3 NMTOs and their Wannier Functions

Ole K. Andersen

Max Planck Institute for Solid State Research

Heisenbergstrasse 1, D-70569 Stuttgart

Contents

1 Introduction

The electronic structure of condensed matter is usually described in terms of one-electron *basis sets*. Basis functions used for computation, or rather, their envelopes are usually mathematically simple functions, plane waves or Gaussians, in particular. A plane wave is a solution of Schrödinger's equation for a flat potential, and products of plane waves are plane waves; as a result, the charge density and its Hartree potential are plane-waves as well. Similarly, a Gaussian is a solution of Schrödinger's equation for a parabolically increasing potential, products of Gaussians are Gaussians, and the Hartree potential for a Gaussian charge density is $1/r$ times the error function. However, in order for such sets to give accurate results, the number of basis functions must be orders of magnitude larger than the number of valence electrons to be described. This is so, because as illustrated at the top of Fig. 1, the potential $V(\mathbf{r})$ in the effective one-electron Schrödinger (Kohn-Sham) equation is neither flat inside the atoms nor parabolically increasing between them. Therefore, a plane-waves basis must include plane waves with energies spread over a region much larger than the one of the one-electron energies ε of interest, and a Gaussian basis must include Gaussians not only with many widths, but also at many positions displaced from those of the atoms. For a discussion of plane-wave basis sets, we refer to the last year's lecture notes [1] by Blöchl [2].

By virtue of solving Schrödinger's equation for a muffin-tin well, the classical linear muffin-tin orbitals (LMTOs) [3, 4] form a comparatively small basis set. But only in the atomic-spheres approximation (ASA) where the MTOs are expanded in partial waves inside atomic spheres, assumed to fill space, do the products $\varphi_l(\varepsilon, r) Y_{lm}(\hat{\mathbf{r}}) \times \varphi_{\bar{l}}(\bar{\varepsilon}, r) Y_{\bar{l}\bar{m}}^*(\hat{\mathbf{r}})$ have the same form, $f_{L'}(r) Y_{l'm'}(\hat{\mathbf{r}})$, as each factor, and this is what makes the LMTO-ASA method exceedingly fast. However, the ASA is only accurate when the atoms are at high-symmetry positions.

For many purposes it is therefore desirable to extract a *small set of intelligible, localized orbitals* spanning merely selected conduction and/or valence states. For instance, if we want to describe the bonding, we need a localized basis set which spans the occupied states only (bond orbitals). If we want to construct models which add interactions to the one-electron Hamiltonian, e.g. electron-electron repulsions, we need a basis set of localized, atomic-like orbitals which describes the one-electron energies and wave functions in a suitable region around the Fermi level.

For an isolated set of energy bands in a *crystal*, $\varepsilon_i(\mathbf{k})$ $(i = 1, ., A)$, this can be done by projecting from their delocalized Bloch eigenstates, $\Psi_i(\mathbf{k}; \mathbf{r})$ $(i = 1, ., A)$ computed with the large basis set, a suitable set of generalized *Wannier functions,* $w_a(\mathbf{r} - \mathbf{t})$. These are enumerated by $a \, (= 1, ., A)$ and the lattice translations, $\mathbf{t}$, of which there are $N \to \infty$, and they form a set of orthonormal functions related to the orthonormal Bloch eigenstates by a unitary transformation:

$$\Psi_i(\mathbf{k}; \mathbf{r}) = N^{-\frac{1}{2}} \sum_{\mathbf{t}} \sum_{a=1}^{A} u_{ia} \exp(i\mathbf{k} \cdot \mathbf{t}) \, w_a(\mathbf{r} - \mathbf{t}), \tag{1}$$

The inverse transformation:

$$w_a(\mathbf{r} - \mathbf{t}) = N^{-\frac{1}{2}} \sum_{\mathbf{k}} \sum_{i=1}^{A} u_{ai}^* \exp(-i\mathbf{k} \cdot \mathbf{t}) \, \Psi_i(\mathbf{k}; \mathbf{r}), \tag{2}$$
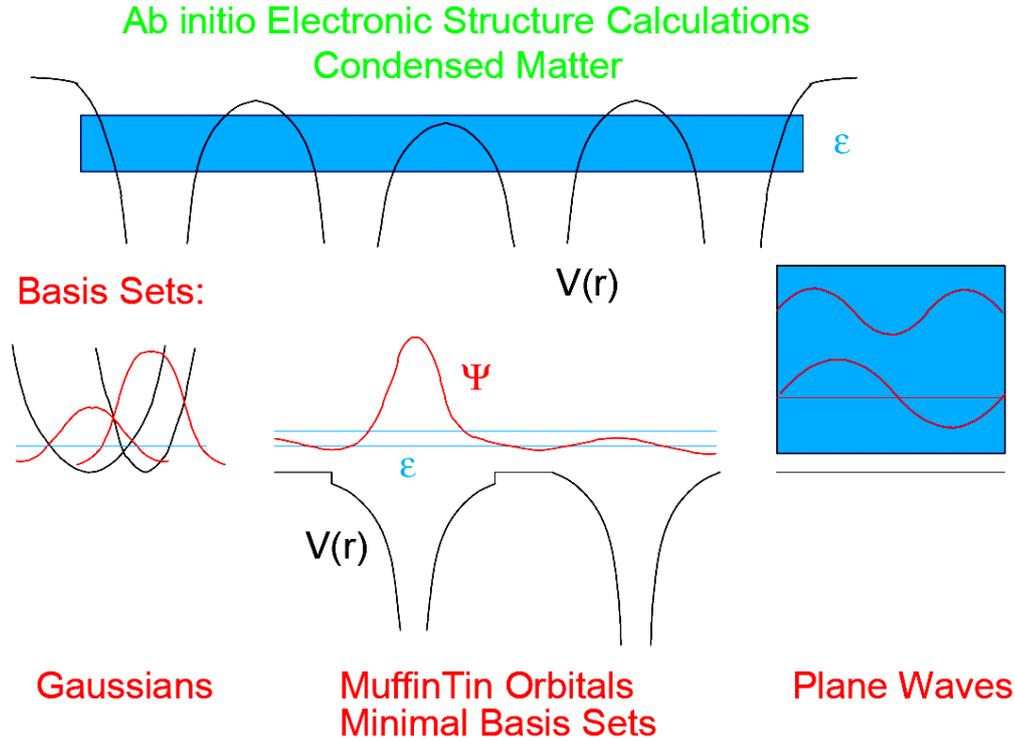
Fig. 1: *Top: Effective one-electron potential in condensed matter (black) and the energy region of interest (blue). Bottom: Basis functions (red), their generating potentials (black), and energies (blue).*

is the Wannier projection. Wannier functions are not unique, because performing a unitary transformation, $W_{a\bar{a};\mathbf{t}-\mathbf{t}'}$, of one set of Wannier functions produces another set which also satisfies Eq. (1), merely with different i and $\mathbf{k}$-dependent phases of the Bloch functions. So the art of Wannier projection from the Bloch states (2) is to choose the i and $\mathbf{k}$-dependent phases of the latter in such a way that the Wannier functions attain desired properties, in particular optimal localization – in some sense. Mazari and Vanderbilt chose to minimize the spread $\langle w \, | |\mathbf{r} - \langle w \, |\mathbf{r}| \, w \rangle|^2 | \, w \rangle$ and developed an – otherwise general – numerical procedure for projecting such "maximally localized" Wannier functions from Bloch states expanded in plane waves [5].

We shall only be interested in generating localized Wannier functions which resemble atomic orbitals, so-called *Wannier orbitals,* or simple linear combinations hereof such as bond orbitals. In this case, it is obvious that the phases in the projection (2) should be chosen such that when summing the Bloch states over i and $\mathbf{k}$, the atomic-orbital characters chosen for the Wannier functions should add up constructively. How localized the resulting Wannier orbitals are, then depends on how well the set of A bands are described by the characters chosen. This procedure was applied –presumably for the first time– by Satpathy and Pawlowska [6] to compute the sp^3 bond orbital in Si. They used the TB-LMTO basis [4] which makes the procedure quite obvious because the LMTO expansion has the same form as the expansion (2) in terms of Wannier functions, except that the unitary $A \times A$ matrix u_{ia} is replaced by the rectangular $A \times (A + P)$

matrix of LMTO eigenvectors. The projection is thus seen to be a *downfolding* in which each Wannier orbital becomes an *active* LMTO dressed by a tail of all the *passive* (P) LMTOs not in the set of active (A) ones. With other local-orbital basis sets, somewhat similar techniques can be used, but unless all basis functions are well localized, the Wannier orbitals obtained may not be sufficiently localized. For a further discussion we refer to last year's lecture notes by Kunes [7].

For molecules, Boys [8] had a long time ago recognized that chemical bonds should be associated with those linear combinations of the occupied molecular orbitals which are most localized, because those linear combinations are most invariant to the surroundings.

The present notes deal with a different kind of basis set, specifically with minimal bases of Nth-order muffin-tin orbitals (NMTOs), also known as 3rd generation MTOs [9–13]. We shall demonstrate that with NMTOs it is possible to generate Wannier functions *directly*, instead of via projection from the delocalized Bloch states. NMTOs are constructed from the partial-wave solutions of Schrödingers equation for a superposition of *overlapping* spherical potential wells (muffin tins, MTs) [14, 15] and NMTO sets are therefore selective in energy. As a consequence, one can construct an NMTO set which picks a specific set of isolated energy bands. Since NMTOs are atom-centered and localized by construction, they do –after symmetric orthonormalization– form a set of localized Wannier functions which, if needed, can be recombined locally to have maximal localization. The NMTO technique is primarily for generating a *localized, minimal basis set* with *specific orbital characters,* and it can therefore be used also to pick a set of bands which overlap other bands outside the energy region of interest [16]. The corresponding NMTOs –orthonormalized or not– we refer to as Wannier-like. Once a computationally efficient representation is implemented for products of NMTOs [17], they should be suitable for full-potential, real-space calculations with a computational effort increasing merely linearly with the size (N) of the system, so-called order-N calculations [18,19].

We start by explaining the LMTO idea of how to construct small basis sets of orbitals, $\chi_{Rlm}(\mathbf{r})$, from partial waves, $\varphi_{Rl}(\varepsilon, r) Y_{lm}(\hat{\mathbf{r}})$, and spherical waves, $h_l(\kappa r) Y_{lm}(\hat{\mathbf{r}})$. Then we define the set of exact, energy-*de*pendent MTOs (EMTOs) [20, 21], also called kinked partial waves (KPWs), which includes *downfolding* and employs *overlapping* MT spheres for their definition. The KPWs are used to derive, first the screened KKR equations, and then the NMTO basis sets. NMTOs with N=1 turn out to have the same form as classical TB-LMTOs in their atomic-spheres approximation (ASA), but without invoking this approximation. Examples of Wannier functions which are orthonormalized NMTOs are given along the way.

2 Classical LMTOs

The idea [22,3,23,24] of how to generate small basis sets of accurate orbitals can be understood by considering first the way in which Wigner and Seitz [25] thought about solving the one-electron eigenvalue problem for a close-packed solid; in case of a crystal, that is the band-structure problem. They divided space into WS cells and assumed the potential to be spherically

symmetric inside each cell,

$$V\left(\mathbf{r}\right) = \sum_{R} v_{R}\left(r_{R}\right).$$ (3)

Here and in the following, $r_{R} \equiv \left|\mathbf{r} - \mathbf{R}\right|$, and R labels the sites, $\mathbf{R}$. With this approximation, Schrödinger's equation (in atomic Rydberg units),

$$\left[\mathcal{H} - \varepsilon\right]\Psi\left(\varepsilon, \mathbf{r}\right) = \left[-\nabla^{2} + V\left(\mathbf{r}\right) - \varepsilon\right]\Psi\left(\varepsilon, \mathbf{r}\right) = 0,$$ (4)

can be treated as a *separable differential equation*: The eigenfunctions must have a *partial-wave expansion* inside each cell,

$$\Psi\left(\varepsilon, \mathbf{r}\right) = \sum_{lm} \varphi_{Rl}\left(\varepsilon, r_{R}\right) Y_{lm}\left(\hat{\mathbf{r}}_{R}\right) c_{Rlm},$$ (5)

and one may therefore proceed by first solving the *radial* Schrödinger equations,

$$-\left[r\varphi_{Rl}\left(\varepsilon, r\right)\right]'' = \left[\varepsilon - v_{R}\left(r\right) - l\left(l+1\right)/r^{2}\right]r\varphi_{Rl}\left(\varepsilon, r\right),$$ (6)

for all R and l and a given energy, ε, and then seek coefficients, c_{Rlm}, for which the partial-wave expansions join together continuously and differentiably at the cell boundaries. The energies for which this is possible are the eigenvalues, $\varepsilon = \varepsilon_{i}$, and $\Psi_{i}\left(\varepsilon_{i}, \mathbf{r}\right)$ the eigenfunctions.

This point of view for instance leads to the *approximate* Wigner-Seitz rules stating that for an elemental, close-packed crystal, where the cell can be substituted by an atomic sphere of the same volume ($\Omega = 4\pi s^{3}/3$), a band of l-character exists between the energies ε_{lB} and ε_{lA} for which respectively the slope and the value of $\varphi_{l}\left(\varepsilon, r\right)$ vanishes at the atomic sphere. These band edges correspond to the bonding and antibonding states of a homonuclear diatomic molecule. In this atomic-spheres approximation (ASA), the input to the band structure from the potential enters exclusively via the dimensionless, *radial logarithmic derivative functions*,

$$D\left\{\varphi_{l}\left(\varepsilon, s\right)\right\} \equiv \partial \ln \left|\varphi_{l}\left(\varepsilon, r\right)\right| / \partial \ln r \big|_{s} = s\varphi_{l}\left(\varepsilon, s\right)' / \varphi_{l}\left(\varepsilon, s\right),$$ (7)

evaluated at the WS sphere, s. These are ever decreasing functions of energy and the bonding/antibonding boundary condition is: $D\left\{\varphi_{l}\left(\varepsilon, s\right)\right\} = 0/\infty$. The shape of the l-partial wave thus changes significantly across a band of predominant l-character (Fig. 2) and, as we shall see in Sect. 3.4, even more across the hybridized sp^{3} valence band of Si.

