npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

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

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

We propose a scheme to construct predictive models for Hamiltonian matrices in atomic orbital representation from ab initio data as a function of 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 Hamiltonian matrix blocks that transform equivariantly with respect to the full rotation group. The approach produces analytical linear models for the Hamiltonian and overlap matrices. Through an application to aluminium, we demonstrate that it is possible to train models from a handful of structures computed with density functional theory, and apply them to produce accurate predictions for the electronic structure. The model generalises well and is able to predict defects accurately from only bulk training data.

npj Computational Materials (2022)8:158 ; https://doi.org/10.1038/s41524-022-00843-2

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 inaccessible 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 Kohn–Sham density functional theory (KS-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 two-centre and three-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], 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 benefit high-throughput materials screening studies[20]. Most commonly, electronic structure simulations of materials are performed in non-atom-centred basis representations such as the pseudopotential plane wave framework, which is not easily amenable to the construction of TB models. TB Hamiltonians are typically constructed via transformation into a maximally localised Wannier function representation[21], which provides a compact atom-centred basis representation with local support[22]. It is also possible to fit Slater–Koster parameters directly to DFT calculations in a data-driven fashion[23,24]. Materials simulations in atom-centred orbital representations as provided by, for example, the FHI-aims code[25] 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]. Examples of such representations include Behler–Parinello symmetry functions[3,26], the SOAP descriptor[27] or the atomic cluster expansion[28,29]. First efforts of direct machine learning prediction of electronic structure have been reported in literature. For example, SchNOrb[30] is a DTNN representation of molecular mean-field electronic structure Hamiltonians, which

[1]Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC V6T 1Z2, Canada. [2]Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, UK. [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, UK. ✉email: J.R.Kermode@warwick.ac.uk

has been used to predict Hamiltonians in local atomic orbital and optimised effective minimal basis representations for organic molecules including up to 13 heavy atoms[30,31]. Hedge and Bowen[32] employed Kernel ridge regression with a bispectrum representation[33] for an analytical representation of a minimal basis DFT Hamiltonian for bulk copper and diamond. Equivariant parameterisations for molecular systems along similar lines to what we describe here have been reported, learning either from the Hamiltonian[34] or from wavefunctions and electronic densities[35]. These works apply linear or nonlinear equivariant models, respectively, to the MD17 molecular dataset, both of which improve on the non-equivariant SchNOrb approach of ref. [30]. However, to our knowledge, the present work is the first to address the specific challenges of learning Hamiltonians in solid state systems.

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 three 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 Fig. 1. We showcase the capabilities of this approach by predicting the band structure of bulk aluminium in different crystal systems.

RESULTS

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

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

where

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

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

$$V_{\text{eff}} = V_{\text{eff}}[\rho], \text{ where} \tag{3}$$

$$\rho = \sum_i f_i |\psi_i|^2, \tag{4}$$

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 by employing a self-consistent field (SCF) algorithm[36,37].

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.

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

$$\chi_a(\boldsymbol{x}) = R_{nl}(r)Y_{lm}(\theta, \phi) \tag{5}$$

where $a = (n, l, m; I)$ is a composite index, and 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 $n = 0, \dots, n_{\max}$, $l = 0, \dots, l_{\max}$, $m = -l_{\max}, \dots, l_{\max}$ characterise 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[25], with the onsite and offsite block structure as illustrated in Fig. 2. With this definition, we can express the overlap between basis functions and the interactions as mediated by the Hamiltonian as follows:

$$H_{ab} = \langle \chi_a | \hat{H} | \chi_b \rangle \text{ and} \tag{6}$$

$$S_{ab} = \langle \chi_a | \chi_b \rangle. \tag{7}$$

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:

$$\mathbf{H}(\boldsymbol{k})\psi_{i\boldsymbol{k}} = \epsilon_{i\boldsymbol{k}}\mathbf{S}(\boldsymbol{k})\psi_{i\boldsymbol{k}}, \quad i = 1, 2, \dots, \tag{8}$$

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 the Methods section IV D, 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 Eqs. (6) and (7), 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

$$V_{\text{eff}} = V_{\text{eff}}(\boldsymbol{x}; \mathbf{R}). \tag{9}$$

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 centred at sites I and J. It is then straightforward to deduce that

$$\boldsymbol{H}_{IJ}(Q\mathbf{R}) = D(Q)^*\boldsymbol{H}_{IJ}(\mathbf{R})D(Q), \tag{10}$$

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

$$D(Q) = \mathrm{Diag}(D^{l_1}(Q), D^{l_2}(Q), \cdots), \tag{11}$$

and $(l_1, l_2, \dots)$ specify the types of orbitals at each site. More details can be found in the "Methods" section "Equivariance of H_{IJ}". Since the focus of the present work is on elemental metallic systems we ignore chemical species information entirely in the present work; this will be addressed in the future either directly as is done for ACE interatomic potentials[38] or using compressed species information[39,40].

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

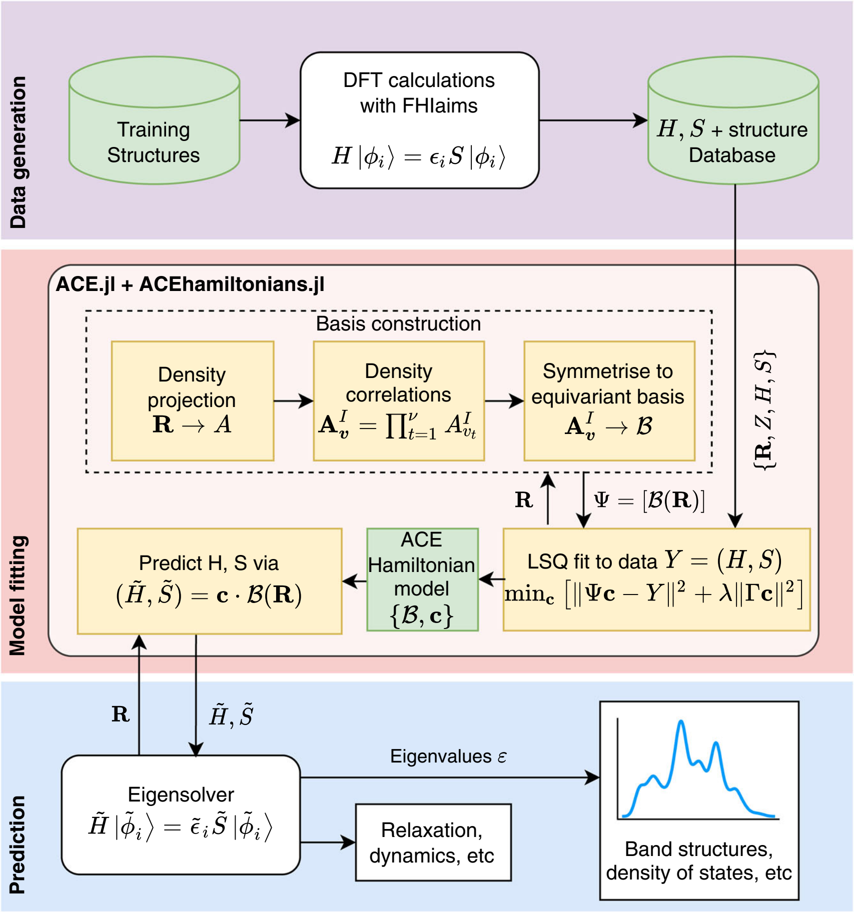
Data generation
Training Structures
DFT calculations with FHIaims
$H|\phi_i\rangle = \epsilon_i S|\phi_i\rangle$
H, S + structure Database
Model fitting
ACE.jl + ACEhamiltonians.jl
Basis construction
Density projection $\mathbf{R} \to A$
Density correlations $\mathbf{A}_{\boldsymbol{v}}^I = \prod_{t=1}^{\nu} A_{v_t}^I$
Symmetrise to equivariant basis $\mathbf{A}_{\boldsymbol{v}}^I \to \mathcal{B}$
$\{\mathbf{R}, Z, H, S\}$
$\mathbf{R}$
$\Psi = [\mathcal{B}(\mathbf{R})]$
Predict H, S via $(\tilde{H}, \tilde{S}) = \mathbf{c} \cdot \mathcal{B}(\mathbf{R})$
ACE Hamiltonian model $\{\mathcal{B}, \mathbf{c}\}$
LSQ fit to data $Y = (H, S)$
$\min_{\mathbf{c}} \left[\|\Psi \mathbf{c} - Y\|^2 + \lambda \|\Gamma \mathbf{c}\|^2 \right]$
Prediction
$\mathbf{R}$
$\tilde{H}, \tilde{S}$
Eigensolver $\tilde{H}|\tilde{\phi}_i\rangle = \tilde{\epsilon}_i \tilde{S}|\tilde{\phi}_i\rangle$
Eigenvalues ε
Relaxation, dynamics, etc
Band structures, density of states, etc

Fig. 1 Schematic of the ACEhamiltonians (atomic cluster expansion for Hamiltonians) workflow. The upper panel shows data generation with the FHI-aims electronic structure theory code, the central panel model fitting with the `ACE.jl` and `ACEhamiltonians.jl` packages, and the lower panel prediction.

