npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Automated high-throughput Wannierisation

Valerio Vitale[1,2]✉, Giovanni Pizzi[3], Antimo Marrazzo[3], Jonathan R. Yates[4], Nicola Marzari[3] and Arash A. Mostofi[2]

Maximally-localised Wannier functions (MLWFs) are routinely used to compute from first-principles advanced materials properties that require very dense Brillouin zone integration and to build accurate tight-binding models for scale-bridging simulations. At the same time, high-throughput (HT) computational materials design is an emergent field that promises to accelerate reliable and cost-effective design and optimisation of new materials with target properties. The use of MLWFs in HT workflows has been hampered by the fact that generating MLWFs automatically and robustly without any user intervention and for arbitrary materials is, in general, very challenging. We address this problem directly by proposing a procedure for automatically generating MLWFs for HT frameworks. Our approach is based on the selected columns of the density matrix method and we present the details of its implementation in an AiiDA workflow. We apply our approach to a dataset of 200 bulk crystalline materials that span a wide structural and chemical space. We assess the quality of our MLWFs in terms of the accuracy of the band-structure interpolation that they provide as compared to the band-structure obtained via full first-principles calculations. Finally, we provide a downloadable virtual machine that can be used to reproduce the results of this paper, including all first-principles and atomistic simulations as well as the computational workflows.

npj Computational Materials (2020)6:66 ; https://doi.org/10.1038/s41524-020-0312-y

INTRODUCTION

The combination of modern high-performance computing, robust and scalable software for first-principles electronic structure calculations, and the development of computational workflow management platforms, has the potential to accelerate the design and discovery of materials with tailored properties using first-principles high-throughput (HT) calculations[1–4].

Wannier functions (WFs) play a key role in contemporary state-of-the-art first-principles electronic structure calculations. First, they provide a means by which to bridge lengthscales by enabling the transfer of information from the atomic scale (e.g., density-functional theory and many-body perturbation theory calculations) to mesoscopic scales at the level of functional nano-devices (e.g., tight-binding calculations with a first-principles-derived WF basis)[5,6]. Second, the compact WF representation provides a means by which advanced materials properties that require very fine sampling of electronic states in the Brillouin zone (BZ) may be computed at much lower computational cost, yet without any loss of accuracy, via Wannier interpolation[7].

Among several variants of WFs[8], maximally-localised Wannier functions (MLWFs), based on the minimisation of the Marzari–Vanderbilt quadratic spread functional Ω, are those most employed in actual calculations in the solid state[8]. One ingredient in the canonical minimisation procedure is the specification of a set of initial guesses for the MLWFs. These are typically trial functions localised in real-space that are specified by the user, based on their experience and chemical intuition. As shall be described in more detail later, in the case of an isolated manifold of bands, the final result for the MLWFs is almost always found to be independent of the choice of initial guess[9]. In the case of entangled bands[10], however, this tends not to be the case and the choice of initial guess strongly affects the quality of the final MLWFs, presenting a challenge to the development of a general-purpose approach to generating MLWFs automatically without user intervention.

Several approaches have been put forward to remove the necessity for user-intervention in generating MLWFs, including the iterative projection method of Mustafa et al.[11], the smooth orthonormal Bloch frames of Levitt et al.[12], and the automated construction of pseudo-atomic orbitals rather than WFs as the local basis to represent the target space, as described by Agapito et al.[13–15]. In addition, some *ad hoc* solutions have been proposed, whose range of applicability is focused onto specific classes of materials[16–19].

A recently proposed algorithm by Damle et al.[20,21], known as the selected columns of the density matrix (SCDM) method, has shown great promise in avoiding the need for user intervention in obtaining MLWFs. Based on QR factorisation with column pivoting (QRCP) of the reduced single-particle density matrix, SCDM can be used without the need for an initial guess, making the approach ideally suited for HT calculations. The method is robust, being based on standard linear-algebra routines rather than on iterative minimisation. Moreover, the authors have proposed an efficient algorithm for the QRCP factorisation that operates on a smaller and numerically more tractable matrix than the full density matrix. Finally, SCDM is parameter-free for an isolated set of composite bands, and requires only two parameters in the case of entangled bands together with the choice of the target dimensionality for the disentangled subspace (i.e., the number of MLWFs required). We emphasise here that the SCDM method can be seen as an extension to solid-state periodic systems of the Cholesky orbitals approach of Aquilante et al.[22], that has been developed from a quantum-chemistry molecular perspective for finite systems. SCDM focuses instead on periodic systems, and it is based on a real-space grid discretisation of the wavefunctions. We discuss in more detail this equivalence in the "The SCDM algorithm and its

[1]Cavendish Laboratory, Department of Physics, University of Cambridge, 19 JJ Thomson Avenue, Cambridge, UK. [2]Departments of Materials and Physics, and the Thomas Young Centre for Theory and Simulation of Materials, Imperial College London, London SW7 2AZ, UK. [3]Theory and Simulation of Materials (THEOS) and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [4]Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK. ✉email: vvitale@ic.ac.uk

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

physical interpretation" section and in the "Methods" section of the Supplementary Material.

In this article, we present a fully-automated protocol based on the SCDM algorithm for the construction of MLWFs, in which the two free parameters are determined automatically (in our HT approach the dimensionality of the disentangled space is fixed by the total number of states used to generate the pseudopotentials in the density functional theory (DFT) calculations). We have implemented the SCDM algorithm in the PW2WANNIER90 interface code between the Quantum ESPRESSO software package[23] and the WANNIER90 code[24]. We have used our implementation as the basis for a complete computational workflow for obtaining MLWFs and electronic properties based on Wannier interpolation of the BZ, starting only from the specification of the initial crystal structure. We have implemented our workflow within the AiiDA[25,26] materials informatics platform, and we used it to perform a HT study on a dataset of 200 materials.

We anticipate here that our scheme works extremely well for our purposes, i.e., band-structure interpolation of both insulating and metallic systems with Wannier functions, but is less suitable for other applications where, for instance, a specific symmetry character of the WFs is required. It is worth mentioning that there are other approaches for constructing Wannier functions, which are based on a minimisation procedure and therefore require an initial guess[27–29] and which could also be automated in a similar fashion. In this work however, we focus only on the automatic generation of maximally-localised Wannier functions. We also note that there exist efficient non-Wannier-based techniques for band-structure interpolation, e.g., Shirley interpolation[30,31]. Whilst these approaches have their own advantages, they do not provide the same insight afforded by a real-space, localised description of the electronic structure, which can often be very helpful for understanding and computing advanced properties.

The manuscript is organised as follows. First, we present a summary of the background theory, starting with MLWFs for isolated and entangled bands followed by the SCDM algorithm, where we focus in particular on providing a physical interpretation of the method. In the "Results and discussions" section, we first provide a preliminary comparison, for a few well-known materials, between MLWFs obtained via the conventional method (i.e., with user-defined initial guesses) and those obtained from SCDM. We then proceed to show the validation of the SCDM method and our workflow for the valence bands of 81 insulating materials. We then discuss our automated protocol to determine the free parameters in the case of entangled bands and validate it on a dataset of 200 semiconducting and metallic materials. Finally, details on the implementation of the SCDM method in PW2WANNIER90 and of the AiiDA workflow are presented in the "Methods" section.

We summarise in this section the main concepts and notations related to maximally-localised Wannier functions that will be useful in the rest of the paper, following the notation in ref. [8].

A Wannier function associated to a band n can be obtained via a unitary transformation of the Bloch state $|\psi_{n\mathbf{k}}\rangle$, known as Wannier transform[32]

$$|w_{\mathbf{R}n}\rangle = \frac{V}{(2\pi)^3}\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}\,|\psi_{n\mathbf{k}}\rangle e^{-i\mathbf{k}\cdot\mathbf{R}}, \tag{1}$$

where V is the real-space primitive cell volume, $\mathbf{R}$ is a Bravais lattice vector, and the integral is over the first BZ. For clarity of notation, we assume spin-degeneracy unless otherwise specified.

The gauge freedom of the Bloch state under multiplication by a k-dependent phase $e^{i\varphi_n(\mathbf{k})}$ results in a non-uniqueness in the definition of the Wannier function. Maximally-localised Wannier functions represent the choice of gauge in which the real-space quadratic spread of the Wannier function is minimised[8,9]. In order to obtain a minimal TB basis set it is therefore beneficial to select the optimal phases that minimise the total spread, so that overlaps and Hamiltonian matrix elements between different Wannier functions decay rapidly to zero as a function of the distance between their centres. Since the integral transformation in Eq. (1) is still a unitary transformation, the resulting $\{|w_{\mathbf{R}n}\rangle\}$ span the same Hilbert space as the original Bloch states $\{|\psi_{n\mathbf{k}}\rangle\}$. Moreover, from the orthogonality of the $|\psi_{n\mathbf{k}}\rangle$ readily follows the orthogonality of the $|w_{\mathbf{R}n}\rangle$, since unitary transformations preserve inner products. Finally, two WFs $|w_{\mathbf{R}n}\rangle$ and $|w_{\mathbf{R}'n}\rangle$ transform into each other under translation by the Bravais lattice vector $\mathbf{R}-\mathbf{R}'$[33].

For an isolated set of J bands describing, e.g., the valence bands of a semiconductor, the most general phase choice for a Wannier transform can be written as

$$|w_{\mathbf{R}n}\rangle = \frac{V}{(2\pi)^3}\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k}\left[\sum_{m=1}^{J}|\psi_{m\mathbf{k}}\rangle U^{(\mathbf{k})}_{mn}\right] e^{-i\mathbf{k}\cdot\mathbf{R}}, \tag{2}$$

where $\mathbf{U}^{(\mathbf{k})}$ is a unitary matrix that, at each wave vector $\mathbf{k}$, mixes Bloch states belonging to different bands, giving as a result a set of J composite WFs. The localisation of the WFs may be improved by choosing the unitary matrices $\mathbf{U}^{(\mathbf{k})}$ such that $|\widetilde{\psi}_{n\mathbf{k}}\rangle = \sum_m |\psi_{m\mathbf{k}}\rangle U^{(\mathbf{k})}_{mn}$ in Eq. (2) is as smooth as possible, i.e., analytic with respect to $\mathbf{k}$ (see, e.g., Duffin[34]). Different approaches have been put forward[35–39] to generate well-localised WFs. In the Marzari–Vanderbilt (MV) approach[9] $\mathbf{U}^{(\mathbf{k})}$ is chosen to minimise the sum of the quadratic spreads of the WFs, given by

$$\Omega = \sum_{n=1}^{J}\left[\langle(\mathbf{r}-\bar{\mathbf{r}}_n)^2\rangle_n\right] = \sum_{n=1}^{J}\left[\langle r^2\rangle_n - \bar{\mathbf{r}}_n^2\right], \tag{3}$$

where $\langle\cdot\rangle_n \equiv \langle w_{n\mathbf{0}}|\cdot|w_{n\mathbf{0}}\rangle$ and $\bar{\mathbf{r}}_n = \langle\mathbf{r}\rangle_n = \langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{0}}\rangle$ is the centre of the n-th Wannier function. The resulting WFs are known as maximally-localised Wannier functions (MLWFs), and are the solid-state equivalent of the Foster–Boys molecular orbitals[40–42] in quantum chemistry.

The total quadratic spread Ω may be separated into two positive-definite terms: $\Omega = \Omega_{\mathrm{I}} + \widetilde{\Omega}$, where

$$\Omega_{\mathrm{I}} = \sum_n\left[\langle r^2\rangle_n - \sum_{m\mathbf{R}}|\langle w_{m\mathbf{R}}|\mathbf{r}|w_{n\mathbf{0}}\rangle|^2\right] \tag{4}$$

and

$$\widetilde{\Omega} = \sum_n\sum_{m\mathbf{R}\neq n\mathbf{0}}|\langle w_{\mathbf{0}n}|\mathbf{r}|w_{\mathbf{R}m}\rangle|^2. \tag{5}$$

It can be shown that[8,9] Ω_{I} is gauge invariant, whereas $\widetilde{\Omega}$ depends on the particular choice of the gauge (i.e., on the choice of $\mathbf{U}^{(\mathbf{k})}$). For an isolated group of bands, therefore, Ω_{I} is evaluated once and for all in the initial gauge and minimising the total spread Ω is equivalent to minimising only the gauge-dependent part $\widetilde{\Omega}$.

For crystalline solids with translational symmetry, it is natural to work in reciprocal space, henceforth referred as k-space. Applying Blount's identities[33] for the representation of the position operator $\mathbf{r}$ and r^2 in k-space and discretising in $\mathbf{k}$ (on a uniform grid) gives[9]

$$\Omega_{\mathrm{I}} = \frac{1}{N_{\mathbf{k}}}\sum_{\mathbf{k},\mathbf{b}} w_b \sum_{m=1}^{J}\left[1 - \sum_{n=1}^{J}\left|M^{(\mathbf{k},\mathbf{b})}_{mn}\right|^2\right], \tag{6}$$

and

$$\widetilde{\Omega} = \frac{1}{N_{\mathbf{k}}}\sum_{\mathbf{k},\mathbf{b}} w_b \left[\sum_{n=1}^{J}\left(-\mathrm{Im}\ln M^{(\mathbf{k},\mathbf{b})}_{nn} - \mathbf{b}\cdot\bar{\mathbf{r}}_n\right)^2 + \sum_{m\neq n}\left|M^{(\mathbf{k},\mathbf{b})}_{mn}\right|^2\right], \tag{7}$$

where the vectors $\{\mathbf{b}\}$ connect a BZ mesh point $\mathbf{k}$ to its nearest neighbours $\mathbf{k}+\mathbf{b}$, the associated weights w_b come from the finite difference representation of the gradient operator in k-space (a result of the change of representation $\mathbf{r}\to i/\hbar\nabla_{\mathbf{k}}$), and $\mathbf{M}^{(\mathbf{k},\mathbf{b})}$ is given by

$$M^{(\mathbf{k},\mathbf{b})}_{mn} = \langle u_{m,\mathbf{k}}|u_{n,\mathbf{k}+\mathbf{b}}\rangle. \tag{8}$$

Since the gradient of Ω with respect to the $U_{mn}^{(\mathbf{k},\mathbf{b})}$ degrees of freedom can be expressed analytically as function of the $M_{mn}^{(\mathbf{k},\mathbf{b})}$, the minimisation of the spread functional may be obtained, for instance, by steepest-descent or conjugate-gradient methods (see refs. [8,9]).

Interestingly, even though the global minimisation of Ω fixes the gauge, a certain degree of non-uniqueness may remain for instance if the minimum is very shallow or flat as in the case of LiCl[9]. This results in different configurations to be degenerate and therefore different solutions (usually related by a global rotation of the MLWFs) can be obtained depending on the initial guess. Moreover, MLWFs are only defined modulo a lattice vector by definition.

In many applications, the group of bands of interest are "entangled", i.e., are not separated by an energy gap from other bands throughout the whole Brillouin zone.

Souza, Marzari, and Vanderbilt[10] (SMV) proposed a "disentanglement" strategy that involves two steps. In the first step, one defines an energy window that encompasses the states of interest and which contains $J_{\mathbf{k}}^{\mathrm{win}}$ bands at each $\mathbf{k}$. This defines a local Hilbert space $\mathcal{F}(\mathbf{k})$ at each k-point, which is spanned by the $J_{\mathbf{k}}^{\mathrm{win}}$ states. Then, for a given number $J \leq \min_{\mathbf{k}} J_{\mathbf{k}}^{\mathrm{win}}$ of target Wannier functions, one finds the optimal set of J-dimensional subspaces $\{\mathcal{S}(\mathbf{k})\}$, with $\mathcal{S}(\mathbf{k}) \subseteq \mathcal{F}(\mathbf{k})$, that have maximum intrinsic smoothness over the BZ, where the intrinsic smoothness of the Hilbert space is measured by Ω_{I}. Heuristically, Ω_{I} represents the "change of character" of the states across the Brillouin zone. (For a rigorous derivation see ref. [9].) The subspaces $\mathcal{S}(\mathbf{k})$ are defined as the span of $\{|u_{n\mathbf{k}}^{\mathrm{opt}}\rangle\}$, which are obtained via a unitary transformation on the $|u_{n\mathbf{k}}\rangle$ that span $\mathcal{F}(\mathbf{k})$:

$$|u_{n\mathbf{k}}^{\mathrm{opt}}\rangle = \sum_{m=1}^{J_{\mathbf{k}}^{\mathrm{win}}} |u_{m\mathbf{k}}\rangle\, U_{mn}^{\mathrm{dis}(\mathbf{k})}, \quad n = 1, \ldots, J. \tag{9}$$

Note that here the $\mathbf{U}^{\mathrm{dis}(\mathbf{k})}$ are rectangular $J_{\mathbf{k}}^{\mathrm{win}} \times J$ matrices, and are unitary in the sense that $(\mathbf{U}^{\mathrm{dis}(\mathbf{k})})^{\dagger}\mathbf{U}^{\mathrm{dis}(\mathbf{k})} = \mathbf{1}_J$ (with $\mathbf{1}_J$ being the $J \times J$ identity matrix), ensuring that $\{|u_{n\mathbf{k}}^{\mathrm{opt}}\rangle\}$ form an orthonormal set. Maximum intrinsic smoothness is achieved by choosing $\mathbf{U}^{\mathrm{dis}(\mathbf{k})}$ to minimise Ω_{I}, which, as discussed earlier, is a measure of the "spillage" between neighbouring subspaces $\mathcal{S}(\mathbf{k})$[10].

