Supplementary information for “Automated effective band structures for defective and mismatched supercells”

Peter Brommer‡ **and David Quigley**

Department of Physics and Centre for Scientific Computing, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

E-mail: p.brommer@warwick.ac.uk, d.quigley@warwick.ac.uk

A. Recap: The supercell EBS method

A.1. Supercell definition and notation

The notation used here is analogous to the one used by Popescu and Zunger [1]. Quantities referring to the primitive cell or pc (supercell or SC) are denoted in small (capital) symbols. The basis vectors $\vec{a}$ ($\vec{A}$) of the pc (SC) are related by $\underline{\vec{A}} = \underline{M} \cdot \underline{\vec{a}}$, or

$$\vec{A}_i = \sum_j m_{ij}\vec{a}_j, \qquad m_{ij} \in \mathbb{Z}, \quad i,j = 1,2,3, \tag{A.1}$$

where the transformation or supercell matrix $\underline{M}$ is nonsingular with integer components, which implies that the SC is commensurate to the pc. The determinant of $\underline{M}$ is the multiplicity N of the SC, i.e. the ratio of the respective volumes $V_{\mathrm{SC}}/v_{\mathrm{pc}}$. The hexagonal pc and orthorhombic SC of the two-dimensional honeycomb net is depicted in figure A.1a.

In reciprocal space, there are consequently two distinct Brillouin zones (see figure A.1b): the primitive cell Brillouin zone (pbz) and the smaller supercell Brillouin zone (SBZ). Their respective basis vectors $\vec{b}_i$ ($\vec{B}_i$) are again connected by the supercell matrix $\underline{M}$:

$$\underline{\vec{B}} = \underline{M}^{-1}\underline{\vec{b}}. \tag{A.2}$$

It should be noted, that the components $(m^{-1})_{ij}$ of $\underline{M}^{-1}$ can be written as an integer divided by the determinant of the transformation matrix $\det\underline{M}$:

$$(m^{-1})_{ij} = \frac{1}{\det\underline{M}} m^*_{ij}, \qquad m^*_{ij} \in \mathbb{Z}, \quad i,j = 1,2,3. \tag{A.3}$$

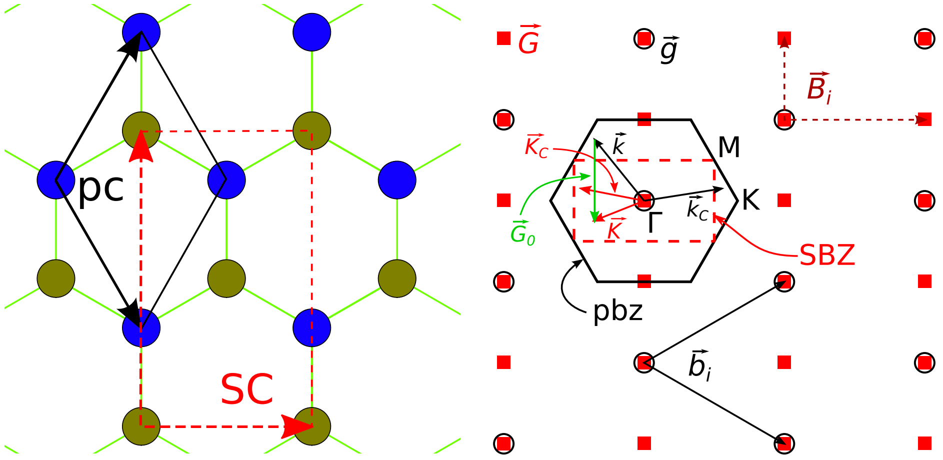
The pbz (SBZ) basis vectors span the infinite set of reciprocal lattice vectors $\{\vec{g}_k\}$ ($\{\vec{G}_k\}$):

$$\vec{g}_k = \sum_i p_i\vec{b}_i, \qquad p_i \in \mathbb{Z}, \quad i = 1,2,3, \tag{A.4}$$

$$\vec{G}_k = \sum_i P_i\vec{B}_i, \qquad P_i \in \mathbb{Z}, \quad i = 1,2,3, \tag{A.5}$$

where obviously $\{\vec{g}_k\} \subset \{\vec{G}_k\}$, i.e. every lattice vector of the pbz is also one of the SBZ, see figure A.1b.

