Equivariant analytical mapping of first principles Hamiltonians to accurate and transferable materials models

Liwei Zhang,[1] Berk Onat,[2] Geneviève Dusson,[3] G. Anand,[4] Reinhard J. Maurer,[5] Christoph Ortner,[1] and James R. Kermode[2, a)]

[1)] *Department of Mathematics, University of Bristish Columbia, 1984 Mathematics Road, Vancouver, BC, Canada, V6T 1Z2*

[2)] *Warwick Center for Predictive Modelling, School of Engineering, University of Warwick, Coventry, CV4 7AL, United Kingdom*

[3)] *Laboratoire de Mathématiques, UMR CNRS 6623, Université Bourgogne Franche-Comté, 16 route de Gray, 25030 Besançon, France*

[4)] *Department of Metallurgy and Materials Engineering, Indian Institute of Engineering Science and Technology-Shibpur, Howrah, WB, India*

[5)] *Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom*

(Dated: 30 November 2021)

We propose a data-driven scheme to construct predictive models of Hamiltonian and overlap matrices in atomic orbital representation from *ab initio* data as a function of local atomic and bond environments. The scheme goes beyond conventional tight binding descriptions as it represents the *ab initio* model to full order, rather than in two-centre or three-centre approximations. We achieve this by introducing an extension to the Atomic Cluster Expansion (ACE) descriptor that represents intraatomic onsite and interatomic offsite blocks of Hamiltonian and overlap matrices that transform equivariantly with respect to the full rotation group in 3 dimensions. The approach produces equivariant analytical maps from first principles data to linear models for the Hamiltonian and overlap matrices. Through an application to FCC and BCC aluminium, we demonstrate that it is possible to train models from a handful of Hamiltonian and overlap matrices computed with density functional theory using the FHIaims code, and apply them to produce accurate predictions for the band structure and density of states in both phases, as well as along the Bain path that connects them.

I. INTRODUCTION

The availability of accurate and highly efficient interatomic potentials is crucial for the atomistic simulation of materials phenomena with intrinsic length and time scales not accessible to first principles electronic structure theory. Examples in materials science include failure processes such as crack propagation[1] and chemical dynamics at reactive surfaces.[2] The advent of machine-learning-based interatomic potentials (MLIPs) has meant that high-fidelity interatomic potentials based on Density Functional Theory (DFT) and beyond have become much more widely available.[3–5] Yet, the effort to generate MLIPs that are both transferable and accurate is still significant and heavily depends on the configurational space spanned by the underlying training data set.[6] Very few MLIPs have been reported that are able to capture different materials phases, surface terminations, and the effects of complex defects on the stability and structure of the material.[5,7,8]

More importantly, MLIPs and conventional interatomic potentials fundamentally neglect explicit electronic degrees of freedom of molecules and materials thereby removing access to the simulation of observables beyond structure and stability, such as electric conductivity and optical response, which depend on the electronic subsystem and electron-phonon coupling. While the ability to predict optical and electronic properties is desirable, the inclusion of electronic degrees of freedom will likely also benefit the transferability of MLIPs.

For decades, semi-empirical and tight-binding (TB) models of electronic structure have sought to combine the efficiency of interatomic potentials with the explicit description of electrons. A plethora of approaches based on 2-centre and 3-centre integral approximations have led to established method frameworks such as the AM1 and PM3 methods,[9,10] the Density Functional Tight-Binding (DFTB) method,[11,12] the Sankey-Niklewski approach as implemented in the FIREBALL code,[13,14] and the xTB approach.[15] Unfortunately, the rigid mathematical form of the integral tabulations in most approaches means that TB parametrizations are limited in accuracy and often do not transfer beyond the materials classes for which they were originally intended.

As ML methods make inroads across a diverse range of molecular simulation workflows,[16] new approaches beyond MLIPs are being pursued that incorporate electronic properties. For molecules, Li et al. have proposed a neural-network-based parametrization pipeline for DFTB,[17] while Stoehr et al. have proposed deep tensor neural networks (DTNNs) to construct beyond-pairwise repulsion potentials.[18] Qiao et al. have shown that the use of symmetry-adapted atomic-orbital features can significantly improve transferability and prediction accuracy of molecular stability.[19]

In the realm of condensed phase materials, the automated construction of tight-binding models from ab initio data has been a topic of great interest as it can

[a)] J.R.Kermode@warwick.ac.uk

benefit high throughput materials screening studies.[20] Most commonly, electronic structure simulations are performed in non-atom-centred basis representations such as the pseudopotential plane wave framework, which is not easily amenible to the construction of TB models. TB Hamiltonians are typically constructed via transformation into a maximally localized Wannier function representation,[21] which provide a compact atom-centred basis representation with local support.[22] It is also possible to fit Slater-Koster parameters directly to DFT bandstructure in a data-driven fashion.Recently, materials simulations in atom-centred orbital representations as provided by, for example, the FHI-aims code[23] are becoming more common, where Wannierization is not necessary and the basis representation provided by the code is directly amenable to machine learning approaches based on local representations of atomic neighbourhoods[6] such as Behler-Parinello symmetry functions,[3,24] the SOAP descriptor,[25] or the Atomic Cluster Expansion.[26,27] Recently, Maurer and coworkers used DTNN representations to predict molecular Hamiltonians from DFT in local atomic orbital and optimized effective minimal basis representations for organic molecules including up to 13 heavy atoms.[28,29] Hedge and Bowen[30] were the first to employ Kernel Ridge Regression with a bispectrum representation[31] for an analytical representation of a minimal basis DFT Hamiltonian for bulk copper and diamond. Very recently, an equivariant parameterisation along similar lines to what we describe here has been proposed for molecular systems.[32]

In this work, we present a completely data-driven approach to analytical model construction based on *ab initio* electronic structure theory. The model is able to faithfully represent electronic structure as a function of atomic configuration and materials composition in nonorthogonal local atomic orbital representation via the Hamiltonian and overlap matrices. This goes beyond conventional TB descriptions as it represents DFT to full order, rather than in two-centre or three-centre approximations. We achieve this by introducing an ACE descriptor to represent intraatomic onsite and interatomic offsite blocks of Hamiltonian and overlap matrices that transform equivariantly with respect to the full rotation group in 3 dimensions. This equivariant descriptor is integrated in an automated data-driven workflow that enables rapid parameterisation of environment dependent TB models directly from DFT data as illustrated in Figure 1. We showcase the capabilities of this approach by predicting the bandstructure of bulk aluminium in different crystal structures.

II. METHODOLOGY

In most electronic structure calculations the ground state of a system is obtained by solving an eigenvalue

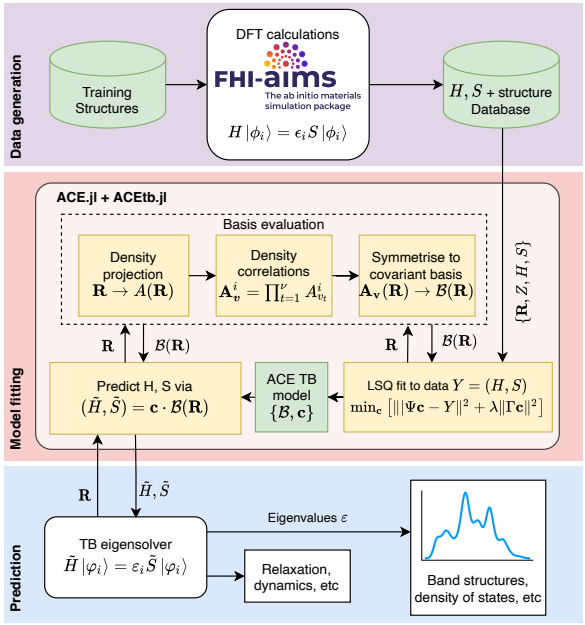

FIG. 1. Schematic of the ACEtb (atomic cluster expansion tight binding) workflow, showing data generation with the FHI-aims density functional theory code, model fitting with the `ACE.jl` and `ACEtb.jl` packages, and prediction.

problem

$$\hat{H}\psi_i = \epsilon_i\psi_i, i = 1, 2, \cdots \tag{1}$$

where

$$\hat{H} = -\tfrac{1}{2}\nabla^2 + V_{\text{eff}}. \tag{2}$$

For example, in the widely used Kohn-Sham DFT model,

$$V_{\text{eff}} = V_{\text{eff}}[\rho], \quad \text{where}$$
$$\rho = \sum_i f_i|\psi_i|^2,$$

and f_i is the occupancy of electronic eigenstate i with wave function ψ_i; i.e., (1) becomes a nonlinear eigenvalue problem, which is extremely computationally demanding and is usually solved employing a Self Consistent Field (SCF) algorithm.[33,34]

In this paper, we are concerned with finding an analytical representation of a self-consistent Hamiltonian operator $\hat{H} = -\frac{1}{2}\nabla^2 + V_{\text{eff}}$ in discrete basis representation.