In the second step, having defined a J-dimensional subspace $|u_{n\mathbf{k}}^{\mathrm{opt}}\rangle$ at each $\mathbf{k}$, one proceeds by minimising $\widetilde{\Omega}$ following the same recipe described in the previous section for the case of an isolated manifold of bands. Further details on the disentanglement procedure can be found in refs. [8,10].

The iterative minimisation of Ω_{I} starts with an initial guess for the subspaces $\mathcal{S}(\mathbf{k})$. However, the spread functional is non-convex and the minimisation may get trapped in a local minimum, often resulting in complex-valued WFs[9] (in the absence of spin–orbit coupling, the WFs at the global spread minimum are expected to be real[43]). For gradient-based minimisation methods, thus, the ability to reach the global minimum strongly depends on the choice of an appropriate starting point, sufficiently close to the final solution. To this aim, if one has a chemical intuition of the target J Wannier functions, an initial guess of J trial localised functions $g_n(\mathbf{r})$ can be defined. These are then projected at every $\mathbf{k}$ onto the $J_{\mathbf{k}}^{\mathrm{win}}$ Bloch states inside the target energy window (for isolated bands, $J_{\mathbf{k}}^{\mathrm{win}} = J, \quad \forall\, \mathbf{k}$), yielding:

$$|\phi_{n\mathbf{k}}\rangle = \sum_{m}^{J_{\mathbf{k}}^{\mathrm{win}}} |\psi_{m\mathbf{k}}\rangle\langle\psi_{m\mathbf{k}}|g_n\rangle \equiv \sum_{m}^{J_{\mathbf{k}}^{\mathrm{win}}} |\psi_{m\mathbf{k}}\rangle A_{mn}^{(\mathbf{k})}, \tag{10}$$

where, at every $\mathbf{k}$, $A_{mn}^{(\mathbf{k})} = \langle\psi_{m\mathbf{k}}|g_n\rangle$ is a $J \times J$ square matrix in the case of an isolated manifold of bands and a $J_{\mathbf{k}}^{\mathrm{win}} \times J$ rectangular matrix in the case of entangled bands. The initial unitary matrix $\mathbf{U}^{\mathrm{dis}(\mathbf{k})}$ can then be obtained by orthonormalising the projected guess orbitals $|\phi_{n\mathbf{k}}\rangle$ through a Löwdin orthogonalisation of $\mathbf{A}^{(\mathbf{k})}$:

$$\mathbf{U}^{\mathrm{dis}(\mathbf{k})} = \mathbf{A}^{(\mathbf{k})}\left(\mathbf{A}^{(\mathbf{k})\dagger}\mathbf{A}^{(\mathbf{k})}\right)^{-1/2}. \tag{11}$$

One possible choice, for instance, is to start from the Bloch states themselves as the projection functions ($g_n(\mathbf{r}) = \psi_{n\mathbf{k}}(\mathbf{r})$), so that the elements of $A^{(\mathbf{k})}$ are the (random) phases of the Bloch states that are computed by the ab initio code. In the case of isolated bands, even a poor initial choice such as this is often sufficient to reach the global minimum of the spread functional (with enough iterations of the minimisation algorithm). Conversely, in the case of entangled bands, the two-step "disentanglement" procedure is usually unable to reach the global minimum of the spread functional unless the initial trial orbitals are already quite close to the final solution.

This strong dependence of the SMV minimisation algorithm on the initial trial functions, and hence on the user's intuition and intervention, has been the main obstruction in the development of fully-automated workflows for generating MLWFs for high-throughput applications.

RESULTS AND DISCUSSIONS

The SCDM algorithm and its physical interpretation

An alternative method to the SMV approach described in the "Introduction" has recently been proposed by Damle et al.[20,21] in the form of the aforementioned selected columns of the density matrix (SCDM) algorithm. The method uses a QR factorisation with column pivoting (QRCP)[44] of the single-particle density matrix (DM),

$$P_{\mathbf{k}} = \sum_{n=1}^{J} |\psi_{n\mathbf{k}}\rangle\langle\psi_{n\mathbf{k}}|, \tag{12}$$

to fix the gauge freedom in a single step, without the need for an iterative minimisation algorithm. In this section, we outline the core concepts of the SCDM method, focusing mainly on the aspects needed to provide a physical interpretation and facilitate its understanding. We refer to the original publications[20,21] for additional details.

For clarity, we start by considering a system sampled at a single k-point, e.g., Γ, and so we drop the index $\mathbf{k}$ from the DM and other quantities; the extension to multiple k-points is given in the next section. We start by considering systems with a finite band-gap between the J valence bands and the conduction bands, e.g., insulators and semiconductors.

Let us first recall that $P = \sum_{n=1}^{J} |\psi_n\rangle\langle\psi_n|$ is gauge-invariant and it is a projector on the space $\mathcal{S}$ spanned by the J valence wavefunctions $\{|\psi_n\rangle\}$. Moreover, in the insulating case, the real-space representation $P(\mathbf{r},\mathbf{r}') \equiv \langle\mathbf{r}|P|\mathbf{r}'\rangle$ of the DM decays exponentially with the distance between two points $\mathbf{r}$ and $\mathbf{r}'$: $P(\mathbf{r},\mathbf{r}') \sim e^{-\gamma|\mathbf{r}-\mathbf{r}'|}$. This is the well-known near-sightedness principle[45–47]. In particular, this means that for a given fixed $\mathbf{r}' = \mathbf{r}_0$, the function

$$\varphi_{\mathbf{r}_0}(\mathbf{r}) \equiv P(\mathbf{r},\mathbf{r}' = \mathbf{r}_0) = \int \mathrm{d}\mathbf{r}' P(\mathbf{r},\mathbf{r}')\delta(\mathbf{r}' - \mathbf{r}_0) \tag{13}$$

represents the projection on the subspace $\mathcal{S}$ of a delta function centred at $\mathbf{r}_0$, and that this projection is an exponentially-localised orbital.

To understand the numerical implementation of the method, we consider from now on the real-space discretised version of the DM. The J valence wavefunctions (or, in the case of periodic systems, the periodic part $u_{n\mathbf{k}}(\mathbf{r})$ of the J valence Bloch states) can be stored on a grid of n_G points in real space $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{n_G}$. We can then define the following $n_G \times J$ matrix $\mathbf{\Psi}$ that contains the values

of the J wavefunctions on the grid points:

$$\Psi = \begin{pmatrix} \psi_{\mathbf{1}}(\mathbf{r}_{\mathbf{1}}) & \dots & \psi_{\mathbf{J}}(\mathbf{r}_{\mathbf{1}}) \\ \vdots & \ddots & \vdots \\ \psi_{\mathbf{1}}(\mathbf{r}_{\mathbf{n_G}}) & \dots & \psi_{\mathbf{J}}(\mathbf{r}_{\mathbf{n_G}}) \end{pmatrix}. \quad (14)$$

With this definition, the orthonormality condition is written as $\mathbf{\Psi}^\dagger\mathbf{\Psi} = \mathbf{1}_J$, while the density matrix (which in discretised form is an $n_G \times n_G$ matrix) can be written as $\mathbf{P} = \mathbf{\Psi}\mathbf{\Psi}^\dagger$, i.e., $P_{ij} = \sum_{n=1}^{J} \psi_n(\mathbf{r}_i)\psi_n^*(\mathbf{r}_j)$.

We can now interpret the j-th column $\mathcal{C}^j$ of the DM, $\mathcal{C}^j_i \equiv P_{ij}$, as the projection on the valence subspace $\mathcal{S}$ of a test orbital ϕ_j that is zero everywhere except at the j-th grid position (i.e., at position $\mathbf{r}_j$). This statement is the discretised version of the projection of a delta function in Eq. (13), i.e., apart from normalisation, ϕ_j is the discretised version of $\delta(\mathbf{r} - \mathbf{r}_j)$. Therefore, thanks to the near-sightedness principle, the orbitals represented by the columns of the DM are localised.

This statement is at the core of the SCDM method. In fact, when searching for Wannier functions, we are looking for a complete and orthogonal basis set of J localised functions that span the subspace $\mathcal{S}$. In our case, the set of all columns $\mathcal{C}^j$ clearly spans the whole subspace $\mathcal{S}$ (since the P operator is the projector on $\mathcal{S}$). However, in essentially all practical situations, $J \ll n_G$ and the set of all these n_G orbitals is redundant. In addition, these orbitals are not orthogonal—intuitively, projecting on delta functions centred at two neighbouring points will typically result in a large overlap between the projected orbitals—and not normalised (e.g., in the limiting case of a delta function centred at a position in space where there is no charge density, the resulting projection will have zero norm). Selecting any set of J linearly-independent columns would form a basis for $\mathcal{S}$, and an initial guess for the Wannier functions could be obtained by orthonormalising these J columns, e.g., with a Löwdin symmetric orthogonalisation. However, if these J columns are not already almost orthogonal, the orthogonalisation will be numerically unstable and, most importantly, will mix them and thereby degrade their localisation. Therefore, the goal of the SCDM method is to select the "most representative" J columns, i.e., the columns that possess the largest norm and that are as orthogonal to each other as possible, i.e., the most "well-conditioned subset", so that the Löwdin orthogonalisation will mix these orbitals as little as possible (Löwdin orthogonalisation minimises the squared difference between the original and orthogonalised functions[48]). Equivalently, as every column is the projection of a delta-like test orbital centred at $\mathbf{r}_j$, we can say that the SCDM algorithm selects J points, from among the original n_G grid points, that define the "most representative" localised projected orbitals.

To achieve this goal, SCDM uses the standard linear algebra QRCP method[44], which factorises a matrix P as $P\Pi = QR$, where Q is a matrix with orthonormal columns, R is a upper-triangular matrix, and Π is a permutation matrix that swaps the columns of P so that the diagonal elements of R are in order of decreasing magnitude $|R_{11}| \geq |R_{22}| \geq \cdots \geq |R_{n_G n_G}|$ (see Methods section of the Supplementary Material for more details). The relevant output of the algorithm is the Π permutation matrix, or more specifically the indexes of the first J columns chosen by the algorithm: these are the "most representative" columns discussed above and, after orthonormalisation, they provide the best guess for the localised Wannier functions of the system. With a slight abuse of notation, in the following we will use the symbol Π also to identify the vector of indexes of the permutation matrix, such that $\Pi(i) = j$ has the following meaning: $\Pi_{ij} = 1$, and all of the other elements in the j-th column are equal to zero.

QRCP (a greedy algorithm) selects columns as follows: since R is triangular (and Q has orthonormal columns), the norm of the first selected column $\mathcal{C}^{\Pi(1)}$ of P is $|R_{11}|^2$ and must be the largest possible, therefore the algorithm will choose the column with the largest norm. The second column $\mathcal{C}^{\Pi(2)}$ is chosen to maximise $|R_{22}|^2$ that, due to the properties of Q and R, is the component of $\mathcal{C}^{\Pi(2)}$ orthogonal to $\mathcal{C}^{\Pi(1)}$, as shown in the Methods section of the Supplementary Material. So, the QRCP algorithm will select as the second vector the one with the largest orthogonal component to the first, and in general will select the k-th vector as the one with the largest orthogonal component to the subspace spanned by the previous $(k-1)$ columns (to be more precise the actual selection process is a heuristic for trying to keep principal sub-matrices of R as well-conditioned as possible). It is worth mentioning that this approach is related to the Cholesky orbitals approach of Aquilante et al.[22], that applies to finite (non-periodic) systems and for a different basis set (a basis of atomic orbitals rather than a real-space grid discretisation). In particular, the Cholesky algorithm used in ref. [22] is a refined version of the original Cholesky decomposition specifically adapted for positive semi-definite matrices, i.e., Cholesky decomposition with full column pivoting (CholCP) $\widetilde{\Pi}^T P \widetilde{\Pi} = L^\dagger L$, where L is an upper triangular matrix and $\widetilde{\Pi}$ is a permutation matrix. In the Methods section of the Supplementary Material, we demonstrate that the selection of the columns in CholCP is the same as in QRCP, at least for the first $J = \mathrm{rank}(P)$ columns, i.e., $(P\Pi)_{:,1:J} = (P\widetilde{\Pi})_{:,1:J}$. This is due to well-known connections between QR factorisations and Cholesky factorisations[44]. Finally, the two methods use undoubtedly related ideas but they are not direct analogues since there are multiple "variants" of SCDM when using localised orbitals.

For an effective practical implementation of the method, a final step is required. In fact, the P matrix can be extremely large, since n_G can be of the order of 100,000 or more (while J is often of the order of 10–100). Therefore, applying the QRCP algorithm directly to P is impractical, both for the memory required to store it ($\mathcal{O}(n_G^2)$), and for the time needed to compute the result ($\mathcal{O}(J \times n_G^2)$). Instead, using the fact that $P = \Psi\Psi^\dagger$ and that the original columns of Ψ are orthonormal, one can prove (see Methods section of the Supplementary Material) that the same permutation matrix Π can be obtained applying the QRCP algorithm directly to the much smaller matrix $\Psi^\dagger$ (of size $J \times n_G$), with a computational cost that scales as $\mathcal{O}(J^2 \times n_G)$. Moreover, the matrix obtained from the first J columns of $(\Psi^\dagger\Pi)$ may be used as the A_{mn} projection matrix of Eq. (10) as a starting point for the usual Wannierisation procedure in order to obtain MLWFs.

Finally, it is worth noting the connection with the "canonical" approach of user-defined initial guesses (e.g., atomic-like orbitals at specified centres): the SCDM method may be thought of as using as initial guesses a set of extremely localised s-like "orbitals" (actually, δ functions), whose centres (located at the points of the real-space grid) are optimally chosen by the SCDM algorithm via the QRCP factorisation.

SCDM for periodic systems: SCDM-k

We now extend the discussion to the case of k-point sampling with more than one k-point (i.e., not only at Γ), still considering an isolated manifold (e.g., the valence bands). The DM $P_{\mathbf{k}} = \sum_n |\psi_{n\mathbf{k}}\rangle\langle\psi_{n\mathbf{k}}|$ is an analytic function of $\mathbf{k}$[43,49], and it is also proven that WFs with an exponential decay exist[50]; numerical studies for the specific case of MLWFs have confirmed this claim for several materials[50,51], and recently there has been a formal proof for 2D and 3D time-reversal-invariant insulators[43]. The SCDM method has been extended also to the case of k-sampling[21] and named in this case "SCDM-k". In summary, the goal is now to select a common set of columns for all the k-dependent density matrices $P_{\mathbf{k}}$. Reference[21] discusses extensively how the method can be extended to a k-point sampling with more than one k point and it shows detailed results of the convergence as a function of the number of k points used in the column-selection algorithm. The final conclusion of the authors is that it is typically sufficient to

select the columns using a single "anchor" k point (typically chosen to be Γ), i.e., it is sufficient to compute the permutation matrix Π using a QRCP on $P_{\mathbf{k}=\Gamma}$ only. Then, this selection of columns can be used for all other k-points.

Extension to entangled bands

Finally, the extension to the entangled case (e.g., for metals or when considering also the conduction bands of insulators and semiconductors) has been proposed in ref. [21]. In this case, a so-called quasi-density matrix is defined,

$$P_{\mathbf{k}} = \sum_n |\psi_{n\mathbf{k}}\rangle f(\epsilon_{n\mathbf{k}}) \langle \psi_{n\mathbf{k}}|, \tag{15}$$

where $f(\epsilon_{n\mathbf{k}})$ is an occupancy function. The isolated-bands case can be recovered by setting $f(\epsilon_{n\mathbf{k}}) = 1$ for energy values $\epsilon_{n\mathbf{k}}$ within the energy range of the isolated bands, and zero elsewhere. For the typical cases of interest of this work (metals, and valence bands and low-energy conduction bands in semiconductors and insulators), one needs bands up to a given energy (typically slightly above the Fermi energy). Then, as suggested in ref. [21], $f(\epsilon)$ can be chosen as the complementary error function:

$$f(\epsilon) = \frac{1}{2}\mathrm{erfc}\left(\frac{\epsilon - \mu}{\sigma}\right). \tag{16}$$

This function depends on two free parameters μ and σ, whose choice is critical to tune the algorithm and obtain a set of Wannier functions that correctly interpolate the low-energy electronic bands of a given material. In the "Entangled bands" section, we describe our protocol to choose the values of μ and σ based on the electronic structure of the material, allowing us to implement a fully automated workflow to construct its Wannier functions via the SCDM method.