‡ Current address: Centre for Predictive Modelling, School of Engineering, University of Warwick, Library Road, Coventry CV4 7AL, UK



(a) real space

(b) reciprocal space

Figure A.1: Real (a) and reciprocal (b) representation of the honeycomb net. (a) Primitive cell (pc, solid lines) and orthorhombic supercell (SC, dashed lines) of the honeycomb net with their respective basis vectors. (b) The pbz of the hexagonal unit cell and the SBZ of the orthorhombic supercell from (a). Γ, M, and K are special points in the pbz. $\vec{b}_i$ ($\vec{B}_i$) are the pc (SC) reciprocal basis vectors spanning the set of reciprocal lattice vectors $\vec{g}$ ($\vec{G}$) represented by black circles (red squares). The wave vector $\vec{k}$ ($\vec{k}_C$) is folded to $\vec{K}$ ($\vec{K}_C$) by the folding vector $\vec{G}_0$ ($\vec{G}_{C,0}$ – not shown). While the pc wave vectors $\vec{k}$ and $\vec{k}_C$ are related by a symmetry operation C of the pbz, this is not necessarily true for the corresponding SC wave vectors.

A.2. Folding and unfolding

The band structure of a periodic solid has the same periodicity as the Brillouin zone. Wave vectors outside the BZ are *folded* into the Brillouin zone (this corresponds to going from the extended-zone scheme to the reduced-zone scheme for electron dispersion [2]). For each wave vector $\vec{k}$ in the pbz, there exists a unique reciprocal lattice vector $\vec{G}_0$ such that

$$\vec{K} = \vec{k} + \vec{G}_0, \quad \text{where } \vec{K} \in \text{SBZ}, \tag{A.6}$$

as shown in figure A.1b. On the other hand, a SBZ wave vector $\vec{K}$ *unfolds* to N distinct $\vec{k}_i \in$ pbz according to

$$\vec{k} = \vec{K} - \vec{G}_i, \quad i = 1, \ldots, N, \tag{A.7}$$

where $N = \det \underline{M}$ is the multiplicity of the supercell. In the 2D model cell of figure A.1, $N = 2$ and $\vec{K}$ unfolds into $\vec{k}_1 = \vec{K}$ ($\vec{G}_1 \equiv \vec{0}$) and $\vec{k}_2 = \vec{k}$ ($\vec{G}_1 \equiv \vec{G}_0$).

The Schrödinger equation of the electronic system can be solved both in pc and SC representation, obtaining the eigenvectors $|\vec{k}n\rangle$ and $|\vec{K}m\rangle$, where n and m are band indices. The objective of the unfolding method of Popescu and Zunger [1] is to recover the $E(\vec{k})$ picture from the much less insightful SC $E(\vec{K})$ (cf. figure 1). By projecting $|\vec{K}m\rangle$ on all pc eigenstates $|\vec{k}_i n\rangle$ of a fixed $\vec{k}_i$, one obtains the spectral weight

$$P_{\vec{K}m}(\vec{k}_i) = \sum_n |\langle \vec{K}m|\vec{k}_i n\rangle|^2, \tag{A.8}$$

which provides a measure of the amount of Bloch character k_i still present in the SC $|\vec{K}m\rangle$. From the $P_{\vec{K}m}(\vec{k}_i)$, a spectral function (SF) is derived as

$$A(\vec{k}_i, E) = \sum_m P_{\vec{K}m}(\vec{k}_i)\delta(E_m - E), \tag{A.9}$$

which is now continuous in energy E and replaces the discrete $E(\vec{k})$. For a perfect SC, the SF corresponds to sequence of δ functions of integer amplitude at each band energy $E_n(\vec{k})$ and thus perfectly recovers the original band structure [1] . Any deviation of the supercell from a perfect structure (defect, disorder, lattice mismatch), will change this picture: The spectral function now has finite width in both $\vec{k}$ and E, and yields the effective band structure if sampled for a path through reciprocal space.