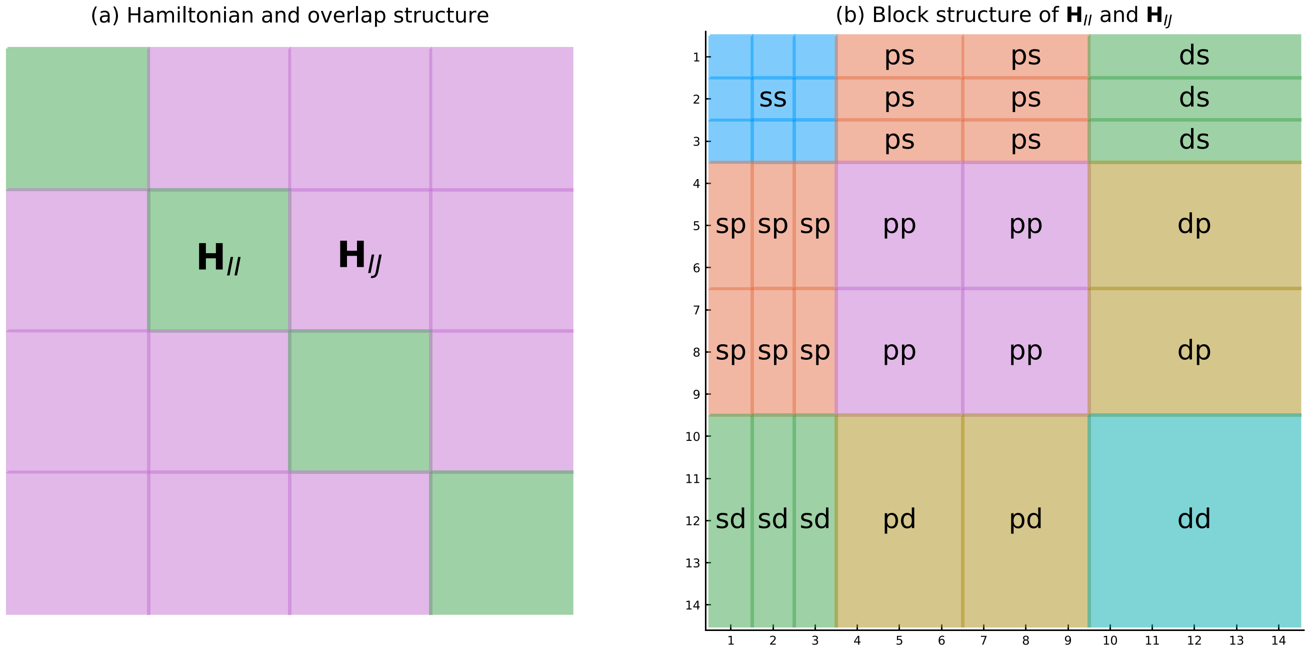
(a) Hamiltonian and overlap structure
$\mathbf{H}_{II}$
$\mathbf{H}_{IJ}$
(b) Block structure of $\mathbf{H}_{II}$ and $\mathbf{H}_{IJ}$
ss
ps
ps
ds
sp
pp
dp
sd
pd
dd
1
2
3
4
5
6
7
8
9
10
11
12
13
14

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.

between orbitals centred 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

$$\begin{aligned}\boldsymbol{H}_{II} &= \boldsymbol{H}_{\rm on}(\mathbf{R}_I), \qquad \text{and}\\ \boldsymbol{H}_{IJ} &= \boldsymbol{H}_{\rm off}(\mathbf{R}_{IJ}),\end{aligned} \tag{12}$$

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

$$\begin{aligned}\mathbf{R}_I &:= \{\boldsymbol{r}_{IK} \,|\, K \neq I\}, \text{ and}\\ \mathbf{R}_{IJ} &:= \left\{\boldsymbol{r}_{IJ}; \{\boldsymbol{r}_K - \tfrac{1}{2}(\boldsymbol{r}_I + \boldsymbol{r}_J) \,|\, K \neq I, J\}\right\},\end{aligned} \tag{13}$$

where $\boldsymbol{r}_{IJ} = \boldsymbol{r}_I - \boldsymbol{r}_J$. In the above definitions, the index K runs over all unit cells N within the crystal volume. According to Eq. (10) the functions $\boldsymbol{H}_{\rm on}$ and $\boldsymbol{H}_{\rm off}$ are equivariant in the sense that

$$\boldsymbol{H}_{\rm on/off}(Q\mathbf{R}) = D(Q)^* \boldsymbol{H}_{\rm on/off}(\mathbf{R}) D(Q). \tag{14}$$

Translation invariance is now built into the dependence of $\boldsymbol{H}_{\rm on/off}$ on relative positions only, while permutation equivariance of $\boldsymbol{H}$ is built into Eq. (12).

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 identity matrices. The off-site blocks follow the same symmetry as the Hamiltonian off-site blocks.

Parameterisation

We parameterise the real-space Hamiltonian and overlap matrix blocks $\boldsymbol{H}_{\rm on}, \boldsymbol{H}_{\rm off}$ and $\boldsymbol{S}_{\rm off}$ using an equivariant ACE basis[28,29,41]. Similar techniques have previously been proposed in other contexts[34,40,42]. In this section, we present a general outline of the ideas, making certain choices of approximation parameters concrete in the "Methods" section "Parameter estimation".

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. 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 in refs. [29,38,41].

1. Parameterisation of H_{on}. We start by choosing a *one-particle* basis,

$$\phi_v(\boldsymbol{x}) := \phi^{\rm on}_{nlm}(\boldsymbol{x}) := P_{nl}(r) Y_{lm}(\hat{\boldsymbol{x}}) f_{\rm cut}(r) \tag{15}$$

where $\boldsymbol{x} = r\hat{\boldsymbol{x}}$ and 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),

$$A^I_v := \sum_{J \neq I} \phi_v(\boldsymbol{r}_{IJ}), \tag{16}$$

$$\boldsymbol{A}^I_{\boldsymbol{v}} := \prod_{t=1}^{\nu} A^I_{v_t} \qquad \text{for } \boldsymbol{v} = (v^1, \ldots, v^\nu), \nu = 1, 2, \ldots . \tag{17}$$

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

$$\boldsymbol{H}_{II} = \boldsymbol{H}_{\rm on}(\mathbf{R}_I) \approx \tilde{\boldsymbol{H}}^{\rm PI}_{\rm on}(\mathbf{R}_I) = \sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^I_{\boldsymbol{v}}, \tag{18}$$

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 centred 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. in the section "Parameter estimation".

The expansion (18) incorporates translation and permutation invariance but not yet the O(3)-equivariance (10). Following the general ACE construction[29] we can achieve this by simply averaging the representation over the group O(3), i.e.,

$$\tilde{\boldsymbol{H}}_{\rm on}(\mathbf{R}_I) = -\!\!\!\!\!\!\int_{\rm O(3)} D(Q) \tilde{\boldsymbol{H}}^{\rm PI}_{\rm on}(Q\mathbf{R}_I) D(Q)^* dQ, \tag{19}$$

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

2. Parameterisation of H_{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

$$\begin{aligned}\phi^{\rm b}_{nlm}(\boldsymbol{r}_{IJ}) &= P^{\rm b}_{nl}(r_{IJ}) Y_{lm}(\hat{\boldsymbol{r}}_{IJ}) f^{\rm b}_{\rm cut}(r_{IJ}),\\ \phi^{\rm e}_{nlm}(\boldsymbol{r}_{IJ,K}) &= P^{\rm e}_{nl}(r_{IJ,K}) Y_{lm}(\hat{\boldsymbol{r}}_{IJ,K}) f^{\rm e}_{\rm cut}(\boldsymbol{r}_{IJ,K}, \boldsymbol{r}_{IJ}).\end{aligned} \tag{20}$$

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

$$A^{IJ}_v := \sum_{K \neq I,J} \phi^{\rm e}_v(\boldsymbol{r}_{IJ,K}), \tag{21}$$

and the product basis becomes

$$\boldsymbol{A}^{IJ}_{\boldsymbol{v}} := \phi^{\rm b}_{v^0}(\boldsymbol{r}_{IJ}) \cdot \prod_{t=1}^{\nu} A^{IJ}_{v^t}, \tag{22}$$

for $\boldsymbol{v} = (v^0, v^1, \ldots, v^\nu)$, with $\nu = 0, 1, 2, \ldots$ 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

$$\boldsymbol{H}_{IJ} = \boldsymbol{H}_{\rm off} \approx \tilde{\boldsymbol{H}}^{\rm PI}_{\rm off}(\mathbf{R}_{IJ}) := \sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^{IJ}_{\boldsymbol{v}}. \tag{23}$$

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

$$\tilde{\boldsymbol{H}}_{\rm off}(\mathbf{R}_{IJ}) := -\!\!\!\!\!\!\int_{\rm O(3)} D(Q) \tilde{\boldsymbol{H}}^{\rm PI}_{\rm off}(Q\mathbf{R}_{IJ}) D(Q)^* dQ. \tag{24}$$

3. Parameterisation of S_{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

$$\tilde{\boldsymbol{S}}_{\rm off}(\boldsymbol{r}_{IJ}) := -\!\!\!\!\!\!\int_{\rm O(3)} D(Q) \left[\sum_v C_v \phi^{\rm b}_v(Q\boldsymbol{r}_{IJ})\right] D(Q)^* \, dQ. \tag{25}$$

This is formally equivalent to a Slater Koster representation of two-centre integrals[43], 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.