To set up this matching problem correctly, however, it is necessary to deal with cells rather than spheres *and* with all the partial waves required to make the one-center expansion (5) converged at the cell-boundary. This takes $l \lesssim 15$ and is *not* practical.

Next, consider the customary and more general way of solving Schrödinger's equation, namely by use of the Raleigh-Ritz variational principle for the Hamiltonian with a trial function expressed as a superposition of basis functions, $\chi_{j}\left(\mathbf{r}\right)$,

$$\Psi\left(\mathbf{r}\right) \approx \sum_{j} \chi_{j}\left(\mathbf{r}\right) b_{j}.$$ (8)

Variation of the coefficients b_{j} leads to the algebraic, generalized eigenvalue problem:

$$\sum_{j} \left(\langle\chi_{\bar{j}}|\mathcal{H}|\chi_{j}\rangle - \varepsilon\langle\chi_{\bar{j}}|\chi_{j}\rangle\right) b_{j} = 0,$$ (9)
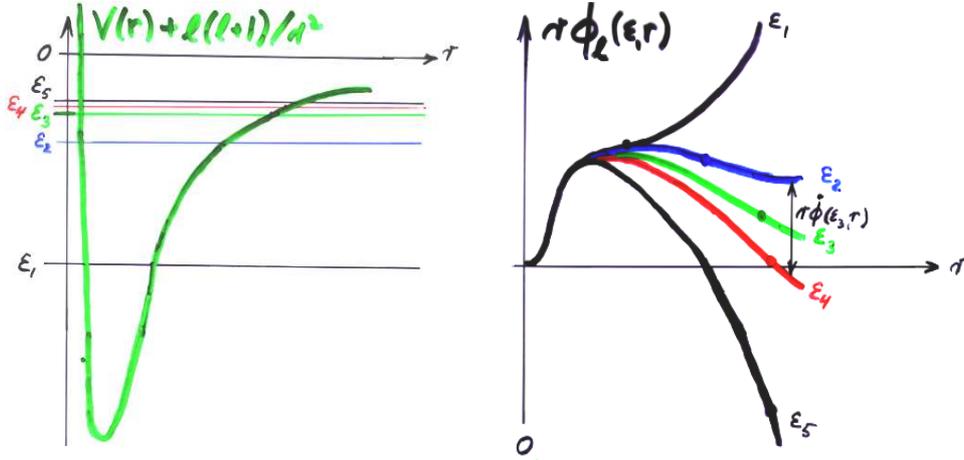
Fig. 2: *Radial potential (green) and energies (colored) together with the corresponding radial wave functions. The latter curve towards the r-axis in the classically allowed r-regions and away from the axis in the classically forbidden regions (schematic).*

for all $\bar{j}$, in terms of Hamiltonian and overlap matrices, $H = \langle \chi | \mathcal{H} | \chi \rangle$ and $O = \langle \chi | \chi \rangle$. The eigenvalues, ε_i, are variational estimates of the one-electron energies, and the eigenvectors, $b_{j,i}$, give the wave functions $\Psi_i(\mathbf{r})$.

The idea is now to construct the basis set in such a way that for the approximate model potential (3), the set solves Schrödinger's equation exactly to linear order in the deviation of the eigenvalue from an energy, ϵ_ν, chosen at the center of interest, i.e. such that the error is

$$\Psi_i(\mathbf{r}) - \Psi_i(\varepsilon_i, \mathbf{r}) \propto (\varepsilon_i - \epsilon_\nu)^2.$$

By virtue of the variational principle, the errors of the eigenvalues will then be of order $(\varepsilon_i - \epsilon_\nu)^4$. Imagine what such linear basis functions must look like if we choose them as atom-centered orbitals, $\chi_{Rlm}(\mathbf{r}_R)$: In order that the linear combination (8) be able to provide the correct eigenfunctions (5) for a spectrum of eigenvalues, ε_i, near ϵ_ν, the *tails* of the orbitals entering a particular cell (R) must, when expanded in spherical harmonics around $\mathbf{R}$, have radial parts which are *energy-derivative functions,* $\dot{\varphi}_{Rl}(\epsilon_\nu, r_R) \equiv \partial\varphi_{Rl}(\varepsilon, r_R)/\partial\varepsilon|_{\epsilon_\nu}$, because then, the sum of the tails added to the head of the orbital will be able to yield the result

$$\varphi_{Rl}(\epsilon_\nu, r_R) + (\varepsilon_i - \epsilon_\nu)\,\dot{\varphi}_{Rl}(\epsilon_\nu, r_R) = \varphi_{Rl}(\varepsilon_i, r_R) + O\left((\varepsilon_i - \epsilon_\nu)^2\right). \qquad (10)$$

Hence, the radial shape of a *head* must be $\varphi_{Rl}(\epsilon_\nu, r_R)$, plus maybe a bit of $\dot{\varphi}_{Rl}(\epsilon_\nu, r_R)$.
The condition that the spherical-harmonics expansion of the tail around site R have the radial behavior $\dot{\varphi}_{Rl}(\epsilon_\nu, r_R)$ for all lm and all R, might seem to determine the shape of the orbital completely and not even allow it to be smooth, but merely continuous. However, adding φ to $\dot{\varphi}$
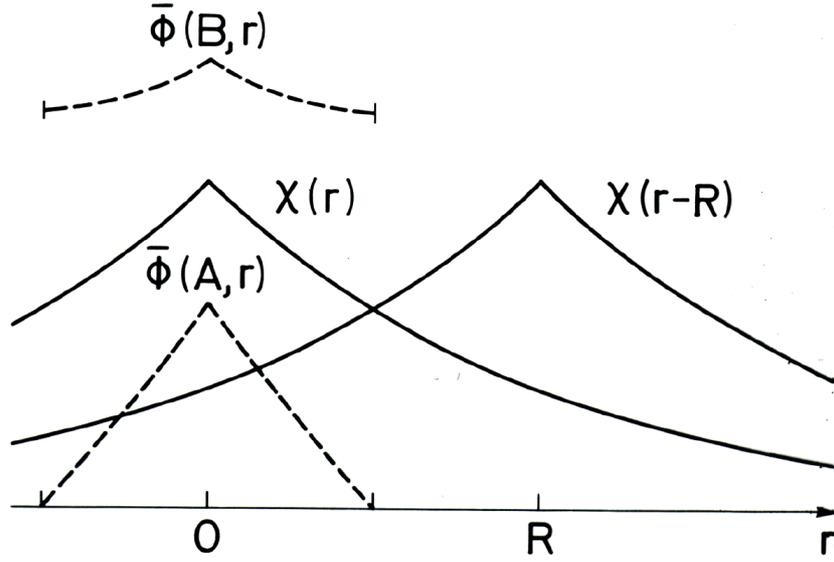
Fig. 3: *Partial waves for the bonding and antibonding states on the central site of a diatomic molecule and the LMTOs on the two sites (schematic). From Ref. [26].*

yields another $\dot{\varphi}$, corresponding to a different energy-dependent normalization, e.g.

$$\partial \left[1 + (\varepsilon - \epsilon_\nu)\, o \right] \varphi\left(\varepsilon, r\right) / \partial \varepsilon \big|_{\epsilon_\nu} = \dot{\varphi}\left(\epsilon_\nu, r\right) + o\varphi\left(\epsilon_\nu, r\right).$$

Hence, $\dot{\varphi}_{Rl}\left(\epsilon_\nu, r\right)$ can be adjusted to have, say, a required value and slope at some radius, a_R, where a linear combination of $\dot{\varphi}_{Rl}\left(\epsilon_\nu, r_R\right) Y_{lm}\left(\hat{\mathbf{r}}_R\right)$ functions can then be matched smoothly onto any given orbital shape.

Fig. 3 illustrates that in order to describe the bonding $\varphi\left(\varepsilon_B, r\right)$ and antibonding $\varphi\left(\varepsilon_A, r\right)$ states for a diatomic molecule with energy-independent orbitals (LMTOs), those orbitals must have heads proportional to $\varphi_l\left(\varepsilon_B, r\right) + \varphi_l\left(\varepsilon_A, r\right)$ and tails proportional to $\varphi_l\left(\varepsilon_B, r\right) - \varphi_l\left(\varepsilon_A, r\right)$. Now, the overbars on the partial waves in the figure indicate that particular normalizations have been chosen. Had we renormalized say $\varphi_l\left(\varepsilon_A, r\right)$ to $o\varphi_l\left(\varepsilon_A, r\right)$, the shapes of the heads and tails of the orbitals would have changed.

More practical than matching the partial waves at the cell boundaries, is therefore to embed the partial waves in a set of *envelope functions* or, from the point of view of the latter, to *augment* the envelope functions with partial waves. In order that the one-center expansions (5) converge in l, the envelope functions must be such that they match $\varphi_{Rl}\left(\varepsilon, r_R\right)$ for high l, whereby augmentation of the high-l waves becomes unnecessary, as long as they are taken into account as the high-l part of the envelopes. As l increases, the centrifugal term of the radial Schrödinger equation (6) drives $\varphi_{Rl}\left(\varepsilon, r\right)$ outwards such that eventually only the outermost, flat part of the potential is being felt. At that point,

$$\varphi_{Rl}\left(\varepsilon, r\right) \to j_l\left(r\kappa_R\right) \to const. \times r^l, \quad \text{where} \quad \kappa_R^2 \equiv \varepsilon - v_R\left(s_R\right). \tag{11}$$

Acceptable envelope functions are therefore decaying solutions of the wave equation:

$$\left[\bigtriangledown^2 + \kappa^2\right] h_l\left(\kappa r\right) Y_{lm}\left(\hat{\mathbf{r}}\right) = 0, \quad \text{with energies} \quad \kappa^2 \sim \varepsilon, \tag{12}$$

and potential zero taken as the average between the atoms. These spherical waves may be linear combined into short-ranged, so-called *screened spherical waves* (SSWs). What makes the orbitals spread out are thus the requirements (*a*) that the orbitals be smooth in all space (hermeticity), (*b*) that partial waves with high l should be treated as tails from low-l orbitals at neighboring sites, and (*c*) that the basis set should span a range of energies. The two latter requirements are to some extent flexible: (*b*) is a choice of how many partial waves to *downfold*, the rest having to be kept as (active) orbitals, and (*c*) is a choice of linearization or N-*ization* of the energy dependence of the partial waves. Finally, it should be noted that any delocalization is enhanced by orthogonalization. But all of this should become clear in the following sections. For all kinds of augmented basis sets, the model potential (3) defining the basis functions is a *superposition* of *spherically symmetric potential wells,* but their range, s_R, varies. LAPWs and LMTOs used in full-potential calculations employ muffin-tin potentials with non-overlapping spheres. Empty spheres –i.e. without nuclei– are included at interstitial sites for open structures. Owing to the sizeable interstitial region ($\sim 0.3\Omega$) and strong discontinuities of the potential at the spheres, such a MT-potential remains a bad approximation to the full potential, whose matrix elements must therefore be included in the Hamiltonian. Nevertheless, such a basis is not optimal and –whenever possible– one uses space-filling potential spheres with a positive radial overlap,

$$\omega^s_{R\bar{R}} \equiv \frac{s_R + s_{\bar{R}}}{\left|\mathbf{R} - \bar{\mathbf{R}}\right|} - 1, \tag{13}$$

not exceeding 20%, and usually neglects the associated errors. For comparison, the overlaps of the WS spheres in face-centered and body-centered cubic packings are respectively 11 and 14%. Nevertheless, if as in the LMTO-ASA method only the ASA potential is included in the one-electron Hamiltonian, the results are too dependent on the choice of sphere radii.

Since distances between close-packed spheres are small compared with the shortest wavelength $2\pi/\kappa_F$ of the valence electrons, the κ^2 dependence of the spherical waves (12) is of far less importance [27] than that of the ε dependence of the logarithmic derivatives (7). For that reason LMTOs of the 1st [22, 3] and 2nd [4] generations used $\kappa^2{\equiv}0$, thus simplifying the decaying Hankel functions to multipole potentials $\propto r^{-l-1}$, which got screened in the 2nd generation. With $\kappa{\equiv}0$ and the ASA, the WS rules for the energies of the band edges in an elementary close-packed solid could be generalized to the unhybridized band structures, $\varepsilon_{li}(\mathbf{k}) = fct(D_l)$, the so-called canonical bands [22, 3, 23, 24, 26, 28].

For the exact, energy-dependent MTOs (EMTOs) [20] with $\kappa^2{=}\varepsilon$, which we shall consider in the following section, the *overlap errors* turn out to be merely of 2nd order in the potential overlap [14] and, as a consequence, EMTOs can handle up to 50% overlap. The overlapping MT approximation (OMTA) is a least-squares fit to the full potential [14, 15] so that the MT discontinuities decrease with increasing overlap.

With EMTOs, also *downfolding* works perfectly [20], which was not the case with the old LM-TOs [29]. However, with increasing downfolding, the range of the EMTOs and herewith their energy dependence increases. So it became necessary to construct not only energy-independent linear basis sets, but also basis sets of *arbitrary order* without increasing the size of the set.

Specifically, for a mesh of $N + 1$ energies, $\epsilon_0, ..., \epsilon_N$, a basis set of Nth order will span the solutions of Schrödinger's equation for the model potential with the error

$$\Psi_i(\mathbf{r}) - \Psi_i(\varepsilon_i, \mathbf{r}) \propto (\varepsilon_i - \epsilon_0)(\varepsilon_i - \epsilon_1) ... (\varepsilon_i - \epsilon_N). \tag{14}$$

These are the so-called NMTO and NAPW basis sets, of which we shall consider the former.