For a plane wave basis set, Popescu and Zunger [1] provide a detailed recipe on how to calculate the spectral weights from (A.8). The eigenfunctions of the SC in this basis are of the form

$$|\vec{K}m\rangle = \left[\sum_{\vec{G}} C_{\vec{K}m}(\vec{G})e^{i\vec{G}\vec{r}}\right] e^{i\vec{K}\vec{r}}, \quad \vec{K} \in \text{SBZ}, \tag{A.10}$$

where the sum runs over the SC reciprocal lattice vectors given by (A.5). The spectral weight is then given by

$$P_{\vec{K}m}(\vec{k}_i) = \sum_{\vec{g}} |C_{\vec{K}m}(\vec{g} + \vec{k}_i - \vec{K})|^2 = \sum_{\vec{g}} |C_{\vec{K}m}(\vec{g} - \vec{G}_i)|^2, \tag{A.11}$$

where the sum now runs over the reciprocal lattice vectors of the primitive cell. As those are a subset of the $\{G_k\}$, all coefficients are well-defined. This process is akin to a Fourier filtering of the plane wave coefficients: Only every Nth coefficient (with the appropriate offset) contributes to the spectral function.

For ultrasoft pseudopotentials (USPP) [3], (A.11) needs to take the relaxed orthonomality of wave functions into account. The right hand side of this equation can be understood as a norm $\langle\phi_i|\phi_i\rangle$ of the fourier-filtered wave functions ϕ_i. For non-normconserving PP this expression has to be replaced by $\langle\phi_i|S|\phi_i\rangle$, where S is the overlap operator [4].

In general, the pc and the SC will not have the same symmetry. Typically, the symmetry of the supercell is reduced either by the choice of supercell (compare figure A.1b, where $\vec{k}$ and $\vec{k}_C$ are related by a rotation C_3, which is a symmetry operator of the hexagonal grid, but not of the orthorhombic grid) or by the decoration; defects may break certain symmetries. As a consequence, $\vec{k}$ vectors equivalent in the pbz may not map to symmetry equivalent vectors in the SBZ. So if for a specified $\vec{k}_i$ point there exist n_C wave vectors $\vec{k}_C$ in the pbz belonging to the same symmetry class $C(\vec{k}_i)$, the resulting spectral function at $\vec{k}$ is obtained as the average:

$$\bar{A}(\vec{k}_i, E) = \frac{1}{n_C} \sum_{\vec{k}_C \in C(\vec{k}_i)} A(\vec{k}_i, E). \tag{A.12}$$

Depending on the nature of the system, more than a single supercell may be required to capture the EBS properly. For example, for an EBS of point defects at constant concentration but random distribution, the cost of simulating a huge cell to account for the disorder may be prohibitively large due to the asymptotic cubic scaling of plane wave DFT. However several smaller systems that sample a number of different arrangements may still be within the scope of the available computing power. In that case, the symmetry-averaged $\bar{A}(\vec{k}_i, E)$ need to be statistically averaged to obtain the final EBS.

B. Implementation details

B.1. CASTEP – plane wave DFT code

There are several reasons for choosing CASTEP as a basis to implement the EBS method. It can cope with the large system sizes of supercells by exploiting MPI parallelisation over $\vec{k}$ points, reciprocal lattice ($\vec{g}$) vectors and bands (the latter is not used in band structure calculations) and scales well up to thousands of atoms and thousands of processors.

Also, CASTEP contains a wealth of powerful features such as automatic determination of cell symmetry operations. As the code is modular Fortran 90, this existing functionality can be accessed with ease.

We note that CASTEP in its unmodified form already implements calculation of band structures, but without any means to project these onto a primitive cell of interest. Starting from a converged, self-consistent electron density, the Schrödinger equation is solved non-selfconsistently at a number of $\vec{k}$ points in the BZ of the current cell, which may or may not be a supercell.

