms and dimers, see Fig. 6. Beyond these qualitative features, the structure is also quantitatively in good agreement with the PBE/*tight* benchmark: The distances between the adatoms shown in deviates on average by 0.37 %, the dimer bond lengths by 1.7 %, and the distances of the rest atoms to the atoms in the top layer deviate on by 2.1 %. Absolute distance errors can be found in the Supplementary Information.

Structural relaxations yield the ground state geometry (or some meta-stable state) of a system at T=0 K. In practice, we are often also interested in finite temperature properties, however. These can be accessed via molecular dynamics (MD) simulations, which by definition also explore non-equilibrium regions of the PES.

To explore the performance of PBE/*min*+*s*/LPC in this setting, we performed MD simulations for bulk Copper at T=100 K, 300 K, 500 K, and 2000 K. For each method, a 108-atom fcc-Cu supercell was first relaxed and subsequently equilibrated for 3 ps in the NVT ensemble (using the Nosé-Hoover thermostat), followed by 3 ps production runs in the NVE ensemble. A 3 fs timestep was used throughout.

The corresponding radial distribution functions (RDFs) are shown in Fig. 7. Due to its underbinding tendencies (and consequently too large unit cell), the PBE/*min*+*s* model significantly overestimates interatomic distances at low to moderate temperatures (100 K, 300 K, 500 K). In contrast, the corrected PBE/*min*+*s*/LPC model faithfully reproduces the PBE/*light* reference, albeit with slightly broadened features at 100 K and 300 K. This trend reverses somewhat as the RDF features broaden with higher temperatures, so that agreement between PBE/*min*+*s*/LPC and the reference is essentially perfect at 500 K, and the RDF is slightly overstructured at 2000 K. Overall, these simulations show that the LPC is also beneficial in molecular dynamics simulations, despite the focus on equilibrium volumes in the parameterization. Furthermore, these simulations indicate that the LPC has no adverse affects on numerical stability of dynamics trajectory, as energy conservation during NVE simulations is unaffected, see SI Fig. 4.