2.1 Crystals

In the above, R runs over all spheres in the system. If it is a crystal with translations $\mathbf{t}$, the wave functions and the basis functions can be chosen to translate according to: $\Psi(\mathbf{r} + \mathbf{t}) = \Psi(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{t})$.

Orbitals can then be Bloch-summed:

$$\chi_{Rlm}(\mathbf{k}; \mathbf{r}) \equiv \sum_{\mathbf{t}} \chi_{Rlm}(\mathbf{r}_R - \mathbf{t}) \exp(i\mathbf{k} \cdot \mathbf{t}),$$

where R now labels the atoms in the primitive cell. Rather than normalizing the Bloch sums over the entire crystal, we have normalized them in the primitive cell. Accordingly, $\sum_{\mathbf{k}}$ must be taken as the average, rather than the sum, over the Brillouin zone. Matrices like the Hamiltonian are translationally invariant, $\langle \chi_{\bar{R}\bar{l}\bar{m}}(\mathbf{r}_{\bar{R}}) | \mathcal{H} | \chi_{Rlm}(\mathbf{r}_R - \mathbf{t}) \rangle = \langle \chi_{\bar{R}\bar{l}\bar{m}}(\mathbf{r}_{\bar{R}} + \mathbf{t}) | \mathcal{H} | \chi_{Rlm}(\mathbf{r}_R) \rangle$, and as a consequence,

$$\langle \chi_{\bar{R}\bar{l}\bar{m}}(\mathbf{k}; \mathbf{r}) | \mathcal{H} | \chi_{Rlm}(\mathbf{k}; \mathbf{r}) \rangle \equiv \sum_{\mathbf{t}} \langle \chi_{\bar{R}\bar{l}\bar{m}}(\mathbf{r}_{\bar{R}}) | \mathcal{H} | \chi_{Rlm}(\mathbf{r}_R - \mathbf{t}) \rangle \exp(i\mathbf{k} \cdot \mathbf{t}).$$

Numerical calculations are often carried out in the $\mathbf{k}$-representation, but since it is trivial to add $\mathbf{k}$ and limit R to the sites in the primitive cell, in formalisms for orbitals it is simpler and more general to use the real-space representation.

3 EMTOs

In this section we define the set of EMTOs (KPWs) [20, 21] and use them to derive the screened KKR equations (21). We first explain what the EMTOs are, starting with their envelope functions, and only thereafter, in Sect. 3.3, how to construct them.

Since EMTOs use *overlapping* MT-potentials for their definition, allow arbitrary downfoldings, enabling the construction of *truly* minimal sets [16], and are usually *localized*, their definition is tricky:

3.1 SSWs

The members, $h^\alpha_{\bar{R}\bar{l}\bar{m}}(\varepsilon, \mathbf{r}_R)$, of the set of envelope functions, $|h^\alpha(\varepsilon)\rangle$, are superpositions of the decaying spherical waves in Eq. (12), at all active sites R and with all active lm, but with the same energy, $\kappa^2 = \varepsilon$. The *set* of SSW envelopes are thus:

$$|h^\alpha(\varepsilon)\rangle = |h(\varepsilon)\rangle M^\alpha(\varepsilon), \tag{15}$$

with a notation in which a set of functions is considered a row-vector: $|h\left(\varepsilon\right)\rangle$, for instance, has the elements $h_l\left(\kappa r_R\right)Y_{lm}\left(\hat{\mathbf{r}}_R\right)\equiv h_{Rlm}\left(\varepsilon,\mathbf{r}\right)$ and $M^{\alpha}\left(\varepsilon\right)$ is a matrix with elements $M^{\alpha}_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right)$. The set of SSWs (15) is characterized by (*a*) the set of $\bar{R}\bar{l}\bar{m}$ values to be included in the set, the *active* values, (*b*) a set of non-overlapping screening spheres, so-called *hard-spheres,* with radii a_R for the active sites, and (*c*) the *phase shifts* $\eta_{Rl}\left(\varepsilon\right)$ of the MT potential for the remaining –the *passive*– Rlm values. With such a partition into active and passive channels, a choice of hard spheres for the former, and the phase shifts for the latter, we can state the *boundary condition* to be satisfied for a member, $h^{\alpha}_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_R\right)$, of the set:

Its spherical-harmonics projections,

$$\mathcal{P}^a_{Rlm}\equiv\int d^3r\,\delta\left(a_R-r_R\right)Y^*_{lm}\left(\hat{\mathbf{r}}_R\right),\qquad(16)$$

onto the hard spheres must *vanish* for all active Rlm values, except for the *own* value, $\bar{R}\bar{l}\bar{m}$, for which we choose to normalize the hard-sphere projection to 1. For the the passive Rlm values, the projection $\mathcal{P}^r_{Rlm}h^{\alpha}_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_{\bar{R}}\right)$ should be a spherical wave phase shifted by $\eta_{Rl}\left(\varepsilon\right)$. As discussed in connection with Eq. (12), this holds automatically for all partial waves with $l>1-3$ because their phase shifts vanish.

With this boundary condition satisfied, the passive channels can be *augmented smoothly* with the appropriate Schrödinger solutions, $\varphi_{Rl}\left(\varepsilon,r_R\right)$, and the active channels, which usually diverge at the origin of r_R, can be *truncated* inside the hard spheres, i.e. for $r_R<a_R$. This truncation of the active channels of $h^{\alpha}_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_{\bar{R}}\right)$ is continuous for $Rlm\neq\bar{R}\bar{l}\bar{m}$, but jumps by 1 in the own channel. In all active channels there is a discontinuity of outwards *slope,*

$$\left.\frac{\partial}{\partial r}\mathcal{P}^r_{Rlm}h^a_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_{\bar{R}}\right)\right|_{a_R}\equiv S^a_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right),\qquad(17)$$

(for the own channel, the derivative should be taken slightly *outside* the sphere), specified by a *slope matrix* whose calculation we shall explain in the section 3.3.

3.2   KPWs

The resulting augmented, truncated, and renormalized SSW, usually denoted $\psi^a_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_{\bar{R}}\right)$, is now ready to have the hole in its own channel (head) filled: The radial filling function is obtained by integrating the radial $\bar{R}\bar{l}$ equation (6) outwards from 0 to $s_{\bar{R}}$ with the proper potential, and from there, smoothly inwards to $a_{\bar{R}}$ with the flat (zero) potential. The solution, $\bar{\varphi}_{\bar{R}\bar{l}}\left(\varepsilon,r\right)$, for the flat potential, and of course the one, $\varphi_{\bar{R}\bar{l}}\left(\varepsilon,r\right)$, for the proper potential, are subsequently normalized such that the value of the former is 1 at $a_{\bar{R}}$. This, we indicate by a superscript a:

$$\varphi^a_{\bar{R}\bar{l}}\left(\varepsilon,r\right)\equiv\varphi_{\bar{R}\bar{l}}\left(\varepsilon,r\right)/\bar{\varphi}_{\bar{R}\bar{l}}\left(\varepsilon,a_{\bar{R}}\right),\quad\bar{\varphi}^a_{\bar{R}\bar{l}}\left(\varepsilon,r\right)\equiv\bar{\varphi}_{\bar{R}\bar{l}}\left(\varepsilon,r\right)/\bar{\varphi}_{\bar{R}\bar{l}}\left(\varepsilon,a_{\bar{R}}\right).\qquad(18)$$

Finally, $\bar{\varphi}^a_{\bar{R}\bar{l}}\left(\varepsilon,r\right)$ is matched continuously, but with a kink to $\psi^a_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}_{\bar{R}}\right)$, and it is truncated outside the interval $a_{\bar{R}}\leq r\leq s_{\bar{R}}$. The Schrödinger solution, $\varphi_{\bar{R}\bar{l}}\left(\varepsilon,r\right)$, is truncated outside the interval $0\leq r\leq s_{\bar{R}}$. Hence, the resulting radial function has been constructed like an

accordion: It starts from the origin as the regular Schrödinger solution which extends all the way out to the radius, $s_{\bar{R}}$, of the potential well. Here, it is matched smoothly to a phase-shifted wave, which then runs inwards to the radius, $a_{\bar{R}}$, of the hard sphere where it matches the SSW wave with a kink. Finally, the SSW continues outwards. The active channels of the SSW are truncated inside all hard spheres with kinks, and the passive channels are substituted smoothly inside all hard spheres with regular Schrödinger solutions. This is illustrated in Fig. 4.

The EMTO, also called the *kinked partial wave* (KPW) is now:

$$\phi^a_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon, r_{\bar{R}}\right) = \left[\varphi^a_{\bar{R}\bar{l}}\left(\varepsilon, r_{\bar{R}}\right) - \bar{\varphi}^a_{\bar{R}\bar{l}}\left(\varepsilon, r_R\right)\right] Y_{\bar{l}\bar{m}}\left(\hat{\mathbf{r}}_{\bar{R}}\right) + \psi^a_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon, \mathbf{r}_{\bar{R}}\right). \quad \blacktriangleleft \qquad (19)$$

Here, the first term is the product of a spherical harmonic times a radial function, which vanishes smoothly at, and outside the own potential-sphere $(s_{\bar{R}})$. The second term is the augmented and truncated SSW, which matches onto the first, pure-angular-momentum term at the central hard sphere with a kink of size $S^a_{\bar{R}\bar{l}\bar{m},\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) - \bar{\varphi}^a_{\bar{R}\bar{l}}\left(\varepsilon, a_{\bar{R}}\right)'$. Although the KPW is everywhere continuous, it has kinks at the hard spheres in all active channels, but is smooth in the passive channels.

We can now try to make a *linear combination*,

$$\sum^A_{Rlm} \phi^a_{Rlm}\left(\varepsilon_i, r_R\right) c^a_{Rlm,i}, \qquad (20)$$

of active (A) KPWs which is *smooth*. This requires that its coefficients satisfy the *kink-cancellation condition*,

$$\sum^A_{\bar{R}\bar{l}\bar{m}} K^a_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) c^a_{\bar{R}\bar{l}\bar{m}} = 0, \qquad (21)$$

for each Rlm. Here we have multiplied each Rlm-equation by a^2_R such that

$$K^a_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) \equiv a^2_R S^a_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) - a_R D\left\{\bar{\varphi}^a_{Rl}\left(\varepsilon, a_R\right)\right\} \delta_{R\bar{R}}\delta_{l\bar{l}}\delta_{m\bar{m}}. \quad \blacktriangleleft \qquad (22)$$

becomes a *Hermitian* matrix. Since the passive channels are smooth by construction, Eqs. (21) must be solved only for the active channels and therefore constitute a set of homogeneous, linear equations. These have a proper solution for those energies, ε_i, which make the determinant of the matrix vanish. Most importantly, the corresponding linear combination is a *solution of Schrödinger's equation* at energy ε_i for the overlapping MT potential to 1st order in the overlap. That this is true, can be seen from the following arguments: The kinks of a KPW are always between two solutions of the same radial wave equation, either partial-wave projections of SSWs, zero, or inwards integrated phase-shifted waves. Since only two linearly independent radial solutions exist, e.g. Bessel and Neumann functions, it follows that if they match without a kink at a_R, as they are required to do for the smooth linear combination of KPWs, then they must be *identical in the entire range* $a_R \le r \le s_R$. This means that throughout the MT-sphere at $\bar{R}$,

$$\sum^A_{Rlm} \psi^a_{Rlm}\left(\varepsilon_i, \mathbf{r}_{\bar{R}}\right) c^a_{Rlm,i} = \sum^A_{\bar{l}\bar{m}} \bar{\varphi}^a_{\bar{R}\bar{l}}\left(\varepsilon_i, r_R\right) Y_{\bar{l}\bar{m}}\left(\hat{\mathbf{r}}_{\bar{R}}\right) c^a_{\bar{R}\bar{l}\bar{m},i} + \sum^P_{\bar{l}\bar{m}} \varphi^a_{\bar{R}\bar{l}}\left(\varepsilon_i, r_R\right) Y_{\bar{l}\bar{m}}\left(\hat{\mathbf{r}}_{\bar{R}}\right) c^a_{\bar{R}\bar{l}\bar{m},i}.$$

Here, the last term comes from the passive (P) channels and the corresponding coefficients, $c^a_{P,i}$ are given by the solutions, $c^a_{A,i}$, of (21), times PA expansion coefficients. If site $\bar{R}$ is passive
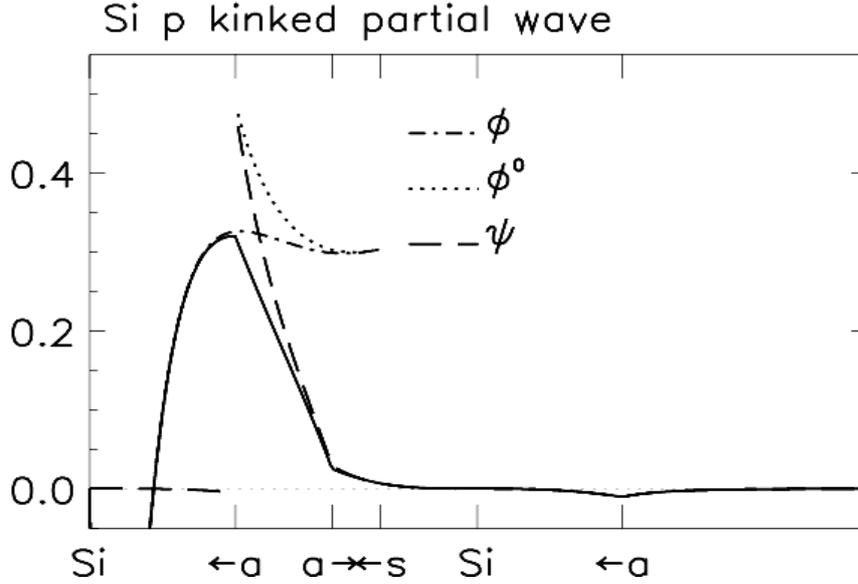
Fig. 4: *Si p_{111} KPW (full) and its constituents: the SSW $\psi_{p_{111}}(\epsilon_0, \mathbf{r})$ (dashed), the partial wave $\varphi_p^a(\epsilon_0, r) Y_{p_{111}}(111)$ labelled ϕ (dot-dashed) for the MT potential, and $\bar{\varphi}_p^a(\epsilon_0, r) Y_{p_{111}}(111)$ labelled ϕ^0 (dotted) for the flat potential. The plot is along the [111]-line from the central Si atom to its nearest neighbor, and from here into the large voids in the diamond structure. The latter were not described by empty-sphere potentials and, as a consequence, the MT overlap ω^s was as large as 50%; see definition (13). The overlap, ω^a, of the hard screening spheres was -25%. Kink are seen at the central and nearest-neighbor a-spheres. This KPW is the member of the 9 orbital/atom set of Si s, p, and d KPWs, so that the partial waves with $l > 2$ were downfolded into the SSWs. This is the reason why ψ does not exactly vanish inside the a-spheres. The value of ψ just outside the own a-sphere is $1 \times Y_{p_{111}}(111) < 1$. The energy of this KPW was chosen slightly above the bottom of the valence band. A 2D plot of this KPW in the $(1\bar{1}0)$-plane may be found in Fig. 5. From Ref. [12].*