The algorithm then proceeds as in the case for isolated bands, computing the QRCP factorisation on the quasi-density-matrix or, in practice, on the matrix $\mathbf{F_k}\mathbf{\Psi}_{\mathbf{k}}^{\dagger}$ at the $\mathbf{k} = \Gamma$ anchor point, with $F_{\mathbf{k}}$ a diagonal matrix with matrix elements $\{f(\epsilon_{1,\mathbf{k}}), \ldots, f(\epsilon_{J^{\mathrm{win}}_{\mathbf{k}},\mathbf{k}})\}$. This approach, therefore, constitutes an alternative to the SMV disentanglement procedure described in the "Introduction" section: matrices obtained from the first J selected columns of $\mathbf{F_k}\mathbf{\Psi}_{\mathbf{k}}^{\dagger}$ at each $\mathbf{k}$ form the projection matrices $\mathbf{A}^{(\mathbf{k})}$, and the $\mathbf{U}^{\mathrm{dis}(\mathbf{k})}$ matrices of Eq. (9) are obtained using the Löwdin transformation of Eq. (11).

SCDM and MLWFs

The SCDM algorithm is able to robustly generate well-localised functions that are used to generate Wannier functions without the need for an initial guess. Whilst this makes the algorithm well-suited for direct integration within HT frameworks, the selection of the columns cannot be controlled by external parameters (at least for isolated bands), and therefore it is not possible to enforce constraints that might be desirable, such as point symmetries. On the contrary, when explicitly specifying atomic-like initial projections, these (if appropriately chosen) provide at least some degree of chemical and symmetry information. In the "SCDM vs MLWFs in well-known materials" section, we discuss how this affects the WFs obtained by the algorithm. Our aim is to leverage on the ability of SCDM to automatically generate a good set of localised functions, and to use these to seed the MV algorithm for the minimisation of the total spread functional, which will give in turn an automated protocol to generate MLWFs. Being able to automatically generate MLWFs will also allow users to seamlessly exploit the set of computational tools that have been developed in recent years for MLWFs and implemented in various codes, such as WANNIER90. In practice, this entails employing the SCDM algorithm to compute the $\mathbf{A}^{(\mathbf{k})}$ matrices of Eq. (10) as follows:

$$A_{mn}^{(\mathbf{k})} = f(\varepsilon_{m\mathbf{k}}) \psi_{m\mathbf{k}}^{*}(\mathbf{r}_n), \tag{17}$$

where the J points $\mathbf{r}_n$ are obtained from the first J columns of the permutation matrix Π, computed at Γ, i.e., $\mathbf{A}^{(\mathbf{k})} = \mathbf{F_k}\mathbf{\Psi}_{\mathbf{k}}^{\dagger}\mathbf{\Pi}_{\Gamma}(\mathbf{J})$, with $\Pi_{\Gamma}(J)$ representing the reduced matrix formed by the first J columns of Π_{Γ}.

SCDM and "disentanglement"

It is worth noting that the SCDM method can be also combined with the SMV disentanglement procedure, as a means of seeding the initial subspace projection. However, this introduces two additional parameters associated with the SMV approach, namely $\varepsilon_{\mathrm{outer}}$, and $\varepsilon_{\mathrm{inner}}$, giving a total of four parameters (together with μ and σ). $\varepsilon_{\mathrm{outer}}$ defines the upper limit of the so-called "outer" energy window discussed in the "Introduction" section, and $\varepsilon_{\mathrm{inner}}$ defines the upper limit of a smaller energy window contained within the outer energy window. This inner window is used to "freeze" the Bloch states within during the minimisation of Ω_{I}, such that they are fully preserved within the selected subspaces $\{\mathcal{S}(\mathbf{k})\}$ (see ref. [10] for a comprehensive description of the outer and inner energy windows). Each additional parameter makes it increasingly difficult to find a robust and automated protocol for obtaining MLWFs. Consequently, when combining SCDM with SMV disentanglement, an optimal selection of all the parameters can be achieved only in an ad hoc, non-automatic fashion (hence only for few materials). As shown in the "The SCDM algorithm and its physical interpretation" section, SCDM employs a generalised form of the density matrix, Eq. (15), which implicitly defines an energy window via the function $f(\varepsilon)$ and selects a smooth manifold by construction. Intuitively, this suggests that SCDM can be used in lieu of the SMV disentanglement procedure. In general, we have found that for the sole purpose of interpolating the energy bands up to a given energy, performing SMV disentanglement step on top of SCDM has at best a marginal improvement on the quality of the interpolation (see "Entangled bands"), and in some cases can even be detrimental due to the case-by-case sensitivity on the choice of energy windows. For this reason, in the "Entangled bands" section we focus exclusively on a protocol for the automatic selection of the free parameters in SCDM, i.e., μ and σ, without considering any additional SMV disentanglement.

SCDM vs MLWFs in well-known materials

As a precursor to the fully-automated high-throughput study on a set of 200 materials that focuses on automatic Wannierisation and band interpolation from SCDM projections and which will be presented in the "Entangled bands" section, in this section we consider in greater depth and detail the performance of the SCDM method on a small set of simple systems with well-known Wannier representations of the electronic structure. Specifically, we compare quadratic spreads, centres and symmetries of the WFs computed from the SCDM gauge (as described in the "The SCDM algorithm and its physical interpretation" section) with the ones computed from carefully chosen initial projections. Comparative studies between SCDM localised functions and MLWFs on well-known materials have recently appeared in the literature[21,29]. However, here we expand on different aspects, focusing in particular on the combination of the SCDM and the MV approaches (SCDM+MLWFs), to better assess its range of applicability, for instance for beyond-DFT methods, e.g., ab initio tight-binding[52,53], DFT+U[54–56], and DMFT[57,58], where the symmetries of the Wannier functions are important.

All DFT calculations have been carried out with Quantum ESPRESSO, using the PBE exchange-correlation functional and Vanderbilt ultrasoft pseudopotentials[59]. MLWFs are generated from Bloch states calculated on a $10 \times 10 \times 10$ Monkhorst–Pack grid of k-points. The SCDM method has been implemented in the PW2WANNIER90 code, which interfaces Quantum ESPRESSO with the WANNIER90 code[24,60], as explained in "Methods". WANNIER90 is used throughout this work to generate the WFs on a real-space

grid and to perform the interpolation of band structures in reciprocal space.

We consider four different schemes for generating Wannier functions: (1) Full minimisation of Ω using the SMV disentanglement algorithm to minimise Ω_I and the MV algorithm to minimise $\widetilde{\Omega}$ (DIS+MLWF); (2) Minimisation of Ω_I only, using the SMV algorithm (DIS); (3) Minimisation of $\widetilde{\Omega}$ only, using the MV algorithm (MLWF); and (4) No minimisation of Ω (proj-ONLY). In each case, the initial J-dimensional subspace at each $\mathbf{k}$ is determined in one of two ways, either by the SCDM method or by projection onto specific atomic-like localised orbitals (Eq. (10)).

We start by studying the Wannierisation of a manifold of bands consisting of the four valence bands plus the four low-lying conduction bands in silicon, the latter being entangled with bands at higher energies. For the SCDM method, we use $\sigma = 2$ eV and $\mu = 10$ eV. This choice is equivalent to that of ref. [21], taking into account a shift in the absolute energy scale, which shifts the value of μ. The outer and inner energy windows (described in the "Introduction"), obtained through convergence tests, are set to $\varepsilon_{\text{outer}} = 17.0$ eV and $\varepsilon_{\text{inner}} = 6.5$ eV.

When using initial projections onto atomic-like orbitals, we find that the spread functional Ω has three minima that are very close to each other and each of which gives eight real MLWFs. The global minimum corresponds to four sp^3-type MLWFs per Si atom in the two-atom unit cell, oriented in a back-bonding (BB) configuration, i.e., with the major lobes of the sp^3-type MLWFs pointing towards the tetrahedral interstitial sites. A representative example of one such BB MLWF is shown in the isosurface plots in the first row of Fig. 1. Intuitively, from an atomic orbital perspective, one might instead expect the sp^3-type MLWFs to be in a front-bonding (FB) configuration, i.e., with the major lobes pointing towards the vertices of the tetrahedra centred on the two non-equivalent Si atoms, as shown in the isosurface plots in the second row of Fig. 1. However, this FB configuration corresponds to a slightly larger value of the total spread Ω and, therefore, constitutes a local minimum of the spread. A third (intermediate) local minimum gives four sp^3-type MLWFs that are in the BB configuration on one Si atom in the unit cell and four sp^3-type in the FB configuration on the other Si atom. At variance with what is stated in ref. [29], all these cases can be found by specifying as initial projections four appropriately oriented sp^3-type orbitals on each Si atom in the unit cell. For the BB configuration: four sp^3-type orbitals centred on the Si atom at (0.0, 0.0, 0.0) (Si1), and four rotated sp^3-type orbitals centred on the other Si atom (Si2) at $(-\frac{1}{4}, \frac{3}{4}, -\frac{1}{4})$ in fractional coordinates with respect to the lattice vectors $\mathbf{a}_1 = (-5.10, 0.00, 5.10)$, $\mathbf{a}_2 = (0.00, 5.10, 5.10)$ and $\mathbf{a}_3 = (-5.10, 5.10, 0.00)$ (in a_0). In the WANNIER90 code this can be specified in the projection block of the input file as: Si1:sp3:z = 0,0, −1:x = 0,1,0; Si2:sp3. For the FB configuration: same as above but with the labels 1 and 2 on the Si atoms interchanged.

With these initial projections, the four different minimisation options described earlier give the same qualitative results. Going from the DIS+MLWF case to DIS to MLWF to proj-ONLY, the spreads of the MLWFs increase, as expected, but the FB/BB character is consistently present (see the top two rows of Fig. 1, the spread of the individual MLWFs (in units of Å^2) is reported underneath each isosurface plot). Performing the SMV disentanglement step results in a reduction of Ω_I from 26.54 to 20.06 Å^2 in both the FB and BB cases, showing that the initial and final selected subspaces from the two different choices of projection have the same intrinsic smoothness.

Instead, starting from SCDM to define the initial subspace, we obtain different qualitative results for the four different minimisation schemes. Wannier functions in the BB configuration are found when a full minimisation is performed (i.e., SCDM followed by SMV and MV minimisation). A representative example of one such WF is shown in the third row and first column of Fig. 1. SCDM selects a less smooth initial subspace ($\Omega_I = 27.54$ Å^2) than specifying atomic

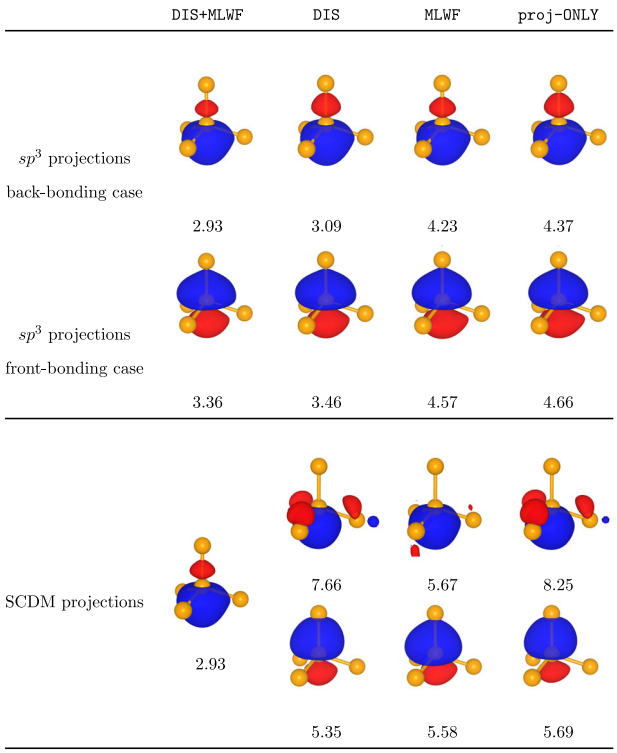

Fig. 1 Wannier functions obtained by Wannierising the four valence bands plus the four low-lying conduction bands in silicon. First row: the initial subspace is defined by projecting the Bloch states $\psi_{n\mathbf{k}}(\mathbf{r})$ on eight appropriately oriented sp^3-type orbitals giving back-bonding (BB) MLWFs in all cases. Second row: as above but with different orientations for the sp^3-type orbitals, resulting in front-bonding (FB) MLWFs in all cases. Third row: the initial subspace is obtained from the SCDM method. Here, the eight sp^3-type WFs are in the BB configuration only when a full minimisation is performed. In all other cases, a mixture of configurations is obtained instead. The values below each WF isosurface (isovalue = ±0.45 Å$^{-3/2}$) is the value of the individual spread in Å^2.

orbital initial projections (26.54 Å^2), but the final spreads are the same as in the equivalent BB case with atomic orbital initial projections. We also observed that in the case of SCDM, the minimisation of both Ω_I and $\widetilde{\Omega}$ required more iterations to achieve the same level of convergence, perhaps reflecting the fact that the initial subspace is less smooth. When using the other minimisation schemes, we find functions of both FB and BB character, all with slightly different individual spreads. Representative isosurfaces are shown in the last three columns of the row labelled "SCDM" in Fig. 1. It is clear that the tetrahedral site symmetry is not preserved in the resulting WFs. Moreover, there is no clear pattern in the individual spreads going from the DIS case to the proj-ONLY case.

When looking at the interpolated band structure, however, a different picture emerges. In the case of choosing atomic orbital projections, the interpolation is very poor if no SMV disentanglement step is included in the minimisation. This shows the importance of disentangling the correct manifold and it is in agreement with what has been previously reported in the literature[8]. On the other hand, in the case of an SCDM-generated initial subspace, the interpolation is only marginally affected by the minimisation scheme employed (see Fig. S1 in Supplementary Note 1).

To summarise, in silicon SCDM performs very well when combined with full spread minimisation, both in terms of the symmetries of the WFs and band interpolation (see Fig. S1). When SCDM is used in isolation, the individual spreads of the resulting WFs are larger than WFs generated from user-defined atomic

orbital projections; the quality of band structure interpolation, however, is almost independent of whether or not subsequent spread minimisation is carried out.

Copper presents a paradigmatic case of a noble metal where a set of bands (e.g., of *d*-orbital character) cross and mix in a narrow energy window around the Fermi energy with a set of broad, nearly-free-electron bands. In this case, the SMV algorithm turns out to be very sensitive to the choice of the initial gauge and a good Wannier representation of the band structure can be achieved only by a careful choice of both initial projections and energy windows. Consequently, the possibility of bypassing these user-intensive steps makes the SCDM an attractive approach. This is particularly important for methodologies such as ab initio tight binding[53], DFT+U[55], and DMFT[58], which deal with strong correlation in a local subspace, e.g., the subspace spanned by *d* orbitals (for transition metals or transition-metal oxides) or *f* orbitals (for rare-earth or actinide intermetallics). For copper, as suggested by Souza et al.[10], in order to generate a faithful representation of the band structure around the Fermi level, we work with a manifold of dimension $J = 7$, which contains one more function than the conventional minimal basis usually employed in tight-binding models. For this system, we focus only on the full minimisation scheme (DIS+MLWF), as it is the most representative when comparing the symmetries of the WFs, as shown in the previous section. For the disentanglement step we set $\varepsilon_{\text{outer}} = 38.0$ eV and $\varepsilon_{\text{inner}} = 19.0$ eV. For SCDM, we set $\mu = 11.40$ eV and $\sigma = 2.0$ eV. The Fermi energy in our calculation is at 12.18 eV. As shown in ref. [10], appropriately selected initial projections are five *d*-type orbitals centred on the Cu atom and two *s*-type orbitals, each centred on one of the two tetrahedral interstitial sites. The resulting seven MLWFs respect the symmetries one would expect from group theory. In fact, the five *d*-like functions give a representation of dimension 3+2 of the O_h point group (which is isomorphic to the site-symmetry group of the origin), with the usual t_{2g} and e_g character (see Fig. 2a, b). The two *s*-like functions give each a one-dimensional representation (a_1) of T_d (which is the site-symmetry group of the tetrahedral interstitial sites), as shown in Fig. 2c.