A. Hamiltonians for extended materials in atomic orbital basis representation

To achieve a finite basis representation, we expand the wave functions ψ_i in a local nonorthogonal atom-centred basis representation

$$\chi_a(\boldsymbol{x}) = R_{nl}(r)Y_{lm}(\theta, \phi) \tag{3}$$

where $a = (n, l, m; I)$ is a composite index, the spatial electron coordinate $\boldsymbol{x}$ and its components r, θ, and ϕ in centrosymmetric coordinates around the atom I are used. Y_{lm} are spherical harmonics that define the angular dependence, and $n = 0, \dots, n_{\max}$, $l = 0, \dots, l_{\max}$, $m = -l_{\max}, \dots, l_{\max}$ characterize the radial and angular nodal structure of the atomic orbital. The choice of $R_{nl}(r)$ varies between different types of atomic orbital basis representations and can involve linear combinations (contractions) of Gaussian functions or numerically tabulated functions. Here we choose the latter as defined in the numeric atom-centred orbital (NAO) basis employed in the FHI-aims code.[23] With this definition, we can express the overlap between basis functions and the interactions as mediated by the Hamiltonian as follows:

$$H_{ab} = \langle \chi_a | \hat{H} | \chi_b \rangle \qquad \text{and} \tag{4}$$
$$S_{ab} = \langle \chi_a | \chi_b \rangle \,. \tag{5}$$

Given a crystal-periodic structure $\mathbf{R} = \{\mathbf{L}_\kappa, \boldsymbol{r}_I, Z_I\}_I$ specified through a set of lattice vectors $\mathbf{L}_{\kappa=1,2,3}$, atom positions $\boldsymbol{r}_I$ and chemical species Z_I, we must consider periodic boundary conditions. As such, a Hamiltonian defined over the whole crystal volume reduces to a block diagonal Hamiltonian where each block corresponds to a vector $\boldsymbol{k}$ in reciprocal space, which can be solved via an independent generalised eigenvalue problem:

$$\mathbf{H}(\boldsymbol{k}) \psi_{i\boldsymbol{k}} = \epsilon_{i\boldsymbol{k}} \mathbf{S}(\boldsymbol{k}) \psi_{i\boldsymbol{k}} \quad i = 1, 2, \dots, \tag{6}$$

where $\psi_{i\boldsymbol{k}}$ are Bloch wave functions and $\mathbf{H}(\boldsymbol{k})$ and $\mathbf{S}(\boldsymbol{k})$ are Hamiltonian and overlap matrices defined in terms of a discrete crystal-periodic basis. In appendix A, we show how $\mathbf{H}(\boldsymbol{k})$ and $\mathbf{S}(\boldsymbol{k})$ can be constructed at arbitrary points $\boldsymbol{k}$ in reciprocal space from real-space representations of Hamiltonian and overlap matrices that span the full crystal volume (typically considered within a certain radius around the central unit cell). As the $\boldsymbol{k}$-dependent matrices and the solution of the set of generalised eigenvalues completely follow from the real-space $\boldsymbol{H}$ and $\boldsymbol{S}$ in (4) and (5), we will go on to develop a representation for those two matrix quantities as a function of the structure $\mathbf{R}$.

Recall that $\hat{H} = -\frac{1}{2}\nabla^2 + V_{\text{eff}}$. The effective potential V_{eff} is not only a function of the spatial electron coordinate $\boldsymbol{x}$ but also of the entire atomic structure, i.e., one should think of

$$V_{\text{eff}} = V_{\text{eff}}(\boldsymbol{x}; \mathbf{R}).$$

For example, in KS-DFT, this dependence arises due to the dependence of V_{eff} on the self-consistent electron density. Our aim will be to construct a general regression scheme for the discretised Hamiltonian exploiting three fundamental, general properties of $\hat{H}$ and in particular V_{eff}: (i) near-sightedness of electronic structure; (ii) smoothness under changes in the atomic structure; and (iii) equivariance of the Hamiltonian. We will discuss in the next section how these properties are to be exploited in the parameterisation.

In preparation, we first make (iii) more precise: let $Q \in \mathrm{O}(3)$ denote an isometry (rotation and reflection) and $Q\mathbf{R} = \{\mathbf{L}_\kappa, Q\boldsymbol{r}_I, Z_I\}_I$ (where we also rotate the cell). Further, let $\boldsymbol{H}_{IJ} = \boldsymbol{H}_{IJ}(\mathbf{R})$ denote the Hamiltonian block corresponding to interactions between orbitals centered at sites I and J. It is then straightforward to deduce that

$$\boldsymbol{H}_{IJ}(Q\mathbf{R}) = D(Q)^* \boldsymbol{H}_{IJ}(\mathbf{R}) D(Q), \tag{7}$$

where $D(Q)$ is a block-Wigner-D matrix,

$$D(Q) = \mathrm{Diag}(D^{l_1}(Q), D^{l_2}(Q), \cdots),$$

and $(l_1, l_2, \dots)$ specify the types of orbitals at each site. More details can be found in Appendix B.

Crucially, there are only two distinct functional relationships that must be "learned" in order to represent the entire Hamiltonian: one for off-site blocks that represent interactions between orbitals centered at two different atoms and one for on-site blocks representing interactions of orbitals at the same atom. More precisely, the translation invariance and permutation equivariance of the Hamiltonian imply that

$$\begin{aligned} \boldsymbol{H}_{II} &= \boldsymbol{H}_{\text{on}}(\mathbf{R}_I), \qquad \text{and} \\ \boldsymbol{H}_{IJ} &= \boldsymbol{H}_{\text{off}}(\boldsymbol{r}_{IJ}, \mathbf{R}_{IJ}), \end{aligned} \tag{8}$$

where $\boldsymbol{r}_{IJ} = \boldsymbol{r}_I - \boldsymbol{r}_J$, $\mathbf{R}_I$ denotes the *atomic environment* of atom I and $\mathbf{R}_{IJ}$ the *bond environment* of the (multiple) bonds between the two atoms i, j. These environments are defined as follows:

$$\begin{aligned} \mathbf{R}_I &:= \big\{ \boldsymbol{r}_{IK} \,|\, K \neq I \big\}, \qquad \text{and} \\ \mathbf{R}_{IJ} &:= \big\{ \boldsymbol{r}_K - \tfrac{1}{2}(\boldsymbol{r}_I + \boldsymbol{r}_J) \,|\, K \neq I, J \big\}. \end{aligned}$$

In the above definitions, the index K runs over all unit cells N within the crystal volume. According to (7) the functions $\boldsymbol{H}_{\text{on}}$ and $\boldsymbol{H}_{\text{off}}$ are equivariant in the sense that

$$\boldsymbol{H}_{\text{on/off}}(Q\mathbf{R}) = D(Q)^* \boldsymbol{H}_{\text{on/off}}(\mathbf{R}) D(Q). \tag{9}$$

Translation invariance is now built into the dependence of $\boldsymbol{H}_{\text{on/off}}$ on relative positions only, while permutation equivariance of $\boldsymbol{H}$ is built into (8).

Several simplifications apply for the treatment of the overlap matrix. For each atom we choose a set of basis functions χ that are orthogonal, which means that the on-site blocks $\boldsymbol{S}_{II}$ are diagonal matrices. The off-site blocks follow the same symmetry as the Hamiltonian off-site blocks.

B. Parameterisation

We parameterise the real-space Hamiltonian and overlap matrix blocks $\boldsymbol{H}_{\text{on}}, \boldsymbol{H}_{\text{off}}$ and $\boldsymbol{S}_{\text{off}}$ using an equivariant ACE basis[26,27,35]. Similar techniques have previously been proposed in other contexts[32,36,37]. In this section,

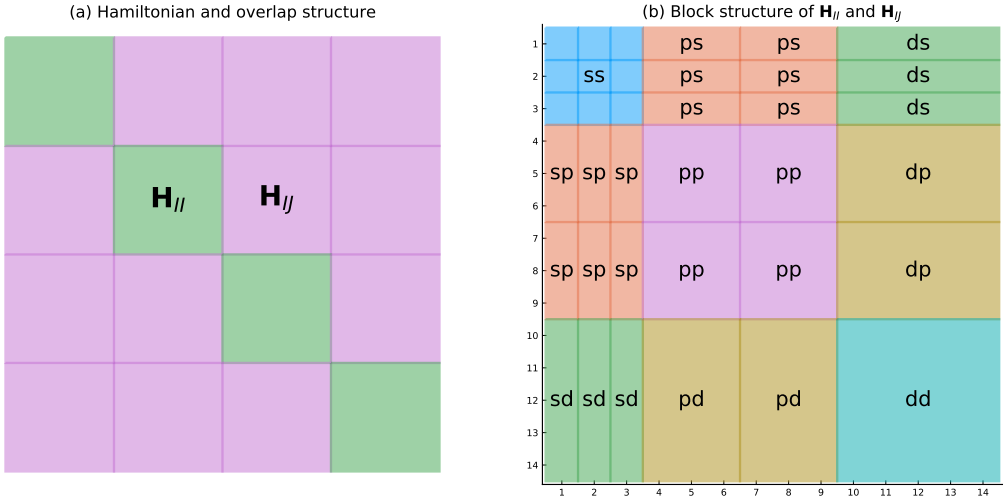

FIG. 2. Block structure and atomic orbital subblocks in the Hamiltonian and overlap matrices used in our models. Each block within panel (a) is a 14×14 matrix with the atomic orbital structure $\boldsymbol{H}_{IJ}$ shown in panel (b). Blocks coloured green in (a) are onsite blocks, while those shown in purple are offsite blocks. Note that the onsite $\boldsymbol{H}_{II}$ are self-adjoint and hence, e.g., only one of the *ps* and *sp* blocks needs to be fitted.

we present a general outline of the ideas, making certain choices of approximation parameters concrete in §II D.