(downfolded), only that term is present on the right-hand side. As a result, the smooth linear combination of KPWs reduces to:

$$\sum_{Rlm}^{A} \phi_{Rlm}^a(\varepsilon_i, r_R) c_{Rlm,i}^a = \sum_{\bar{l}\bar{m}}^{A+P} \varphi_{\bar{R}\bar{l}}^a(\varepsilon_i, r_{\bar{R}}) Y_{\bar{l}\bar{m}}(\hat{\mathbf{r}}_{\bar{R}}) c_{\bar{R}\bar{l}\bar{m},i}^a \tag{23}$$
$$+ \sum_{R \neq \bar{R}}^{A} \sum_{lm}^{A} [\varphi_{Rl}^a(\varepsilon_i, r_R) - \bar{\varphi}_{Rl}^a(\varepsilon_i, r_R)] Y_{lm}(\hat{\mathbf{r}}_R) c_{Rlm,i}^a,$$

near site $\bar{R}$. This is a solution of Schrödinger's equation, $\Psi_i(\varepsilon_i, \mathbf{r})$, plus an error consisting of *tongues* from the overlap of the neighboring muffin tins.

Now, the radial part of such a tongue is

$$\frac{1}{2}(s_R - r_R)^2 v_R(s_R) \varphi_{Rl}^a(\varepsilon, s_R),$$

to lowest order in $s_R - r_R$, as may be seen from the radial equation (6). Here, $v_R(s_R)$ is the MT-discontinuity. Operating finally with $\mathcal{H} - \varepsilon_i$ on the smooth linear combination (20), of

which (23) is the expansion around site $\bar{R}$, yields the error:

$$\sum_{\bar{R}} v_{\bar{R}} \left(r_{\bar{R}}\right) \sum_{R \neq \bar{R}}^{A} \sum_{lm}^{A} \left[\varphi_{Rl}^{a}\left(\varepsilon_i, r_R\right) - \bar{\varphi}_{Rl}^{a}\left(\varepsilon_i, r_R\right)\right] Y_{lm}\left(\hat{\mathbf{r}}_R\right) c_{Rlm,i}^{a}$$

$$\sim \frac{1}{2} \sum_{R\bar{R}}^{\text{pairs}} v_{\bar{R}}\left(r_{\bar{R}}\right) \left[\left(s_{\bar{R}} - r_{\bar{R}}\right)^2 + \left(s_R - r_R\right)^2\right] v_R\left(s_R\right) \Psi_i\left(\mathbf{r}\right),$$

which is obviously of 2nd order in the potential overlap. Q.E.D.

The set of homogeneous linear equations (21) are the *screened KKR equations,* albeit in radial-derivative gauge (denoted by a Latin superscript, e.g. a) rather than in phase-shift gauge (denoted by the corresponding greek superscript, α). For other uses of screened KKR –or multiple scattering– theory see e.g. Refs. [31] and [32].

Before we use this to derive NMTOs, let us explain how the slope-matrix is computed.

3.3   Structure matrix

The bare Hankel function to be used in the construction (15) of the SSW envelopes, is a spherical harmonics times the radial function,

$$\kappa^{l+1}\left[n_l\left(\kappa r\right) - i j_l\left(\kappa r\right)\right] = \left[\kappa^{l+1} n_l\left(\kappa r\right)\right] - i\kappa^{2l+1}\left[\kappa^{-l} j_l\left(\kappa r\right)\right] \tag{24}$$

$$\rightarrow -\frac{(2l-1)!!}{r^{l+1}}\left[1 + \frac{\varepsilon r^2}{2(2l-1)}\dots\right] - i\kappa\frac{(\varepsilon r)^l}{(2l+1)!!}\left[1 - \frac{\varepsilon r^2}{2(2l+3)}\dots\right],$$

for $\varepsilon \to 0$. Here, $(2l+1)!! \equiv (2l+1)(2l-1) \cdot \dots \cdot 3 \cdot 1$ and $(-1)!! \equiv 1$. For $\varepsilon = \kappa^2 \leq 0$, this Hankel function is real and decays asymptotically as $e^{-r|\kappa|}/r$. The spherical Neumann and Bessel functions, normalized as respectively $\kappa^{l+1} n_l\left(\kappa r\right)$ and $\kappa^{-l} j_l\left(\kappa r\right)$, are real for all real energies and they are respectively irregular and regular at the origin. For $\varepsilon > 0$, the Hankel function therefore has an imaginary part, which is the solution for the homogeneous problem. The energy region of interest for the valence and low-lying conductions bands is $\varepsilon \sim 0$, and the advantage of using *screened* Hankel functions (15), is that in this region there are no solutions to the homogeneous hard-sphere problem; they start at higher energies. The screened Hankel functions are therefore localized and real.

In order to obtain explicit expressions for the transformation and slope matrices, $M^a\left(\varepsilon\right)$ and $S^a\left(\varepsilon\right)$, we first need to expand a bare spherical wave centered at $\bar{R}$ in spherical harmonics around $R \neq \bar{R}$. Since the wave is regular around R, its expansion is in terms of Bessel functions and is:

$$n_{\bar{l}}\left(\kappa r_{\bar{R}}\right) Y_{\bar{l}\bar{m}}\left(\hat{\mathbf{r}}_{\bar{R}}\right) = \sum_{lm} j_l\left(\kappa r_R\right) Y_{lm}\left(\hat{\mathbf{r}}_R\right) \times$$

$$\sum_{l'} 4\pi i^{-\bar{l}+l-l'} C_{\bar{l}\bar{m},lm,l'}\, n_{l'}\left(\kappa \left|\bar{\mathbf{R}} - \mathbf{R}\right|\right) Y_{l'\, m-\bar{m}}^{*}\left(\widehat{\bar{\mathbf{R}} - \mathbf{R}}\right).$$

Here, $n_{\bar{l}}$ and $n_{l'}$ can be any l-independent linear combination of a Neumann and a Bessel function. For a pure Bessel function, the expansion holds for all r_R, while for an irregular function, it holds for $r_R < \left|\bar{\mathbf{R}} - \mathbf{R}\right| \neq 0$. The l'-summation runs over $\left|l - \bar{l}\right|$, $\left|l - \bar{l}\right| + 2$, ..., $l + \bar{l}$ whereby

$i^{-\bar{l}+l-l'}$ is real, and $C_{\bar{l}\bar{m},lm,l'} \equiv \int Y_{\bar{l}\bar{m}}(\hat{r}) Y^*_{lm}(\hat{r}) Y_{l'\,m-\bar{m}}(\hat{r}) d\hat{r}$. Now, since we shall renormalize the Bessel and Neumann functions when changing to radial-derivative gauge, we can start out in phase-shift gauge and use these functions without prefactors which make them real, and for the Hankel function use:

$$\kappa\left[n_l\left(\kappa r\right) - i j_l\left(\kappa r\right)\right] \equiv h_l^{\alpha=0}\left(\kappa r\right).$$

The conventional bare structure matrix is then

$$B^{\alpha=0}_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) \equiv \sum_{l'} \left(-\right)^{\frac{-\bar{l}+l-l'}{2}} 4\pi C_{\bar{l}\bar{m},lm,l'} h^0_{l'}\left(\kappa\left|\bar{\mathbf{R}}-\mathbf{R}\right|\right) Y^*_{l'\,m-\bar{m}}\left(\widehat{\bar{\mathbf{R}}-\mathbf{R}}\right) \quad\blacktriangleleft \qquad (25)$$

and if we define the on-site part of the structure matrix as $B^0_{Rlm,\bar{R}\bar{l}\bar{m}}\left(\varepsilon\right) = -i\kappa\delta_{l\bar{l}}\delta_{m\bar{m}}$, the one-center expansions may be written as:

$$\left|h^0\right\rangle = \left|\kappa n\right\rangle + \left|j\right\rangle B^0. \qquad (26)$$

Here and in the following we drop the common energy argument.

This screening transformation (15) is now defined by the requirement that the set of screened Hankel functions have one-center expansions formally similar to (26):

$$\left|h^\alpha\right\rangle = \left|h^0\right\rangle M^\alpha = \left|\kappa n\right\rangle + \left|j^\alpha\right\rangle B^\alpha, \qquad (27)$$

but with modified radial tail-functions:

$$j^\alpha_{Rlm}\left(\varepsilon,r\right) \equiv j_l\left(\kappa r\right) - n_l\left(\kappa r\right)\tan\alpha_{Rlm}\left(\varepsilon\right). \qquad (28)$$

For the active channels, these should vanish at the hard sphere and for the passive channels, they should join onto the proper Schrödinger solutions. Hence, $\alpha_{Rlm}\left(\varepsilon\right)$ is the *hard-sphere phase shift* when Rlm is *active* and the *proper phase shift* when Rlm is *passive*, i.e.:

$$\tan\alpha_{Rlm}\left(\varepsilon\right) = \frac{j_l\left(\kappa a_R\right)}{n_l\left(\kappa a_R\right)} \frac{D\left\{j^\alpha_{Rlm}\left(\varepsilon,a\right)\right\} - D\left\{j_l\left(\kappa a_R\right)\right\}}{D\left\{j^\alpha_{Rlm}\left(\varepsilon,a\right)\right\} - D\left\{n_l\left(\kappa a_R\right)\right\}} \qquad (29)$$

with

$$D\left\{j^\alpha_{Rlm}\left(\varepsilon,a\right)\right\} \equiv \begin{cases} \infty & Rlm \in A \\ D\left\{\bar{\varphi}_{Rl}\left(\varepsilon,a_R\right)\right\} & Rlm \in P \end{cases}.$$

$\alpha\left(\varepsilon\right)$ depends on m, only if the division into active and passive channels is m-dependent. This is the case, say, if one wants to select the Cu $d_{x^2-y^2}$ conduction band in a high-temperature superconducting cuprate [9, 12].

If we now insert Eqs. (28) and (26) in (27) and compare the coefficients of $\left|\kappa n\right\rangle$ and $\left|j\right\rangle$, we obtain the following expressions for the screening transformation and the screened structure matrix:

$$M^\alpha = 1 - \frac{\tan\alpha}{\kappa} B^\alpha, \qquad (30)$$

$$B^\alpha = \kappa\cot\alpha - \kappa\cot\alpha\left[B^0 + \kappa\cot\alpha\right]^{-1}\kappa\cot\alpha.$$

Here, all matrices are square with the high-l blocks neglected (see Eq. (11)) and $\kappa \cot \alpha$ is a diagonal matrix. We see that the amount of lm-multipole charge at site R which screens the $\bar{l}\bar{m}$-multipole at site $\bar{R}$, is $(\tan \alpha_{Rl}/\kappa) B^\alpha_{Rlm,\bar{R}\bar{l}\bar{m}}$. By taking the radial derivatives at the hard spheres, we can find the desired expression for the slope matrix:

$$a^2 S^a (\varepsilon) = aD \{j (\kappa a)\} + \frac{1}{j (\kappa a)} \left[B^0 (\varepsilon) + \kappa \cot \alpha (\varepsilon) \right]^{-1} \frac{1}{j (\kappa a)}. \quad \blacktriangleleft \qquad (31)$$

Note that $\kappa \cot \alpha (\varepsilon)$ is real for all real energies and that

$$j (\kappa a) \kappa \cot \alpha (\varepsilon) j (\kappa a) \to -\frac{1}{(2l + 1) a} \frac{D \{j^\alpha (\varepsilon, a)\} + l + 1}{D \{j^\alpha (\varepsilon, a)\} - l}, \quad \text{for } \varepsilon \to 0.$$

For most purposes, the hard screening spheres can be taken to depend only on the type of atom, and it turns out that for respectively $spdf$-, spd-, sp-, and s-sets, the shortest range of the spherical waves is obtained for radial overlaps (13) of $\omega^a_{R\bar{R}} = $ –15, –23, –36, and –52%. In the first two cases, the range of the structure matrix is so short that it can be generated by inversion of $B^0 (\varepsilon) + \kappa \cot \alpha (\varepsilon)$ in real space, using clusters of 20-100 sites, depending on the hard-sphere packing. Whereas a bare Hankel function has *pure* $\bar{l}\bar{m}$ character, and the bare structure matrix therefore transforms according to the Slater-Koster scheme, the screened structure matrix does *not,* because a screened Hankel function merely has *dominant* $\bar{l}\bar{m}$-character and tends to avoid the surrounding hard spheres.

Downfolding of channels with attractive potentials increases the range and energy dependence of the structure matrix (31). Downfolding is therefore usually performed as a second, **k**-space step, after the strongly screened structure matrix has been generated in real space and subsequently Bloch-summed to **k**-space.