When using SCDM projections, the symmetries of the *d*-type MLWFs are not fully recovered. This can clearly be seen in Fig. 2d, e, where the *d*-type functions show mixed t_{2g}/e_g character (this is a feature of all five *d*-type functions).

Isolated bands

Until here, we have looked into the details of the Wannier functions that can be obtained from SCDM projections, by focusing on the paradigmatic examples of silicon and copper (see "SCDM vs MLWFs in well-known materials"). We focused on comparing Wannier functions as obtained by adopting different initial projections, given that good atomic-like projections can often be easily identified through chemical intuition. Now we take a complementary perspective, by considering any given crystal structure, where we face the problem of finding good initial projections without any prior chemical knowledge of the system. This is particularly relevant for high-throughput studies, where crystal-structure databases are systematically screened with first-principles simulations. In order to produce high-throughput Wannier functions, it is fundamental to provide an algorithm that does not require human interaction in the choice of the initial projections. In addition, such an algorithm must be able to use only information that is either contained in the crystal structure and the pseudopotential, or that can be computed by a simple first-principles simulation, such as the projected density of states. To this aim, human-specified atomic-like projections are not suitable, and we propose the SCDM method as the workhorse for the automated choice of the initial projections.

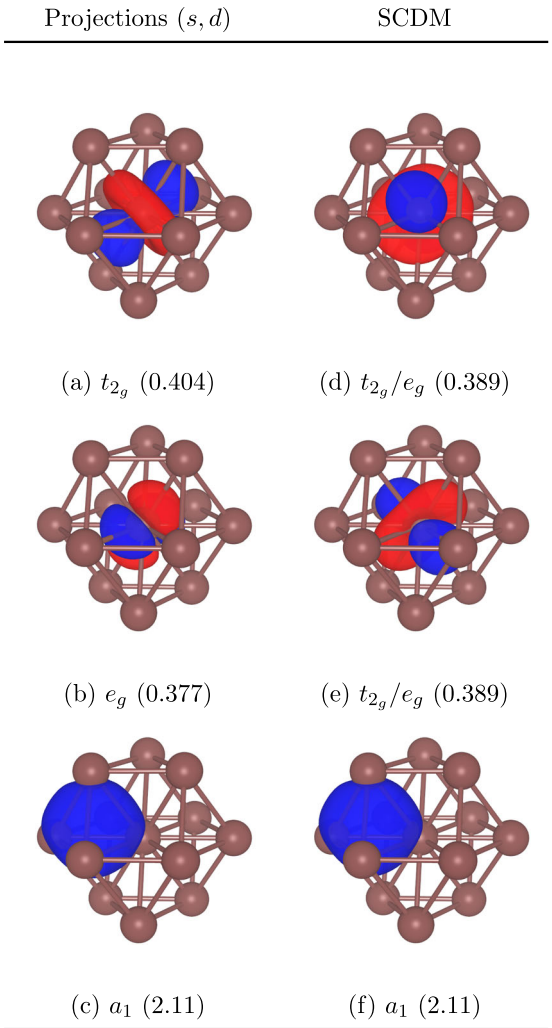

Fig. 2 MLWFs obtained by Wannierising the *s*–*d* complex in copper. First column: three representative MLWFs obtained from using atomic orbital projections to define the initial subspace (see main text for description). Panel (**a**) shows one of the three MLWFs with t_{2_g} character; panel (**b**) shows one of the two MLWFs with e_g character; panel (**c**) shows one of the two broad *s*-like orbitals centred on a tetrahedral-interstitial site. Second column: three representative MLWFs obtained from using SCDM to define the initial subspace. Panels (**d**) and (**e**) show two of the five MLWFs with mixed t_{2_g}/e_g character; panel (**f**) shows one of the two broad *s*-like orbitals centred on an tetrahedral-interstitial site. Below each function its individual spread in Å^2 is reported. Isosurfaces are plotted with an isovalue of ±0.45 Å$^{-3/2}$.

In order to ascertain the effectiveness of the SCDM method in generating well-localised Wannier functions in an automated way, we start by testing the algorithm for isolated manifolds. We compare Wannier interpolations and direct DFT calculations for the band structure of the valence bands of a set of 81 insulating bulk crystalline materials spanning a wide range of chemical and structural space, for the full list the Reader is referred to ref. [61]. We quantify the differences between two band structures by introducing a simple metric that is inspired by the so-called "bands distance" introduced in ref. [62]. Here we define the distance between DFT and Wannier-interpolated bands as:

$$\eta = \sqrt{\sum_{n\mathbf{k}} \left(\varepsilon_{n\mathbf{k}}^{\text{DFT}} - \varepsilon_{n\mathbf{k}}^{\text{Wan}}\right)^2}, \tag{18}$$

where $\varepsilon_{n\mathbf{k}}^{\text{DFT}}$ and $\varepsilon_{n\mathbf{k}}^{\text{Wan}}$ are respectively the DFT and Wannier-interpolated

band structures, and the summation runs over the occupied bands only. Later in the "Entangled bands" section, we will introduce a finite smearing to deal with conduction-band states and metallic systems. As in ref. [62], to take into account the possibility that significant differences between band structures may occur only in sub-regions of the Brillouin zone or in small energy ranges, we also compute

$$\eta^{\max} = \max_{n\mathbf{k}} \left(\left| \varepsilon^{\mathrm{DFT}}_{n\mathbf{k}} - \varepsilon^{\mathrm{Wan}}_{n\mathbf{k}} \right| \right) \tag{19}$$

where, essentially, we select the point ($n\mathbf{k}$) with the worst interpolation, which is responsible for the largest contribution to η. We use η and $\eta^{\max}$ to assess the effect of iteratively minimising the spread $\tilde{\Omega}$ to obtain maximally-localised Wannier functions ("SCDM+MLWF"), compared to the one-shot Wannier orbitals that are obtained by using the SCDM projections only ("SCDM-only"). We note that in the following MLWF might refer either to a maximally-localised WF or to the maximal localisation procedure itself, the meaning being always clear from the context.

For each of the 81 structures of the benchmark set, we first perform a variable-cell optimisation and we then compute the band structure on a high-symmetry path using DFT. The cell and the path are standardised using seekpath according to the prescription of ref. [63]. The ground-state charge density is obtained using a k-point spacing of 0.2 $\mathrm{Å}^{-1}$ in the irreducible Brillouin zone (unless otherwise stated). Band structures are then calculated using the charge density frozen from the earlier calculation and sampling the high-symmetry path with a spacing of 0.01 $\mathrm{Å}^{-1}$. Then we compute the WFs and the real-space Hamiltonian with WANNIER90, starting from a non-self-consistent field (NSCF) DFT calculation performed on a possibly different k-point grid on the full BZ and employing the ground-state charge density computed earlier. At this point, the bands distance is then calculated by diagonalising the Wannier Hamiltonian using the TBMODELS code[64] on the same k-points used in the DFT bands calculation.

All DFT calculations are carried out using the Quantum ESPRESSO distribution[23], employing the PBE functional[65] and a beta version of the SSSP v1.0 efficiency pseudopotential library[62,66–70], where the norm-conserving ONCV pseudopotentials[71] are recompiled using version 3.3.1 of the code, and the pseudopotentials for Ba and Pb are replaced by `Ba.pbe-spn-kjpaw_psl.1.0.0.UPF` and `Pb.pbe-dn-kjpaw_psl.0.2.2.UPF` of the pslibrary. In Fig. 3, we report histograms of η and $\eta^{\max}$ for four different k-point spacings, namely $\rho_k = 0.15$, 0.2, 0.3, and 0.4 $\mathrm{Å}^{-1}$, used in the NSCF step to construct Wannier functions. We stress that for an isolated set of bands, such as for the valence bands of an insulator, the SCDM method involves no free parameters and the only parameter to set is the k-point grid spacing ρ_k of a uniform grid that is used to diagonalise the Hamiltonian. Hence it is fundamental to elaborate a strategy for the choice of ρ_k, as this finally removes every free parameter from the construction of Wannier functions for isolated bands.

The SCDM method is found to work well for all of the 81 systems studied, with the exception of two that have very poor interpolation. Notably, these two structures (three if we consider the SCDM-only method) are the ones that exhibit the highest initial spread Ω per Wannier function. Although a large initial spread does not necessarily imply poor interpolation, it certainly correlates with a potential risk of poor Wannierisation and it could be used as a marker for triggering a check on the quality of bands interpolation within the calculation workflow. We postpone the discussion on the causes of the poor performance of the SCDM method in these systems until the end of this section, where we also provide possible solutions that can be automated.

To get a sense of the typical quality of a good SCDM+MLWF interpolation, we report in Fig. 4 the comparison between direct-DFT and SCDM+MLWF interpolated band structures for CaO ($\eta = 0.06$ meV, $\eta^{\max} = 0.23$ meV) and C_3Mg_2 ($\eta = 0.4$ meV, $\eta^{\max} = 5.6$ meV) run with a k-point spacing $\rho_k = 0.2$ $\mathrm{Å}^{-1}$; the direct and interpolated band structures are essentially indistinguishable (e.g., the largest difference in energy between the bands in the case of CaO is of $\eta^{\max} = 0.23$ meV).

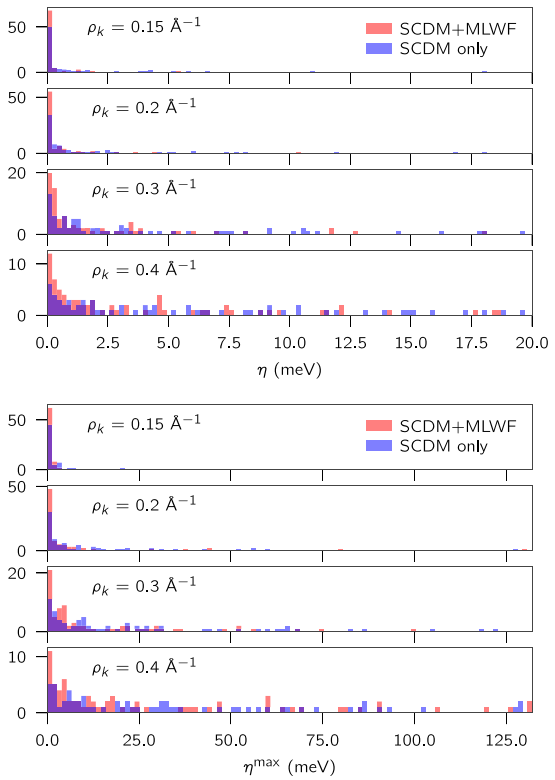

Fig. 3 Average and max band distance η using SCDM-only and SCDM+MLWF for the valence bands of 81 insulating materials. Top (bottom) panel: average (max) band distance η using SCDM-only (blue) and SCDM+MLWF (red) obtained using four different k-point grids with spacing ρ_k. The MLWF procedure improves the interpolation accuracy, although SCDM-only Wannier functions perform already remarkably well. The histograms focus on the most relevant interval and few outliers are not shown, in particular at $\rho_k = 0.2$ $\mathrm{Å}^{-1}$ 98% (79/81) of the SCDM+MLWF bands and 96% (78/81) of the SCDM-only bands exhibit $\eta < 20$ meV, while 98% (79/81) of the SCDM+MLWF bands and 93% (75/81) of the SCDM-only bands exhibit $\eta^{\max} < 130$ meV.

Figure 3 shows the distribution of η and $\eta^{\max}$ across the whole set of insulators for the four different k-point grids. We find that a grid with spacing $\rho_k = 0.2$ $\mathrm{Å}^{-1}$ is typically sufficient to provide accurate interpolated band structures, in particular 96% of the materials (78/81) for SCDM-only and 98% (79/81) for SCDM +MLWF show $\eta < 20$ meV, and 93% (75/81) of the SCDM+MLWF bands and 74% (60/81) of the SCDM-only bands display $\eta < 2$ meV. As shown in Fig. 3, $\eta^{\max}$ follows a similar trend, with 95% (77/81) of the SCDM+MLWF bands and 86% (70/81) of the SCDM-only bands showing an $\eta^{\max} < 50$ meV, and 90% (73/81) of SCDM+MLWF bands and 77% (62/81) of the SCDM-only bands showing an $\eta^{\max} < 20$ meV.

Those systems with $\eta > 20$ meV or, in other words, interpolated bands that are significantly less accurate with respect to the majority of the sample, are considered to be outliers. In Table 1, we report the number of outliers for the four different k-point spacings, both in the case of SCDM-only and SCDM+MLWF. Clearly, increasing the k-point density produces fewer outliers and, in this respect, the SCDM+MLWF seems to converge slightly faster than SCDM-only, in agreement with the results shown in Fig. 3.

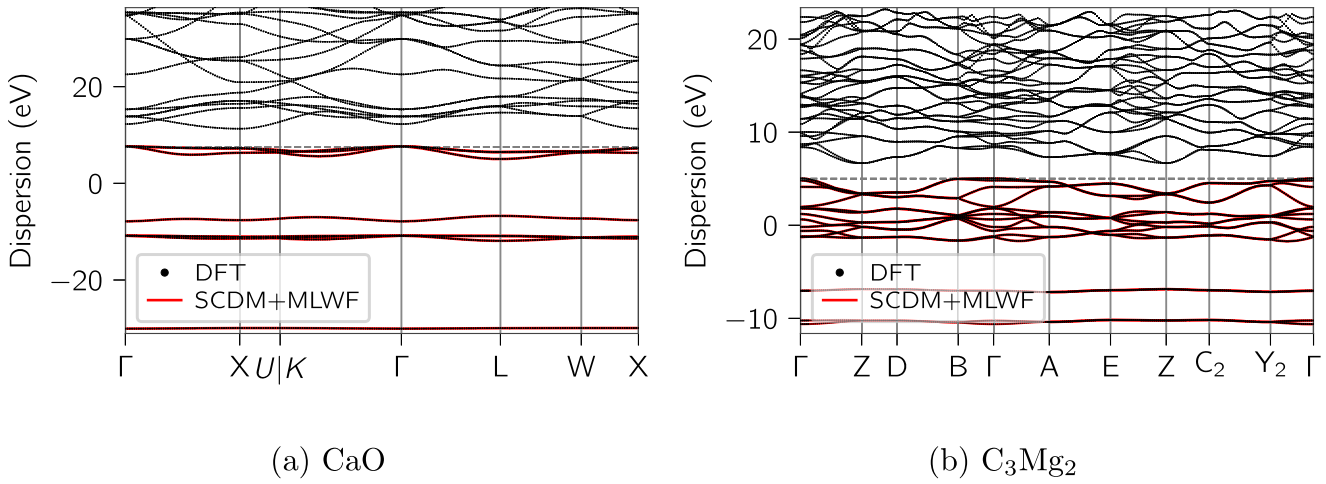

Fig. 4 Comparison between Wannier-interpolated valence bands and the full direct-DFT band structure. Wannier-interpolated (solid red) and full DFT band structure (black dots), using the MLWF procedure on SCDM projections and $\rho_k = 0.2$ Å^{-1}. The dashed line labels the valence band maximum (VBM). **a** Band structure of CaO ($\eta = 0.06$ meV, $\eta^{max} = 0.23$ meV, VBM = 7.52 eV). **b** Band structure of C_3Mg_2 ($\eta = 0.4$ meV, $\eta^{max} = 6.35$ meV, VBM = 5.0 eV).

Table 1. Number of interpolated bands showing $\eta > 20$ meV, i.e., outliers, with different k-point spacings ρ_k.

ρ_k [Å^{-1}]	SCDM-only	SCDM+MLWF
0.15	3	2
0.2	3	2
0.3	6	2
0.4	16	8

As we will discuss shortly, the superior performance of SCDM +MLWF is linked with the increased localisation associated with the MLWF procedure. As mentioned before, localisation is also related to the poor interpolation of the outliers: at all k-point spacings, outliers are among the systems with the largest initial spreads. On one hand, a larger initial spread signals a potential problem with the SCDM projections, on the other hand it requires a denser k-point grid for convergence (the less localised the Wannier functions are, the more long-range the Wannier Hamiltonian is).