We denote the parameterised Hamiltonian and overlap by $\tilde{\boldsymbol{H}}, \tilde{\boldsymbol{S}}$. For the sake of simplicity we focus the presentation on $\tilde{\boldsymbol{H}}$ and remark on the relevant modification for $\tilde{\boldsymbol{S}}$ at the end. Since the focus of the present work is on elemental metallic systems we ignore chemical species information entirely. All procedures are straightforward to generalise for multiple species with the only effect being an increased number of $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{S}}$ blocks that have to be considered as element combinations increase. In the present case, $\tilde{\boldsymbol{H}}_{\rm on}$ is invariant under permutations of $\mathbf{R}_I$ and $\tilde{\boldsymbol{H}}_{\rm off}$ is invariant under permutations of $\mathbf{R}_{IJ}$. Both can therefore be parameterised by the ACE model. Here, we closely follow the procedures introduced by Dusson *et al.*[27], Drautz[35], Lysogorskiy *et al.*[38].

1. Parameterisation of $\boldsymbol{H}_{\rm on}$: We start by choosing a *one-particle* basis,

$$\phi_v(\boldsymbol{x}) := \phi^{\rm on}_{nlm}(\boldsymbol{x}) := P_{nl}(r) Y_l^m(\hat{\boldsymbol{x}}) f_{\rm cut}(r) \qquad (10)$$

where we have identified the composite index $v \equiv (nlm)$. The radial cutoff or envelope function $f_{\rm cut}(r)$ ensures that only interactions of nearby atoms are taken into account, exploiting the near-sightedness of electronic structure.

Given the one-particle basis we can form the density projection and projected ν-correlations (product basis),

$$A^I_v := \sum_{J \neq I} \phi_v(\boldsymbol{r}_{IJ}), \qquad \text{and}$$
$$\boldsymbol{A}^I_{\boldsymbol{v}} := \prod_{t=1}^{\nu} A^I_{v_t} \qquad \text{for } \boldsymbol{v} = (v^1, \dots, v^\nu), \nu = 1, 2, \dots.$$

The $\boldsymbol{A}^I_{\boldsymbol{v}}$ form a complete basis of permutation-invariant polynomials, hence we can approximate

$$\boldsymbol{H}_{II} = \boldsymbol{H}_{\rm on}(\mathbf{R}_I) \approx \tilde{\boldsymbol{H}}^{\rm PI}_{\rm on}(\mathbf{R}_I) = \sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^I_{\boldsymbol{v}}, \qquad (11)$$

where $\boldsymbol{A}^I_{\boldsymbol{v}}$ are scalar and the parameters $C_{\boldsymbol{v}} = (C^{\alpha_1\alpha_2}_{\boldsymbol{v}})^{N_{\rm orb}}_{\alpha_1,\alpha_2=1}$ have the same dimensionality as $\boldsymbol{H}_{II}$ i.e., $N_{\rm orb} \times N_{\rm orb}$ (recall that $\boldsymbol{H}_{II}$ denotes the onsite Hamiltonian block corresponding to orbitals centered at atom I). The summation over $\boldsymbol{v}$ will be restricted to a finite set, the choice of which is a crucial aspect of the model accuracy; cf. § II D.

The expansion (11) incorporates translation and permutation invariance but not yet the O(3)-equivariance (7). Following the general ACE construction[27] we can achieve this by simply averaging the representation over the group O(3), i.e.,

$$\tilde{\boldsymbol{H}}_{\rm on}(\mathbf{R}_I) = \fint_{\rm O(3)} D(Q) \tilde{\boldsymbol{H}}^{\rm PI}_{\rm on}(Q\mathbf{R}_I) D(Q)^* dQ,$$

In step 4. we will review how this integration is explicitly resolved.

2. Parameterisation of $\boldsymbol{H}_{\rm off}$: The procedure for parameterising $\boldsymbol{H}_{\rm off}$ is similar to that of $\boldsymbol{H}_{\rm on}$, the main difference being that the presence of a bond rather than a site changes the permutation-invariance. Specifically, we now need to define one-particle basis functions for the bond variable and for the environment variables

$$\phi^{\rm b}_{nlm}(\boldsymbol{r}_{IJ}) = P^{\rm b}_{nl}(r_{IJ}) Y^m_l(\hat{\boldsymbol{r}}_{IJ}) f^{\rm b}_{\rm cut}(r_{IJ}),$$
$$\phi^{\rm e}_{nlm}(\boldsymbol{r}_{IJ,K}) = P^{\rm e}_{nl}(r_{IJ,K}) Y^m_l(\hat{\boldsymbol{r}}_{IJ,K}) f^{\rm e}_{\rm cut}(\boldsymbol{r}_{IJ,K}, \boldsymbol{r}_{IJ}). \qquad (12)$$

where $\boldsymbol{r}_{IJ} = r_{IJ}\hat{\boldsymbol{r}}_{IJ}$ and $\boldsymbol{r}_{IJ,K} := \boldsymbol{r}_K - \frac{1}{2}(\boldsymbol{r}_I + \boldsymbol{r}_J)$. Note in particular that the cutoff function for the environment,

$f^{\rm e}_{\rm cut}$, no longer depends only on the radius but may be more general: we require only that $f^{\rm e}_{\rm cut}(\boldsymbol{r}_{IJ,K}, \boldsymbol{r}_{IJ})$ is invariant under joint rotation of both arguments which allows, e.g., ellipsoidal or cylindrical cutoff geometries.

The density projection for the bond environment $\mathbf{R}_{IJ}$ is now given by

$$A^{IJ}_v := \sum_{K\neq I,J} \phi^{\rm e}_v(\boldsymbol{r}_{IJ,K}),$$

and the product basis becomes

$$\boldsymbol{A}^{IJ}_{\boldsymbol{v}} := \phi^{\rm b}_{v^0}(\boldsymbol{r}_{IJ}) \cdot \prod_{t=1}^{\nu} A^{IJ}_{v^t},$$

for $\boldsymbol{v} = (v^0, v^1, \dots, v^\nu)$, with $\nu = 0, 1, 2, \dots$ the correlation order of the bond environment. As in the on-site case, the $A^{IJ}_{\boldsymbol{v}}$ form a complete basis of polynomials that are invariant under permutations of $\mathbf{R}_{IJ}$ and we may therefore approximate

$$\boldsymbol{H}_{IJ} = \boldsymbol{H}_{\rm off} \approx \tilde{\boldsymbol{H}}^{\rm PI}_{\rm off}(\boldsymbol{r}_{IJ}, \mathbf{R}_{IJ}) := \sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^{IJ}_{\boldsymbol{v}}. \qquad (13)$$

which we finally symmetrize to obtain also the $O(3)$-equivariance,

$$\tilde{\boldsymbol{H}}_{\rm off}(\boldsymbol{r}_{IJ}, \mathbf{R}_{IJ}) := \fint_{\rm O(3)} D(Q) \tilde{\boldsymbol{H}}^{\rm PI}_{\rm off}(Q\boldsymbol{r}_{IJ}, Q\mathbf{R}_{IJ}) D(Q)^* dQ. \qquad (14)$$

3. Parameterisation of $\boldsymbol{S}_{\rm off}$: The environment-dependence of $\boldsymbol{H}_{\rm off}$ enters only through the effective potential $V_{\rm eff}$ which is not present in the overlap matrix definition. Therefore, we simply parameterise $\boldsymbol{S}_{\rm off}$ by

$$\tilde{\boldsymbol{S}}_{\rm off}(\boldsymbol{r}_{IJ}) := \fint_{\rm O(3)} D(Q) \Big[\sum_v C_v \phi^{\rm b}_v(Q\boldsymbol{r}_{IJ}) \Big] D(Q)^* \, dQ. \qquad (15)$$

This is formally equivalent to a Slater Koster representation of 2-centre integrals,[39] which is exact in the case of the overlap. For our ACE parameterisation, this means that we only need to use correlation order $\nu = 0$, i.e. no environment-dependence of the bond integral needs to be considered.

4. Recursive symmetrisation: In all three cases $\tilde{\boldsymbol{H}}_{\rm on}, \tilde{\boldsymbol{H}}_{\rm off}, \tilde{\boldsymbol{S}}_{\rm off}$ we have reduced the parameterisation to an integral over the symmetry group $O(3)$, i.e.,

$$\tilde{\boldsymbol{K}}(\mathbf{R}_\bullet) = \fint_{\rm O(3)} D(Q) \Big[\sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^\bullet_{\boldsymbol{v}}(Q\mathbf{R}_\bullet) \Big] D(Q)^*, \qquad (16)$$

where $\tilde{\boldsymbol{K}}$ denotes one of the three model components $\tilde{\boldsymbol{H}}_{\rm on}, \tilde{\boldsymbol{H}}_{\rm off}, \tilde{\boldsymbol{S}}_{\rm off}$ and $\mathbf{R}_\bullet$ denotes an atom environment $\mathbf{R}_I$ or bond environment $\mathbf{R}_{IJ}$. In particular, for off-site overlap $\boldsymbol{S}_{\rm off}$,

$$\boldsymbol{A}^{IJ}_{\boldsymbol{v}}(\mathbf{R}_{IJ}) = \phi^{\rm b}_v(Q\boldsymbol{r}_{IJ}).$$