3.4   Example: sp^3 bonded Si

As an example, let us now consider the effects of downfolding on the Si p_{111} member of the spd set of KPWs shown in Fig. 4. This KPW set was for an energy ϵ_0 near the bottom of the valence band. Using also the KPW set for an energy ϵ_1 at the middle of the valence band, plus the one for ϵ_3 near the top of the valence band, an spd NMTO set with $N = 2$ can be formed. How, will be explained in the following section. The p_{111} KPWs at the three energies, the NMTO, and the band structure obtained with the set of nine spd NMTOs per atom are all shown in the left-hand double column of Fig. 5. The middle column shows the same for the sp set, that is the one where also the Si d partial waves are downfolded, and therefore only contains four NMTOs per atom. With 2 atoms per cell, the first set yields 18 bands and the latter 8. Those bands are seen to be identical. In order to describe merely the filled bands, the 4 valence bands, we have to construct a set with merely 4 orbitals per cell and this we do by starting from the symmetry-breaking, completely ionic description $Si^{4-}Si^{4+}$ and put the s and p orbitals on Si^{4-} and *none* on Si^{4+}. That is, all partial waves are downfolded on every second Si atom. The corresponding KPW and NMTO sets are shown in the last column. We see that this NMTO set does give the

valence band only, and that it does so very well. Such a set which picks merely the occupied bands, we call truly minimal [16].

The pictures KPW(E0) show how for the spd set the p_{111} KPW has a kink at its own hard sphere and vanishes inside the neighboring sphere, except for the f and higher partial waves which are allowed to penetrate. Allowing also the d partial waves inside the neighbor has fairly little effect, but allowing all partial waves makes the KPW spill smoothly onto the site. Going now to higher energies, the KPW(E1) and KPW(E2) pictures show how the central kink vanishes as the p radial function bends increasingly toward the axis and how the KPW spreads increasingly around the neighboring hard sphere. Without any confinement in that sphere, the delocalization increases dramatically near the top of the valence band. Nevertheless, the NMTO valence band Wannier orbitals are correct and their sp^3 hybrid shown in Fig. 6 is the well-known bond orbital, which is as localized as can be. Examples for graphite sp^2 σ-bonds, as well as for p_z π-bonds and anti-bonds, may be found in Ref. [16].

4   NMTOs

Finally, we have come to construct energy-*in*dependent orbitals [10]. Specifically, we want to make a superposition of the set of KPWs (19), evaluated at a mesh of energies, $\epsilon_0, .., \epsilon_N$, such that the resulting set of NMTOs,

$$\left|\chi^{(0..N)}\right\rangle = \sum_{n=0}^{N} \left|\phi\left(\epsilon_n\right)\right\rangle L_n^{(0..N)} = \qquad (32)$$
$$\left|\phi\left[0\right]\right\rangle + \left|\phi\left[01\right]\right\rangle \left(E^{(0..N)} - \epsilon_0\right) + \; . \; + \left|\phi\left[0..N\right]\right\rangle \left(E^{(N-1,N)} - \epsilon_{N-1}\right) . \left(E^{(0..N)} - \epsilon_0\right), \; \blacktriangleleft$$

spans the solutions of Schrödinger's equation for the model potential to within an error given by Eq. (14). This is *discrete polynomial approximation* for a *Hilbert space,* and $L_n^{(0..N)}$ are Lagrangian matrices, whose sum is the unit matrix. For $N = 0$, the NMTO set is the set of EMTOs evaluated at the energy $\epsilon_0 \equiv \epsilon_\nu$. The second, rearranged series is the ascending Newton interpolation formula in terms of divided differences, e.g. $\phi\left[0\right] \equiv \phi\left(\epsilon_0\right)$ and $\phi\left[01\right] \equiv \frac{\phi(\epsilon_0) - \phi(\epsilon_1)}{\epsilon_0 - \epsilon_1}$. In general, they are defined by:

$$\phi\left[0..N\right] \equiv \sum_{n=0}^{N} \frac{\phi\left(\epsilon_n\right)}{\prod_{m=0, \neq n}^{N} \left(\epsilon_n - \epsilon_m\right)}. \qquad (33)$$

Moreover, $\left(E^{(N-1,N)} - \epsilon_{N-1}\right) . \left(E^{(0..N)} - \epsilon_0\right)$ is a product of N energy *matrices,* which are generally not Hermitian and do not commute. The NMTO is independent of the order of the energy points, but the individual terms in the Newton series are not, and only when the energies are ordered according to size does this series have a clear interpretation. If the energy mesh condenses onto ϵ_ν, then $\phi\left[0..N\right] \to \overset{(N)}{\phi}\left(\epsilon_\nu\right)/N!$ and the Newton series becomes a truncated Taylor series. In order to be able to pick bands which overlap other bands, it is necessary to put the energies where only active bands are present. With only one energy point at disposal, there is little flexibility, so the Taylor series is not practical.
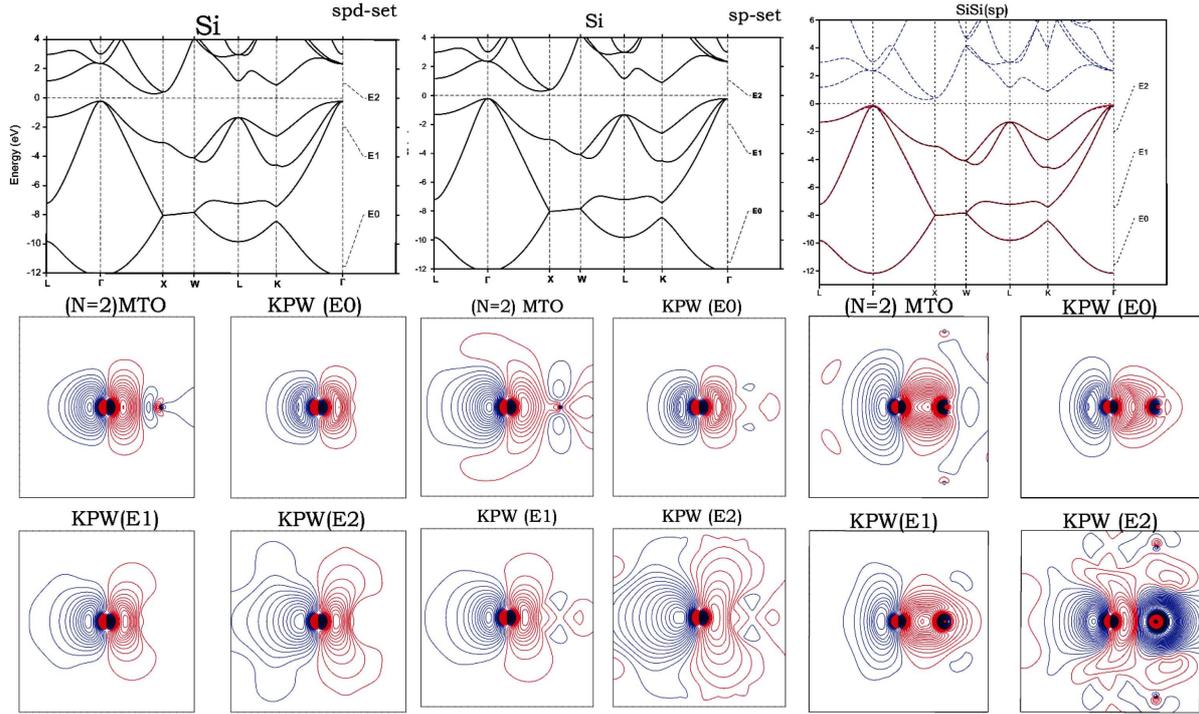
Fig. 5: *Band structure of Si obtained with $(N{=}2)$MTO sets of increasing downfolding. The first set contains the 9 s, p, and d NMTOs per atom, the second the 4 s and p NMTOs per atom, and the last merely the s and p NMTOs on every second atom. The Si p_{111} members of the corresponding NMTO sets, as well as of the constituting KPW sets at the three energies, ϵ_0, ϵ_1, and ϵ_2, indicated to the right of the band structures, are shown in the $(1\bar{1}0)$ plane containing a Si and its nearest neighbor along $[111]$. The NMTO bands are red and the exact ones blue. For the first two sets, no difference can be seen. The last set is seen to give merely the valence bands, and that very well. After orthonormalization, this NMTO set is thus a set of Wannier functions for the valence band. By being placed only on every second atom, this NMTO set breaks the symmetry, but does spills onto the other atoms correctly because the sp^3 hybrid of the orthonormalized NMTOs yields the well-known, symmetric bond orbital shown in Fig. 6. Hence, by starting from the ionic $Si^{4-}Si^{4+}$ picture, which gets the electron count right, the NMTO method creates the correct covalency. From Ref. [12]*

We have dropped all superscripts a because, from now on they do not change; screening and downfolding is done at the level of forming the EMTOs. Note that, in contrast to LMTO sets of the 2nd generation [4], NMTO sets for different screenings span different Hilbert spaces; the factor in front of the error term (14) depends on $s - a$ [10]. It is obvious that for N given, the error must increase with the degree of downfolding, because downfolding decreases the size of the basis set. This, on the other hand, makes it necessary to go beyond *linear* basis sets if one wants to generate truly minimal basis sets picking merely the *occupied* bands (see Fig. 6).

NMTOs can be used to generate Wannier functions directly, because with an appropriate choice of active channels, one can generate an NMTO set for the isolated set of bands in question. Upon making the mesh finer, the NMTO set will converge to the proper Hilbert space spanned by any set of Wannier functions. After orthonormalization, the NMTO set will therefore *be* a set
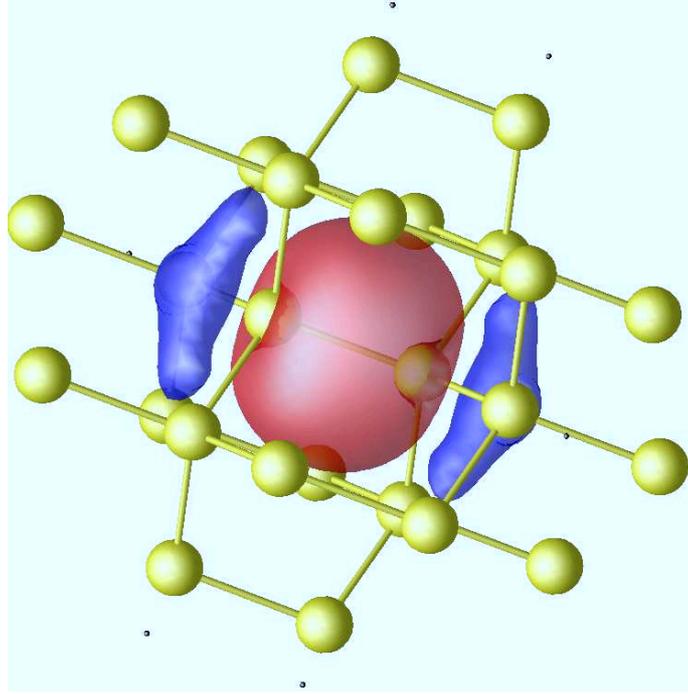
Fig. 6: *Si sp^3 bond orbital computed as the sp^3 directed orbital from the set of s and p NMTOs on every second atom. See Fig. 5. In order that no asymmetry of the bond orbital could be seen, it was necessary to use $N = 6$. (Courtesy A. Alam).*

of Wannier functions. NMTOs are localized *a priory* by virtue of the hard-sphere confinement of the constituent EMTOs, and since NMTOs are not orthonormal, they can –but must not– be more localized than maximally localized Wannier functions.

We shall now see that the Lagrangian matrices as well as the Hamiltonian and overlap matrices for the model potential, are all expressed solely in terms of the kink- or KKR matrix (22) and its first energy derivative matrix evaluated at points of the energy mesh. In fact, the NMTO formalism is much simpler if expressed in terms of the Green matrix,

$$G\left(\varepsilon\right) \equiv K\left(\varepsilon\right)^{-1}, \qquad \blacktriangleleft \tag{34}$$

also called the resolvent or scattering path operator [31, 32].

Since a single KPW (19) solves Schrödinger's differential equation for the model potential, except at the kinks, operation with the Hamiltonian gives a series of delta-functions at the hard spheres in the active channels:

$$\left(\varepsilon - \mathcal{H}\right)\phi_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon, \mathbf{r}\right) = \sum_{Rlm}^{A} \delta\left(r_R - a_R\right) Y_{lm}\left(\hat{\mathbf{r}}_R\right) K_{Rlm, \bar{R}\bar{l}\bar{m}}\left(\varepsilon\right). \tag{35}$$

Solving for $\delta\left(r_R - a_R\right) Y_{lm}\left(\hat{\mathbf{r}}_R\right),$ leads to:

$$\delta\left(r_R - a_R\right) Y_{lm}\left(\hat{\mathbf{r}}_R\right) = \left(\varepsilon - \mathcal{H}\right) \sum_{\bar{R}\bar{l}\bar{m}}^{A} \phi_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon, \mathbf{r}\right) G_{\bar{R}\bar{l}\bar{m}, Rlm}\left(\varepsilon\right) \tag{36}$$

which shows that the linear combinations,

$$\gamma_{Rlm}\left(\varepsilon,\mathbf{r}\right) = \sum_{\bar{R}\bar{l}\bar{m}}^{A} \phi_{\bar{R}\bar{l}\bar{m}}\left(\varepsilon,\mathbf{r}\right) G_{\bar{R}\bar{l}\bar{m},Rlm}\left(\varepsilon\right), \tag{37}$$

of KPWs –all with the same energy and screening– may be considered a Green function, $G\left(\varepsilon,\bar{\mathbf{r}},\mathbf{r}\right)$, which has $\bar{\mathbf{r}}$ *confined* to the hard spheres, i.e. $\bar{\mathbf{r}} \to Rlm$. Considered a function of $\mathbf{r}$, this Green function is a solution with energy ε of the Schrödinger equation, except at its own sphere and for its own angular momentum, where it has a kink of size unity. This kink becomes negligible when ε is close to a one-electron energy, because the Green function has a pole there. In Eq. (37), the confined Green function is factorized into a vector of KPWs, $|\phi\left(\varepsilon\right)\rangle$, which has the full spatial dependence and a weak energy dependence, and a Green matrix, $G\left(\varepsilon\right)$, which has the full energy dependence. Now, we want to factorize the $\mathbf{r}$ and ε-dependences *completely* and, hence, to approximate the confined Green function, $|\phi\left(\varepsilon\right)\rangle G\left(\varepsilon\right)$, by $\left|\chi^{(0..N)}\right\rangle G\left(\varepsilon\right)$.