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.,

$$\tilde{\boldsymbol{K}}(\mathbf{R}_{\bullet}) = \fint_{\mathrm{O}(3)} D(Q)\left[\sum_{\boldsymbol{v}} C_{\boldsymbol{v}} \boldsymbol{A}^{\bullet}_{\boldsymbol{v}}(Q\mathbf{R}_{\bullet})\right] D(Q)^{*}, \tag{26}$$

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

$$\boldsymbol{A}^{IJ}_{\boldsymbol{v}}(\mathbf{R}_{IJ}) = \phi^{\mathrm{b}}_{v}(\boldsymbol{r}_{IJ}). \tag{27}$$

In order to make our description clearer, we denote $\boldsymbol{v} \equiv \boldsymbol{nlm}$ with $\boldsymbol{n}, \boldsymbol{l}, \boldsymbol{m}$ being the lists of corresponding indices in $\boldsymbol{A}^{IJ}_{\boldsymbol{v}}$. Thus, we can deduce that

$$\boldsymbol{A}^{\bullet}_{\boldsymbol{nlm}}(Q\mathbf{R}_{\bullet}) = \sum_{\boldsymbol{\mu}} \boldsymbol{D}^{\boldsymbol{l}}_{\boldsymbol{\mu m}}(Q) \boldsymbol{A}^{\bullet}_{\boldsymbol{nl}\mu}(\mathbf{R}_{\bullet}), \tag{28}$$

where $\boldsymbol{D}^{\boldsymbol{l}}_{\boldsymbol{\mu m}}(Q) = \prod_{t} D^{l_t}_{\mu_t m_t}(Q)$ since the angular dependence of the one-particle basis functions in all cases is in terms of spherical harmonics Y_{lm}. Furthermore, we write

$$C_{\boldsymbol{v}} = \sum_{\alpha,\beta=1}^{N_{\mathrm{orb}}} c^{\alpha\beta}_{\boldsymbol{v}} E^{\alpha\beta}, \tag{29}$$

where $E^{\alpha\beta} \in \mathbb{R}^{N_{\mathrm{orb}} \times N_{\mathrm{orb}}}$ with $E^{\alpha\beta}_{\alpha'\beta'} = \delta_{\alpha\alpha'}\delta_{\beta\beta'}$. Inserting these two identities into Eq. (26) yields

$$\begin{aligned} \tilde{\boldsymbol{K}}(\mathbf{R}_{\bullet}) &= \sum_{\boldsymbol{n},\boldsymbol{l},\boldsymbol{m},\alpha,\beta} c^{\alpha\beta}_{\boldsymbol{v}} \sum_{\boldsymbol{\mu}} \mathcal{U}^{\alpha\beta}_{\boldsymbol{l\mu m}} \boldsymbol{A}^{\bullet}_{\boldsymbol{nl\mu}}(\mathbf{R}_{\bullet}) \\ &=: \sum_{\boldsymbol{n},\boldsymbol{l},\boldsymbol{m},\alpha,\beta} c^{\alpha\beta}_{\boldsymbol{nlm}} \mathcal{B}^{\alpha\beta}_{\boldsymbol{nlm}}(\mathbf{R}_{\bullet}), \end{aligned} \tag{30}$$

where the “generalised coupling coefficients” are given by

$$\mathcal{U}^{\alpha\beta}_{\boldsymbol{l\mu m}} = \fint_{\mathrm{O}(3)} \mathbf{D}^{\mathbf{l}}_{\boldsymbol{\mu m}}(Q) D(Q) E^{\alpha\beta} D(Q)^{*} dQ. \tag{31}$$

Their definition involves an integral over products of Wigner-D matrices which can be precomputed explicitly (i.e., without the need for quadrature which would incur a discretisation error) using the recursion proposed by Dusson et al. [29] and independently by Nigam et al. [34].

Note that Eq. (30) parameterises $\tilde{K}$ in terms of the scalar parameters $c^{\alpha\beta}_{\boldsymbol{v}}$, while the basis functions are now matrix-valued

$$\mathcal{B}^{\alpha\beta}_{\boldsymbol{nlm}}(\mathbf{R}_{\bullet}) = \sum_{\boldsymbol{\mu}} \mathcal{U}^{\alpha\beta}_{\mathbf{l\mu m}} \boldsymbol{A}^{\bullet}_{\boldsymbol{nl\mu}}(\mathbf{R}_{\bullet}). \tag{32}$$

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 overcomplete. This linear dependence arises exactly within fixed $\boldsymbol{nl}$ blocks. In a straightforward adaption of the general procedures outlined by Dusson et al. [29] 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

$$\mathcal{B}_{\boldsymbol{nl}k}(\mathbf{R}_{\bullet}) := \sum_{\boldsymbol{\mu}} \mathcal{U}^{\boldsymbol{nl}}_{k\boldsymbol{\mu}} \boldsymbol{A}^{\bullet}_{\boldsymbol{nl\mu}}(\mathbf{R}_{\bullet}). \tag{33}$$

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

$$\tilde{\boldsymbol{H}}_{\mathrm{on}} := \mathbf{c}^{\mathrm{on}} \cdot \mathcal{B}^{\mathrm{on}}, \tag{34}$$

$$\tilde{\boldsymbol{H}}_{\mathrm{off}} := \mathbf{c}^{\mathrm{off}} \cdot \tilde{\mathcal{B}}^{\mathrm{off}}, \tag{35}$$

$$\tilde{\boldsymbol{S}}_{\mathrm{off}} := \mathbf{c}^{\mathrm{S}} \cdot \tilde{\mathcal{B}}^{\mathrm{S}}, \tag{36}$$

all of which inherit exactly the translation and permutation invariance as well as O(3)-equivariance of $\boldsymbol{H}_{\mathrm{on}}, \boldsymbol{H}_{\mathrm{off}}, \boldsymbol{S}_{\mathrm{off}}$. 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*. After imposing the symmetries outlined above we still need to ensure self-adjointness of the assembled Hamiltonian and overlap operators which we achieve by simply substituting $\tilde{\boldsymbol{H}} \leftarrow \frac{1}{2}(\tilde{\boldsymbol{H}} + \tilde{\boldsymbol{H}}^{*})$, and analogously for the overlap.

Validation

We generated DFT data for FCC and BCC aluminium, and followed the procedure outlined above to construct ACE models for the Hamiltonian and overlap using several choices of basis sets. Full details of data generation, parameter estimation and prediction procedures are given in the “Methods” section.

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.

Each basis set is defined by three parameters: the correlation order v and the maximum polynomial degrees $n_{\mathrm{max}}, l_{\mathrm{max}}$ used in both the radial basis functions $P_{nl}(r)$ and the angular basis function $Y_{lm}(\hat{\boldsymbol{x}})$ of Eqs. (15) and (20). In all our tests, the polynomial degrees are truncated in the manner of total degree, i.e., we let $n + l \le d_{\mathrm{max}}$ for a given d_{max}.

For the onsite models, the body order is one more than the correlation order, i.e. $v = 1$ corresponds to two body and $v = 2$ 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 v particles from the environment). The offsite model has further flexibility in that one can choose different d_{max} for bond and environment, say, $d^{\mathrm{b}}_{\mathrm{max}}$ and $d^{\mathrm{e}}_{\mathrm{max}}$. To avoid overemphasising the impact of environment, we set $d^{\mathrm{e}}_{\mathrm{max}} = \lceil d^{\mathrm{b}}_{\mathrm{max}}/2 \rceil$ 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 $v = 2$ (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. 3a. 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 ~2% relative error.