Since the angular dependence of the one-particle basis functions in all cases is in terms of spherical harmonics Y^m_l we can deduce that

$$\boldsymbol{A}^\bullet_{\boldsymbol{nlm}}(Q\mathbf{R}_\bullet) = \sum_{\boldsymbol{\mu}} \boldsymbol{D}^{\boldsymbol{l}}_{\boldsymbol{\mu m}}(Q) \boldsymbol{A}^\bullet_{\boldsymbol{nl\mu}}(\mathbf{R}_\bullet),$$

where $\boldsymbol{D}^{\boldsymbol{l}}_{\boldsymbol{\mu m}}(Q) = \prod_t D^{l_t}_{\mu_t m_t}(Q)$. Furthermore, we write

$$C_{\boldsymbol{v}} = \sum_{\alpha,\beta=1}^{N_{\rm orb}} c^{\alpha\beta}_{\boldsymbol{v}} E^{\alpha\beta},$$

where $E^{\alpha\beta} \in \mathbb{R}^{N_{\rm orb}\times N_{\rm orb}}$ with $E^{\alpha\beta}_{\alpha'\beta'} = \delta_{\alpha\alpha'}\delta_{\beta\beta'}$. Inserting these two identities into (16) yields

$$\begin{aligned} \tilde{\boldsymbol{K}}(\mathbf{R}_\bullet) &= \sum_{\boldsymbol{n},\boldsymbol{l},\boldsymbol{m},\alpha,\beta} c^{\alpha\beta}_{\boldsymbol{v}} \sum_{\boldsymbol{\mu}} \mathcal{U}^{\alpha\beta}_{\boldsymbol{l\mu m}} \boldsymbol{A}^\bullet_{\boldsymbol{nl\mu}}(\mathbf{R}_\bullet) \\ &=: \sum_{\boldsymbol{n},\boldsymbol{l},\boldsymbol{m},\alpha,\beta} c^{\alpha\beta}_{\boldsymbol{nlm}} \mathcal{B}^{\alpha\beta}_{\boldsymbol{nlm}}(\mathbf{R}_\bullet), \end{aligned} \qquad (17)$$

where the "generalized coupling coefficients" are given by

$$\mathcal{U}^{\alpha\beta}_{\boldsymbol{l\mu}m} = \int_{\rm O(3)} \boldsymbol{D}^{\mathbf{l}}_{\boldsymbol{\mu}\mathbf{m}}(Q) D(Q)^* E^{\alpha\beta} D(Q) dQ.$$

Their definition involves an integral over products of Wigner-D matrices which can be precomputed explicitly (i.e., without need for quadrature which would incur a discretisation error) using the recursion proposed by Dusson *et al.*[27] and independently by Nigam, Willatt, and Ceriotti[32].

Note that (17) parameterises $\tilde{K}$ in terms of the scalar parameters $c^{\alpha\beta}_{\boldsymbol{v}}$, while the basis functions are now matrix-valued,

$$\mathcal{B}^{\alpha\beta}_{\boldsymbol{nlm}}(\mathbf{R}_\bullet) = \sum_{\boldsymbol{\mu}} \mathcal{U}^{\alpha\beta}_{\mathbf{l\mu m}} \boldsymbol{A}^\bullet_{\boldsymbol{nl\mu}}(\mathbf{R}_\bullet).$$

Since the coupling coefficients $\mathcal{U}$ are extremely sparse, the operation to obtain $\mathcal{B}$ from $\boldsymbol{A}^\bullet$ is relatively cheap.

Due to the coupling, the basis $\mathcal{B}^{\alpha\beta}_{\boldsymbol{nlm}}$ is normally over-complete. This linear dependence arises exactly within fixed $\boldsymbol{nl}$ blocks. In a straightforward adaption of the general procedures outlined by Dusson *et al.*[27] we use elementary linear algebra techniques to reduce the basis in a block-by-block fashion by constructing reduced coupling coefficients $\mathcal{U}^{\boldsymbol{nl}}_{k\boldsymbol{\mu}}$ and defining

$$\mathcal{B}_{\boldsymbol{nl}k}(\mathbf{R}_\bullet) := \sum_{\boldsymbol{\mu}} \mathcal{U}^{\boldsymbol{nl}}_{k\boldsymbol{\mu}} \boldsymbol{A}^\bullet_{\boldsymbol{nl\mu}}(\mathbf{R}_\bullet). \qquad (18)$$

In summary, after dropping the detailed multi-index notation and replacing it with a simple enumeration of the basis, we obtain *linear models* for

$$\begin{aligned} \tilde{\boldsymbol{H}}_{\rm on} &:= \mathbf{c}^{\rm on} \cdot \mathcal{B}^{\rm on}, \\ \tilde{\boldsymbol{H}}_{\rm off} &:= \mathbf{c}^{\rm off} \cdot \mathcal{B}^{\rm off}, \\ \tilde{\boldsymbol{S}}_{\rm off} &:= \mathbf{c}^{\rm S} \cdot \mathcal{B}^{\rm S}, \end{aligned}$$

all of which inherit exactly the translation and permutation invariance as well as O(3)-equivariance of $\boldsymbol{H}_{\rm on}, \boldsymbol{H}_{\rm off}, \boldsymbol{S}_{\rm off}$. In the limit of infinite basis size and infinite cutoff radius these models can (in principle) be converged to within arbitrary accuracy. In this sense, they are *universal*.

C. Data generation

The datasets used in this work are constructed for face-centred cubic (FCC) and body-centred cubic (BCC) phases of Al. Our data was generated through electronic structure calculations with the all-electron numeric atomic orbital code FHI-aims (version 190530).[23] We used the Perdew-Burke-Enzerhof (PBE) generalized gradient approximation[40] to the exchange-correlation within the Kohn-Sham formulation of density functional theory and neglect spin in our treatment. The convergence criteria for charge density, sum of eigenvalues, and total energy of the self-consistent cycles were set to 10^{-5} e/a$_0^3$, $5{\times}10^{-5}$ eV, and 10^{-6} eV, respectively. The default *tight* FHI-aims basis set and integration grid definitions were used, which use a basis set confinement with a maximum radial basis function extent of 6 Å. We modify the set of atomic basis functions that we employ to achieve optimal computational efficiency. Systematic convergence tests showed that band energies are converged up to 10 eV above the Fermi level when using a minimal basis plus a single d orbital from Tier 1. Therefore, we used a basis set comprising s and p orbitals of the minimal basis set plus one d orbital from the Tier 1 setting, yielding the 14 atomic basis functions for Al illustrated in Fig. 2(b).

The optimal equilibrium lattice constants for FCC and BCC Al were determined in primitive cells with a $9{\times}9{\times}9$ Monkhorst-Pack k-point mesh[41] to be 4.05 Å and 3.29 Å respectively. To sample a variety of distorted atomic configurations for Al, we carried out molecular dynamics simulations at a temperature of 500 K using $9 \times 9 \times 9$ supercells of the primitive FCC and BCC unit cells. 500 molecular dynamics simulations in the NPT ensemble were performed for each phase using a 5 fs timestep using the Embedded Atom Method (EAM) potential proposed by Zhou *et. al.*[42]. Single point DFT total energy calculations were carried out on the final configurations of each of these 500 MD simulations using FHI-aims with the parameters described above and a single k-point at Γ. We stored the resulting H and S matrices giving a dataset

$$\begin{aligned}&\big\{(H_{II}, \mathbf{R}_I)\big\},\\ &\big\{(H_{IJ}, \mathbf{R}_{IJ})\big\},\\ &\big\{(S_{IJ}, \mathbf{R}_{IJ})\big\}.\end{aligned}$$

where II, IJ indicate on- and off-site blocks of the Hamiltonian and overlap matrices while $\mathbf{R}$ are the corresponding atomic structure data as defined at II A. For the onsite part for the Hamiltonian, we used 730 training and 728 test blocks and for the offsite part 2000 training and 1369 test blocks.

D. Parameter Estimation

We have defined three linear models for equivariant components of Hamiltonian and overlap matrices (up to the choice of approximation parameters). It remains to specify a parameter estimation procedure to determine the model parameters which typically number in the thousands to tens of thousands. There are essentially two choices we can make: (i) fit the models to observed properties such as band structure, energies, forces; or (ii) fit the models directly to match a reference Hamiltonian. Both approaches have advantages and disadvantages. We have chosen to follow route (ii) which is particularly attractive from both theoretical and numerical perspectives as it results in a linear least squares problem.

Let $\tilde{\boldsymbol{K}} = \boldsymbol{c} \cdot \mathcal{B}$ be one of the three linear models, and $\{(\boldsymbol{K}_*^{(\tau)}, \mathbf{R}_\bullet^{(\tau)}\}_\tau$ the corresponding training set, then we set up the loss function

$$L_0(\boldsymbol{c}) = \sum_\tau \big|\boldsymbol{K}_*^{(\tau)} - \tilde{\boldsymbol{K}}(\mathbf{R}_\bullet^{(\tau)})\big|^2.$$

Since $\tilde{\boldsymbol{K}}$ is linear in $\boldsymbol{c}$ it follows that L can be rewritten as

$$L_0(\boldsymbol{c}) = \big\|\Psi\boldsymbol{c} - \boldsymbol{y}\big\|^2,$$

where Ψ is the design matrix and $\boldsymbol{y}$ contains the reference model values. To prevent overfitting, we regularise the least squares system with a generalised Tychonov term,

$$L_\lambda(\boldsymbol{c}) := \big\|\Psi\boldsymbol{c} - \boldsymbol{y}\big\|^2 + \lambda\big\|\Gamma\boldsymbol{c}\big\|^2, \tag{19}$$

where $\Gamma = \mathrm{diag}(\Gamma_{kk})$ with Γ_{kk} an estimate for the curvature of the kth basis function which enforces smoothness of the model[27,38] and λ is a regularisation parameter. Throughout this work, we define Γ_{kk} by

$$\Gamma_{kk} = \begin{cases}\sum_\nu (n_\nu^2 + l_\nu^2 + m_\nu^2), & \text{onsite}\\ (n_0^2 + l_0^2 + m_0^2) + \sum_\nu (n_\nu^2 + l_\nu^2 + m_\nu^2), & \text{offsite}\end{cases}$$

with λ always set to be 10^{-7}. We then solve the regularised least squares system (19) using an iterative LSQR algorithm with termination tolerance 10^{-6}.