Note that subtracting from the Green function a function which is analytical in energy and remains in the Hilbert space spanned by the set $|\phi\left(\epsilon_n\right)\rangle$ produces an equally good Green function, in the sense that both yield the same solutions of Schrödinger's equation. We therefore first define a set $\left|\chi^{(0..N)}\left(\varepsilon\right)\right\rangle$ by:

$$\left|\gamma\left(\varepsilon\right)\right\rangle = \left|\phi\left(\varepsilon\right)\right\rangle G\left(\varepsilon\right) \equiv \left|\chi^{(0..N)}\left(\varepsilon\right)\right\rangle G\left(\varepsilon\right) + \sum_{n=0}^{N} \left|\phi\left(\epsilon_n\right)\right\rangle G\left(\epsilon_n\right) F_n^{(0..N)}\left(\varepsilon\right), \tag{38}$$

and then determine the analytical functions, $F_n^{(0..N)}\left(\varepsilon\right)$, in such a way that $\left|\chi^{(0..N)}\left(\varepsilon\right)\right\rangle$ takes the *same* value, $\left|\chi^{(0..N)}\right\rangle$, at all mesh points. If that can be done, then

$$\left|\chi^{(0..N)}\left(\varepsilon\right)\right\rangle = \left|\chi^{(N)}\right\rangle + O\left(\left(\varepsilon - \epsilon_0\right)..\left(\varepsilon - \epsilon_N\right)\right),$$

and $\left|\chi^{(0..N)}\right\rangle$ is the set of NMTOs. Now, since

$$\left|\chi^{(0..N)}\left(\epsilon_0\right)\right\rangle = .. = \chi^{(0..N)}\left(\epsilon_N\right),$$

the Nth divided difference of $\left|\chi^{(0..N)}\left(\varepsilon\right)\right\rangle G\left(\varepsilon\right)$ equals $\left|\chi^{(0..N)}\right\rangle$ times the Nth divided difference of $G\left(\varepsilon\right)$. Moreover, if we let $F_n^{(0..N)}\left(\varepsilon\right)$ be a polynomium of $(N$-1)st degree (Nth degree yields zero-solutions for the NMTOs), their Nth divided difference on the mesh will vanish. As a result

$$\left|\gamma\left[0..N\right]\right\rangle = \left(|\phi\rangle G\right)\left[0..N\right] = \left|\chi^{(0..N)}\right\rangle G\left[0..N\right],$$

and we have therefore found the solution:

$$\left|\chi^{(0..N)}\right\rangle = \left(|\phi\rangle G\right)\left[0..N\right] G\left[0..N\right]^{-1} \tag{39}$$

for the NMTO set. The divided difference of the product is easily evaluated using (33):

$$\left(|\phi\rangle G\right)\left[0..N\right] = \sum_{n=0}^{N} \frac{|\phi\left(\epsilon_n\right)\rangle G\left(\epsilon_n\right)}{\prod_{m=0,\neq n}^{N}\left(\epsilon_n - \epsilon_m\right)},$$

in terms of the values of the KPWs and the Green matrix on the energy mesh. This expression, together with the similar one for $G\left[0..N\right]$, are those needed to determine the Lagrange matrices in Eq. (32).

4.1   Smoothness and products of NMTOs

(N=0)MTOs are of course the *kinked* partial waves at ϵ_0, but (N>0)MTOs are *smooth* because according to (36), the kinks of $|\gamma(\varepsilon)\rangle$ are independent of ε. This does however not imply that for a single NMTO, the KPW accordion is completely compressed, like for a smooth linear combination of KPWs (20) with the same energy. The linear combinations making up an NMTO have *different* energies and, as a consequence, discontinuities remain in ($2N$+1)st radial derivatives at the hard spheres. Projecting an NMTO onto an active channel, leads to a radial function of the type $\varphi(r) - \bar{\varphi}(r) + \mathcal{P}^r \psi(\mathbf{r})$, where $\varphi(r) - \bar{\varphi}(r) \propto (s-r)^2$ near s and

$$\mathcal{P}^r \psi(\mathbf{r}) - \bar{\varphi}(r) \propto (r-a)^{2N+1} (\varepsilon - \epsilon_0)\,..\,(\varepsilon - \epsilon_N)$$

near a. Since the latter error is of the same order as (14), it should be included there. This means that *cross-terms* between φ, $\bar{\varphi}$, and $\mathcal{P}^r \psi$ can be *neglected*, and that leads to the following simple prescription for evaluating the *product* of two KPWs with different energies:

$$|\phi\rangle \langle\phi| \;=\; |\varphi Y\rangle \langle Y\varphi| \;-\; |\bar{\varphi}Y\rangle \langle Y\bar{\varphi}| \;+\; |\psi\rangle \langle\psi|, \tag{40}$$

occurring in the expression for the product $\left|\chi^{(N)}\right\rangle \left\langle\chi^{(N)}\right|$ of two NMTOs as needed for evaluation of matrix elements and the *charge density.* The sum of the first two terms in (40) is simply a finite sum of spherical harmonics times radial functions which vanish smoothly outside the MT spheres. The third term is more complicated because the SSWs do not have pure lm-character but merely short range. What we know about the SSWs is the structure matrix which specifies the spherical-harmonics expansions of the radial derivatives at the hard spheres. It is therefore practical to *interpolate* a *product* of *strongly* screened spherical waves across the hard-sphere interstitial by a *sum* of SSWs. Specifically, we fit –at all spheres and for all spherical-harmonics with $l \lesssim 6-$ the radial values plus first 3 derivatives of the product (e.g. the charge-density) to those of a sum of SSWs with 4 different energies. The so-called *value-and-derivative functions,* each one vanishing in all channels except its own, are purely structural and exceedingly well localized because the value and first 3 derivatives vanish at all *other* spheres. We are currently writing an efficient self-consistent, full-potential NMTO code using this interpolation technique [17].

In order to figure out how the Hamiltonian operates on an NMTO, we use Eq (35) for N=0 and obtain: $(\mathcal{H} - \varepsilon_0)\left|\chi^{(0)}\right\rangle = -|\delta\rangle K(\epsilon_0)$. For the smooth NMTOs with N>0 we can neglect the kink terms when operating on (38), and then take the Nth divided difference to get rid of the polynomials:

$$\mathcal{H}\,|\gamma[0..N]\rangle = |(\varepsilon\gamma(\varepsilon))\,[0..N]\rangle = |\gamma[0.N-1]\rangle + \epsilon_N\,|\gamma[0..N]\rangle. \tag{41}$$

Using the definition (39) of the NMTO we multiply by $G\,[0..N]^{-1}$ from the right and obtain:

$$(\mathcal{H} - \epsilon_N)\left|\chi^{(0..N)}\right\rangle = |\gamma[0.N-1]\rangle\,G\,[0..N]^{-1} = \left|\chi^{(0.N-1)}\right\rangle \left(E^{(0..N)} - \epsilon_N\right), \tag{42}$$

where $\left|\chi^{(0.N-1)}\right\rangle$ is the set obtained by omitting the last point on the mesh and

$$E^{(0..N)} - \epsilon_N \equiv G\,[0..N-1]\,G\,[0...N]^{-1}$$

is the coefficient of all but the first term of the descending Newton series analogous to the ascending one in (32). The energy matrices are in general given by:

$$E^{(0..M)} = (\varepsilon G) \, [0..M] \; G \, [0..M]^{-1} \tag{43}$$

Expression (42) shows that increasing N increases the smoothness of the NMTOs and also their range, unless $E^{(0..N)}$ converges as is the case for a set of isolated bands. If $E^{(0..N)}$ is converged, so is the NMTO basis, and so is the Newton series. This series expresses the NMTO as a kinked partial wave at the same site and with the same angular momentum, plus a smoothing cloud of energy-derivative functions centered at all sites and with all angular momenta.

4.2   Hamiltonian and overlap matrices

With the aim of obtaining the expressions for the overlap and Hamiltonian matrices needed in a variational calculation (9), we first find expressions involving $|\phi(\varepsilon)\rangle$ and $|\gamma(\varepsilon)\rangle$.
Multiplication of (35) from the left by $\langle \phi(\varepsilon)|$ and using (40), together with the facts that $\bar{\varphi}(\varepsilon, a) = 1$, that $\mathcal{P}^a \psi(\varepsilon, \mathbf{r}) = 1$ in the own channel, 0 in the other active channels, and solves the radial Schrödinger equation in the passive channels, leads to the result:

$$\langle \phi(\varepsilon) \, |\mathcal{H} - \varepsilon| \, \phi(\varepsilon) \rangle = -K(\varepsilon). \qquad \blacktriangleleft \tag{44}$$

Here again we have resorted to matrix notation. The Hamiltonian matrix for the $N{=}0$ set is thus

$$\langle \chi^{(0)} \, |\mathcal{H} - \epsilon_0| \, \chi^{(0)} \rangle = -K(\epsilon_0). \tag{45}$$

In a similar way, and with the use of Green's second theorem, one finds that the overlap matrix between two EMTOs with different energies is:

$$\langle \phi(\bar{\varepsilon}) \, |\phi(\varepsilon) \rangle = \frac{K(\bar{\varepsilon}) - K(\varepsilon)}{\bar{\varepsilon} - \varepsilon} \rightarrow \dot{K}(\varepsilon), \quad \text{for } \bar{\varepsilon} \rightarrow \varepsilon. \tag{46}$$

Note that by virtue of the definition of $|\psi\rangle$, there are no 3-center terms here. Hence, the overlap matrix for the $N{=}0$ set is simply:

$$\langle \chi^{(0)} | \chi^{(0)} \rangle = \langle \phi(\epsilon_0) \, |\phi(\epsilon_0) \rangle = \dot{K}(\epsilon_0). \tag{47}$$

From Eqs. (44), (46), and (37) one finds:

$$\langle \gamma(\bar{\varepsilon}) \, |\gamma(\varepsilon) \rangle = -\frac{G(\bar{\varepsilon}) - G(\varepsilon)}{\bar{\varepsilon} - \varepsilon} \rightarrow \dot{G}(\varepsilon) = G(\varepsilon) \, \dot{K}(\varepsilon) \, G(\varepsilon), \; \text{for } \bar{\varepsilon} \rightarrow \varepsilon. \tag{48}$$

If we now take the Mth divided difference with respect to $\bar{\varepsilon}$ and the Nth with respect to ε, both on the mesh, then use (33) and order such that $M \leq N$, we find a double sum. If reordered to a single sum, with due care taken for the terms where $\bar{\varepsilon}{=}\varepsilon$, it reduces to the expression

$$\langle \gamma \, [0..M] \, |\gamma \, [0...N] \rangle = -G \, [[0..M] \, ..N], \tag{49}$$

where the right-hand side is minus the highest derivative of that polynomium of degree $M + N + 1$ which coincides with $G(\varepsilon)$ at the points $\epsilon_0, ..., \epsilon_N$ *and* has the same first derivatives $\dot{G}(\varepsilon)$ at the points $\epsilon_0, .., \epsilon_M$ (Hermit interpolation) [10]. For the matrix element of the Hamiltonian, expressions (41) and (49) yield:

$$\langle \gamma\,[0..N]\,|\mathcal{H} - \epsilon_N|\,\gamma\,[0..N]\rangle = \langle \gamma\,[0..N]\,|\gamma\,[0.N-1]\rangle = -G\,[[0.N-1]\,N]\,.$$

The NMTO Hamiltonian and overlap matrices are thus given by the following, most elegant expression which involves nothing but the values and first derivatives of the KKR Green matrix, $G(\varepsilon)$, on the energy mesh:

$$G\,[0..N]\,\langle \chi^{(0..N)}\,|\mathcal{H} - \varepsilon|\,\chi^{(0..N)}\rangle\,G\,[0..N] = -G\,[[0.N-1]\,N] + (\varepsilon - \epsilon_N)\,G\,[[0..N]]\,, \quad (50)$$

i.e.

$$\langle \chi^{(0..N)}\,|\mathcal{H} - \epsilon_N|\,\chi^{(0..N)}\rangle = -G\,[0..N]^{-1}\,G\,[[0.N-1]\,N]\,G\,[0..N]^{-1} \quad\blacktriangleleft \quad (51)$$

and

$$O^{(0..N)} \equiv \langle \chi^{(0..N)}\,|\,\chi^{(0..N)}\rangle = G\,[0..N]^{-1}\,G\,[[0..N]]\,G\,[0..N]^{-1}\,. \quad\blacktriangleleft \quad (52)$$

The variational calculation will give eigenvalues, which for the model potential has errors proportional to $(\varepsilon_i - \epsilon_0)^2\,(\varepsilon_i - \epsilon_1)^2\,..\,(\varepsilon_i - \epsilon_N)^2$.