For offsite blocks $\boldsymbol{H}_{IJ}$ we considered models with correlation orders of both $v = 1$ (body order 3), Fig. 3b, and $v = 2$ (body order 4), Fig. 3c. Both approaches show good convergence in the accuracy of the training set as the maximum degree is increased. However, for sub-blocks 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* sub-blocks 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 order 2 models on the training set, so to avoid issues of overfitting we use order 1 only for $\boldsymbol{H}_{IJ}$, and also limit the maximum polynomial degree for individual sub-blocks as discussed in more detail in the section “Cross-validation and model selection”.

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 overfitting for the offsite overlap so we fixed the maximum polynomial degree for $\boldsymbol{S}_{IJ}$ at 16, the highest value we tried.

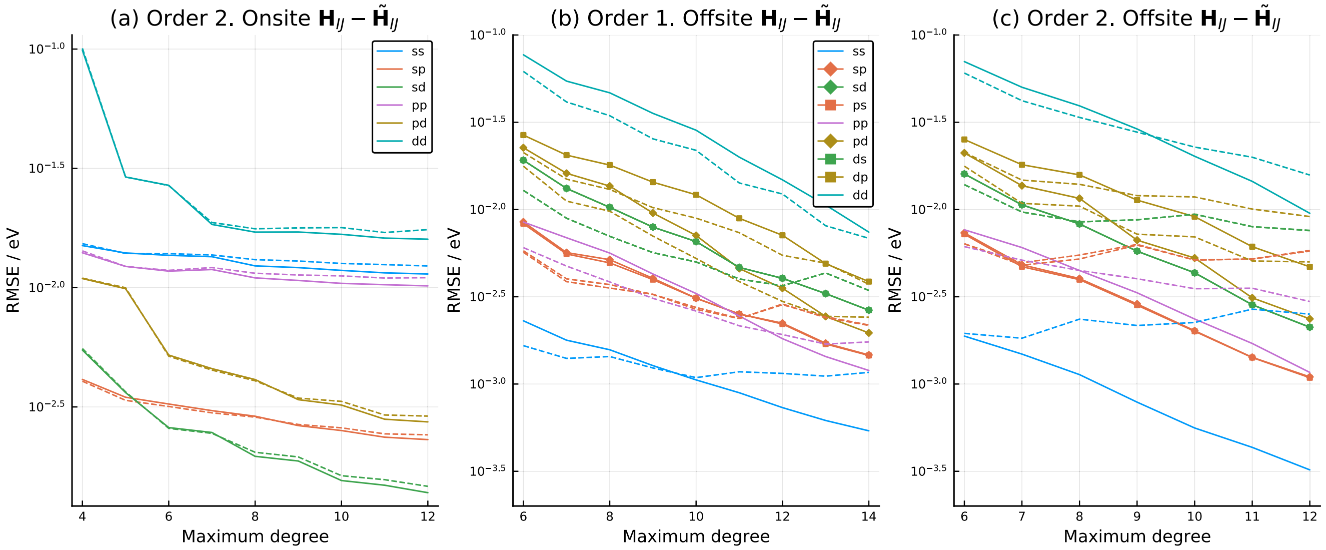

Fig. 3 Convergence of Hamiltonian and overlap blocks with respect to the order and maximum degree of the ACE basis set. a Onsite Hamiltonian blocks $\boldsymbol{H}_{II}$ fitted with order 2 models of varying maximum degree. **b** Offsite Hamiltonian blocks with order 1 ACE models. **c** Offsite Hamiltonian 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.

Table 1. ACE basis set parameters for our optimised models for $\boldsymbol{H}_{II}$, $\boldsymbol{H}_{IJ}$ and $\boldsymbol{S}_{IJ}$.

Onsite Hamiltonian H_{II}		
Correlation order ν		2
Cutoff radius $r_{\rm cut}$		10 Å
Maximum polynomial degree $d_{\max}$		9
Regularisation λ		10^{-7}
Offsite Hamiltonian $\boldsymbol{H}_{IJ}$		
Correlation order ν		1
Bond cutoff radius $r^{\rm b}_{\rm cut}$		10 Å
Env. cutoff radius $r^{\rm e}_{\rm cut}$		5 Å
Env. cutoff radius $z^{\rm e}_{\rm cut}$		5 Å
Maximum polynomial degree $d^{\rm b}_{\max}$		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 $r_{\rm cut}$		10 Å
Maximum polynomial degree $d_{\max}$		16
Regularisation λ		10^{-7}

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.

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 1. 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, $3 \times 2 = 6sp$ models, and $2 \times 2 = 4pp$ models. For the $3 \times 3 = 9ss$ 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 ~ 1% or less).

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 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 $9 \times 9 \times 9k$-point mesh and also shows excellent agreement

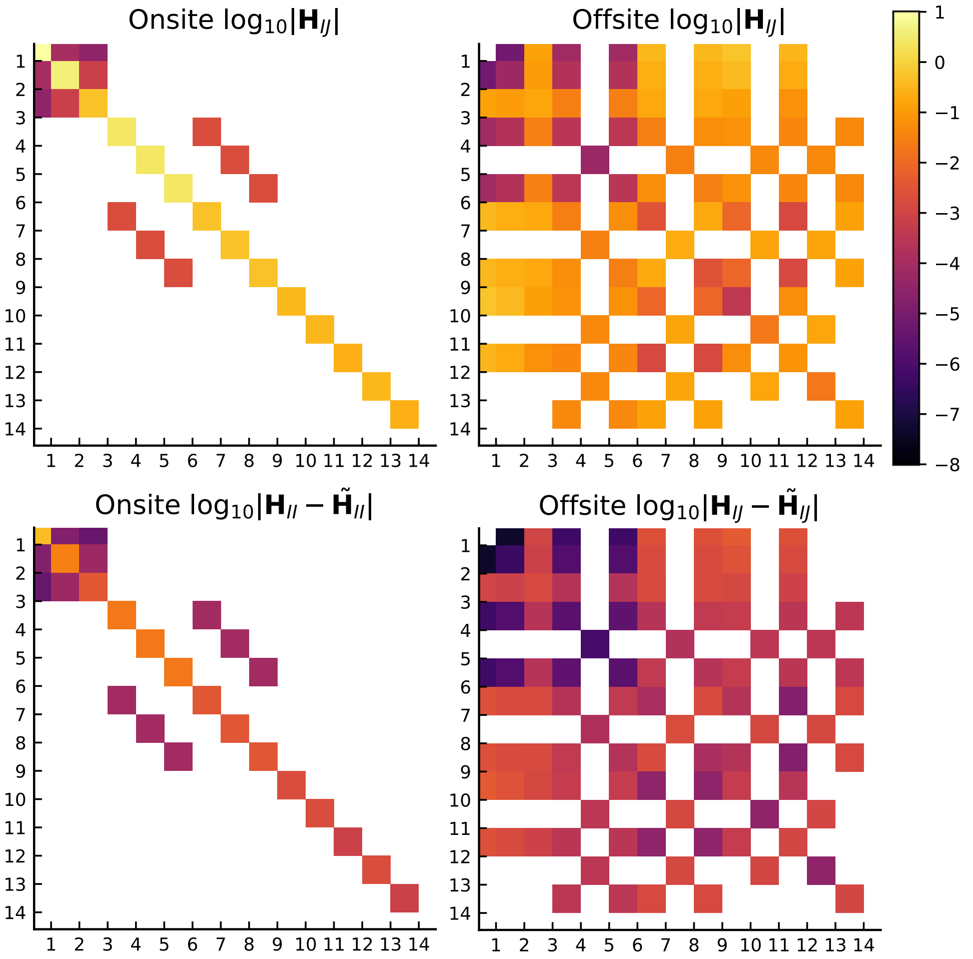

Fig. 4 Accuracy of predicted Hamiltonian blocks for the FCC crystal. Magnitudes (above) and errors (below) for onsite (left) and offsite (right) $\tilde{\boldsymbol{H}}$ for prediction on the FCC ground state unit cell (not included in the training set).

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 predict electronic observables.

The figure also shows confidence intervals for the predicted band structures. These have been estimated to leading order using a simple a priori error analysis to propagate errors in the Hamiltonian $\Delta\boldsymbol{H} = \tilde{\boldsymbol{H}} - \boldsymbol{H}$ and overlap $\Delta\boldsymbol{S} = \tilde{\boldsymbol{S}} - \boldsymbol{S}$ to expected errors in the bands using the result[44]

$$\tilde{\epsilon} - \epsilon \sim \langle \phi | \Delta\boldsymbol{H} - \epsilon \Delta\boldsymbol{S} | \phi \rangle \tag{37}$$

in the limit as $\Delta H, \Delta S \to 0$, where ϕ, ϵ and $\tilde{\phi}$, $\tilde{\epsilon}$ are eigenfunctions and eigenvalues of the reference and approximated systems, respectively. 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.