For the radial basis set P_{nl} we used

$$\begin{aligned}\xi(r) &= \Big(\frac{1 + r_0}{1 + r}\Big)^2\\ P_{nl}(r) &= Q_n(\xi(r))\end{aligned}$$

where Q_n is a polynomial of degree n such that $\int_{\xi_0}^{\xi_1} Q_n(\xi) Q_{n'}(\xi) d\xi = \delta_{nn'}$ and $[\xi_0, \xi_1] = \xi([0, r_{\rm cut}])$; see Dusson *et al.*[27] for full details.

The envelope function for both on-site term and off-site environment basis function is defined as

$$f_{\rm cut}(r; r_{\rm cut}) = f^{\rm b}_{\rm cut}(r; r_{\rm cut}) = \begin{cases} (r^2/r_{\rm cut}^2 - 1)^2, & r \le r_{\rm cut}, \\ 0, & r > r_{\rm cut}, \end{cases}$$

and that for offsite environment is given by a bond-related cylindrical cutoff function

$$\begin{aligned} & f^{\rm e}_{\rm cut}(z, r; z_{\rm cut}, r_{\rm cut}) \\ =& \begin{cases} \left(\frac{r^2}{r_{\rm cut}^2} - 1\right)^2 & \left(\frac{z^2}{(z_{\rm cut} + l_{\rm bond}/2)^2} - 1\right)^2, \\ & r \le r_{\rm cut}, |z| \le z_{\rm cut} + l_{\rm bond}/2, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

where (z, r, θ) are the cylindrical coordinates of an environment atom (though θ is not used in this definition) and $l_{\rm bond}$ is the length of the corresponding bond. Note that both $f_{\rm cut}$ and $f^{\rm b}_{\rm cut}$ are rotation invariant, they will not influence the equivariance of the basis at all. Meanwhile, though the cylindrical curoff function $f^{\rm e}_{\rm cut}$ is bond-dependent, it can be easily checked that it do no harm to rotation symmetry as well.

In our implementation, the on-site cutoff $r_{\rm cut}$ is chosen to be 9.0 Å for $\boldsymbol{H}_{\rm on}$ and the off-site bond cutoff is set to be 10.0 Å. We set $r^{\rm e}_{\rm cut} = z^{\rm e}_{\rm cut} = 5.0$ Å for the off-site environment.

As noted above, we used correlation order $\nu = 0$ for the offsite overlap $\boldsymbol{S}_{II}$ since these blocks are not environment dependent. For $\boldsymbol{H}_{II}$ we used correlation order $\nu = 2$ throughout, while for $\boldsymbol{H}_{IJ}$ we tested correlation orders of both $\nu = 1$ and $\nu = 2$. The maximum polynomial degree was chosen on a case-by-case basis to control the balance between accuracy and transferability through a cross-validation procedure as discussed in more details in § III below.

E. Prediction

The software implementation of our method follows the workflow illustrated in Figure 1. The Julia packages `ACE.jl`[43] and `ACEtb.jl` implement the general Atomic Cluster Expansion basis sets and the specialisation to fitting and predicting Hamiltonians, respectively. We predict the real space matrices $\boldsymbol{H}(\boldsymbol{R})$ and $\boldsymbol{S}(\boldsymbol{R})$

Given an input configuration $\boldsymbol{R}$ we use the scheme described above to predict $\tilde{\boldsymbol{H}}_{on}(\boldsymbol{R}), \tilde{\boldsymbol{H}}_{\rm off}(\boldsymbol{R})$ and $\tilde{\boldsymbol{S}}_{\rm off}(\boldsymbol{R})$. We can assemble complete approximate Cartesian Hamiltonian and overlap matrices $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{S}}$ from the predicted blocks. We can construct k-dependent variants and associated bandstructures via a standalone Julia implementation contained within the `ACEtb.jl` pacakge.

Using either the reference or the predicted matrices we can solve the generalised eigenproblems of the form

$$\boldsymbol{H}(\mathbf{k})\phi_i = \epsilon_i \boldsymbol{S}(\mathbf{k})\phi_i \tag{20}$$

$$\tilde{\boldsymbol{H}}(\mathbf{k})\tilde{\phi}_i = \tilde{\epsilon}_i \tilde{\boldsymbol{S}}(\mathbf{k})\tilde{\phi}_i \tag{21}$$

to obtain k-dependent band energies $\epsilon_i, \tilde{\epsilon}_i$ and orbitals (eigenfunctions) $\phi_i, \tilde{\phi}_i$ for the reference and predicted systems respectively, where $i = 1, \ldots, N_{\rm orb}$ and in this work $N_{\rm orb} = 14$. Band structures, density of states (DoS) and other derived quantities can be computed by post-processing the band energies following standard practices.

III. RESULTS

A. Validation

The ACE basis sets need to be carefully chosen for a particular application. The larger the basis, the higher the achievable accuracy, but larger basis sets also carry a risk of loss of transferability through overfitting. For a given orbital angular momentum l, each basis set is defined by two parameters: the correlation order ν and the maximum polynomial degree $n_{\rm max}$ used in the radial basis functions $P_{nl}(r)$ of (10) and (12). For the onsite models, the body order is one more than the correlation order, i.e. $\nu = 1$ corresponds to two body and $\nu = 2$ to three body, while for the offsite models the body order is two more than the correlation order (since each term in the body order expansion depends on the bond in addition to ν particles from the environment). The offsite model has further flexibility in that one can choose different $n_{\rm max}$ for bond and environment, say, $n^{\rm b}_{\rm max}$ and $n^{\rm e}_{\rm max}$. To avoid over emphasising the impact of environment, we set $n^{\rm e}_{\rm max} = \lceil n^{\rm b}_{\rm max}/2 \rceil$ in our implementation.

We tested the accuracy of the fitted Hamiltonian and overlap matrices using different choices of these basis set parameters. The results are illustrated in Fig.3. For the onsite blocks $\boldsymbol{H}_{II}$, we can obtain accurate and transferable results for all sub-blocks with correlation order $\nu = 2$ (body order 3), with no significant overfitting as can be seen from the close agreement of prediction accuracies on the training and test datasets in Fig. 3(a). The largest errors are on the *dd* subblock, which also has the largest matrix entries; the RMSE of ~ 10 meV on this sub-block corresponds to a $\sim$2% relative error.

For offsite blocks $\boldsymbol{H}_{IJ}$ we considered models with correlation orders of both $\nu = 1$ (body order 3), Fig. **??**(b), and $\nu = 2$ (body order 4), Fig. 3(c). Both approaches show good convergence in the accuracy on the training set as the maximum degree is increased. However, for sublocks that include interaction with s orbitals, we observe that overfitting occurs at lower degrees for the order 2 models than for the order 1 case. We speculate that this might result from the higher order basis sets providing too much flexibility for functions that have relatively simple functional behaviour. Since s orbitals have no intrinsic rotational dependence, all rotational equivariance behaviour in *sp* and *sd* sublocks comes from how the p or d orbitals are positioned with respect to the environment.

We find the correlation order 1 models provide sufficient accuracy, in fact closely comparable to that of the

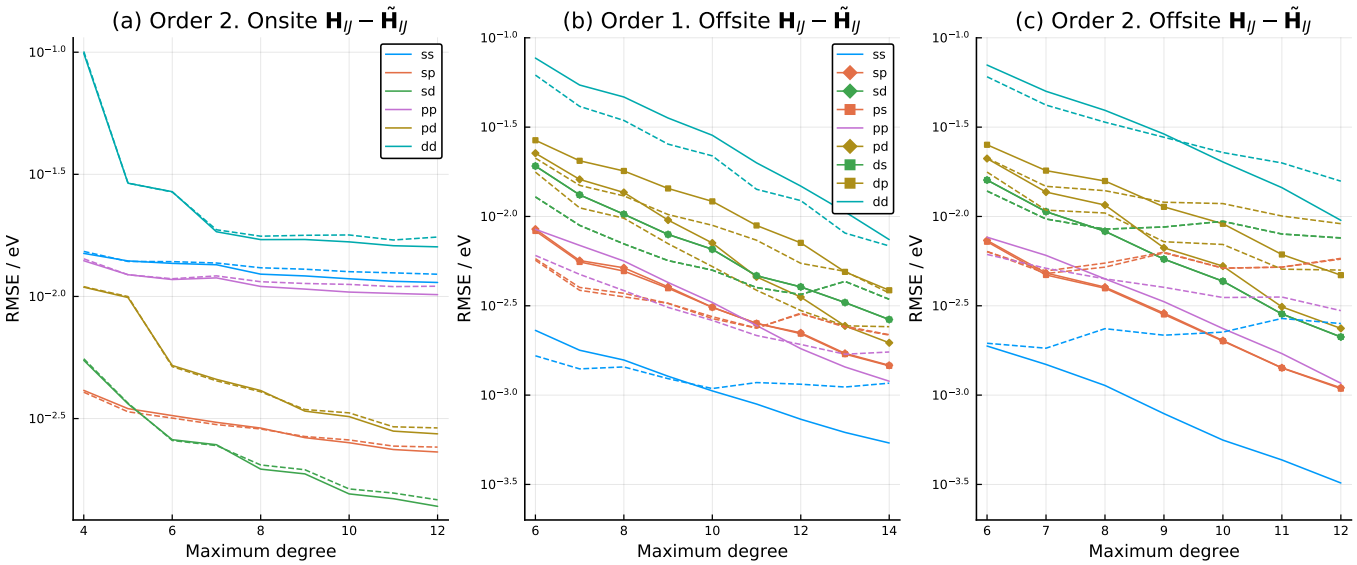

FIG. 3. Convergence of Hamiltonian and overlap blocks with respect to the order and maximum degree of the ACEtb basis set. (a) Onsite Hamiltonian blocks $\boldsymbol{H}_{II}$ fitted with order 2 models of varying maximum degree. (b) Offsite Hamiltonian blocks with order 1 ACE models.(c) Offsite Hamiltonian blocks with order 2 ACE models. In all plots solid lines show errors on training data and dashed lines errors on test data. Colours match the block structure of Fig. 2. Note the distinct markers that distinguish the non-adjoint entries in the offsite Hamiltonian and overlap blocks.

order 2 models on the training set, so to avoid issues of overfitting we use order 1 only for $\boldsymbol{H}_{IJ}$, and also to limit the maximum polynomial degree for individual sub-blocks as discussed in more detail in § III B below.