4.3 Orthonormal NMTOs (Wannier orbitals)

In many cases one would like to work with a set of *orthonormal N*MTOs, e.g. Wannier orbitals, and preserve the Rlm-character of each NMTO. In order to arrive at this, we should – in the language of Löwdin – perform a *symmetrical orthonormalization* of the NMTO set. According to (52), such a representation is obtained by the following transformation:

$$\left|\check{\chi}^{(0..N)}\right\rangle = \left|\chi^{(0..N)}\right\rangle\,G\,[0..N]\,\sqrt{-G\,[[0..N]]}^{\,-1} = \left|\chi^{(0..N)}\right\rangle\,\sqrt{O^{(0..N)}}^{\,-1}\,, \quad\blacktriangleleft \quad (53)$$

because it yields:

$$\left\langle \check{\chi}^{(0..N)}\,|\,\check{\chi}^{(0..N)}\right\rangle = -\sqrt{-G\,[[0..N]]}^{\,-1\dagger}\,G\,[[0...N]]\,\sqrt{-G\,[[0..N]]}^{\,-1} = 1.$$

Note that this means: $-G\,[[0..N]] = \sqrt{-G\,[[0..N]]}^{\,\dagger}\sqrt{-G\,[[0..N]]}$. In this orthonormal representation, the Hamiltonian matrix becomes:

$$\left\langle \check{\chi}^{(0..N)}\,|\mathcal{H} - \epsilon_N|\,\check{\chi}^{(0..N)}\right\rangle = -\sqrt{-G\,[[0..N]]}^{\,-1\dagger}\,G\,[[0.N-1]\,N]\,\sqrt{-G\,[[0..N]]}^{\,-1}\,. \quad\blacktriangleleft \quad (54)$$

To find an efficient way to *compute* the *square root* of the Hermitian, positive definite matrix $-G\,[[0...N]]$ may be a problem. Of course one may diagonalize the matrix, take the square root of the eigenvalues, and then back-transform, but this is time consuming. Cholesky decomposition is a better alternative, but that usually amounts to staying in the original representation. Löwdin orthogonalization works if the set is nearly orthogonal, because then the overlap matrix is nearly diagonal, and Löwdin's solution was to normalize the matrix such that it becomes 1 along the diagonal and then expand in the off-diagonal part, Δ :

$$\sqrt{1 + \Delta}^{\,-1} = 1 - \frac{1}{2}\Delta + \frac{3}{8}\Delta^2 - ... \quad (55)$$

This should work for the NMTO overlap matrix (52) when the NMTOs are nearly orthogonal.

4.4 LMTOs

For $N=0$, we have the results: $\left|\chi^{(0)}\right\rangle = \left|\phi\left(\epsilon_0\right)\right\rangle$, (45), and (47).

For comparison with classical TB-LMTOs [4, 26, 33, 34], now consider the case $N=1$ with the two-point mesh condensed onto ϵ_ν. From Eq. (43) we find the following energy matrix:

$$E^{(1)} = \epsilon_\nu + G\dot{G}^{-1} = \epsilon_\nu - \dot{K}^{-1}K = \epsilon_\nu + \langle\phi|\phi\rangle^{-1}\langle\phi\,|\mathcal{H} - \epsilon_\nu|\,\phi\rangle,$$

Here and in the following an omitted energy argument means that $\varepsilon=\epsilon_\nu$. Insertion in the Taylor series (32), yields:

$$\left|\chi^{(1)}\right\rangle = |\phi\rangle - \left|\dot{\phi}\right\rangle \dot{K}^{-1}K, \tag{56}$$

which shows that the LMTO is smooth and has the form anticipated in Sect. 2. The Hamiltonian and overlap matrices are from Eq. (50):

$$\left\langle\chi^{(1)}\,|\mathcal{H} - \epsilon_\nu|\,\chi^{(1)}\right\rangle = -\dot{G}^{-1}\frac{\ddot{G}}{2!}\dot{G}^{-1} = -K + K\dot{K}^{-1}\frac{\ddot{K}}{2!}\dot{K}^{-1}K,$$

$$\left\langle\chi^{(1)}\,|\,\chi^{(1)}\right\rangle = -\dot{G}^{-1}\frac{\dddot{G}}{3!}\dot{G}^{-1} = \dot{K} - K\dot{K}^{-1}\frac{\ddot{K}}{2!} - \frac{\ddot{K}}{2!}\dot{K}^{-1}K + K\dot{K}^{-1}\frac{\dddot{K}}{3!}\dot{K}^{-1}K.$$

Had we instead used the Taylor series (56) to compute the overlap matrix, we would of course have obtained the same result and as consequences, $\ddot{K}=2!\left\langle\phi|\dot{\phi}\right\rangle$ and $\dddot{K}=3!\left\langle\dot{\phi}|\dot{\phi}\right\rangle$. This may also be obtained from the general relation (49). Had we used the Taylor series to compute the Hamiltonian matrix, we would have used Eq. (42) with $N=1$, to obtain the same result.

In order to make $E^{(1)}$ Hermitian and, hence, to transform it into a 1st-order *Hamiltonian:*

$$\dot{K}^{\frac{1}{2}}E^{(1)}\dot{K}^{-\frac{1}{2}} = \epsilon_\nu - \dot{K}^{-\frac{1}{2}}K\dot{K}^{-\frac{1}{2}} \equiv H^\alpha,$$

one must symmetrically orthonormalize the 0th-order set, which now becomes:

$$\left|\check{\chi}^{(0)}\right\rangle = |\phi\rangle\,\dot{K}^{-\frac{1}{2}} = |\phi\rangle\,\langle\phi|\phi\rangle^{-\frac{1}{2}} \equiv |\phi^\alpha\rangle.$$

Here and in the following, the superscript α is the one used in the classical LMTO [4] –not the new NMTO– formalism. After applying the same transformation to the LMTO set (56), it becomes:

$$\left|\chi^{(1)}\right\rangle\dot{K}^{-\frac{1}{2}} = |\phi^\alpha\rangle + \left|\dot{\phi}^\alpha\right\rangle(H^\alpha - \epsilon_\nu) = |\chi^\alpha\rangle, \quad \blacktriangleleft$$

where $\left|\dot{\phi}^\alpha\right\rangle = \left|\dot{\phi}\right\rangle\dot{K}^{-\frac{1}{2}}$. This expression for the LMTO is the one envisaged in expression (10) of Sect. 2: The tail-functions are $\dot{\phi}^\alpha(\mathbf{r})$ and the head of the $\bar{R}\bar{l}\bar{m}$-orbital is

$$\phi^\alpha_{\bar{R}\bar{l}\bar{m}}(\mathbf{r}_{\bar{R}}) + \sum_{lm}\dot{\phi}^\alpha_{\bar{R}lm}(\mathbf{r}_{\bar{R}})(H^\alpha - \epsilon_\nu)_{\bar{R}lm,\bar{R}\bar{l}\bar{m}}.$$

In order to show explicitly how the solutions of Schrödinger's equation for the solid can be described through overlap of orbitals, we may simply diagonalize H^α. Naming its eigenvectors and eigenvalues respectively $u_{Rlm,i}$ and ε_i, the linear combination of orbitals given by an eigenvector is:

$$\begin{aligned}|\chi^\alpha\rangle u_i &= |\phi^\alpha\rangle u_i + \left|\dot{\phi}^\alpha\right\rangle H^\alpha u_i = \left[|\phi^\alpha\rangle + \left|\dot{\phi}^\alpha\right\rangle(\varepsilon_i - \epsilon_\nu)\right]u_i \\ &= |\phi^\alpha(\varepsilon_i)\rangle u_i + O\left((\varepsilon_i - \epsilon_\nu)^2\right),\end{aligned}$$

as anticipated. LMTOs thus naturally describe the way in which the overlap of orbitals leads to broadening of levels into bands. It is however worth pointing out that although this $(N=1)$MTO formalism has been brought to the same form as the classical TB-LMTO formalism in the ASA, the $(N=1)$MTO formalism employs *no* ASA.

4.5   Example: NiO

The Mott insulator NiO has the NaCl structure so that each Ni is surrounded by an O octahedron and vice versa. In the ionic picture, the configuration is Ni $3d^8$ with the 2 electrons of highest energy in the e_g orbitals. LDA band structures are shown on the left-hand side of Fig. 7 with the corresponding sets of $(N=1)$MTO Wannier orbitals on the right-hand side.

Starting from the bottom, we see the 3 O p bands and the 3 congruent Wannier orbitals which span those green bands. No discrepancy can be seen between the green and the highly accurate LAPW bands (black). We remember that, except for the effects of linearization and orthogonalization, which are small in this case of a rather narrow band, this O p_z orbital can have *no* O p character on any other O site. However, it has all other characters downfolded and we see, in particular, $pd\sigma$ bonds to two Ni e_g d_{3z^2-1} orbitals and $pd\pi$ bonds to the four Ni t_{2g} orbitals.

Going now to the 5 Ni d bands seen in the middle panel, we see the corresponding $pd\sigma$ anti-bonds for the Ni e_g orbitals to the appropriate O p orbitals and the corresponding $pd\pi$ anti-bonds for the Ni t_{2g} orbitals. Since $pd\sigma$ hopping is stronger than $pd\pi$, the reddish e_g-like band lies above the blueish t_{2g}-like band, which is full in the LDA. Like for a member of the O p set, the d_{3z^2-1} member of the Ni d set can have *no* d character on any other Ni atom, and this is seen to localize the orbitals quite well. The electronic configuration with respect to this set is $p^6 d^8 = t_{2g}^6 e_g^2$.

The members of the 8-orbital O p Ni d set describing the 10 eV wide pd band structure are shown at the top of the figure, to the right. By virtue of having neither O p nor Ni d character in their tails, these orbitals are more localized and atomic-like. Merely the O p orbitals have a bit of bonding Ni sp character due to covalency with the band seen above 2 eV. Due to the O p character in the empty part of the e_g the configuration and the concomitant e_g character in the full O p band, the configuration with respect to this set is $p^{5.4} d^{8.6}$, i.e. with holes in the p orbitals and more than 8 d electrons!

As is evident from the figure, NiO is a metal in the LDA, which is completely wrong. Nevertheless, LDA Wannier orbitals form very reasonable one-electron basis sets for many-electron calculations such as LDA+DMFT and multiplet ligand-field theory (MLFT) cluster calculations for x-ray spectroscopies [35]. Due to the complexities of many-electron calculations, one is tempted to use a small basis set, e.g. for NiO, the set of 5 Ni d orbitals, or even a set of merely 2 Ni e_g orbitals. This is however inaccurate, because the Coulomb repulsion is strong between two d electrons on the same site and one cannot neglect the d character in the p band. A small computational bonus for using the larger pd set, is that the d orbitals have simpler shapes so that the dd Coulomb repulsion with good approximation can be described by the 3 Slater integrals [35]. On the other hand, the pictures of the pd set tell little, whereas those of the smaller sets bring
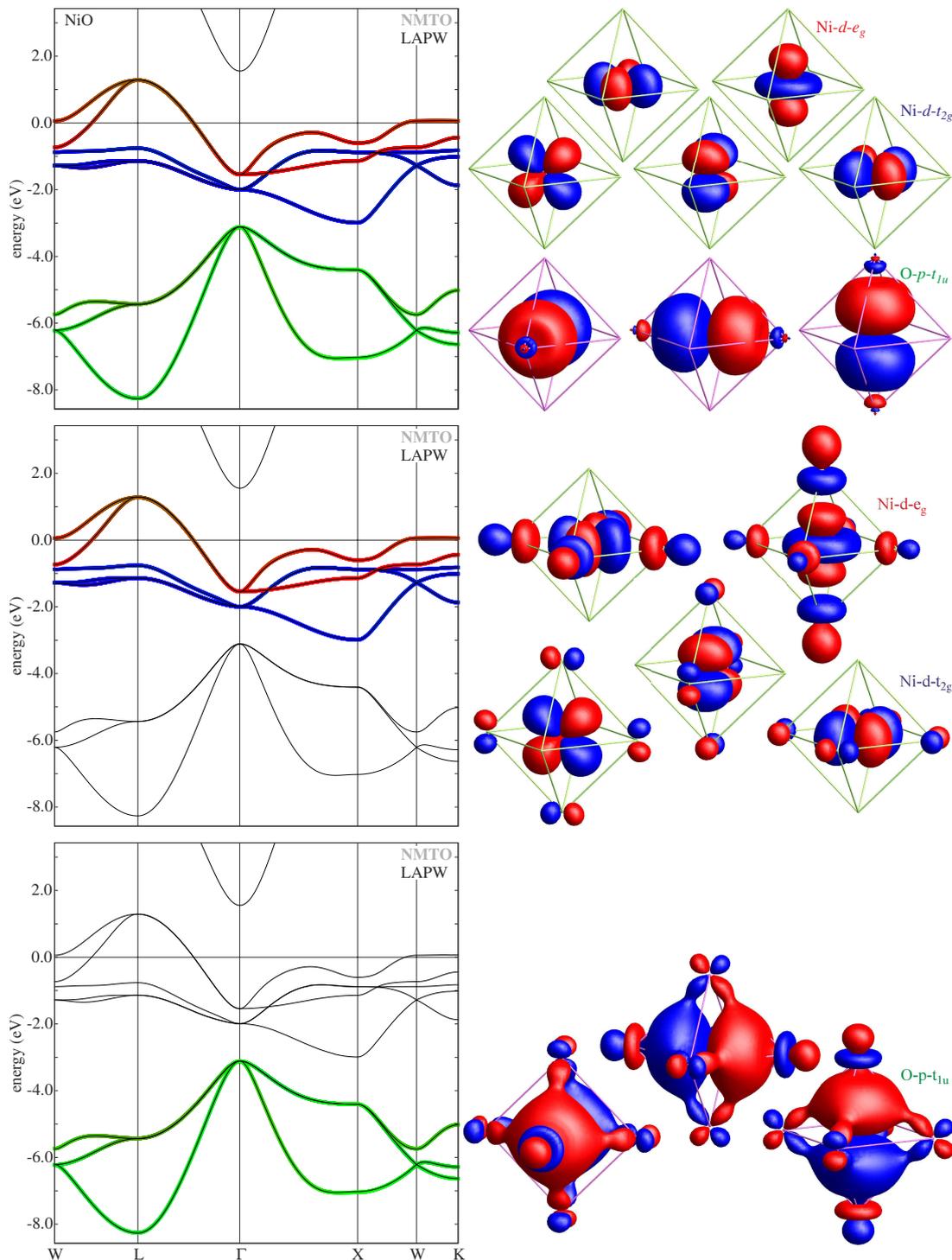
Fig. 7: *LDA band structures of NiO calculated with a large set of LAPWs (black) and three different $(N=1)$MTO basis sets (colored) whose Wannier orbitals are shown to the right of the bands. From the bottom and up: The 3 O p bands (green), the 5 Ni d bands with t_{2g} character blue and e_g character red, and the 8 O p Ni d bands. The Wannier orbitals are shown as $w(\mathbf{r}) = \pm const$ surfaces with the $\pm$ sign indicated by red/blue and $const$ determined by the condition that 90% of the probability density is inside the surface. From Ref. [35].*

out covalency effects very clearly. This becomes particularly relevant when symmetry-lowering lattice distortions take place. Examples may be found in Refs. [36], [37], [38, 39], and [40].