Figure 6 shows the convergence of band structures and DoS with respect to the maximum polynomial degree used in the ACE basis set, and for two choices of correlation order $v = 1$ and $v = 2$.

The error in the DoS is computed using the first Wasserstein (or 'earthmover') distance between the reference and predicted DoS, which is a natural metric for comparing densities of states since it is a distance between probability distributions (see, e.g., ref. [45]). The error in band structures is defined as the RMSE in the k-dependent band energies

$$E_{\rm band}(\mathbf{k}) = \sum_{i=1}^{N_{\rm orb}} f\left(\frac{\epsilon_i - \varepsilon_F}{\sigma}\right)\epsilon_i(\mathbf{k}) \tag{38}$$

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

The models with untuned parameters shown with the solid lines and dashed lines in Fig. 6 are already sufficiently accurate to produce good band structures and densities of states. However, when increasing the maximum degree used for all subblocks simultaneously, some overfitting can be seen, similar to that observed in the direct validation results of Fig. 3, and once again this arises at lower degrees of 9–12 with $v = 2$ than with $v = 1$, where maximum 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 the section "Cross-validation and model selection", shown with the horizontal dotted lines in the figure to produce band structures with a RMSE of <0.4 eV for both phases.

BCC to FCC transition

As a 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 and compared them to reference values computed with FHI-aims for the same systems. As can be seen in Fig. 7, the predicted electronic structure agrees well at all points along the path, suggesting good extrapolative behaviour beyond

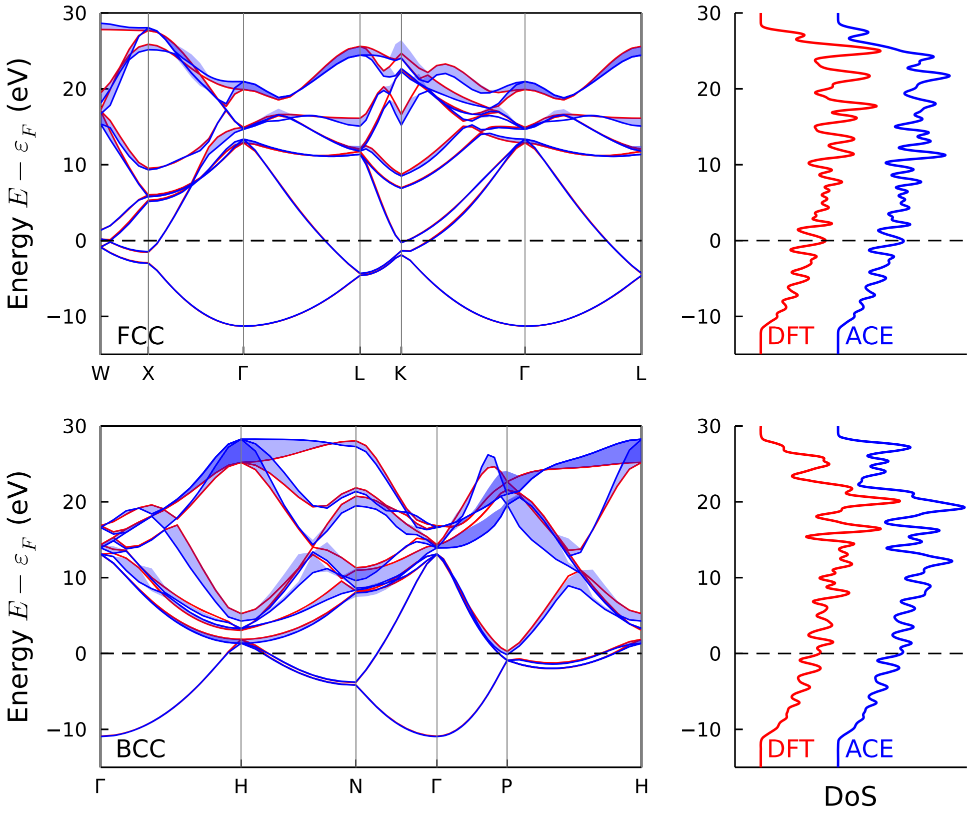

Fig. 5 FCC and BCC band structures obtained with DFT (red) and predicted by an ACE model with onsite H order 2, and offsite H and S order 1 (blue). Confidence intervals shown with blue ribbons are from a priori analysis of the errors in band spectrum expected to result from known errors in $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{S}}$ (see text). Energies are shown relative to the DFT Fermi level.

the training set, which includes only environments accessible from the two minima at moderate temperatures during MD.

Notably, nowhere along the path is the accuracy of the ACE model worse than it is for FCC or BCC, although accuracy drops off outside the BCC–FCC range $1/\sqrt{2}<c/a<1$. We interpret this as meaning that along the Bain path we see different global structures, but similar local environments, whereas to the left of BCC and to the right of FCC we go outside the range of local environments included in the training set.

Restricted training databases

To further test how well the model generalises across crystal systems, we carried out two further fits using the same optimal parameters as for the final model presented above, but with the training database restricted to either FCC only or BCC only configurations (using subsets of the same MD-generated structures as above). We then checked the ability of the resulting ACE Hamiltonian models to predict the DFT electronic structure of both crystals. The results, illustrated in Fig. 8 and summarised in Table 2 convincingly demonstrate the approach has excellent transferability, since the FCC DoS (and also the associated full band structure) can be accurately predicted using only BCC training data, and vice versa.

Defected structures

As a final test of our models' ability to predict outside of the domain of the training sets, we predicted the electronic structure of a 728 atom $9\times9\times9$ FCC Aluminium supercell containing a single vacancy. The structure was obtained by deleting an atom from the supercell and performing a geometry optimisation with FHI-aims until the maximum force was $<5\times10^{-3}$ eV/Å. We then compared the projected DoS (PDoS) for the atomic orbitals neighbouring the vacancy as obtained with DFT with the predictions of our optimal ACE Hamiltonian model, without refitting. The DFT and ACE PDoS are shown in Fig. 9 and demonstrate convincingly that our model is able to capture the changes in the local electronic structure associated with the introduction of a defect, indicating that it correctly predicts the self-consistent field (SCF) relaxations of the Hamiltonian without a need for an explicit SCF loop in the approximate scheme.

DISCUSSION

We have reported a 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 our methodology to produce accurate predictions for the band structure in aluminium in both FCC and BCC phases from limited training data. The 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 the door to the high-throughput computation of quantities which depend on electronic properties, such as photoemission spectra, transport coefficients, and electron–phonon coupling constants, all of which can currently only be accurately computed with first principles methods[46].

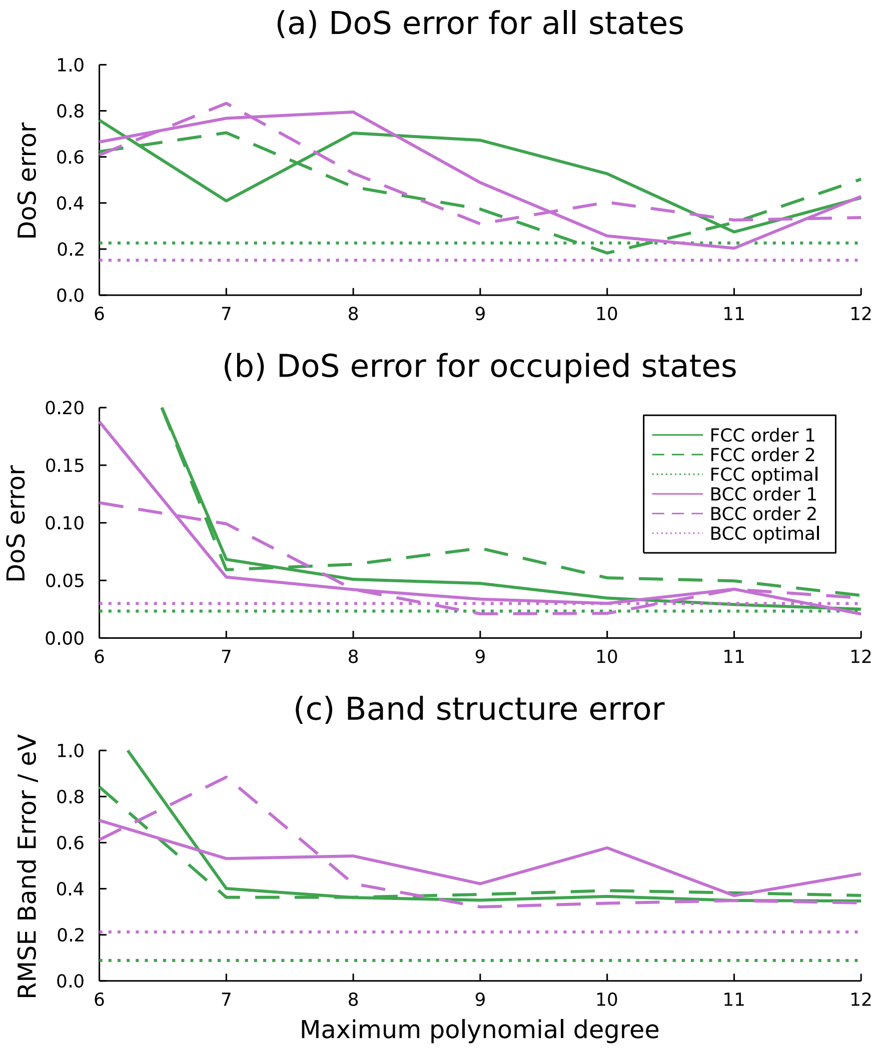

Fig. 6 Convergence of FCC and BCC band structures and DoS with respect to the correlation order v and maximum polynomial degree d_{max} used in the ACE basis set. Dotted lines show the optimised model of the section "Cross-validation and model selection". **a** Error in the full DoS. **b** Error in the occupied states, i.e. those below the Fermi level. **c** Band error computed with Eq. (38).