As expected from the lack of environment dependence, the offsite overlap $\boldsymbol{S}_{IJ}$ is very well reproduced at correlation order 0 (body order 2), with a RMSE of 10^{-4}. We do not observe any over-fitting for the offsite overlap so we fixed the maximum polynomial degree for $\boldsymbol{S}_{IJ}$ at 16, the highest value we tried.

B. Cross-validation and Model Selection

To eliminate overfitting we used the cross-validation results illustrated in Fig. 3 to select a customised basis set for each sub-block, as set out in Table I. Note that the maximum polynomial degree can be chosen for each individual sub-block model shown in the schematic in Fig. 1, i.e. there are 9 ss models, $3 \times 2 = 6$ sp models, and $2 \times 2 = 4$ pp models. For the 9 ss sub-blocks of the offsite Hamiltonian we found it necessary to reduce the degree only for the $3s - 3s$ entry, which arises from the fact that the FHI-aims basis set features two s orbitals in the valence shell of Al.

We used our optimised model to predict the Hamiltonian and overlap for the FCC and BCC equilibrium crystal geometries. These were not included in the training set, which comprises only perturbed structures from molecular dynamics, so can be viewed as a test of its transferability. The magnitudes and associated errors in the onsite and one of the nearest-neighbour offsite blocks of the Hamiltonian matrix are illustrated in Fig. 4 for the FCC case; BCC results are of comparable accuracy. These results demonstrate the correct equivariance of the predictions with matrix entries, i.e. entries which should be zero by symmetry being correctly captured. Comparing the upper and lower panels also illustrates that the errors are always orders of magnitude smaller than the corresponding magnitudes, ensuring that the relative error is well controlled (typically $\sim$1% or less).

C. Prediction of band structures and DoS

So far we have assessed only errors made on the quantities used in fitting the models, i.e. the Hamiltonian and overlap matrix elements. While it is reassuring that these are accurately captured, a stronger test of the predictive power of our new formulation is to use it to predict electronic observables such as the band structure and DoS. Figure 5 compares predictions of these quantities for FCC and BCC aluminium with those computed from the reference FHI-aims Hamiltonian and overlap matrices. There is excellent agreement for all occupied bands, and also bands within 10 eV of the Fermi level (which is itself in close agreement between the reference and predicted systems). The DoS was integrated on a dense $9 \times 9 \times 9$ k-point mesh and also shows excellent agreement for the occupied states for both FCC and BCC, with significant errors only arising well above the Fermi level, giving confidence in the ability of our model to be predict electronic observables.

The figure also shows confidence intervals for the pre-

Onsite Hamiltonian $\boldsymbol{H}_{II}$	
Correlation order ν	2
Cutoff radius r_{cut}	10 Å
Maximum polynomial degree n_{max}	9
Regularisation λ	10^{-7}

Offsite Hamiltonian $\boldsymbol{H}_{IJ}$		
Correlation order ν		1
Bond cutoff radius $r^{\text{b}}_{\text{cut}}$		10 Å
Env. cutoff radius $r^{\text{e}}_{\text{cut}}$		5 Å
Env. cutoff radius $z^{\text{e}}_{\text{cut}}$		5 Å
Maximum polynomial degree $n^{\text{b}}_{\text{max}}$		14 14 14
	ss	14 14 14
		14 14 9
	sp	14 14 12
		14 14 10
	sd	14 14 11
	pp	13 13
		13 13
	pd	14 14
	dd	14
Regularisation λ		10^{-7}

Offsite overlap $\boldsymbol{S}_{IJ}$	
Correlation order ν	0
Cutoff radius r_{cut}	10 Å
Maximum polynomial degree n	16
Regularisation λ	10^{-7}

TABLE I. ACE basis set parameters for our optimised models for $\boldsymbol{H}_{II}$, $\boldsymbol{H}_{IJ}$ and $\boldsymbol{S}_{IJ}$. Maximum polynomial degree can be specified independently for each component model shown in Fig. 2. The maximum polynomial degrees for the adjoint blocks *ps*, *ds* and *dp* of $\boldsymbol{H}_{IJ}$ are the transposes of those shown for *sp*, *sd* and *pd*, respectively.

dicted band structures. These have been estimated using a simple *a posteriori* error analysis to propagate errors in the Hamiltonian $\Delta\boldsymbol{H} = \boldsymbol{H} - \tilde{\boldsymbol{H}}$ and overlap $\Delta\boldsymbol{S} = \boldsymbol{S} - \tilde{\boldsymbol{S}}$ to expected errors in the bands using the result[44]

$$|\tilde{\epsilon} - \epsilon| = \langle\phi|\Delta\boldsymbol{H} - \epsilon\Delta\boldsymbol{S}|\phi\rangle$$

where ϕ and ϵ are the eigenfunctions and eigenvalues of the reference system. Repeating this for each k-point leads to the error bounds shown. The error estimates prove reliable: the DFT bands, shown in red, are almost always contained within the blue shaded region.

Fig. 6 shows the convergence of bandstructures and DoS with respect to the maximum polynomial degree used in the ACE basis set, and for two choices of correlation order $\nu = 1$ and $\nu = 2$.

The error in the DoS is computed using the first Wasserstein (or 'earthmover') distance between the reference and predicted DoS, an appropriate metric for comparing probability distributions. The error in band structures is defined as the RMSE in the k-dependent band energies

$$E_{\text{band}}(\mathbf{k}) = \sum_{i=1}^{N_{\text{orb}}} f\left(\frac{\epsilon_i - \varepsilon_F}{\sigma}\right)\epsilon_i(\mathbf{k})$$

along the high-symmetry k-paths shown in Fig. 5, where $f(\bullet)$ is the Fermi function, ε_F is the Fermi level of system and the smearing width is taken to be $\sigma = 0.086$ eV, corresponding to an electronic temperature of 1000 K.

When increasing the maximum degree used for all subblocks simultaneously, similar overfitting can be seen in the direct validation results of Fig. 3, and once again this arises at lower degrees of 9-12 with $\nu = 2$ than with $\nu = 1$, where degrees of up to 13–14 are possible without overfitting. Errors in the DoS and the band structure for both FCC and BCC are further reduced when using the optimised model of § III B, shown with the horizontal dashed lines in the figure to produce band structures with a RMSE of less than 0.4 eV for both phases.

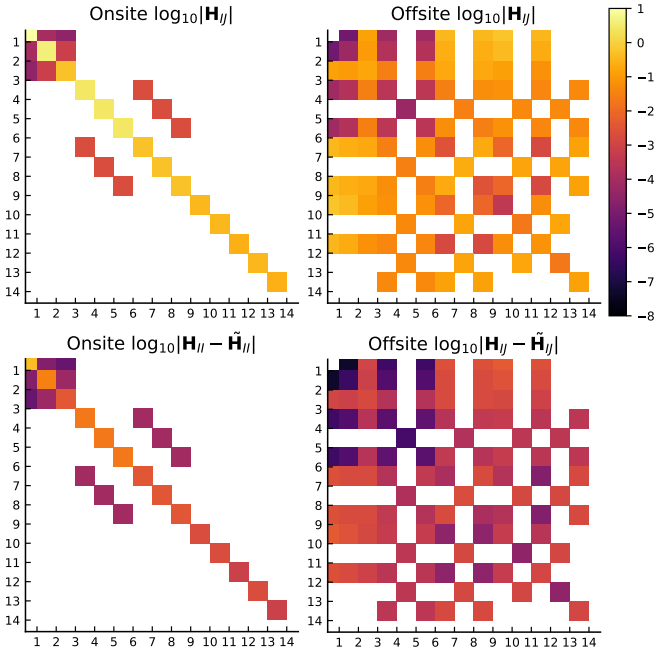

FIG. 4. Magnitudes (above) and errors (below) for onsite and offsite $\tilde{\boldsymbol{H}}$ for prediction on the FCC ground state unitcell (not included in training set).