5   Standard Löwdin downfolding and N-ization

In the NMTO method we first construct the set of energy-dependent, downfolded KPWs (EMTOs) from multiple scattering theory, i.e. we compute the structure matrix (31) in a strongly screened (e.g. spd) representation and then downfold this matrix to the desired degree for each energy. *Thereafter* we N-ize the EMTOs to form the energy-independent NMTO basis set. This is different from standard Löwdin downfolding which partitions a given, large set of energy-independent, strongly localized orbitals into active and passive subsets, $|\chi\rangle = |\chi_A\rangle + |\chi_B\rangle$, and then eliminates the latter. Had one chosen this large basis set to be one of strongly screened NMTOs, N-ization would have come *before* downfolding, and this is also the sequence in which LMTO downfolding was first done [29]. Below, we shall first review Löwdin downfolding because it is similar to, but much more familiar than screened multiple scattering theory, and then indicate that *subsequent* use of the N-ization technique might be useful.
Partitioning the generalized eigenvalue equations (9) yields:

$$(H - \varepsilon O)_{AA}\, b_A + (H - \varepsilon O)_{AP}\, b_P = 0$$
$$(H - \varepsilon O)_{PA}\, b_A + (H - \varepsilon O)_{PP}\, b_P = 0$$

in block notation. Solving the bottom equations for b_P,

$$b_P = -\left[(H - \varepsilon O)_{PP}\right]^{-1} (H - \varepsilon O)_{PA}\, b_A, \tag{57}$$

and inserting in the upper equations, yields the well-known set of Löwdin-downfolded secular equations:

$$\left\{(H - \varepsilon O)_{AA} - (H - \varepsilon O)_{AP} \left[(H - \varepsilon O)_{PP}\right]^{-1} (H - \varepsilon O)_{PA}\right\} b_A = 0. \tag{58}$$

These, together with the "upfolding" (57) give the exact eigenfunction coefficients $b_I = (b_A, b_P)$, as long as the proper energy dependences are kept. But in order for the secular matrix to have the desirable $H - \varepsilon O$ form, the energy dependence of the complicated matrix

$$(H - \varepsilon O)_{AP} \left[(H - \varepsilon O)_{PP}\right]^{-1} (H - \varepsilon O)_{PA}$$

is either neglected or linearized.
We are interested in the set of downfolded *orbitals* giving rise to this secular matrix. This is the energy-dependent set:

$$|\phi_A(\varepsilon)\rangle \equiv |\chi_A\rangle - |\chi_P\rangle \left[(H - \varepsilon O)_{PP}\right]^{-1} (H - \varepsilon O)_{PA} \equiv |\chi_A\rangle + |\chi_P\rangle\, D_{PA}(\varepsilon), \quad \blacktriangleleft \tag{59}$$

with each member $|\phi_a(\varepsilon)\rangle$ being the active orbital $|\chi_a\rangle$, *dressed* by an *energy-dependent* linear combination of passive orbitals. How well localized $|\phi_a(\varepsilon)\rangle$ is, depends on how well the chosen set $|\chi_A\rangle$ reproduces the eigenstates at ε.

That $\mathcal{H} - \varepsilon$ represented in this set is the matrix in (58), is seen by first operating on (59) with $\mathcal{H} - \varepsilon$, and then projecting onto the active and passive subsets:

$$\langle \chi_A | \mathcal{H} - \varepsilon | \phi_A (\varepsilon) \rangle = (H - \varepsilon O)_{AA} - (H - \varepsilon O)_{AP} \left[(H - \varepsilon O)_{PP} \right]^{-1} (H - \varepsilon O)_{PA}$$
$$\langle \chi_P | \mathcal{H} - \varepsilon | \phi_A (\varepsilon) \rangle = 0.$$

Forming finally the linear combination (59) yields the desired result:

$$\left\langle \phi_A (\varepsilon) \left| \hat{H} - \varepsilon \right| \phi_A (\varepsilon) \right\rangle = (H - \varepsilon O)_{AA} - (H - \varepsilon O)_{AP} \left[(H - \varepsilon O)_{PP} \right]^{-1} (H - \varepsilon O)_{PA}. \quad \blacktriangleleft$$

One can show that this equals $-G_{AA} (\varepsilon)^{-1}$, exactly as equations (44) and (34) in MTO theory. In fact, the entire N-ization procedure could be used to remove the energy dependence of the Löwdin-downfolded set (59). The result for the dress is:

$$D_{PA}^{(0..N)} = G_{PA} [0..N] \, G_{AA} [0..N]^{-1} \approx G_{PA} (\varepsilon) \, G_{AA} (\varepsilon)^{-1} = D_{PA} (\varepsilon) ,$$

and therefore the major cause for delocalization seems to be the Löwdin downfolding (59). This procedure is computationally more demanding than the one we have described, and yields less localized downfolded obitals. It certainly only works for orbital basis sets with merely one radial function per Rlm [41].

6   Localization

It seems to me, that the NMTO construction in which one first generates a set of most localized solutions (KPWs) of Schrödinger's equation at a each energy, and then interpolates both the local (from the radial functions) and the global (from the downfolding) energy dependencies in one, common N-ization step, leads to a set of Wannier orbitals which are at least as localized as those obtained by Löwdin downfolding of the set of classical LMTOs as explained above, and have a spread close to the minimal one [5]. Computed rms values, $\left\langle w \left| |\mathbf{r} - \langle w | \mathbf{r} | w \rangle|^2 \right| w \right\rangle^{1/2}$, of the spread support this:

For the vanadium t_{2g} Wannier orbitals in V_2O_3, we [38] find $1.30\,\text{Å}$ for the a_{1g} and $1.40\,\text{Å}$ for the e_g^π orbitals. These values are significantly smaller than those, $1.35\,\text{Å}$ and $1.57\,\text{Å}$ obtained from TB-LMTOs [42]. This is consistent with results for the t_{2g} Wannier orbitals in the cubic perovskite $SrVO_3$ where the rms spread, $1.38\,\text{Å}$, of the NMTO Wannier orbital [43] is significantly smaller than the one, $1.54\,\text{Å}$, obtained by downfolding plus (N=0)-ization of the classical, nearly-orthonormal LMTO set [44]. The $1.38\,\text{Å}$ rms spread of this NMTO Wannier orbital is, in fact, only marginally above the minimum value found to be $1.3\dot{6}\,\text{Å}$ or $1.37\,\text{Å}$, depending on whether a mixed-basis-pseudopotential scheme or the FP-LAPW scheme was used for the LDA calculation of the Bloch functions in (2) [43]. Hence, at least for these t_{2g} systems, the NMTO Wannier functions seem to be close to those maximally localized in the sense of Marzari and Vanderbilt [5].

References

[1] E. Pavarini, E. Koch, A. Lichtenstein, D. Vollhardt (eds.)
The LDA+DMFT approach to strongly correlated materials,
Reihe Modeling and Simulation, Vol. 1 (Forschungszentrum Jülich, 2011)
`http://www.cond-mat.de/events/correl11`

[2] P. Blöchl: *Theory and Practice of Density-Functional Theory*, in [1]

[3] O.K. Andersen, Phys.Rev. B **12**, 3060 (1975)

[4] O.K. Andersen and O. Jepsen, Phys. Rev. Lett **53**, 2571 (1984)

[5] N. Marzari and D. Vanderbilt, Phys. Rev **B 56** 12847 (1997)

[6] S. Satpathy and Z. Pawlowska, Phys. Stat. Solidi (B), **145**, 555 (1988)

[7] J. Kunes: *Wannier functions and construction of model Hamiltonians*, in [1]

[8] S.F. Boys, Rev. Mod. Phys. **32**, 296 (1960)

[9] O.K. Andersen and T. Saha-Dasgupta, Phys. Rev. B **62**, R16219 (2000)

[10] O.K. Andersen, T. Saha-Dasgupta, R.W. Tank, C. Arcangeli, O. Jepsen, G. Krier in: *Electronic Structure and Physical Properties of Solids. The Uses of the LMTO Method*, ed. H. Dreyssé. Springer Berlin/Heidelberg (2000) [Springer Lecture notes in Physics, vol. 535]

[11] R.W. Tank and C. Arcangeli, phys. stat. sol. **217**, 89 (2000)

[12] O.K. Andersen, T. Saha-Dasgupta, S. Ezhov, L. Tsetseris, O. Jepsen, R.W. Tank, C. Arcangeli, G. Krier, Psi-k Newsletter **45** (June 2001), 86-119
`ftp://psi-k.dl.ac.uk./newsletters/news_45/Highlight_45.pdf`

[13] O.K. Andersen, T. Saha-Dasgupta, S. Ezhov, Bull. Mater. Sci. **26**, 19 (2003).

[14] O.K. Andersen, C. Arcangeli, R.W. Tank, T. Saha-Dasgupta, G. Krier, O. Jepsen, I. Dasgupta in: *Tight-Binding Approach to Computational Materials Science*, eds. L. Colombo, A. Gonis, P. Turchi. MRS Symposium Proceedings Series, vol 491 (1998)

[15] M. Zwierzycki and O.K. Andersen, Acta Physica Polonica A **115**, 64 (2009)

[16] E. Zurek, O. Jepsen, O.K. Andersen, ChemPhysChem **6**, 1934 (2005)

[17] Y. Nohara and O.K. Andersen, unpublished

[18] A. J. Williamson, R. Q. Hood, J. C. Grossman, Phys. Rev. Lett. **87** ,246406 (2001)

[19] A. J. Williamson, J. C. Grossman, R. Q. Hood, A. Puzder, G. Galli,
Phys.Rev. Lett. **89**, 196803 (2002)

[20] O.K. Andersen, O. Jepsen, and G. Krier in: *Lectures in Methods of Electronic Structure Calculations,* Eds.: V. Kumar, O.K. Andersen, A. Mookerjee, World Sci. Publ.Co., Singapore (1994)

[21] L. Vitos: *Computational Quantum Mechanics for Materials Engineers; The EMTO Method and Applications (*Springer, London, 2007)

[22] O.K. Andersen, Solid State Commun. **13**, 133 (1973)

[23] O.K. Andersen and O. Jepsen, Physica B+C **91**, 317 (1977)

[24] A.R. Mackintosh and O.K. Andersen in *Electrons at the Fermi Surface,* Ed. M. Springford, (Cambridge University Press, 1980)

[25] E. Wigner and F. Seitz, Phys. Rev. **43**, 804 (1933); *ibid* **46**, 509 (1934)

[26] O.K. Andersen, O. Jepsen, D. Glötzel in *Highlights of Condensed-Matter Theory*, Course LXXXIX, Eds. F. Bassani, F. Fumi and M.P. Tosi, "International School of Physics "Enrico Fermi", Varenna, Italy, (North-Holland, 1985)

[27] O.K. Andersen, Phys. Rev. Lett. **27**, 1211 (1971)

[28] H.L. Skriver: *The LMTO Method* (Springer, Berlin, Heidelberg, NewYork, Tokyo 1984)

[29] W.R.L. Lambrecht and O.K. Andersen, Phys. Rev. B **34**, 2439 (1986)

[30] W. Kohn and J. Rostoker, Phys. Rev., **94** 111, (1954)

[31] H. Ebert: *Multiple-Scattering Formalism for Correlated Systems: a KKR-DMFT Approach*, in [1]

[32] See lecture of R. Zeller

[33] O.K. Andersen, Z. Pawlowska, and O. Jepsen, Phys.Rev. B **34**, 5253 (1986)

[34] O.K. Andersen, O. Jepsen, and M. Sob in *The Electronic Band Structure and its Application*, Lecture Note Series, Vol. **283**, Ed. M.Yussouff (Springer-Verlag, 1987)

[35] M.W. Haverkort, M. Zwierzycki, and O.K. Andersen, Phys. Rev. B **85**, 165113 (2012)

[36] E. Pavarini, S. Biermann, A.P. Poteryaev, A.I. Lichtenstein, A. Georges, O.K. Andersen, Phys. Rev. Lett. **92**, 176403 (2004)

[37] E. Pavarini, A. Yamasaki, J. Nuss, O.K. Andersen, New Journal of Physics **7**, 188 (2005)

[38] T. Saha-Dasgupta, O. K. Andersen, J. Nuss, A. I. Poteryaev, A. Georges, A. I. Lichtenstein, arXiv: 0907.2841 v1

[39] A.I. Poteryaev, J.M. Tomczak, S. Biermann, A. Georges, A.I. Lichtenstein, A.N. Rubtsov, T. Saha-Dasgupta, O.K. Andersen, Phys. Rev. B **76**, 085127 (2007)

[40] E. Pavarini: *The LDA+DMFT Approach*, in [1]

[41] E. Zurek, J. Autschbach, O.K. Andersen, AIP Conf. Proc. **963**, 1421 (2007)

[42] I.V. Solovyev, Z.V. Pchelkina and V.I. Anisimov, Phys. Rev. B **75**, 0145110 (2007)

[43] F. Lechermann, A. Georges, A. Poteryaev, S. Biermann, M. Posternak, A. Yamasaki, and O.K. Andersen, Phys. Rev. B **74**, 125120 (2006)

[44] I. V. Solovyev, Phys. Rev. B **73**, 155177 (2006)