Our implementation of the EBS method is a separate executable named `bs_sc2pc` compiled from a CASTEP main program file that is stripped down to only perform a band structure calculation, but using all the functionality provided by different modules (from handling MPI communications to reading input files to performing the actual calculation). EBS-specific computational tasks were then added during setup and output phases. Details of the implementation are described in the following section.

B.2. Constructing an EBS with `bs_sc2pc`

B.2.1. Required data. To determine an EBS, one requires the following input data, readily generated using the existing software:

- Atomic positions, symmetry operations and lattice vectors for the supercell.
- Symmetry operations and lattice vectors of the primitive cell. Alternatively, atomic positions for the primitive cell can be used to determine primitive cell symmetry at runtime.
- Electronic structure of the supercell computed on a standard k-point mesh. In practice this is read from the restart file generated by a typical DFT calculation. This file contains the converged wave function of the supercell, which is used to initialize the electron density for the band structure calculation. While in principle it would be possible to determine the self-consistent electron density during a run, the computational requirements of a self-consistent energy calculation and a band structure calculation are typically very different for many-atom supercells (the former uses few k-points, the latter may use many), which makes it difficult to choose a parallelisation strategy that works efficiently for both.
- $\vec{k}$ points at which to perform the EBS calculation.
- Optionally, range and sampling interval of the spectral function.

All other quantities required for an EBS (such as the supercell matrix $\underline{M}$, cf. (A.1)) are determined by the code at runtime. Note that the electronic structure of the primitive cell is not required; the sum over primitive cell eigenstates in (A.8) is not performed explicitly. In the following sections we describe the main modifications to CASTEP during setup (section B.2.2) and output (section B.2.3).

B.2.2. Setup: transformation matrix, symmetries, and folding vectors. After startup, the primitive cell is read; this includes the $n_{\vec{k}}$ pbz $\vec{k}$ at which the band energies are to be evaluated. With this information, the supercell matrix $\underline{M}$ can be determined. If no matrix with integer components can be found, the program aborts. This implies that for supercells whose geometry deviates from the ideal case, the primitive cell needs to be adjusted accordingly, usually by deforming the primitive cell in such a way, that the supercell is again an integer multiple. As any modifications to the primitive cell geometry may have consequences for underlying pc band structure, the EBS of the supercell needs to be interpreted accordingly; any deviations from the primitive cell band structure may be introduced either by the strained primitive cell or by the supercell. Also, any distortion can reduce the primitive cell symmetry. To account for this, it may become necessary to average an EBS over multiple differently oriented paths in the strained pc which are equivalent in the original high-symmetry cell. While pc and SC basis vectors need not be collinear (see figure A.1a), it is required that pc and SC have the same *orientation*.

Some consideration must be given to the n_s pbz symmetry operations either specified in the pc cell file or determined on the fly, and if/how these map to SC symmetry operations. In the primitive cell, duplicate points generated from the same initial point are eliminated (e.g. barring any translational symmetry, the Γ point is mapped to itself under any operation, but it is retained only once). This locates the wave vectors $\vec{k}_C$ belonging to the same symmetry class $C(\vec{k})$ as $\vec{k}$ (compare (A.12)). At the end of this process, there is a list of n_C $\vec{k}_C$ points with $n_{\vec{k}} \leq n_C \leq n_s n_{\vec{k}}$.

The resulting full $\vec{k}$ point list is then mapped to the SBZ and the folding vector $\vec{G}_{0,i}, i = 1,\ldots,n_C$ (A.6) for each of them is stored. The folding process might cause several $\vec{k}$ to result in identical $\vec{K}$ (or at a set of $\vec{K}$ that are equivalent under a symmetry operation of the SBZ), albeit with different $\vec{G}_0$. Eigenvalue calculations at each $\vec{K}$ are independent of this construction, and hence careful consideration of symmetry can be used to significantly reduce computational load requirements by eliminating redundant $\vec{K}$ points. To this end, it is checked whether any two of the n_C $\vec{K}$ points, irrespective of their original $\vec{k}$, are identical under any of the N_S symmetry operations (including the identity operation) of the SBZ, which is again read from the supercell file or determined on the fly. The band structure calculation is then carried out on the remaining $N_{\vec{K},\text{unique}}$ unique $\vec{K}$ points only, with $1 \leq N_{\vec{K},\text{unique}} \leq n_C$. This $\vec{K}$ point elimination and consequent reduction in computational cost can be deliberately exploited by adjusting the density of $\vec{k}$ points along a path: If the supercell dimension and the $\vec{k}$ sampling are in register, most points will actually be redundant (see also section 4.3 for an example calculation).