D. BCC to FCC transition

As a final challenging test, we used our optimised model to predict the Hamiltonian and overlap matrices along the Bain transformation path from BCC to FCC. We then diagonalised the predicted matrices to obtain the eigenvalues and hence the DOS at each point along the path. As can be seen in Fig. 7, the predicted DoS

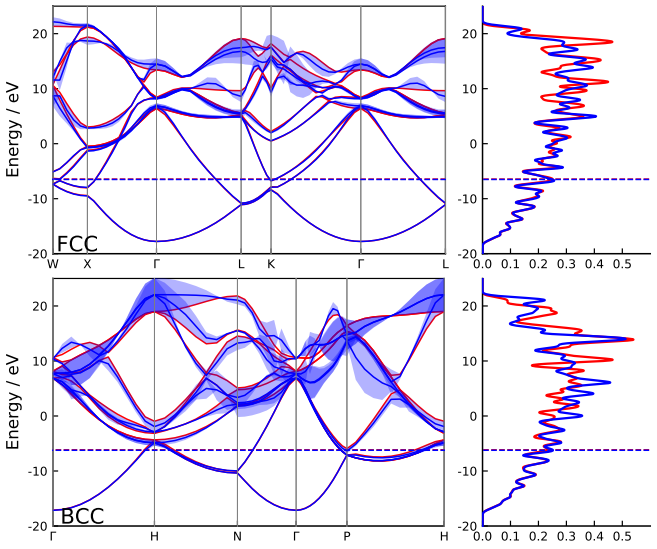

FIG. 5. FCC and BCC band structures obtained with DFT (red) and predicted by an ACEtb model with onsite H order 2, and offsite H and S order 1. Confidence interval shown with blue ribbons are from *a posteriori* analysis of the errors in band spectrum expected to result from known errors in $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{S}}$. The Fermi levels of the two methods are shown with the red and blue dashed lines (which are in close agreement).

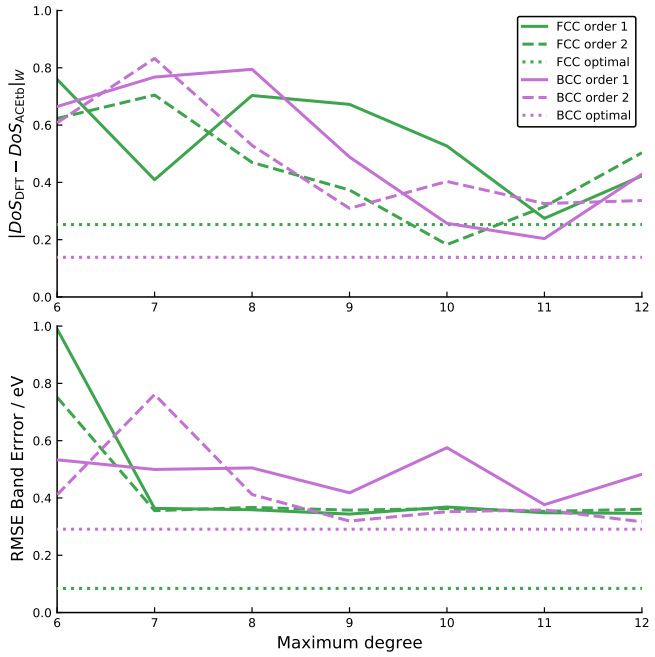

FIG. 6. Convergence of FCC and BCC band structures and DoS with respect to the correlation order and maximum polynomial degree used in the ACE basis set. Dotted lines show the optimised model of § III B.

agree well at the two ends of the path, and vary smoothly along it, suggesting good extrapolative behaviour beyond the training set, which includes only environments accessible from the two minima at moderate temperatures during MD.

IV. DISCUSSION AND CONCLUSION

We have reported a novel data-driven scheme to construct predictive models of Hamiltonian and overlap matrices from *ab initio* data. Our scheme incorporates all relevant symmetry operations, giving an equivariant analytical map from first principles data to linear models for the Hamiltonian and overlap matrices as a function of the atomic and bond environments. We have shown that it is possible to apply the new methodology to produce accurate predictions for the band structure in aluminium in both FCC and BCC phases from limited training data. The new approach has huge potential for delivering comparable accuracy to DFT while at the same time reaching time and length scales far beyond its' capabilities. For example, it opens to door to high-throughput computation of quantities which depend on electronic properties, such as transport coefficients or anharmonic contributions to phonon modes that can currently only be accurately computed with first principles methhods.[45]

Our results are extremely encouraging, and there are a number of avenues open for further exploration. From a computational performance perspective, we note the evaluation of Hamiltonian and overlap blocks is trivially parallelisable with perfect scaling. Performance enhancements would also come from further optimisation of the ACE basis used to represent the Hamiltonian and overlap matrices, e.g. by sparsifying to reduce the basis set size, or by incorporating non-linearity to reduce the maximum degree required.[46] Moreover, Bayesian ap-

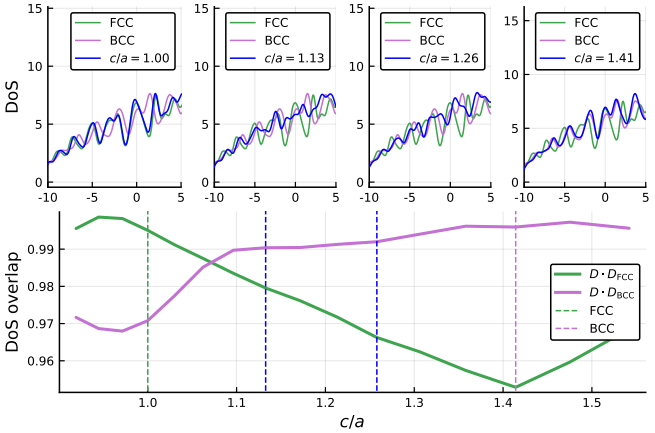

FIG. 7. Densities of states along the transition from BCC $c/a = 1$ to FCC $c/a = \sqrt{2} \approx 1.41$. Upper panels: comparisons of the ACEtb-predicted DoS (blue) with the reference values for the FCC (green) and BCC (purple) ground states obtained with FHIaims. Lower panel: projection of the ACEtb DoS along the Bain path onto the FCC and BCC references DoS. In both cases, note the good agreement with FCC on the left and with BCC on the right and the smooth variation in between.

proaches to model selection could be used instead of cross-validation. This would lead to more efficient model construction, as well as the possibility of *a priori* error estimates on the accuracy of model predictions using a Bayesian linear regression framework.

Further comprehensive studies of the dependence of accuracy and transferability of models on quantity and type of training data, as well as extension to materials and system with more complex bonding environments are also necessary. In future we will expand this approach to explore multi-component systems. A further extension will be to fit a potential $\tilde{E}$ to allow total energy and forces to be predicted by adding a correction to the band energy. For example, $\tilde{E}$ could be represented by an ACE potential determined from the local atomic environments.

ACKNOWLEDGMENTS

This work was financially supported by a Leverhulme Trust Research Project Grant (RPG-2017-191), the EPSRC (EP/R043612/1), the NOMAD Centre of Excellence (European Commission grant agreement ID 951786), and the UKRI Future Leaders Fellowship programme (MR/S016023/1). We acknowledge computational resources provided by the Scientific Computing Research Technology Platform of the University of Warwick, the EPSRC-funded HPC Midlands+ computing centre (EP/P020232/1) and on ARCHER2 (https://www.archer2.ac.uk/) via the UK Car-Parinello consortium (EP/P022065/1).

[1]E. Bitzek, J. R. Kermode, and P. Gumbsch, "Atomistic aspects of fracture," Int. J. Fract. **191**, 13–30 (2015).

[2]B. Jiang and H. Guo, "Dynamics in reactions on metal surfaces: A theoretical perspective," **150**, 180901, publisher: AIP Publishing LLC.

[3]J. Behler, "Four generations of high-dimensional neural network potentials," (), 10.1021/acs.chemrev.0c00868, publisher: American Chemical Society.

[4]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," **121**, 10142–10186, publisher: American Chemical Society.

[5]V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," **121**, 10073–10141, publisher: American Chemical Society.

[6]F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials," **121**, 9759–9815, publisher: American Chemical Society.

[7]Y. Mishin, "Machine-learning interatomic potentials for materials science," Acta Mater. **214**, 116980 (2021).

[8]J. Behler and G. Csányi, "Machine learning potentials for extended systems: a perspective," Eur. Phys. J. B **94**, 142 (2021).

[9]M. J. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. Stewart, "AM1: A new general purpose quantum mechanical molecular model1," **107**, 3902–3909, publisher: American Chemical Society.

[10]J. J. P. Stewart, "Optimization of parameters for semiempirical methods i. method," **10**, 209–220, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540100208.

[11]D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon," **51**, 12947.

[12]M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties," **58**, 7260–7268.

[13]O. F. Sankey and D. J. Niklewski, "Ab initio multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems," **40**, 3979–3995.

[14]J. P. Lewis, K. R. Glaesemann, G. A. Voth, J. Fritsch, A. A. Demkov, J. Ortega, and O. F. Sankey, "Further developments in the local-orbital density-functional-theory tight-binding method," **64**, 195103.

[15]C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," **15**, 1652–1671, publisher: American Chemical Society.

[16]J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, "Perspective on integrating machine learning into computational chemistry and materials science," **154**, 230903, publisher: American Institute of Physics.

[17]H. Li, C. Collins, M. Tanha, G. J. Gordon, and D. J. Yaron, "A density functional tight binding layer for deep learning of chemical hamiltonians," **14**, 5764–5776, publisher: American Chemical Society, 1808.04526.

[18]M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko, "Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks," **11**, 6835–6843, publisher: American Chemical Society.

[19]Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. MillerIII, "OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features," **153**, 124111, publisher: AIP Publishing LLCAIP Publishing.

[20]A. R. Supka, T. E. Lyons, L. Liyanage, P. D'Amico, R. Al Rahal Al Orabi, S. Mahatara, P. Gopal, C. Toher, D. Ceresoli, A. Calzolari, S. Curtarolo, M. B. Nardelli, and M. Fornari, "AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians," **136**, 76–84.

[21]K. F. Garrity and K. Choudhary, "Database of wannier tight-binding hamiltonians using high-throughput density functional theory," **8**, 106, bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Electronic structure;Topological matter Subject_term_id: electronic-structure;topological-matter.