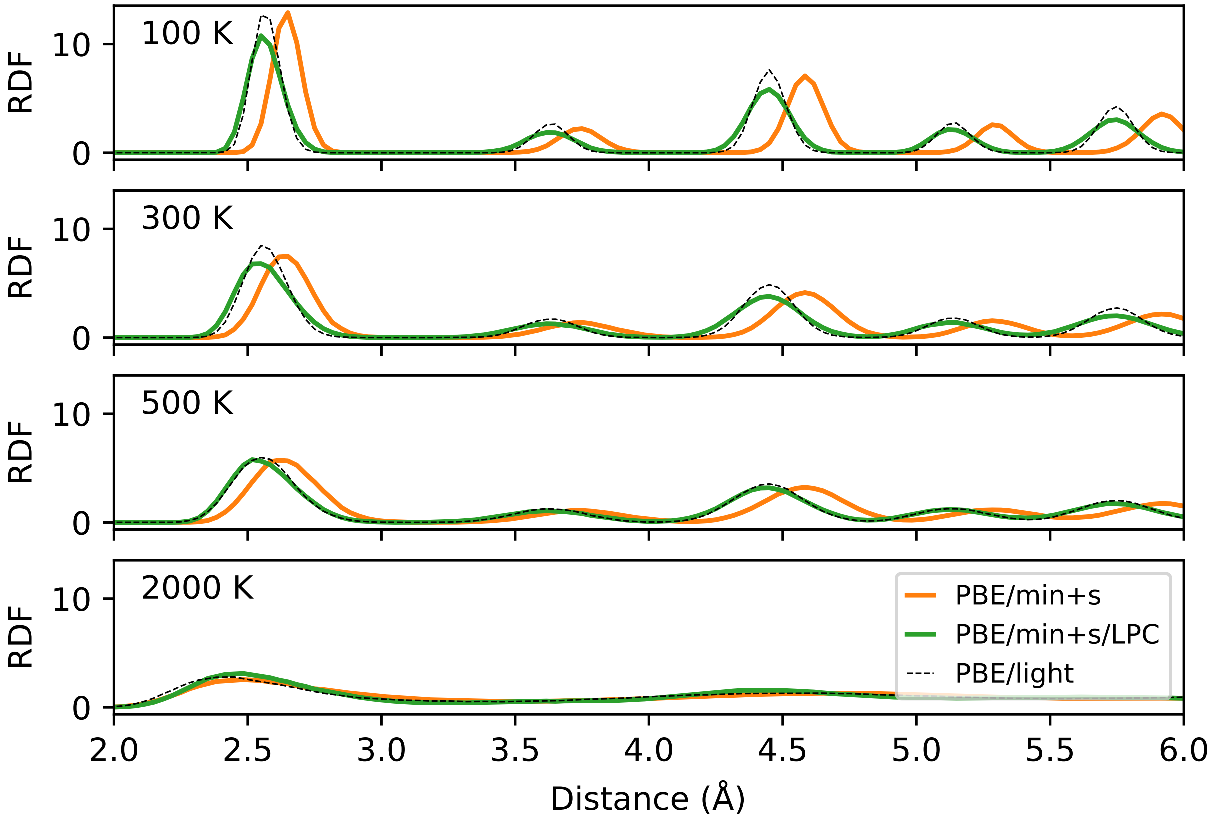


Figure 7: Performance of PBE/*min*+*s*/LPC for MD simulations of bulk Copper compared to PBE/*min*+*s* and PBE/*light*. Radial distribution functions for 108 atom supercells in the NVE ensemble after equilibrating at T=100 K, 300 K, 500 K, 2000 K averaged over 600 snapshots after equilibration.

IV. Conclusion

In this paper, we proposed a semiempirical small basis set density functional method for cost-efficient, large-scale material simulations denoted PBE/*min*+*s*/LPC. The method is parameterized for elements up to radon (Z = 1-86, excluding the lanthanide series). The method employs a well-balanced, near-minimal *min*+*s* NAO basis set, which leads to significant savings in terms of computational cost and memory demand, compared to fully converged calculations. In order to address the systematic overestimation of bond lengths caused by basis set incompleteness, a minimally-invasive pairwise correction is used. The resulting method reliably provides accurate equilibrium volumes for mono- and poly-elemental crystals in diverse bonding situations. Despite focusing on equilibrium structures for the parameterization, PBE/*min+s*/LPC does not deteriorate the quality of the baseline method when out of equilibrium (*e.g.* for distorted structures or in MD simulations). While the proposed method is geared towards use in the FHI-aims code, the underlying concepts could easily be transferred to other codes using atom centered basis functions.

We envision that methods like PBE/*min+s*/LPC will be useful in the space between fully converged first principles methods (which offer high accuracy at high comutational cost) and efficient ML potentials (which are computationally efficient but not always reliable, depending on the availability of adequate training data). For example, they can be used for initial relaxations or MD trajectories to generate realistic atomistic configurations for training an ML potential. In this case, the systematic volume errors of pure PBE/*min+s* would be problematic, because they would bias the configurations away from the target region. PBE/*min+s*/LPC can also be useful as a pre-relaxation method, when fully converged DFT structures are required.

Acknowledgements

The authors acknowledge the Max Planck Computing and Data Facility (MPCDF) for computation time. We thank Noam Bernstein for providing the initial geometries of the monoelemental fitting set. E.K. acknowledges support by the IMPRS for Elementary Processes in Physical Chemistry.

Data availability

The reference data used in this study is available in the supporting information to this article.

Author declarations

The authors have no conflicts to disclose.

References

(1) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(2) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B Condens. Matter* **2009**, *79*, 155107.

(3) Marek, A.; Blum, V.; Johanni, R.; Havu, V.; Lang, B.; Auckenthaler, T.; Heinecke, A.; Bungartz, H.-J.; Lederer, H. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *J. Phys. Condens. Matter* **2014**, *26*, 213201.

(4) Hernandez, V.; Roman, J. E.; Vidal, V. SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Softw.* **2005**, *31*, 351–362.

(5) Dawson, W.; Nakajima, T. Massively parallel sparse matrix function calculations with NTPoly. *Comput. Phys. Commun.* **2018**, *225*, 154–165.

(6) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modifica-

tion of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(7) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(8) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(9) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B Condens. Matter* **1995**, *51*, 12947–12957.

(10) Seifert, G.; Porezag, D.; Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

(11) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B Condens. Matter* **1998**, *58*, 7260–7268.

(12) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

(13) Caldeweyher, E.; Brandenburg, J. G. Simplified DFT methods for consistent structures and energies of large systems. *J. Phys. Condens. Matter* **2018**, *30*, 213001.