Figure 3 also shows that, when considering valence bands only, the MLWF procedure moderately improves the quality of band interpolation with respect to SCDM-only, resulting in narrower η and η^{max} distributions, although band interpolation is often already excellent using an SCDM-only approach. We emphasise, however, that it is known that for the valence bands of gapped systems, a set of randomly-centred Gaussian functions can be often used as starting projections leading to good MLWFs. We compare, therefore, the performance of SCDM projections vs randomly-centred Gaussian orbital projections as a starting point for the MLWF procedure (which we refer to as the "random +MLWF" scheme), assessing their comparative robustness and accuracy of band interpolation. Figure 5 reports the distribution of η and η^{max} with k-point spacing $\rho_k = 0.2$ Å^{-1}. The SCDM projections are found to perform better, leading to narrower distributions: 98% of the materials (79/81) show $\eta < 20$ meV for SCDM+MLWF against the 89% (72/81) for random+MLWF, and 93% (75/81) of the SCDM+MLWF bands display $\eta < 2$ meV against 75% (61/81) of random+MLWF bands. As shown in Fig. 5, η^{max} follows a similar trend, with 95% (77/81) of the SCDM+MLWF bands and 81% (66/81) of the random+MLWF bands showing an $\eta^{max} < 50$ meV, and 90% (73/81) of SCDM+MLWF bands and 74% (60/81) of the random+MLWF bands showing an $\eta^{max} < 20$ meV.

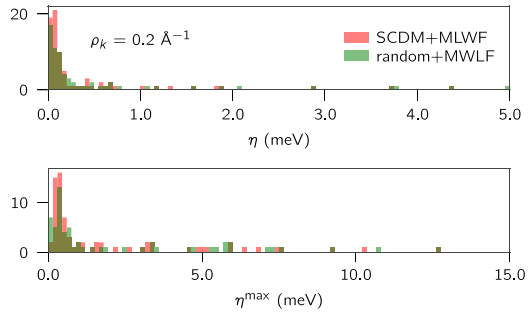

Fig. 5 Average and max band distance η using random+MLWF and SCDM+MLWF for the valence bands of 81 insulating materials. Top (bottom) panel: average (max) band distance η using random+MLWF (green) and SCDM+MLWF (red) obtained using $\rho_k = 0.2$ Å^{-1}. SCDM projections perform better than random projections when used in conjunction with the MLWF procedure. The histograms focus on the most relevant interval and few outliers are not shown, in particular the 96% (78/81) of the SCDM+MLWF bands and the 83% (67/81) of the random+MLWF bands exhibit an $\eta < 5$ meV, while the 90% (73/81) of the SCDM+MLWF bands and the 74% (60/81) of the random+MLWF bands exhibit an $\eta^{max} < 15$ meV.

Therefore, while SCDM is able to provide WFs resulting in a more accurate band interpolation, we emphasise here that for isolated manifolds the minimisation procedure is quite robust also when providing randomly-centred s-like Gaussian orbital projections.

We now elaborate on the differences between random and SCDM initial projections. First, random projections typically generate a much higher initial spread (7.5 Å^2 per WF) compared to SCDM (1.0 Å^2 per WF). We find that the MLWF procedure is often sufficient to localise Wannier functions even in the case of large initial spreads: for 63 out of 81 materials the MLWF procedure brings both the random projections and the SCDM projections cases to the same minimum spread value. Notably, it never happens that the spread is similar and the quality of the interpolation is very different, while the opposite happens only in the case of He, a pathological case (1 atom and 2 electrons per cell) where random projections give a poorly localised Wannier function while still being able to provide a very good interpolation. For 15 materials (16 if we include He), random projections provide a very poor starting point and the MLWF procedure remains trapped in a local minimum with large spread. In these

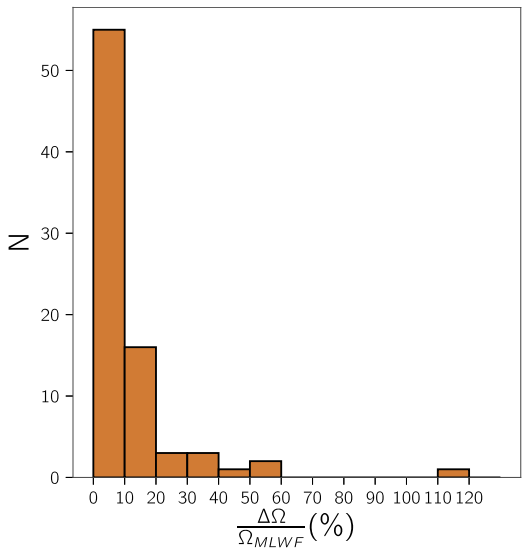

Fig. 6 Histogram of the relative variation of the total quadratic spread Ω before and after the MLWF procedure. The data has been obtained considering the valence bands of our set of 81 insulators, with $\rho_k = 0.2\,\text{Å}^{-1}$. The SCDM+MLWF procedure provides Wannier functions that are moderately more localised with respect to SCDM-only, with a relative variation within 10–20% for most materials.

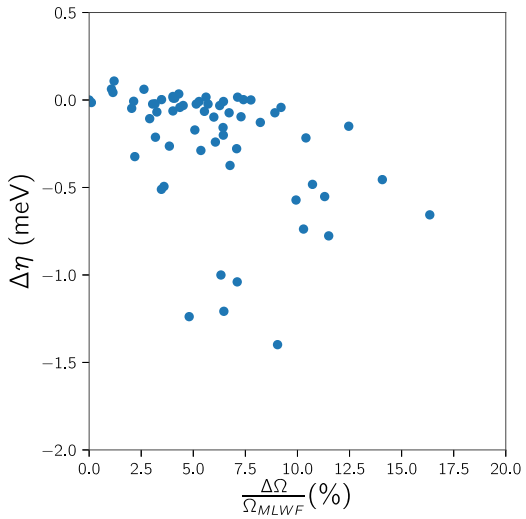

Fig. 7 $\Delta\eta$ vs $\Delta\Omega/\Omega^{MLWF}$ scatter plot (valence bands only). The dataset consists of the 81 insulators described in the main text (only 61 out of 81 visible in the axes range). $\Delta\eta$ and $\Delta\Omega/\Omega^{MLWF}$ represent the quantitative deviation between SCDM+MLWF and SCDM-only in terms of band structures and total spreads, respectively. Maximally-localised Wannier functions give comparable and often more accurate interpolated bands.

cases, instead, SCDM projections are a good starting point with low spread and the MLWF procedure further reduces it and a higher-quality interpolation is achieved, as demonstrated by the lower η values. Finally, there are two materials for which both SCDM-only and SCDM+MLWF do not perform well, but where random+MLWF happens to perform better than SCDM+MLWF. For one of these cases, Al_2Os, we have checked that excluding the semi-core states greatly improves the performance and the quality of the interpolated bands. We believe that the reason lies in the fact that, if semi-core states are present, then there are some projections, centred on the same site, that possess the same symmetry character, e.g., *p*-like projections with different principal quantum numbers (for instance 1*p*- and 2*p*-like). With a relatively low plane-wave energy cutoff, the real-space grid is too coarse and there are not enough degrees of freedom for the column selection in the QRCP step to distinguish or describe sufficiently well these same-symmetry-character states.

In the other case, Se_2Sn, there are no semi-core states. Here instead, some SCDM projections show an initial value of Ω_D—the sum of the diagonal elements of $\widetilde{\Omega}$ in Eq. (5)—that is not zero or very close to zero ($\Omega_D > 0.5\,\text{Å}^2$), which could be used as a diagnostic indicator for problematic systems. In particular, SCDM+MLWF seems to get trapped in a state in which there are a number of well-localised WFs and two that are diffuse and spread over multiple sites. This set of WF are real with a total spread of 28 Å^2 and Ω_D of 2 Å^2. We found that a possible solution to recover a good interpolation is to add some noise (adding small random numbers to the search direction components, as implemented in WANNIER90) during the minimisation to help the algorithm escape from the unwanted local minimum.

We propose some technical solutions that could be easily added to a workflow:

- Automatically detect and exclude semi-core states (if any). This is generally a safe choice as these states are not physically interesting for most applications. Alternatively, one could retain the semi-core states and increase the cutoff energy (or equivalently the density of the real-space grid).
- If the problem is not in describing semi-core states, then check the value of Ω_D, if it is above a given threshold (e.g., >1.0 Å^2) for one or more initial projections, introduce some noise in the minimisation.
- If none of the above work, switch to random+MLWF projections, which may give a better final result.

To study now more in detail the effect of minimising the spread, we start by comparing the total spread Ω obtained using SCDM+MLWF and SCDM-only, by computing:

$$\frac{\Delta\Omega}{\Omega^{\text{MLWF}}} = \frac{\Omega^{\text{SCDM}} - \Omega^{\text{MLWF}}}{\Omega^{\text{MLWF}}} \tag{20}$$

where Ω^{SCDM} and Ω^{MLWF} are the total spreads obtained with SCDM-only and SCDM+MLWF, respectively. As reported in Fig. 6, the SCDM-only Wannier functions are already well localised and $\frac{\Delta\Omega}{\Omega^{\text{MLWF}}}$ is less than 10% for 68% (55/81) of systems, and less than 20% for 88% of them (71/81).

An interesting question is whether the difference in spread due to the MLWF procedure correlates with the difference in the quality of the interpolation. To assess this, we compute the quantity

$$\Delta\eta = \eta^{\text{MLWF}} - \eta^{\text{SCDM}}, \tag{21}$$

where η^{SCDM} and η^{MLWF} are the band distances obtained with SCDM-only and SCDM+MLWF respectively. Figure 7 shows a scatter plot of $\Delta\eta$ vs $\Delta\Omega/\Omega^{\text{MLWF}}$, showing that a reduction in the spread typically implies an improvement in the quality of the interpolation ($\Delta\eta < 0$). These findings highlight that SCDM-only Wannier functions are already sufficiently localised and represent well the valence manifold, and the subsequent MLWF procedure (starting from a very good guess) safely refines the initial choice of SCDM, improving the accuracy of the Wannier Hamiltonian by increasing localisation. In general, the greatest benefit from the MLWF procedure is visible in the interpolation of the almost-flat semi-core states. In fact often, when using SCDM-only Wannier functions for the interpolation of these states, the interpolated bands show an oscillatory behaviour, with the maximum absolute difference with respect to the DFT bands of the order of a few meV (comparable to the spread of those bands). From our results, a smoother and more accurate interpolation is usually recovered after a MLWF procedure.

Before discussing the case of entangled bands, we summarise here the main conclusions that can be drawn for isolated bands. All the results we obtained, displayed in Figs 3, 4, and 6, consistently support the effectiveness of adopting SCDM projections for the Wannier interpolation of the valence bands of insulators. The quality of the interpolation is very high for 98% of the structures, with only 2 (out of 81) cases showing a poor interpolation. Although SCDM-only Wannier functions are shown to provide already accurate band structures, the MLWF procedure appears to improve both the quality of interpolation (lower η) and localisation (lower spread). Hence, we suggest the SCDM+MLWF method with $\rho_k = 0.2$ Å^{-1} as the standard protocol for producing accurate and efficient Wannier Hamiltonians describing the valence bands of bulk insulating crystals.

Entangled bands

We now consider the case of entangled bands. With the intent of describing a fully automatic protocol, we limit ourselves to the case of Wannier interpolation of all states up to a given energy (excluding, if appropriate, manifolds of low-lying semicore states that are isolated in energy from the rest of the band structure) and we do not consider the case of computing Wannier functions for a manifold of bands of given symmetry within a narrow energy window (e.g., d states in copper or t_{2g}/e_g states in a transition-metal oxide, see "SCDM vs MLWFs in well-known materials") that is entangled with bands above and below in energy.

In the case of entangled bands, the SCDM method demands the choice of three free parameters: μ and σ, as described at the end of "The SCDM algorithm and its physical interpretation" section, as well as J, the target number of Wannier functions. These parameters play a fundamental role in the selection of the columns of the quasi-DM and hence greatly affect the overall quality of the subspace selection and, consequently, the bands interpolation. In particular, since there is no equivalent definition of an inner energy window[10] in the SCDM method, it is not guaranteed that a subspace that includes the physically-relevant lowest-lying bands will be selected because the greedy QRCP algorithm, owing to an inappropriate choice of μ and σ, might favour states that are higher in energy. It is, therefore, key to the success of the automation process to have a protocol that automatically chooses these parameters in a robust and systematic way. We will now describe such a protocol, and in the "High-throughput verification" section we show its effectiveness on a large set of chemically diverse materials.

Protocol

To identify appropriate values of μ, σ, and J, we first compute the "projectability" $p_{n\mathbf{k}}$, which measures how well each Bloch state $|\psi_{n\mathbf{k}}\rangle$ is represented in a Hilbert space $\mathcal{A}$ defined by a given set of localised functions. Indeed, in the entangled case, WFs contain contributions from the valence states plus specific conduction states, typically corresponding to the anti-bonding partners of the valence states. The selection of these specific conduction states—out of the very many—can be challenging, because they are not necessarily the lowest energy ones. This idea motivates the use of projectability as a measure to see which conduction states might be more important.

Similarly to Agapito et al.[14], we choose as our localised functions the set of N_{PAO} pseudo-atomic orbitals (PAO) $\phi_{Ilm}(\mathbf{r})$ employed in the generation of the pseudopotentials, where I is an index running over the atoms in the cell and lm define the usual angular momentum quantum numbers. We then construct Bloch sums $\phi_{\mu\mathbf{k}}(\mathbf{r}) = \frac{1}{N_\mu}\sum_{\mathbf{R}} e^{-i\mathbf{k}\cdot\mathbf{R}}\phi_\mu(\mathbf{r}-\mathbf{R})$, where $\mu = \{Ilm\}$ and N_μ is the number of lattice vectors $\mathbf{R}$ contained in the Born-von Karman cell (which is equal to the number of k-points sampled in the BZ). Finally, a Hilbert space $\mathcal{A}^{\mathbf{k}}$ at each k-point in the BZ is defined as the space spanned by the Löwdin-orthogonalised functions $\widetilde{\phi}_{\mu\mathbf{k}}(\mathbf{r}) = \sum_\nu \left(\left(\mathbf{S}^{\mathbf{k}}\right)^{-1/2}\right)_{\mu\nu}\phi_{\nu\mathbf{k}}(\mathbf{r})$, with $S^{\mathbf{k}}_{\mu\nu} = \langle\phi_{\mu\mathbf{k}}(\mathbf{r})|\phi_{\nu\mathbf{k}}(\mathbf{r})\rangle$, and $\mathcal{A}$ is given by the direct sum $\mathcal{A} = \bigoplus_{\mathbf{k}}\mathcal{A}^{\mathbf{k}}$.

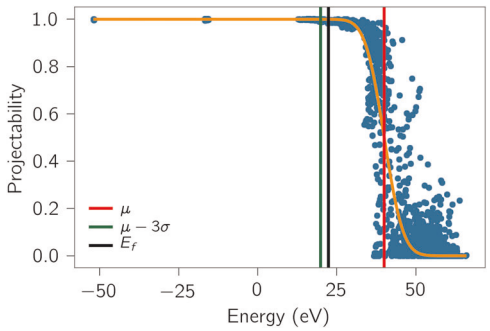

Fig. 8 Projectability of the state $|n\mathbf{k}\rangle$ as a function of the corresponding energy $\varepsilon_{n\mathbf{k}}$ for tungsten. Each blue dot represents the projectability as defined in Eq. (22). The yellow line shows the fitted complementary error function. The vertical red line represents the value of μ_{fit} while the vertical green line represents the optimal value of μ, i.e., $\mu_{\mathrm{opt}} = \mu_{\mathrm{fit}} - 3\sigma_{\mathrm{fit}}$. The value of the Fermi energy is also shown for reference (black line).

The projectability of each Bloch state onto $\mathcal{A}$ is then defined as

$$p_{n\mathbf{k}} = \sum_{I,l,m} |\langle\psi_{n\mathbf{k}}|\phi_{\mathbf{k}Ilm}\rangle|^2, \tag{22}$$

where $0 \le p_{n\mathbf{k}} \le 1$. The projections $\langle\psi_{n\mathbf{k}}|\phi_{\mathbf{k}Ilm}\rangle$ are computed straightforwardly using the projwfc.x code from Quantum ESPRESSO. In particular, for the pseudopotentials considered in this work, the number of valence electrons and the atomic orbitals included in the pseudopotential files may be found in Table S1 in Supplementary Note 2.

As the first step of our protocol, we choose J as the total number of projections N_{PAO} considered in the sum of Eq. (22). Since we aim to interpolate the bands up to a given energy above the Fermi level, fixing $J = N_{\mathrm{PAO}}$ is a conservative choice, as the number of PAOs is usually greater or equal to the number of valence bands plus few conduction bands.