A standard CASTEP band structure calculation is then performed using the set of unique $\vec{K}$ points remaining. This ensures that our tool can remain agnostic to any further DFT implementation choices like exchange-correlation functionals or pseudopotentials, as this is all accounted for in the underlying band structure code.

B.2.3. Evaluation: spectral weights and spectral function. After the band energies are calculated at the $N_{\vec{K},\text{red}}$ SBZ points, the spectral weights ((A.8) and (A.11)) and with those the spectral function (A.9) can be calculated. For each of the n_C symmetrized pbz $\vec{k}_C$ points, first the mapped unique $\vec{K}$ is identified. For each band of that $\vec{K}$, we loop over the plane waves contributing to the eigenstates of this band m. We then check, whether the plane wave at $\vec{G}$ should be included in the sum over $\vec{g}$ in (A.11). If this is the case, we keep this coefficient and reject it otherwise. At the end, we calculate the spectral weight $P_{\vec{K}m}(\vec{k}_i)$ as the S-norm of the filtered wave function with coefficients $C_{\vec{K}m}(\vec{G})$.

The SF is realized as a histogram. The number of bins and their width are either user-specified or calculated from minimal and maximal eigenvalue and a default number of bins. Once all plane waves of a certain band $|\vec{K},m\rangle$ have been evaluated, the appropriate spectral function bin value is increased by $P_{\vec{K}m}(\vec{k}_i)$ multiplied by the weighting factor $n_C(\vec{k}_i)^{-1}$. This whole procedure is then repeated for all pbz symmetry-equivalent $\vec{k}_C$ retained for a certain $\vec{k}_i$, and we obtain the symmetry-averaged spectral function $\bar{A}(E,\vec{k}_i)$, which is written to file. These steps are duplicated for all $\vec{k}$ of the pc (and, in case of a spin-polarized calculation, for the second spin channel as well).

B.2.4. Postprocessing: averaging and graphical output over several supercell realisations. Averaging over independent band structure calculations is performed in post-processing. In this way it is possible to account for occupational disorder beyond a single supercell or to average over multiple paths in the pbz whose equivalence is broken by a strained primitive cell.

The resulting spectral function $\bar{A}(E,\vec{k}_i)$ is then a energy histogram of spectral intensities at a sequence of $\vec{k}$-points. An intuitive graphical representation of this dataset is obtained by plotting the individual histograms as lines vertically next to each other along the $\vec{k}$-point path (cf. figure 2 of Ref. 1, and figures 4 and 6 in the main article). There, the individual histogram lines are only plotted for values where the spectral intensity is non-zero. Alternatively, $\bar{A}(E,\vec{k}_i)$ can be represented in a contour plot, as demonstrated in figure 2(a).

References

[1] Popescu V and Zunger A 2012 Extracting E versus $\vec{k}$ effective band structure from supercell calculations on alloys and impurities *Phys. Rev. B* **85** 085201

[2] Ashcroft N W and Mermin N D 1976 *Solid state physics* Brooks/Cole, Belmont, CA, USA

[3] Vanderbilt D 1990 Soft self-consistent pseudopotentials in a generalized eigenvalue formalism *Phys. Rev. B* **41** 7892–5

[4] Laasonen K, Pasquarello A, Car R, Lee C and Vanderbilt D 1993 Car-parrinello molecular dynamics with vanderbilt ultrasoft pseudopotentials *Phys. Rev. B* **47** 10142–53