(14) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, 054107.

(15) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.* **2021**, *154*, 064103.

(16) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.

(17) Koepernik, K.; Eschrig, H. Full-potential nonorthogonal local-orbital minimum-basis band-structure scheme. *Phys. Rev. B Condens. Matter* **1999**, *59*, 1743–1757.

(18) Yu, V. W.-z.; Moussa, J.; Blum, V. Accurate frozen core approximation for all-electron density-functional theory. *J. Chem. Phys.* **2021**, *154*, 224107.

(19) Lejaeghere, K.; Van Speybroeck, V.; Van Oost, G.; Cottenier, S. Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals. *Crit. Rev. Solid State Mater. Sci.* **2014**, *39*, 1–24.

(20) Huhn, W. P.; Blum, V. 103-Compound Band Structure Benchmark of Post-SCF Spin-Orbit Coupling Treatments in Density-Functional Theory. *Phys. Rev. Mater.* **2017**, *1*, 033803.

(21) Vuckovic, S.; Burke, K. Quantifying and Understanding Errors in Molecular Geometries. *J. Phys. Chem. Lett.* **2020**, *11*, 9957–9964.

(22) Murnaghan, F. D. The Compressibility of Media under Extreme Pressures. *Proc. Natl. Acad. Sci. U.S.A.* **1944**, *30*, 244–247.

(23) Birch, F. Finite Elastic Strain of Cubic Crystals. *Phys. Rev. Journals Archive* **1947**, *71*, 809–824.

(24) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(25) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.

(26) White, A. D.; Voth, G. A. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *J. Chem. Theory Comput.* **2014**, *10*, 3023–3030.

(27) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(28) Pyykkö, P.; Atsumi, M. Molecular Single-Bond Covalent Radii for Elements 1-118. *Chem. Eur. J.* **2009**, *15*, 186–197.

(29) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832–2838.

(30) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(31) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. Consistent van der Waals Radii for the Whole Main Group. *J. Phys. Chem. A* **2009**, *113*, 5806–5812.

(32) Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; Csanyi, G. Mapping Materials and Molecules. *Acc. Chem. Res.* **2020**, *53*, 1981–1991.

(33) Gao, F.; Han, L. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* **2012**, *51*, 259–277.

(34) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272.

(35) Kruse, H.; Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **2012**, *136*, 154101.

(36) Castelli, I. E.; García-Lastra, J. M.; Thygesen, K. S.; Jacobsen, K. W. Bandgap calculations and trends of organometal halide perovskites. *APL Mater.* **2014**, *2*, 081514.

(37) Pandey, M.; Jacobsen, K. W. Promising quaternary chalcogenides as high-band-gap semiconductors for tandem photoelectrochemical water splitting devices: A computational screening approach. *Phys. Rev. Mater.* **2018**, *2*, 105402.

(38) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-principles-based multiscale modelling of heterogeneous catalysis. *Nat Catal* **2019**, *2*, 659–670.

(39) Nocedal, J.; Wright, S. J. *Numerical optimization*, second edition ed.; Springer series in operation research and financial engineering; Springer: New York, NY, 2006.

(40) Pfrommer, B. G.; Côté, M.; Louie, S. G.; Cohen, M. L. Relaxation of Crystals with the Quasi-Newton Method. *J. Comput. Phys.* **1997**, *131*, 233–240.

(41) Stewart, J. J. P. Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. *J. Mol. Model.* **2004**, *10*, 155–164.

(42) Janes, R. W.; Palmer, R. A. Practical limitations observed using the AM1, MNDO and MINDO/3 semi-empirical methods for charge calculation and structure optimization in 1,2,4-triazine ring-containing compounds. *J. Mol. Struct.* **1995**, *339*, 95–101.

(43) Shen, Y.; Morozov, S. I.; Luo, K.; An, Q.; Goddard Iii, W. A. Deciphering the Atomistic Mechanism of Si(111)-7 × 7 Surface Reconstruction Using a Machine-Learning Force Field. *J. Am. Chem. Soc.* **2023**, *145*, 20511–20520.

(44) Takayanagi, K.; Tanishiro, Y.; Takahashi, S.; Takahashi, M. Structure analysis of Si(111)-7 × 7 reconstructed surface by transmission electron diffraction. *Surf. Sci.* **1985**, *164*, 367–392.