We then use the values of the projectability to inform the choice of μ and σ. First, we plot the projectability for all Bloch states as a function of the corresponding band energy $\epsilon_{n\mathbf{k}}$, as shown in Fig. 8 (to illustrate the procedure, we show plots for one prototypical material, namely crystalline tungsten (W), but similar plots and trends also hold for the other materials considered in this work). The general trend is that $p_{n\mathbf{k}} \sim 1$ for low-energy states, which are well-represented by the chosen pseudo-atomic orbitals, and $p_{n\mathbf{k}} \sim 0$ for high-energy states that originate either from free-electron-like states or from localised states with an orbital character that is not included in the set listed in Table S1 in Supplementary Note 2, e.g., atomic orbitals with principal quantum number $n > 3$ (i.e., more than two radial nodes). We then fit this plot to a complementary error function as in Eq. (16), extracting the two parameters μ_{fit} and σ_{fit}. The core of our protocol lies on the actual choice of the μ and σ parameters used as input for the SCDM method by setting

$$\mu = \mu_{\mathrm{fit}} - 3\sigma_{\mathrm{fit}}, \quad \sigma = \sigma_{\mathrm{fit}}. \tag{23}$$

Let us now motivate this choice. We observe that σ_{fit} measures the typical energy spread of the bands originating from states within $\mathcal{A}$, and therefore is a good physical guess also for σ. The naive choice $\mu = \mu_{\mathrm{fit}}$, however, produces extremely poor interpolation of the bands for most of the materials that we have tested, see "High-throughput verification". The reason is that it gives too great a weight in Eq. (15) to states that have relatively small projectability ($p_{n\mathbf{k}} < 1$). As a consequence the SCDM algorithm

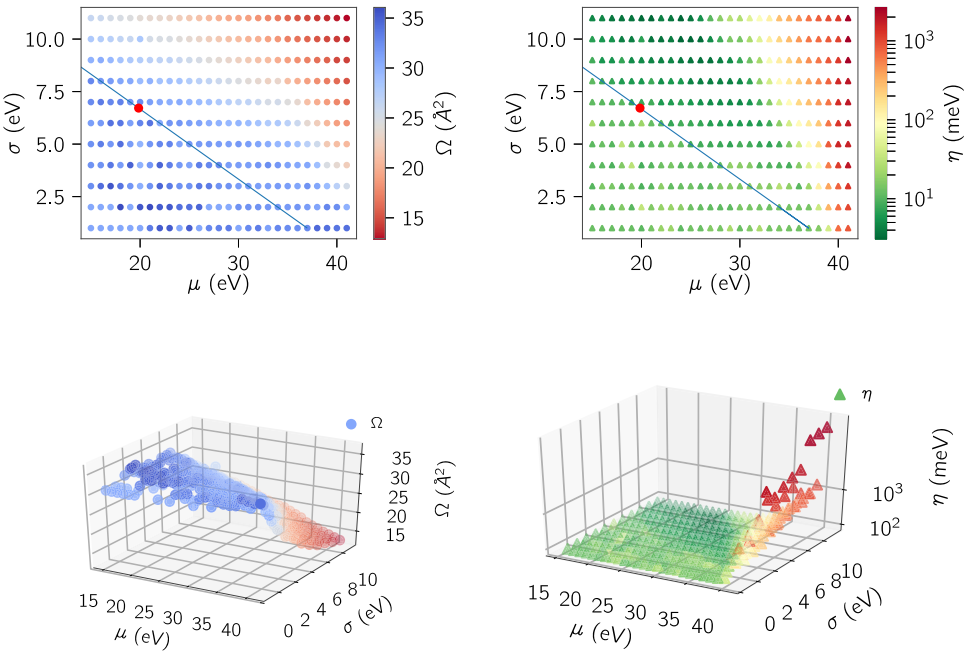

Fig. 9 Assessment of the SCDM+MLWF method for tungsten (W) as a function of the SCDM input parameters μ and σ. Left panel: bands distance η. Right panel: total position spread Ω. The blue line represents $\mu = \mu_{\text{fit}} - 3\sigma$ where the red dot corresponds to the choice dictated by our protocol $\mu = \mu_{\text{fit}} - 3\sigma_{\text{fit}}$. The smearing function to compute η has smearing $\tau = 0.1$ eV and ν is set to 1 eV above the Fermi energy.

might select columns representing better these states rather than those with projectability close to 1 at low energy, that are essential and physically relevant to include. In these cases, the corresponding band interpolation shows large oscillations and has large errors with respect to the DFT band structure in large portions of the BZ. We need therefore to choose a smaller value $\mu < \mu_{\text{fit}}$. On the other hand, however, we note that the weight of states much above μ becomes numerically zero in Eq. (15), i.e., these states become completely unknown to the algorithm. Therefore, by choosing a too low value of μ, i.e., discarding too many relevant states, the SCDM algorithm will fail because it will have to choose J columns within a matrix of smaller rank.

We need, therefore, a general and automatic recipe for choosing an appropriate, intermediate value of μ. Our choice $\mu = \mu_{\text{fit}} - \kappa\sigma_{\text{fit}}$ is guided by the consideration that states that start to have a significant component of their character outside $\mathcal{A}$ should be weighted in SCDM by Eq. (16) with a small weight, that is still though not exactly zero, giving the algorithm some freedom to pick up some of their character (for instance, states at energy $\epsilon \geq \mu_{\text{fit}}$ have more than 50% of their character outside $\mathcal{A}$ and are weighted in SCDM with a factor $\leq \frac{1}{2}\text{erfc}(\kappa)$; e.g., $\kappa = 3$ gives $\frac{1}{2}\text{erfc}(3) \approx 10^{-5}$).

In order to explain better our specific choice of $\kappa = 3$, we consider again the case of tungsten for the SCDM+MLWF case and we report in Fig. 9 the final total spread Ω (left-hand side) and the band distance η (right-hand side) as a function of a range of values of μ and σ. In particular, in the case of entangled bands, we generalise the definition of η by introducing a smearing, as we have mentioned in the previous section. More specifically, we extend the definition of the distance between DFT and Wannier-interpolated bands to:

$$\eta = \sqrt{\frac{\sum_{n\mathbf{k}} \left(\varepsilon_{n\mathbf{k}}^{\text{DFT}} - \varepsilon_{n\mathbf{k}}^{\text{Wan}}\right)^2 \tilde{f}_{n\mathbf{k}}}{\sum_{n\mathbf{k}} \tilde{f}_{n\mathbf{k}}}}, \tag{24}$$

where

$$\tilde{f}_{n\mathbf{k}} = \sqrt{f_{n\mathbf{k}}^{\text{DFT}}(\nu,\tau) f_{n\mathbf{k}}^{\text{Wan}}(\nu,\tau)}, \tag{25}$$

and $f_{n\mathbf{k}}^{\text{DFT(Wan)}}(\nu,\tau)$ is the Fermi–Dirac distribution for the state at energy $\varepsilon_{n\mathbf{k}}^{\text{DFT(Wan)}}$, ν is a fictitious chemical potential and τ is a smearing width computed on the direct ($\varepsilon_{n\mathbf{k}}^{\text{DFT}}$) and Wannier-interpolated ($\varepsilon_{n\mathbf{k}}^{\text{Wan}}$) band structures. As in the "Isolated bands" section, we take into account the possibility that significant differences between band structures may occur only in sub-regions of the Brillouin zone or in small energy ranges, so we also compute

$$\eta^{\max} = \max_{n\mathbf{k}} \left(\tilde{f}_{n\mathbf{k}} \left| \varepsilon_{n\mathbf{k}}^{\text{DFT}} - \varepsilon_{n\mathbf{k}}^{\text{Wan}} \right| \right). \tag{26}$$

In particular, the value of ν in $\tilde{f}_{n\mathbf{k}}(\nu,\tau)$ is set to 1 eV above the Fermi energy and the smearing width τ is 0.1 eV. In this way, only states up to slightly more than 1 eV above the Fermi level have a weight significantly different from zero when comparing band structures. In both panels of Fig. 9, we also show the line representing $\mu = \mu_{\text{fit}} - 3\sigma$ to discuss our choice of $\kappa = 3$, as well as the point ($\mu_{\text{fit}} - 3\sigma_{\text{fit}}$, σ_{fit}) on this line. Our target is to have η as small as possible, indicating a good interpolation of the band structure. As visible in Fig. 9, and as mentioned in the previous two paragraphs, large values of μ and σ degrade significantly the quality of the band interpolation: in this case there are many states at high energy with a non-negligible weight and the QRCP, being a greedy algorithm, might select a subspace that better represents these states rather than the lowest energy states. It can also be seen that a larger μ, which results in more states with higher weight, gives the SCDM algorithm more freedom in the choice of the subspace, which in turn results in a lower total spread Ω (at the expenses of a potentially worse interpolation).

On the other hand, also moving to the region of small μ and σ is detrimental for the quality of the band interpolation (and partially also for the value of Ω). Even if the values of η in this region are not so large as in the region of large μ and σ, the quality of the interpolation is much less robust and both η and Ω depend strongly on the precise values of the two parameters. In this case, we are discarding relevant states from the initial space used for the column selection of $\mathbf{F}_{\mathbf{k}}\boldsymbol{\Psi}_{\mathbf{k}}$, therefore removing important information needed by the method for a good interpolation.

Our choice of $\kappa = 3$, thus, together with $\sigma = \sigma_{\text{fit}}$, allows us to locate our choice of (μ, σ) in the intermediate region where η is small and both η and Ω are relatively insensitive to small variations

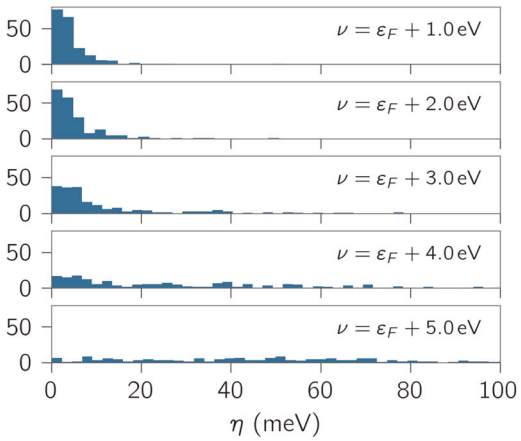

Fig. 10 Distribution of the band distance η for different values of the fictitious chemical potential ν. The chemical potential is defined as $\nu = \varepsilon_F + \Delta$ ($\Delta = 1, 2, 3, 4, 5$ eV) and the smearing τ in the Fermi–Dirac distribution is 0.1 eV. All calculations have been performed with a k-point spacing of $\rho_k = 0.2$ Å^{-1}.

of the two parameters. Ultimately, this specific choice for κ will be justified and validated in our high-throughput study of "High-throughput verification", where we show that the automated algorithm resulting from this choice is robust when tested on 200 chemically and structurally different materials, whose full list is available in ref. [61].

We also emphasise here that the choice of μ and σ plays two different roles: the first is to give a relative weight to the states at the anchor point, namely Γ, that are used for the SCDM column selection; the second is to have a smooth dependence of the subspace as a function of **k**, therefore resulting in a small Ω_I.

High-throughput verification

In this section, we present the results of the high-throughput calculations for the general case of 200 materials that have been chosen to cover a large region of structural (12 different Bravais lattices) and chemical (67 different elements) space. The free parameters in the SCDM method have been chosen by the automatic procedure outlined in the previous section. The structure of this section parallels the one for isolated bands; in particular, we make use of the bands distance η introduced in Eq. (24) to quantitatively assess the Wannier interpolation. In the case of metals, we also need to appropriately select the value of the fictitious chemical potential ν and of the smearing width τ in the distribution $\tilde{f}_{n\mathbf{k}}(\nu, \tau)$ of Eq. (25) (the final values used in this work are reported in the previous section), in order for η and $\eta^{\max}$ to be reliable measures for the interpolation quality of the bands of physical interest. Indeed, the Wannier-interpolated bands are not expected to reproduce accurately the dispersion of the DFT bands at high energies; and the energy up to which the Wannier-interpolated bands may be deemed to be accurate depends mainly on the number of target WFs J which, in turn, is determined in our procedure by the number of PAOs in the pseudopotentials. In most applications, however, the high-energy bands are not of interest; therefore, ν and τ should be chosen so as to define a bands distance that only takes into account the relevant low-energy bands. For most practical applications, this means for states up to a small amount (usually a few eV) above the Fermi energy.

To verify up to which energy the interpolation is accurate (for the number of PAOs in the pseudopotentials chosen in this work, see Table S1 in Supplementary Note 2) we show in Fig. 10 the distribution of band distances for different values of $\nu = \varepsilon_F + \Delta$, with $\Delta = 1, 2, 3, 4, 5$ eV, and τ fixed at 0.1 eV in order to have a smooth but sharp-edged Fermi–Dirac distribution. When ν is set at 4 eV or more above the Fermi energy ($\Delta \geq 4$ eV, bottom panels in Fig. 10), the distribution is very broad and with a long tail. In this case states much above the Fermi energy, where the Wannier interpolation does not reproduce any more the DFT band structure, are given a non-negligible weight $\tilde{f}_{n\mathbf{k}}$ which significantly increases the value of the band distance. The distribution becomes much narrower and closer to $\eta = 0$ eV for $\Delta \leq 3$ eV; in particular, for $\nu = \varepsilon_F + 1.0$ eV, 98% of the materials have $\eta <$ 50 meV. Since for many applications having a good interpolation up to 1 eV above the Fermi energy is sufficient, in the rest of this work we choose $\nu = \varepsilon_F + 1.0$ eV (for entangled bands) as a reliable measure of the quality of the interpolation in the energy region of interest.

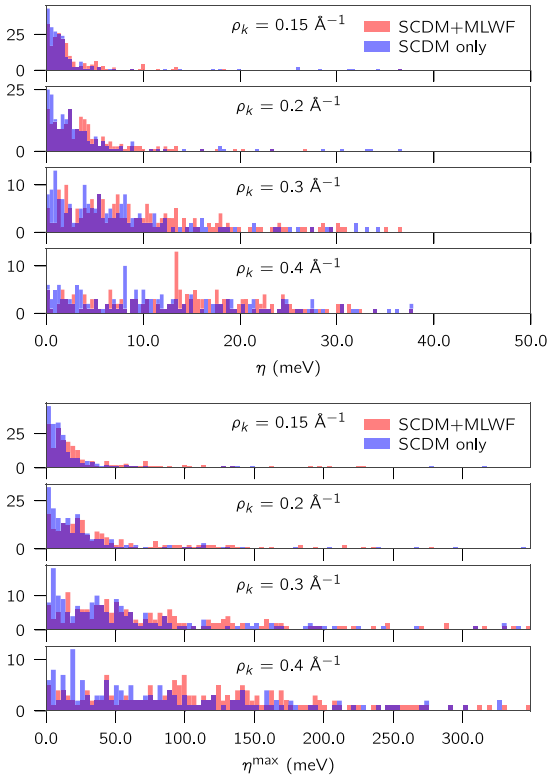

Fig. 11 Average and max band distance for the valence and few conduction bands of 200 materials. Top (bottom) panel: histogram of average (max) band distance η ($\eta^{\max}$) in meV using SCDM-only (blue) and SCDM+MLWF (red) obtained using four different **k**-point grids with spacing ρ_k. The MLWF procedure slightly worsens the accuracy of the interpolation when compared to SCDM-only Wannier functions. The histograms focus on the most relevant interval and few outliers are not shown, in particular at $\rho_k = 0.2$ Å^{-1} 98% (196/200) of the SCDM+MLWF bands and 99.5% (199/200) of the SCDM-only bands exhibit $\eta < 50$ meV, while 98% (195/200) of the SCDM+MLWF bands and 94% (188/200) of the SCDM-only bands exhibit $\eta^{\max} < 350$ meV.