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[38]. Moreover, Bayesian approaches 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 through uncertainty propagation.

Further comprehensive studies of the dependence of accuracy and transferability of models on quantity and type of training data, as well as an extension to materials and systems 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.

METHODS

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)[25]. We used the Perdew–Burke–Enzerhof (PBE) generalised gradient approximation[47] to the exchange-correlation energy within the KS-DFT formulation, and neglected 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$_0^3$, 5×10^{-5}, and 10^{-6} eV, respectively. The default *tight* FHI-aims basis set and integration grid definitions were used, which uses 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 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. 9b.

The optimal equilibrium lattice constants for FCC and BCC Al were determined in primitive cells with a 9 × 9 × 9 Monkhorst–Pack **k**-point mesh[48] to be 4.05 and 3.29 Å, respectively. To sample a variety of distorted atomic configurations for Al, we carried out molecular dynamics (MD) simulations at a temperature of 500 K using

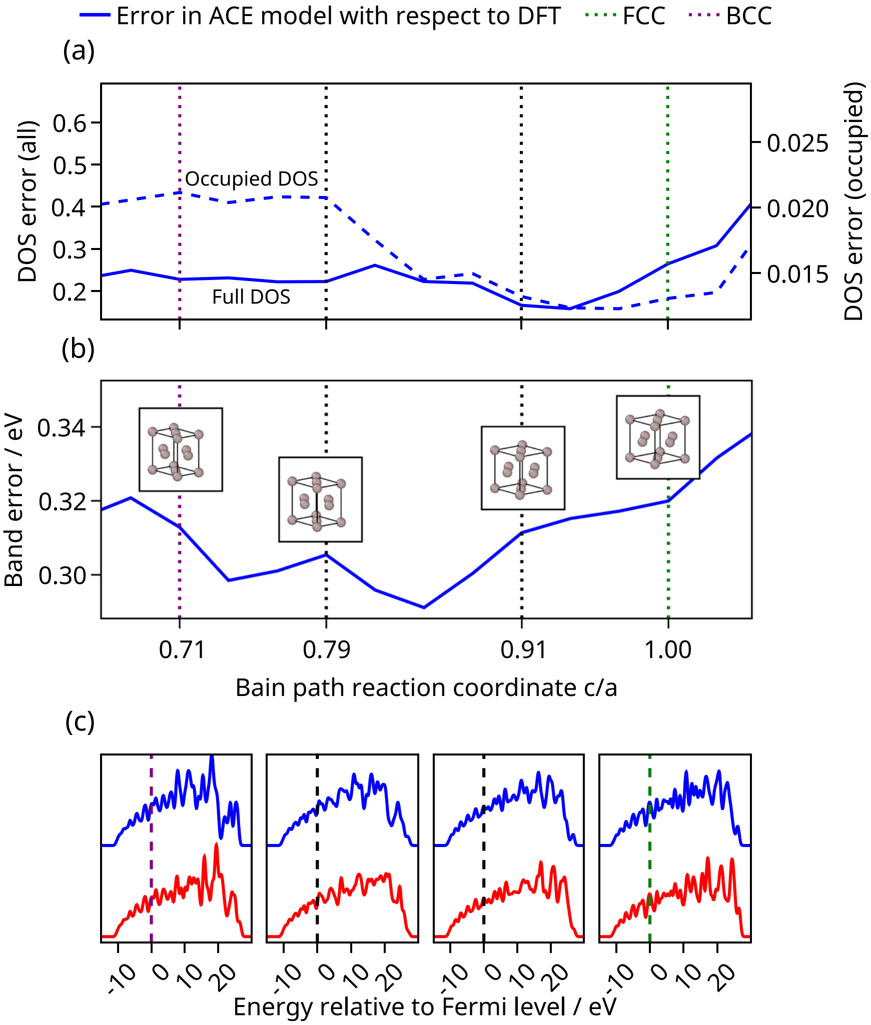

Fig. 7 Electronic structure along the transition from BCC $c/a = 1/\sqrt{2} \approx 0.71$ to FCC $c/a = 1$. **a** Error in the density of states made by our ACE model with respect to the DFT reference (measured with the Wasserstein distance) as a function of *c*/*a* along the Bain path. The solid line shows the full error in the DoS (right vertical axis), while the dashed line shows the error in the occupied states (left vertical axis; note the change of scale). **b** RMSE error in the electronic band structure (along high-symmetry *k*-path for the BCC structure) as a function of *c*/*a* along the Bain path. Insets illustrate the structure of the cubic cell at points along the path. **c** Comparisons of densities of state for the ACE model (blue) and DFT (red) at four points along the path, including the BCC (left) and FCC (right) structures.

$9 \times 9 \times 9 = 729$ atom supercells of the primitive FCC and BCC unit cells. MD simuations for each phase were performed in the *NPT* ensemble using a 5 fs timestep and the embedded atom method (EAM) potential proposed by Zhou et al. [49]. 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

$$\{(H_{II}, \mathbf{R}_I)\}, \tag{39}$$

$$\{(H_{IJ}, \boldsymbol{r}_{IJ}, \mathbf{R}_{IJ})\}, \tag{40}$$

$$\{(S_{IJ}, \boldsymbol{r}_{IJ}, \mathbf{R}_{IJ})\}. \tag{41}$$

where *II*, *IJ* indicate on- and off-site blocks of the Hamiltonian and overlap matrices while $\boldsymbol{r}_{\bullet}$ and $\mathbf{R}_{\bullet}$ are the corresponding atomic structure data as defined in the section "Hamiltonians for extended materials in atomic orbital basis representation". For the optimised model reported in the section "Cross-validation and model selection", we used 1000 training and 1000 test blocks for the onsite part of the Hamiltonian and 2000 training and 2000 test blocks for the offsite Hamiltonian and overlap matrices (with more offsite than onsite data to reflect the far greater number of offsite blocks in the target matrices). Equal numbers of samples were taken from the FCC and BCC MD data.

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

$$L_0(\boldsymbol{c}) = \sum_\tau |\boldsymbol{K}_*^{(\tau)} - \tilde{\boldsymbol{K}}(\mathbf{R}_{\bullet}^{(\tau)})|^2. \tag{42}$$

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

$$L_0(\boldsymbol{c}) = \|\Psi \boldsymbol{c} - \boldsymbol{y}\|^2, \tag{43}$$

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,

$$L_\lambda(\boldsymbol{c}) := \|\Psi \boldsymbol{c} - \boldsymbol{y}\|^2 + \lambda \|\Gamma \boldsymbol{c}\|^2, \tag{44}$$

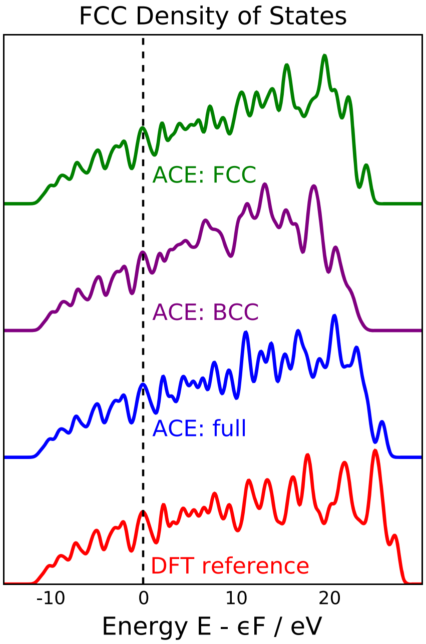

Fig. 8 Comparison of FCC and BCC DoS predicted with ACE models for full and restricted training databases. The reference DFT DoS is shown in red. A vertical shift has been applied to separate the DoS for each ACE model.