[22]N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, "Maximally localized wannier functions: Theory and applications," **84**, 1419–1475, publisher: American Physical Society, 1112.5411.

[23]V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "Ab initio molecular simulations with numeric atom-centered orbitals," **180**, 2175–2196, ISBN: 0010-4655.

[24]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," **134**, 074106 (), ISBN: 0021-9606.

[25]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," **87**, 184115, 1209.3140v2.

[26]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B Condens. Matter **99**, 014104 (2019).

[27]G. Dusson, M. Bachmayr, G. Csanyi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, "Atomic cluster expansion: Completeness, efficiency and stability," ArXiv:1911.03550.

[28]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chem-

istry with a deep neural network for molecular wavefunctions," **10**, 5024.
[29]M. Gastegger, A. McSloy, M. Luya, K. T. Schütt, and R. J. Maurer, "A deep neural network for molecular wave functions in quasi-atomic minimal basis representation," **153**, 044123, publisher: American Institute of Physics Inc., 2005.06979.
[30]G. Hegde and R. C. Bowen, "Machine-learned approximations to density functional theory hamiltonians," **7**, 42669, tex.ids= hegdeMachinelearnedApproximationsDensity2017, hegdeMachinelearnedApproximationsDensity2017a number: 1 publisher: Nature Publishing Group.
[31]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," **104**, 136403, publisher: American Physical Society.
[32]J. Nigam, M. Willatt, and M. Ceriotti, "Equivariant representations for molecular hamiltonians and n-center atomic-scale properties," (2021), arXiv:2109.12083.
[33]E. Cancès, G. Kemlin, and A. Levitt, "Convergence analysis of direct minimization and self-consistent iterations," SIAM J. Matrix Anal. Appl. **42**, 243–274 (2021).
[34]N. D. Woods, M. C. Payne, and P. J. Hasnip, "Computing the self-consistent field in kohn–sham density functional theory," **31**, 453001, publisher: IOP Publishing.
[35]R. Drautz, "Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer," Phys. Rev. B Condens. Matter **102**, 024104 (2020).
[36]A. Grisafi and M. Ceriotti, "Incorporating long-range physics in atomic-scale machine learning," The Journal of Chemical Physics **151**, 204105 (2019).
[37]J. Nigam, S. Pozdnyakov, and M. Ceriotti, "Recursive evaluation and iterative contraction of n-body equivariant features," The Journal of Chemical Physics **153**, 121101 (2020).
[38]Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, and R. Drautz, "Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon," (2021).
[39]J. C. Slater and G. F. Koster, "Simplified LCAO method for the periodic potential problem," **94**, 1498–1524.
[40]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).
[41]H. J. Monkhorst and J. D. Pack, "Special points for brillouin zone integration," **13**, 5188–5192.
[42]X. W. Zhou, R. A. Johnson, and H. N. G. Wadley, "Misfit-energy-increasing dislocations in vapor-deposited cofe/nife multilayers," Phys. Rev. B **69**, 144113 (2004).
[43]C. O. et a., "ACE.jl: Approximation of symmetric functions with polynomials and spherical harmonics,".
[44]M. G. Crandall and P. H. Rabinowitz, "Bifurcation, perturbation of simple eigenvalues and linearized stability," (1973).
[45]F. Knoop, T. A. R. Purcell, M. Scheffler, and C. Carbogno, "Anharmonicity measure for materials," Phys. Rev. Materials **4**, 083809 (2020).
[46]Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, *et al.*, "Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon," npj Computational Materials **7**, 1–12 (2021).

Appendix A: Transformation of H and S from real to reciprocal space representation

According to Bloch's theorem, in crystal-periodic structures, the Hamiltonian and overlap matrices defined in terms of real-space atomic orbitals can be transformed into a block-diagonal form and solved via a set of N_k independent generalised eigenvalue problems where each block corresponds to a vector $\boldsymbol{k}$ within the reciprocal unit cell:

$$\mathbf{H}(\boldsymbol{k})\psi_{i\boldsymbol{k}} = \epsilon_{i\boldsymbol{k}}\mathbf{S}(\boldsymbol{k})\psi_{i\boldsymbol{k}} \quad i = 1, 2, \cdots \tag{A1}$$

where $\psi_{\nu\boldsymbol{k}}$ are Bloch wave functions and $\mathbf{H}(\boldsymbol{k})$ and $\mathbf{S}(\boldsymbol{k})$ are Hamiltonian and overlap matrices defined in terms of a discrete crystal-periodic basis.

For this, we define crystal-periodic generalised basis functions $\chi_{a,\boldsymbol{k}}$ from real-space basis functions as follows:

$$\chi_{a\boldsymbol{k}}(\boldsymbol{x}) = \sum_N \exp\{i\boldsymbol{k}\cdot\mathbf{NL}\}\chi_a(\boldsymbol{x}+\mathbf{NL}). \tag{A2}$$

In (A2), $\mathbf{L}$ refers to the column matrix of lattice vectors and $\mathbf{N} = (N_1, N_2, N_3)$ is an index vector that specifies the position of the unit cell (in multiples of the lattice vectors) in which orbital χ_a is located.

The matrix elements of $\mathbf{H}(\boldsymbol{k})$ and $\mathbf{S}(\boldsymbol{k})$, respectively, are constructed via

$$H_{ab}(\boldsymbol{k}) = \langle\chi_{a\boldsymbol{k}}|\hat{H}|\chi_{b\boldsymbol{k}}\rangle = \tag{A3}$$

$$\sum_{N,N'} \exp\{i\boldsymbol{k}\cdot(\mathbf{N}'-\mathbf{N})\cdot\mathbf{L}\}\underbrace{\langle\chi_{a,N'}|\hat{H}|\chi_{b,N}\rangle}_{=H_{ab}(\mathbf{N},\mathbf{N}')} \tag{A4}$$

and

$$S_{ab}(\boldsymbol{k}) = \langle\chi_{a\boldsymbol{k}}|\chi_{b\boldsymbol{k}}\rangle = \tag{A5}$$

$$\sum_{N,N'} \exp\{i\boldsymbol{k}\cdot(\mathbf{N}'-\mathbf{N})\cdot\mathbf{L}\}\underbrace{\langle\chi_{a,N'}|\chi_{b,N}\rangle}_{=S_{ab}(\mathbf{N},\mathbf{N}')} \tag{A6}$$

where $H_{ab}(\mathbf{N},\mathbf{N}')$ and $S_{ab}(\mathbf{N},\mathbf{N}')$ are as defined in (4) and (5) for atomic orbitals defined in different unit cells $\mathbf{N}$ and $\mathbf{N}'$.

In this work, we use this transformation to map the real-space matrices to arbitrarily dense $\boldsymbol{k}$-grids as is common practice for localised basis sets such as atomic orbitals or maximally localized Wannier functions. We then calculate eigenvalues $\epsilon_{\nu\boldsymbol{k}}$ at arbitrary points in reciprocal space to calculate converged electronic densities-of-state and band structures.

Appendix B: Equivariance of $\boldsymbol{H}_{IJ}$

For the real space Hamiltonian $\mathbf{H}(\mathbf{0})$, we decompose it as $\mathbf{H}(\mathbf{0}) = \left(\boldsymbol{H}_{IJ}\right)_{I,J=1}^{N_{\text{atom}}}$ (c.f., Fig. 2). Denote $a = (n,l,m;I) := (\alpha;I), b = (n',l',m';J) := (\beta;J)$, we may then write

$$\boldsymbol{H}_{IJ}^{\alpha\beta}(\mathbf{R}) = \langle\chi_a|\hat{H}|\chi_b\rangle$$

In the definition of χ_a, the radial basis $R_{nl}(r)$ is invariant under rotation and $Y_l^m(Q(\theta,\phi))$ can be expressed as linear combination of $Y_l^\mu(\theta,\phi)$, i.e.,

$$\chi_{(n,l,m;I)}(Q\boldsymbol{x};Q\mathbf{R}) = \sum_\mu D^l_{\mu m}\chi_{(n,l,\mu;I)}(\boldsymbol{x};\mathbf{R}). \tag{B1}$$

Besides,

$$\begin{aligned}
&\boldsymbol{H}_{IJ}^{\alpha\beta}(Q\mathbf{R}) \\
&= \int_{\mathbb{R}^3} \chi_a(\boldsymbol{x};\mathbf{R})^* V_{\text{eff}}(\boldsymbol{x},Q\mathbf{R})\chi_b(\boldsymbol{x};Q\mathbf{R})d\boldsymbol{x} \\
&= \int_{\mathbb{R}^3} \chi_a(Q\boldsymbol{x};Q\mathbf{R})^* V_{\text{eff}}(Q\boldsymbol{x},Q\mathbf{R})\chi_b(Q\boldsymbol{x};Q\mathbf{R})d\boldsymbol{x} \\
&= \int_{\mathbb{R}^3} \chi_a(Q\boldsymbol{x};Q\mathbf{R})^* V_{\text{eff}}(\boldsymbol{x},\mathbf{R})\chi_b(Q\boldsymbol{x};Q\mathbf{R})d\boldsymbol{x}.
\end{aligned} \tag{B2}$$

Combining (B1) and (B2), we see immediately that

$$\boldsymbol{H}_{IJ}(Q\mathbf{R}) = D(Q)^*\boldsymbol{H}_{IJ}(\mathbf{R})D(Q),$$

where

$$D(Q) = \text{Diag}(D^{l_1}(Q), D^{l_2}(Q), \cdots),$$

and D^{l_i} indicate the Wigner-D matrices.