As in the case of isolated bands, the first step is to study the effect of the k-point grid density on the interpolation, to fix the last free parameter in the calculations. As shown in Fig. 11, a grid with spacing $\rho_k = 0.2$ Å^{-1} is typically sufficient to provide accurate interpolated band structures: in particular, 94% of the materials (187/200) for SCDM-only and 97% (193/200) for SCDM+MLWF show $\eta < 20$ meV, and 72% (144/200) of the SCDM+MLWF bands and 79% (157/200) of the SCDM-only bands display $\eta < 5$ meV. Moreover, $\eta^{\max}$ follows a similar trend, with 72% (143/200) of the SCDM+MLWF bands and 82% (163/200) of the SCDM-only bands showing an $\eta^{\max} < 50$ meV, and 35% (70/200) of SCDM+MLWF

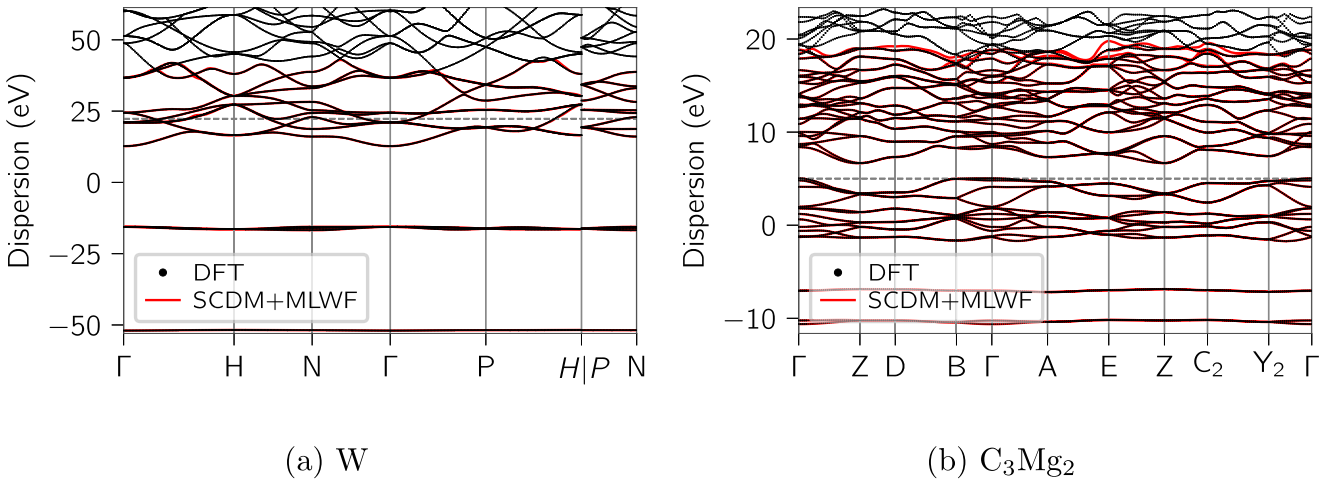

Fig. 12 Comparison between Wannier-interpolated valence bands plus few conduction bands and the full direct-DFT band structure. Wannier-interpolated bands are in solid red and full DFT bands are in solid black. Panel (**a**), $\eta = 20$ meV, $\eta^{\max} = 415$ meV, $\mu = 19.85$ eV and $\sigma = 6.71$ eV and C_3Mg_2. Panel (**b**), $\eta = 2$ meV, $\eta^{\max} = 11$ meV, $\mu = 0.86$ eV and $\sigma = 5.63$ eV using the MLWF procedure on SCDM projections and $\rho_k = 0.2$ Å^{-1}. Note that, while we show all Wannier-interpolated bands, the band distance η considers only bands up to about 1 eV above the Fermi level (see text).

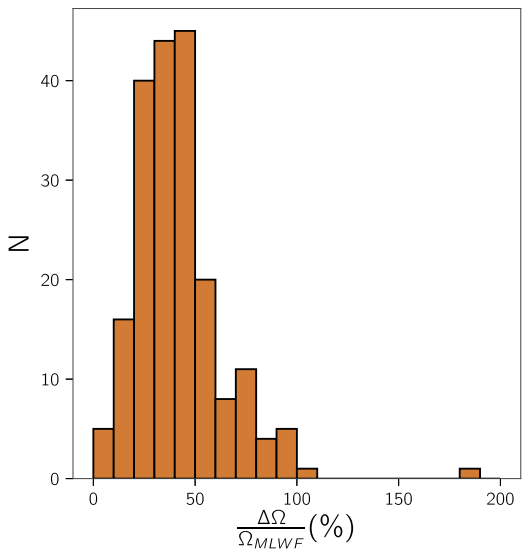

Fig. 13 Average and max band distance for the valence and few conduction bands of 200 materials. Histogram of the relative variation of the total quadratic spread Ω before and after the MLWF procedure for the band structures of our set of 200 materials, obtained for $\rho_k = 0.2$ Å^{-1}. The SCDM+MLWF procedure provides Wannier functions that are substantially more localised with respect to SCDM-only, with a relative variation between 20 and 60% for most materials.

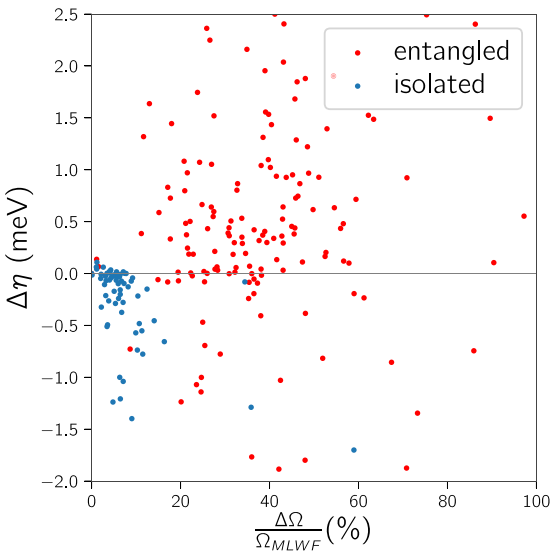

Fig. 14 $\Delta\eta$ vs $\Delta\Omega/\Omega^{\mathrm{MLWF}}$ scatter plot (valence and few conduction bands). The dataset consists of all $200 + 81$ materials, with entangled bands (red dots, 148 out of 200 visible in the axes range) and with isolated bands (blue dots, 64 out of 81 visible) showing $\Delta\eta$ vs $\Delta\Omega/\Omega^{\mathrm{MLWF}}$, that is the quantitative deviation between SCDM+MLWF and SCDM-only in terms of band structures and total spreads, respectively. Maximally-localising Wannier functions give potentially more accurate interpolated bands for valence bands only, whereas for entangled bands the trend is reversed.

bands and 52% (104/200) of the SCDM-only bands showing an $\eta^{\max}$ <20 meV, as shown in Fig. 11. We therefore set ρ_k to 0.2 Å^{-1} for further analysis in this section.

Figure 12a shows the Wannier-interpolated bands (red lines) for tungsten (W), a metallic system, and Fig. 12b shows the Wannier-interpolated valence bands plus few conduction bands (in red) for the insulator C_3Mg_2 (and these can be compared with Fig. 4b for the interpolation of the valence bands only).

Unlike the case of isolated bands, for entangled bands the MLWF procedure substantially increases the localisation of the resulting Wannier functions from SCDM projections, giving for instance a $\frac{\Delta\Omega}{\Omega^{\mathrm{MLWF}}}$ between 20 and 60% for 75% (149/200) of materials, with 30 materials showing a 60% or more increase in $\frac{\Delta\Omega}{\Omega^{\mathrm{MLWF}}}$, see Fig. 13.

We now look at how the difference in spread due to the MLWF procedure correlates with the difference in the quality of the interpolated band structures. Although the correlation is not as strong as in the case of isolated bands, it can be seen (Fig. 14) that the trend is almost reversed: reducing the spread tends to worsen the quality of the band interpolation. In fact, the majority of systems (71%, 142/200) show a positive change in $\Delta\eta$, meaning that SCDM-only provides better interpolation. The main reason behind this effect is that, in the selection of the optimal manifold $\mathcal{S}(\mathbf{k})$, the SCDM algorithm might include contributions from

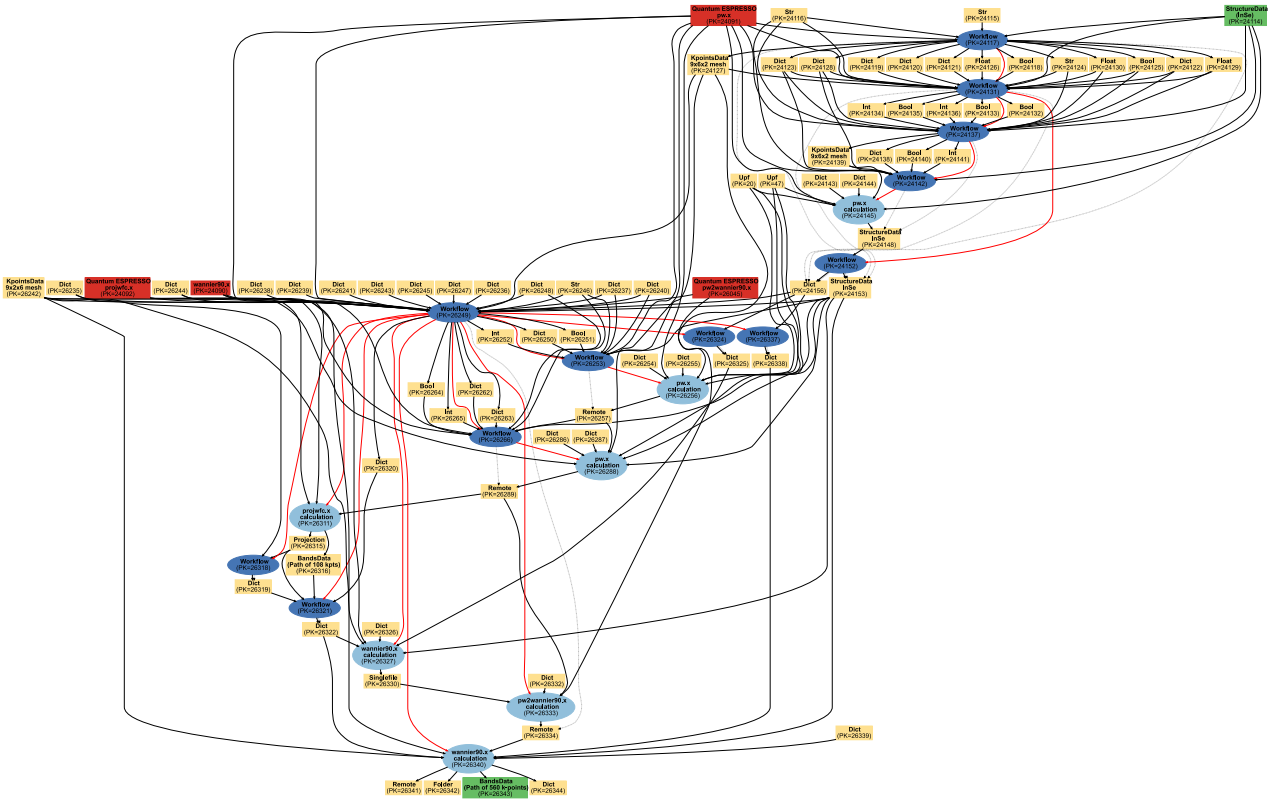

Fig. 15 Provenance graph automatically generated by AiiDA. The graph has been generated by running a WANNIER90 calculation using Quantum ESPRESSO as the input code for an InSe crystal, top green node (link labels have been removed for clarity). Red arrows represent caller–called relationships between a workflow and a subworkflow or a calculation; continuous lines connect calculations on a supercomputer (light blue ellipses) to their inputs and to the outputs they create, while dotted lines connect workflows (dark blue ellipses) to the data they return. Other data nodes are represented as yellow rectangles. In the top-right part of the graph, a set of workflows drive variable-cell relaxations of the initial structure via Quantum ESPRESSO; the central part contains the self-consistent, non-self-consistent and band-structure Quantum ESPRESSO calculations; in the bottom-left part are located the calculations computing the projection of the wavefunctions on a localised atomic basis set. At the bottom of the graph, we can find the WANNIER90 calculation, producing a set of output nodes that includes the Wannier-interpolated band structure (bottom green node).

higher energy states. The subsequent MLWF step does not use the information on the target band structure. Therefore, while mixing the states via the U matrix to minimise the spread, such spurious contributions can be distributed on the lower-energy states and, as a consequence, worsen the interpolation quality. However, we emphasise that in most cases, even when the MLWF algorithm increases the value of η, it does so only marginally: in 182 out of 200 systems (91%) the MLWF scheme either increases η by less than 5 meV or reduces it. More in detail, 163 out of these 182 materials show a variation $|\Delta\eta|$ within only 5.0 meV, and only one system among these exhibits $\eta^{\mathrm{MLWF}} > 20$ meV. Moreover, for the remaining 19 (out of 182) systems the MLWF procedure improves the bands interpolation, notably yielding $\eta^{\mathrm{MLWF}} < 20$ meV for all of them. Finally, for the remaining 18 systems (9%), the MLWF scheme worsens the results with $|\Delta\eta| > 5$ meV and only in six cases the interpolation quality is quite poor ($\eta^{\mathrm{MLWF}} > 20$ meV). In all these cases, a possible reason for failure might be related to the choice of columns in the SCDM algorithm, which is performed only at Γ (see discussion in "SCDM for periodic systems: SCDM-k"), for materials where the relative order of electronic states at Γ and at the BZ boundary is inverted. In this situation, spurious contributions might enter into the QR decomposition as discussed above.

We have presented an approach to generate a set of maximally localised Wannier functions in an automated way that has the advantage of being simple, robust and applicable also in the more general case of so-called entangled bands. The high sensitivity of iterative minimisation algorithms to the initial conditions, which was a long-standing problem in particular for the entangled-band case, is overcome by employing the selected columns of the density matrix[20,21] (SCDM) algorithm to automatically choose the initial subspace. For the Wannierisation of isolated bands, SCDM is a parameter-free method, whereas for entangled bands two real numbers μ and σ must be specified, whose appropriate choice is critical for the success of the method, in addition to the target dimensionality of the manifold to be described (i.e., the number of Wannier functions). We have proposed and validated a protocol to choose these parameters by leveraging information encoded in the projectability of the Bloch states on pseudo-atomic orbitals. We found that the SCDM method works very well for band-structure interpolations, but does not perform as well for other kind of applications where, for instance, a specific symmetry character of the WFs is desirable.

To make the method available to any researcher, we have implemented the SCDM algorithm in PW2WANNIER90, part of the open-source Quantum ESPRESSO distribution, and added corresponding functionality to the open-source WANNIER90 code. We have also discussed how the full procedure is implemented as AiiDA[25,26] workflows, encoding the knowledge that is needed to perform all steps (DFT simulations, selection of the parameters, Wannierisation) into an automated software. This enables MLWFs to be obtained and used to calculate material properties by providing the crystal structure of a material as the only input. Furthermore, we are distributing publicly and freely all codes and workflows discussed in this work within a virtual machine[61] preconfigured with the open source codes AiiDA, Quantum

ESPRESSO, and WANNIER90. This VM allows anyone to explore and reproduce straightforwardly the present results without the need to install or configure anything, and without the need of implementing again workflows and algorithms, in the true spirit of FAIR data sharing[72] and Open Science. In addition, interested researchers are not constrained to re-run the calculations performed in this work, but can perform their own simulations, either with different parameters or on new materials. To the best of our knowledge, this is the first time that such level of reproducibility is offered accompanying a scientific paper in the field of DFT simulations.

We have demonstrated the robustness of the present approach by carrying out high-throughput calculations on a dataset of 200 bulk crystalline materials, of which 81 are insulators, spanning a wide chemical and structural space. The main metric we used to assess the results is the so-called band distance[62], quantifying the difference between the Wannier-interpolated band structures and the corresponding direct DFT band structures. In particular, we obtain excellent interpolations: for entangled bands, 97% of the materials show an average bands distance $\eta < 20$ meV and 72% show $\eta < 5$ meV. For the insulating subset, when limiting to valence bands only, 93% show $\eta < 2$ meV.

We believe that this work is a significant step forward towards completely automated high-throughput calculations of advanced materials properties exploiting Wannier functions.

METHODS

AiiDA[25,26] is a python materials' informatics platform to automate, manage and coordinate simulations and workflows, and to encourage sharing of both the resulting data and the workflow codes used to generate them. While general in its design, its plugins cover many materials science codes, including Quantum ESPRESSO[73] and WANNIER90[74].

Our implementation of the SCDM method inside the open-source code Quantum ESPRESSO makes it available to any researcher. Moreover, our protocol for the choice of the SCDM parameters discussed in "Protocol" describes an effective procedure to automatically compute the Wannier functions of any material. However, the actual computation starting only from the crystal coordinates is non-trivial. The choice of numerical parameters (cutoffs, *k*-point grid density, convergence parameters) requires some prior knowledge and experience. Moreover, the full simulation for each material involves a complex sequence of steps, requiring a user to run over 10 different executables. Therefore, we have implemented the full procedure as AiiDA workflows, making it thus possible to repeat seamlessly the calculations for many different materials with minimal effort.

Furthermore, AiiDA keeps track of the provenance of the data generated in the simulations in a fully automated way, in the form of a directed graph (see Fig. 15 for an example of the provenance tracked for one material), where nodes can be calculations, workflows or data. This means that any researcher accessing the AiiDA database can inspect not only the final data, but also explore which calculation generated it, its relevant (raw and parsed) outputs and the complete set of its input parameters, and see how these input data were, in turn, obtained as output of previous calculations, traversing the graph up to the original input crystal structure.