Table 2. Errors in the FCC and BCC DoS predicted with ACE models for full and restricted training databases.

Crystal	Training database	DoS error (all)	DoS error (occ.)
FCC	FCC+BCC	0.424	0.015
FCC	BCC	0.930	0.081
FCC	FCC	0.732	0.044
BCC	FCC+BCC	0.308	0.023
BCC	BCC	0.550	0.041
BCC	FCC	0.311	0.025

The error in the full DoS and in the occupied states below the Fermi level are reported.

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[29,38] and λ is a regularisation parameter. Throughout this work, we define Γ_{kk} by

$$\Gamma_{kk} = \sum_v (n_v^2 + l_v^2 + m_v^2), \tag{45}$$

and λ is always set to be 10^{-7}. We then solve the regularised least squares system (44) using an iterative LSQR algorithm with termination tolerance 10^{-6}.

For the radial basis set P_{nl} we used

$$\xi(r) = \left(\frac{1+r_0}{1+r}\right)^2 \tag{46}$$

$$P_{nl}(r) = Q_n(\xi(r)) \tag{47}$$

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 $[\xi_0, \xi_1] = \xi([0, r_{\mathrm{cut}}])$; see ref. [29] for full details.

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

$$f_{\mathrm{cut}}(r; r_{\mathrm{cut}}) = f^{\mathrm{b}}_{\mathrm{cut}}(r; r_{\mathrm{cut}}) = \begin{cases} (r^2/r_{\mathrm{cut}}^2 - 1)^2, & r \le r_{\mathrm{cut}}, \\ 0, & r > r_{\mathrm{cut}}, \end{cases} \tag{48}$$

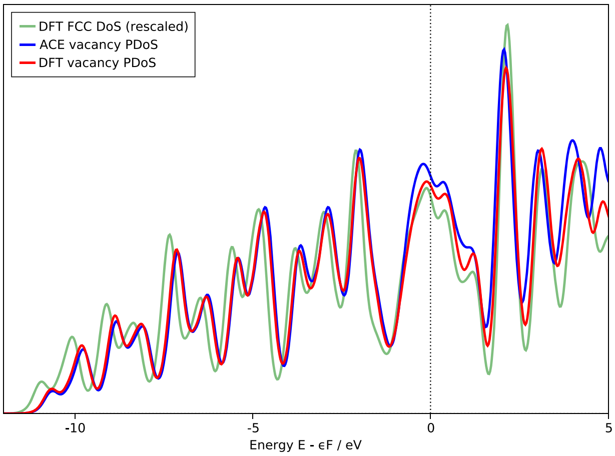

Fig. 9 Comparison of PDoS for a 728-atom Al vacancy supercell between the reference DFT results (red) and those predicted by our ACE Hamiltonian model (blue), which was not trained on vacancy data. PDoS includes orbitals associated with the nearest neighbours of the vacancy. The PDoS for the perfect FCC structure is shown in green to allow the changes in the electronic structure due to the defect to be assessed.

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

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

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

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

As noted above, we used correlation order $v = 0$ for the offsite overlap $\boldsymbol{S}_{II}$ since these blocks are not environment-dependent. For $\boldsymbol{H}_{II}$ we used correlation order $v = 2$ throughout, while for $\boldsymbol{H}_{IJ}$ we tested correlation orders of both $v = 1$ and $v = 2$. 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 detail in the section "Methods" in the main text.

Prediction

The software implementation of our method follows the workflow illustrated in Fig. 1. The Julia packages `ACE.jl`[50] and `ACEhamiltonians.jl` implement the general Atomic Cluster Expansion basis sets and the specialisation to fitting and predicting Hamiltonians, respectively. Given an input configuration **R** we use the scheme described above to predict $\tilde{\boldsymbol{H}}_{\mathrm{on}}(\mathbf{R}), \tilde{\boldsymbol{H}}_{\mathrm{off}}(\mathbf{R})$ and $\tilde{\boldsymbol{S}}_{\mathrm{off}}(\mathbf{R})$. We then 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 `ACEhamiltonians.jl` package.

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

$$\boldsymbol{H}(\mathbf{k})\phi_i = \epsilon_i \boldsymbol{S}(\mathbf{k})\phi_i \tag{49}$$

$$\tilde{\boldsymbol{H}}(\mathbf{k})\tilde{\phi}_i = \tilde{\epsilon}_i \tilde{\boldsymbol{S}}(\mathbf{k})\tilde{\phi}_i \tag{50}$$

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 $i = 1, \ldots, N_{\mathrm{orb}}$ and in this work $N_{\mathrm{orb}} = 14$. Band structures, the density of states (DoS) and other derived quantities can be computed by post-processing the band energies following standard practices.

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:

$$\mathbf{H}(\boldsymbol{k})\psi_{i\boldsymbol{k}} = \epsilon_{i\boldsymbol{k}}\mathbf{S}(\boldsymbol{k})\psi_{i\boldsymbol{k}} \quad i = 1, 2, \cdots \tag{51}$$

where $\psi_{v\boldsymbol{k}}$ 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:

$$\chi_{a\boldsymbol{k}}(\boldsymbol{x}) = \sum_N \exp\{i\boldsymbol{k} \cdot \mathbf{NL}\}\chi_a(\boldsymbol{x} + \mathbf{NL}). \tag{52}$$

In Eq. (52), **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

$$H_{ab}(\boldsymbol{k}) = \langle \chi_{a\boldsymbol{k}} | \hat{H} | \chi_{b\boldsymbol{k}} \rangle = \tag{53}$$

$$\sum_{N,N'} \exp\{i\boldsymbol{k} \cdot (\mathbf{N}' - \mathbf{N}) \cdot \mathbf{L}\} \underbrace{\langle \chi_{a,N'} | \hat{H} | \chi_{b,N} \rangle}_{= H_{ab}(\mathbf{N},\mathbf{N}')} \tag{54}$$

and

$$S_{ab}(\boldsymbol{k}) = \langle \chi_{a\boldsymbol{k}} | \chi_{b\boldsymbol{k}} \rangle = \tag{55}$$

$$\sum_{N,N'} \exp\{i\boldsymbol{k} \cdot (\mathbf{N}' - \mathbf{N}) \cdot \mathbf{L}\} \underbrace{\langle \chi_{a,N'} | \chi_{b,N} \rangle}_{= S_{ab}(\mathbf{N},\mathbf{N}')} \tag{56}$$

where $H_{ab}(\mathbf{N}, \mathbf{N}')$ and $S_{ab}(\mathbf{N}, \mathbf{N}')$ are as defined in Eqs. (6) and (7) for atomic orbitals defined in different unit cells **N** and **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 localised Wannier functions. We then calculate eigenvalues $\epsilon_{v\boldsymbol{k}}$ at arbitrary points in reciprocal space to calculate converged electronic densities-of-state and band structures.

Equivariance of $\boldsymbol{H}_{IJ}$

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

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

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

$$\chi_{(n,l,m;I)}(Q\boldsymbol{x}; Q\mathbf{R}) = \sum_\mu D^l_{\mu m}\chi_{(n,l,\mu;I)}(\boldsymbol{x}; \mathbf{R}). \tag{57}$$

Here, $\chi_{\boldsymbol{\cdot}}$ is **R**-dependent since it is atom-centred. Besides,

$$\begin{aligned}
&\boldsymbol{H}^{\alpha\beta}_{IJ}(Q\mathbf{R}) \\
&\quad = \int_{\mathbb{R}^3}\chi_a(\boldsymbol{x}; Q\mathbf{R})^* V_{\mathrm{eff}}(\boldsymbol{x}, Q\mathbf{R})\chi_b(\boldsymbol{x}; Q\mathbf{R})d\boldsymbol{x} \\
&\quad = \int_{\mathbb{R}^3}\chi_a(Q\boldsymbol{x}; Q\mathbf{R})^* V_{\mathrm{eff}}(Q\boldsymbol{x}, Q\mathbf{R})\chi_b(Q\boldsymbol{x}; Q\mathbf{R})d\boldsymbol{x} \\
&\quad = \int_{\mathbb{R}^3}\chi_a(Q\boldsymbol{x}; Q\mathbf{R})^* V_{\mathrm{eff}}(\boldsymbol{x}, \mathbf{R})\chi_b(Q\boldsymbol{x}; Q\mathbf{R})d\boldsymbol{x}.
\end{aligned} \tag{58}$$

Combining Eqs. (57) and (58), we see immediately that

$$\boldsymbol{H}_{IJ}(Q\mathbf{R}) = D(Q)^* \boldsymbol{H}_{IJ}(\mathbf{R})D(Q), \tag{59}$$

where

$$D(Q) = \mathrm{Diag}(D^{l_1}(Q), D^{l_2}(Q), \cdots), \tag{60}$$

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

Sometimes, the angular term in Eq. (5) is chosen to use real spherical harmonics rather than complex ones, i.e.,

$$\chi_a(\boldsymbol{x}) = R_{nl}(r)S_{lm}(\theta, \phi), \text{ and} \tag{61}$$

$$S_{lm} = \sum_{m'} C_{mm'}Y_{lm}, \tag{62}$$

where $\{C_{mm'}\}$ are the corresponding transforming coefficients. Equivalently, we may go through all possible indices m with respect to a fixed l and obtain the following matrix form:

$$\mathbf{S}_l(\mathbf{R}) = C_l\mathbf{Y}_l(\mathbf{R}). \tag{63}$$

In this case, the equivariance of $\boldsymbol{H}_{IJ}$ simply follows, just with the varied equivariant matrix

$$\tilde{D}(Q) = \mathrm{Diag}(\tilde{D}^{l_1}(Q), \tilde{D}^{l_2}(Q), \cdots), \tag{64}$$

and $\tilde{D}^{l_i}(Q) = C_{l_i}D^{l_i}(Q)$.