The AiiDA workflows that we have written start by calling existing subworkflows available in the AiiDA-quantumespresso[73] plug-in that, given a crystal structure, perform a variable-cell atomic relaxation to obtain the converged DFT charge density. These workflows also contain useful heuristics and recovery mechanisms to reach convergence in case of common problems (e.g., by changing the diagonalisation algorithm) as well as automatic selection of parameters, including pseudopotentials and cutoffs from the SSSP library[62]. Once the charge density is computed, the workflow first standardises the cell using the symmetry-detection library spglib[75] and the seekpath[63] library that, in addition, provide a standardised band-structure path. Then, it proceeds along two parallel branches: on one side, it computes the DFT band structure along the suggested path. In parallel, it computes the Wannier functions: if first computes wavefunctions on a full uniform grid using a non-self-consistent Quantum ESPRESSO calculation, and then computes the PDOS, the projectabilities, and fits them to obtain the μ and σ parameters for the SCDM. Using these data, it prepares the WANNIER90 input file and runs it in pre-processing mode to generate the input file needed by the code interfacing Quantum ESPRESSO with WANNIER90 (PW2WANNIER90). The latter is then run to compute quantities needed by WANNIER90, including the $\mathbf{A}^{(\mathbf{k})}$ matrices obtained with the SCDM method. Finally, the workflow drives the execution of WANNIER90 to compute the (maximally-localised) Wannier functions and produce the output quantities of interest (spreads, interpolated band structure on the same path of the DFT code, plots of the Wannier functions, etc.).

In an effort to improve the verification and dissemination of computational results, and in order to make the present work available to all, we are distributing all codes and workflows discussed here within a preconfigured virtual machine (VM)[61] based on the Quantum Mobile VM[76] available on the Materials Cloud[77]. The relevant quantum codes (Quantum ESPRESSO, WANNIER90) and the informatics' platform AiiDA come pre-installed and configured in the VM, ready to run through the workflows described above. A simple README file guides new users in the installation of the VM and in the execution of the workflow, to compute—with essentially no user intervention—the interpolated band structure of a material of choice.

DATA AVAILABILITY

All data generated for this work can be obtained by downloading the publicly available Virtual Machine (VM) on the Materials Cloud (https://doi.org/10.24435/materialscloud:2019.0044/v2). The VM contains the AiiDA workflow, the structures of the 200 materials (in XSF format) and the simulation codes (Quantum ESPRESSO and WANNIER90). The latter have been pre-installed and, once configured, the VM is ready to be used. Inside, a README file explains in detail how to retrieve all data. In addition, the VM also contains the Ansible scripts to regenerate the VM from scratch.

CODE AVAILABILITY

All codes used for this work are open-source and hence available to any researcher. In particular, the latest stable version of WANNIER90 can be downloaded at http://www.wannier.org/download.

The latest stable version of Quantum ESPRESSO can be found at https://www.quantum-espresso.org/download.

Likewise, for the AiiDA code the latest stable version can be found at http://www.aiida.net/download.

Received: 30 August 2019; Accepted: 18 March 2020;
Published online: 01 June 2020

REFERENCES

1. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191 (2013).
2. Oba, F. & Kumagai, Y. Design and exploration of semiconductors from first principles: a review of recent advances. *Appl. Phys. Express* **11**, 060101 (2018).
3. Marzari, N. The frontiers and the challenges. *Nat. Mater.* **15**, 381 (2016).
4. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
5. Calzolari, A., Marzari, N., Souza, I. & Nardelli, M. B. Ab initio transport properties of nanostructures from maximally localized Wannier functions. *Phys. Rev. B* **69**, 035108 (2004).
6. Gresch, D. et al. Automated construction of symmetrized Wannier-like tight-binding models from ab initio calculations. *Phys. Rev. Mater.* **2**, 103805 (2018).
7. Yates, J. R., Wang, X., Vanderbilt, D. & Souza, I. Spectral and Fermi surface properties from Wannier interpolation. *Phys. Rev. B* **75**, 195121 (2007).
8. 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).
9. Marzari, N. & Vanderbilt, D. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B* **56**, 12847–12865 (1997).
10. Souza, I., Marzari, N. & Vanderbilt, D. Maximally localized Wannier functions for entangled energy bands. *Phys. Rev. B* **65**, 035109 (2001).
11. Mustafa, J. I., Coh, S., Cohen, M. L. & Louie, S. G. Automated construction of maximally localized Wannier functions: optimized projection functions method. *Phys. Rev. B* **92**, 165134 (2015).
12. Cancès, E., Levitt, A., Panati, G. & Stoltz, G. Robust determination of maximally localized Wannier functions. *Phys. Rev. B* **95**, 075114 (2017).
13. Agapito, L. A., Ferretti, A., Calzolari, A., Curtarolo, S. & Nardelli, M. B. Effective and accurate representation of extended Bloch states on finite Hilbert spaces. *Phys. Rev. B* **88**, 165127 (2013).

14. Agapito, L. A., Ismail-Beigi, S., Curtarolo, S., Fornari, M. & Nardelli, M. B. Accurate tight-binding Hamiltonian matrices from ab initio calculations: minimal basis sets. *Phys. Rev. B* **93**, 035104 (2016).
15. Agapito, L. A. & Bernardi, M. Ab initio electron–phonon interactions using atomic orbital wave functions. *Phys. Rev. B* **97**, 235146 (2018).
16. Rajen, N. & Coh, S. What can one learn about material structure given a single first-principles calculation? *Phys. Rev. Mater.* **2**, 053606 (2018).
17. Zhang, Z. et al. High-throughput screening and automated processing toward novel topological insulators. *J. Phys. Chem. Lett.* **9**, 6224–6231 (2018).
18. Olsen, T. et al. Discovering two-dimensional topological insulators from high-throughput computations. *Phys. Rev. Mater.* **3**, 024005 (2019).
19. Gresch, D. et al. Automated construction of symmetrized Wannier-like tight-binding models from ab initio calculations. *Phys. Rev. Mater.* **2**, 103805 (2018).
20. Damle, A., Lin, L. & Ying, L. Compressed representation of Kohn–Sham orbitals via selected columns of the density matrix. *J. Chem. Theory Comput.* **11**, 1463–1469 (2015).
21. Damle, A. & Lin, L. Disentanglement via entanglement: a unified method for Wannier localization. *Multiscale Model. Simul.* **16**, 1392–1410 (2018).
22. Aquilante, F., Pedersen, T. B., Sánchez de Merás, A. & Koch, H. Fast noniterative orbital localization for large molecules. *J. Chem. Phys.* **125**, 174101 (2006).
23. Giannozzi, P. et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter* **29**, 465901 (2017).
24. Mostofi, A. A. et al. An updated version of Wannier90: a tool for obtaining maximally-localised Wannier functions. *Comput. Phys. Commun.* **185**, 2309–2310 (2014).
25. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
26. Huber S. P. et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Preprint at https://arxiv.org/abs/2003.12476 (2020).
27. Thygesen, K. S., Hansen, L. B. & Jacobsen, K. W. Partly occupied Wannier functions. *Phys. Rev. Lett.* **94**, 026405 (2005).
28. Thygesen, K. S., Hansen, L. B. & Jacobsen, K. W. Partly occupied Wannier functions: construction and applications. *Phys. Rev. B* **72**, 125119 (2005).
29. Damle, A., Levitt, A. & Lin, L. Variational formulation for Wannier functions with entangled band structure. *Multiscale Model. Simul.* **17**, 167–191 (2019).
30. Shirley, E. L. Optimal basis sets for detailed Brillouin-zone integrations. *Phys. Rev. B* **54**, 16464–16469 (1996).
31. Prendergast, D. & Louie, S. G. Bloch-state-based interpolation: an efficient generalization of the Shirley approach to interpolating electronic structure. *Phys. Rev. B* **80**, 235126 (2009).
32. Wannier, G. H. The structure of electronic excitation levels in insulating crystals. *Phys. Rev.* **52**, 191–197 (1937).
33. Blount, E. *Formalisms of Band Theory*, Vol. 13 (Elsevier, 1962).
34. Duffin, R. J. Discrete potential theory. *Duke Math. J.* **20**, 233–251 (1953).
35. Stephan, U., Martin, R. M. & Drabold, D. A. Extended-range computation of Wannier-like functions in amorphous semiconductors. *Phys. Rev. B* **62**, 6885–6888 (2000).
36. Ku, W., Rosner, H., Pickett, W. E. & Scalettar, R. T. Insulating ferromagnetism in $La_4Ba_2Cu_2O_{10}$: an ab initio Wannier function analysis. *Phys. Rev. Lett.* **89**, 167204 (2002).
37. Lu, W. C., Wang, C. Z., Chan, T. L., Ruedenberg, K. & Ho, K. M. Representation of electronic structures in crystals in terms of highly localized quasiatomic minimal basis orbitals. *Phys. Rev. B* **70**, 041101 (2004).
38. Qian, X. et al. Quasiatomic orbitals for ab initio tight-binding analysis. *Phys. Rev. B* **78**, 245112 (2008).
39. Andersen, O. K. & Saha-Dasgupta, T. Muffin-tin orbitals of arbitrary order. *Phys. Rev. B* **62**, R16219–R16222 (2000).
40. Boys, S. F. Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another. *Rev. Mod. Phys.* **32**, 296–299 (1960).
41. Foster, J. M. & Boys, S. F. Canonical configurational interaction procedure. *Rev. Mod. Phys.* **32**, 300–302 (1960).
42. Foster, J. M. & Boys, S. F. A quantum variational calculation for HCHO. *Rev. Mod. Phys.* **32**, 303–304 (1960).
43. Panati, G. & Pisante, A. Bloch bundles, Marzari-Vanderbilt functional and maximally localized Wannier functions. *Commun. Math. Phys.* **322**, 835–875 (2013).
44. Golub, G. & Van Loan, C. *Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences* (Johns Hopkins University Press, 1996).
45. Cloizeaux, J. D. Analytical properties of *n*-dimensional energy bands and Wannier functions. *Phys. Rev.* **135**, A698–A707 (1964).
46. Prodan, E. & Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 11635–11638 (2005).
47. Benzi, M., Boito, P. & Razouk, N. Decay properties of spectral projectors with applications to electronic structure. *SIAM Rev.* **55**, 3–64 (2013).
48. Carlson, B. C. & Keller, J. M. Orthogonalization procedures and the localization of Wannier functions. *Phys. Rev.* **105**, 102–103 (1957).
49. Nenciu, G. Dynamics of band electrons in electric and magnetic fields: rigorous justification of the effective Hamiltonians. *Rev. Mod. Phys.* **63**, 91–127 (1991).
50. Brouder, C., Panati, G., Calandra, M., Mourougane, C. & Marzari, N. Exponential localization of Wannier functions in insulators. *Phys. Rev. Lett.* **98**, 046402 (2007).
51. He, L. & Vanderbilt, D. Exponential decay properties of Wannier functions and related quantities. *Phys. Rev. Lett.* **86**, 5341–5344 (2001).
52. Horsfield, A. P. & Bratkovsky, A. M. Ab initio tight binding. *J. Phys.: Condens. Matter* **12**, R1–R24 (1999).
53. Fang, S. et al. Ab initio tight-binding hamiltonian for transition metal dichalcogenides. *Phys. Rev. B* **92**, 205108 (2015).
54. Anisimov, V. I., Aryasetiawan, F. & Lichtenstein, A. I. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+U method. *J. Phys.: Condens. Matter* **9**, 767–808 (1997).
55. Schnell, I., Czycholl, G. & Albers, R. C. Hubbard-*U* calculations for Cu from first-principle Wannier functions. *Phys. Rev. B* **65**, 075103 (2002).
56. Novoselov, D., Korotin, D. M. & Anisimov, V. I. Hellmann–Feynman forces within the DFT+U in Wannier functions basis. *J. Phys.: Condens. Matter* **27**, 325602 (2015).
57. Georges, A., Kotliar, G., Krauth, W. & Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.* **68**, 13–125 (1996).
58. Lechermann, F. et al. Dynamical mean-field theory using Wannier functions: a flexible route to electronic structure calculations of strongly correlated materials. *Phys. Rev. B* **74**, 125120 (2006).
59. Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* **41**, 7892–7895 (1990).
60. Mostofi, A. A. et al. Wannier90: A tool for obtaining maximally-localised Wannier functions. *Comput. Phys. Commun.* **178**, 685–699 (2008).
61. Vitale, V. et al. Automated high-throughput wannierisation. *Materials Cloud Archive*. https://doi.org/10.24435/materialscloud:2019.0044/v2 (2019).
62. Prandini, G., Marrazzo, A., Castelli, I. E., Mounet, N. & Marzari, N. Precision and efficiency in solid-state pseudopotential calculations. *npj Comput. Mater.* **4**, 72 (2018).
63. Hinuma, Y., Pizzi, G., Kumagai, Y., Oba, F. & Tanaka, I. Band structure diagram paths based on crystallography. *Comput. Mater. Sci.* **128**, 140–184 (2017).
64. Gresch, D. et al. Automated construction of symmetrized Wannier-like tight-binding models from ab initio calculations. *Phys. Rev. Mater.* **2**, 103805 (2018).
65. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
66. Garrity, K. F., Bennett, J. W., Rabe, K. M. & Vanderbilt, D. Pseudopotentials for high-throughput DFT calculations. *Comput. Mater. Sci.* **81**, 446–452 (2014).
67. Corso, A. D. Pseudopotentials periodic table: from H to Pu. *Comput. Mater. Sci.* **95**, 337–350 (2014).
68. Schlipf, M. & Gygi, F. Optimization algorithm for the generation of ONCV pseudopotentials. *Comput. Phys. Commun.* **196**, 36–44 (2015).
69. Topsakal, M. & Wentzcovitch, R. Accurate projected augmented wave (PAW) datasets for rare-earth elements (RE = La–Lu). *Comput. Mater. Sci.* **95**, 263–270 (2014).
70. van Setten, M. et al. The PseudoDojo: training and grading a 85 element optimized norm-conserving pseudopotential table. *Comput. Phys. Commun.* **226**, 39–54 (2018).
71. Hamann, D. R. Optimized norm-conserving Vanderbilt pseudopotentials. *Phys. Rev. B* **88**, 085117 (2013).
72. Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*,**3** (2016).
73. AiiDA Plugin for the Quantum ESPRESSO Codes. http://github.com/aiidateam/aiida-quantumespresso (accessed 27 June 2019).
74. AiiDA Plugin for the Wannier90 Code. http://github.com/aiidateam/aiida-wannier90 (accessed 27 June 2019).
75. Togo, A. & Tanaka, I. Spglib: a software library for crystal symmetry search. Preprint at http://arxiv.org/abs/1808.01590 (2018).
76. Quantum Mobile on the Materials Cloud. https://www.materialscloud.org/work/quantum-mobile (accessed 27 June 2019).
77. Talirtz, L. et al. Materials Cloud, a platform for open computational science. Preprint at https://arxiv.org/abs/2003.12510 (2020).

ACKNOWLEDGEMENTS

V.V. acknowledges support from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 676531 (project E-CAM). G.P., A. M., and N.M. acknowledge support by the NCCR MARVEL of the Swiss National Science Foundation and the European Union's Centre of Excellence MaX "Materials design at the Exascale" (Grant No. 824143). G.P., A.M., and N. M. acknowledge PRACE for awarding us simulation time on Piz Daint at CSCS (project ID 2016153543) and Marconi at CINECA (project ID 2016163963). V.V. and A.A.M. acknowledge support from the Thomas Young Centre under grant TYC-101. J.R.Y. is grateful for

computational support from the UK national high performance computing service, ARCHER, for which access was obtained via the UKCP consortium and funded by EPSRC Grant Ref EP/P022561/1. V.V. acknowledges Prof. Mike Payne for support, and Prof. Lin Lin and Dr. Anil Damle for useful discussions. G.P. acknowledges Dr. Francesco Aquilante for useful discussions. A.M. acknowledges Prof. Ivo Souza for useful comments on the manuscript. The authors acknowledge Norma Rivano for testing the virtual machine and Dr. Sebastiaan P. Huber for the implementation of the AiiDA-Quantum ESPRESSO workflow for geometry relaxations.

AUTHOR CONTRIBUTIONS

V.V. implemented and tested the SCDM method on selected materials. G.P. and A.M. developed the automation protocols and the workflows, run the high-throughput simulations, and generated the Virtual Machine. N.M., A.A.M., and J.R.Y supervised the project. All authors analysed the results and contributed to writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-0312-y.

Correspondence and requests for materials should be addressed to V.V.