DATA AVAILABILITY

Supporting data for this manuscript comprising the electronic structure training data, ACE Hamiltonian and overlap models, prediction results and an archived copy of the source code is available from https://doi.org/10.5281/zenodo.6561452.

CODE AVAILABILITY

The `ACE.jl` package which implements the Atomic Cluster Expansion basis sets used here is available from https://github.com/acesuit/ACE.jl. The `ACEhamiltonians.jl` package which extends its capabilities to learning Hamiltonian and overlap matrices is available from https://github.com/ACEsuit/ACEhamiltonians.jl/tree/arXiv.2111.13736; with examples provided at https://github.com/ACEsuit/ACEhamiltoniansExamples.

Received: 20 January 2022; Accepted: 5 July 2022;
Published online: 22 July 2022

REFERENCES

1. Bitzek, E., Kermode, J. R. & Gumbsch, P. Atomistic aspects of fracture. *Int. J. Fract.* **191**, 13–30 (2015).
2. Jiang, B. & Guo, H. Dynamics in reactions on metal surfaces: a theoretical perspective. *J. Chem. Phys.* **150**, 180901 (2019).
3. Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
4. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).

5. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
6. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
7. Mishin, Y. Machine-learning interatomic potentials for materials science. *Acta Mater.* **214**, 116980 (2021).
8. Behler, J. & Csányi, G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* **94**, 142 (2021).
9. Dewar, M. J., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. AM1: a new general purpose quantum mechanical molecular model1. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).
10. Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **10**, 209–220 (1989).
11. 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* **51**, 12947 (1995).
12. Elstner, M. et al. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **58**, 7260–7268 (1998).
13. Sankey, O. F. & Niklewski, D. J. Ab initio multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems. *Phys. Rev. B* **40**, 3979–3995 (1989).
14. Lewis, J. P. et al. Further developments in the local-orbital density-functional-theory tight-binding method. *Phys. Rev. B* **64**, 195103 (2001).
15. 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.* **15**, 1652–1671 (2019).
16. Westermayr, J., Gastegger, M., Schütt, K. T. & Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *J. Chem. Phys.* **154**, 230903 (2021).
17. Li, H., Collins, C., Tanha, M., Gordon, G. J. & Yaron, D. J. A density functional tight binding layer for deep learning of chemical hamiltonians. *J. Chem. Theory Comput.* **14**, 5764–5776 (2018).
18. Stöhr, M., Medrano Sandonas, L. & Tkatchenko, A. Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks. *J. Phys. Chem. Lett.* **11**, 6835–6843 (2020).
19. Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & MillerIII, T. F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
20. Supka, A. R. et al. AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians. *Comput. Mater. Sci.* **136**, 76–84 (2017).
21. Garrity, K. F. & Choudhary, K. Database of wannier tight-binding hamiltonians using high-throughput density functional theory. *Sci Data* **8**, 106 (2021).
22. Marzari, N., Mostofi, A. A., Yates, J. R., Souza, I. & Vanderbilt, D. Maximally localized wannier functions: theory and applications. *Rev. Mod. Phys.* **84**, 1419–1475 (2012).
23. Barzdajn, B., Garrett, A. M., Whiting, T. M. & Race, C. P. Development of data-driven spd tight-binding models of fe-parameterisation based on qsgw and dft calculations including information about higher-order elastic constants. *Model. Simul. Mater. Sci. Eng.* **29**, 085006 (2021).
24. Jenke, J., Ladines, A. N., Hammerschmidt, T., Pettifor, D. G. & Drautz, R. Tight-binding bond parameters for dimers across the periodic table from density-functional theory. *Phys. Rev. Materials* **5**, 023801 (2021).
25. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Commun.* **180**, 2175–2196 (2009).
26. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
27. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
28. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
29. Dusson, G. et al. Atomic cluster expansion: completeness, efficiency and stability. *J. Comp. Phys.* **454**, 110946 (2022).
30. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
31. Gastegger, M., McSloy, A., Luya, M., Schütt, K. T. & Maurer, R. J. A deep neural network for molecular wave functions in quasi-atomic minimal basis representation. *J. Chem. Phys.* **153**, 044123 (2020).
32. Hegde, G. & Bowen, R. C. Machine-learned approximations to density functional theory hamiltonians. *Sci. Rep.* **7**, 42669 (2017).
33. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
34. Nigam, J., Willatt, M. J. & Ceriotti, M. Equivariant representations for molecular hamiltonians and n-center atomic-scale properties. *J. Chem. Phys.* **156**, 014115 (2022).
35. Unke, O. et al. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *NeurIPS* **34**, 14434–14447 (2021).
36. Cancès, E., Kemlin, G. & Levitt, A. Convergence analysis of direct minimization and self-consistent iterations. *SIAM J. Matrix Anal. Appl.* **42**, 243–274 (2021).
37. Woods, N. D., Payne, M. C. & Hasnip, P. J. Computing the self-consistent field in kohn-sham density functional theory. *J. Phys.: Condens. Matter* **31**, 453001 (2019).
38. Lysogorskiy, Y. et al. Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *npj Comput. Mater.* **7**, 1–12 (2021).
39. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661–29668 (2018).
40. Nigam, J., Pozdnyakov, S. & Ceriotti, M. Recursive evaluation and iterative contraction of *n*-body equivariant features. *J. Chem. Phys.* **153**, 121101 (2020).
41. Drautz, R. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B* **102**, 024104 (2020).
42. Grisafi, A. & Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **151**, 204105 (2019).
43. Slater, J. C. & Koster, G. F. Simplified LCAO method for the periodic potential problem. *Phys. Rev.* **94**, 1498–1524 (1954).
44. Crandall, M. G. & Rabinowitz, P. H. *Bifurcation, Perturbation of Simple Eigenvalues and Linearized Stability* (University of Wisconsin-Madison, Mathematics Research Center, 1973).
45. Ben Mahmoud, C., Anelli, A., Csányi, G. & Ceriotti, M. Learning the electronic density of states in condensed matter. *Phys. Rev. B* **102**, 235130 (2020).
46. Knoop, F., Purcell, T. A. R., Scheffler, M. & Carbogno, C. Anharmonicity measure for materials. *Phys. Rev. Mater.* **4**, 083809 (2020).
47. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
48. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin zone integration. *Phys. Rev. B* **13**, 5188–5192 (1976).
49. Zhou, X. W., Johnson, R. A. & Wadley, H. N. G. Misfit-energy-increasing dislocations in vapor-deposited CoFe/NiFe multilayers. *Phys. Rev. B* **69**, 144113 (2004).
50. Ortner, C. et al. ACE.jl: Approximation of symmetric functions with polynomials and spherical harmonics. https://github.com/ACEsuit/ACE.jl

ACKNOWLEDGEMENTS

This work was financially supported by a Leverhulme Trust Research Project Grant (RPG-2017-191), the Engineering and Physical Science Research Council (EPSRC) under grant 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+ consortium (EP/P020232/1, EP/T022108/1) and on ARCHER2 via the UK Car-Parinello consortium (EP/P022065/1). For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

AUTHOR CONTRIBUTIONS

R.J.M., C.O. and J.R.K. designed the research. G.A. and B.O. generated the training data. B.O. and A.M. developed the workflow infrastructure linking from DFT calculations to parametrised Hamiltonians. L.Z., G.D., and C.O. designed and implemented the equivariant basis sets. L.Z. and J.R.K. generated and validated the models for aluminium with input from all authors. All authors analysed the results, contributed to the manuscript and approved its final version.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to James R. Kermode.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.