Compact representation of one-particle wavefunctions and scalar fields obtained from electronic-structure calculations

Sergey V. Levchenko[a,b,c], Matthias Scheffler[b]

[a] *Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, 3 Nobel Street, 143026 Moscow, Russia*
[b] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany*
[c] *National University of Science and Technology "MISIS", 119049 Moscow, Russia*

Abstract

We present a code-independent compact representation of one-electron wavefunctions and other volumetric data (electron density, electrostatic potential, etc.) produced by electronic-structure calculations. The compactness of the representation insures minimization of digital storage requirements for the computational data, while the code-independence makes the data ready for "big data" analytics. Our approach allows to minimize differences between original and the new representation, and is in principle information-lossless. The procedure for obtaining the wavefunction representation is closely related to construction of natural atomic orbitals, and benefits from the localization of Wannier functions. Thus, our approach fits perfectly any infrastructure providing a code-independent tool set for electronic-structure data analysis.

Keywords: wavefunction storage, electronic-structure calculations, materials database
PACS: 31.15.A-, 07.05.Kf

1. Introduction

Self-consistent one-electron wavefunctions (WFs) contain detailed information on the electronic structure (ES) of a system in the mean-field approximation. A quick access to WFs would allow for an exhaustive analysis of many ES properties. WFs can be either calculated on demand or stored on disk (by disk we mean any device for storing information). Nowadays, some ES calculations are acceptably accurate and fast, so that storage of WFs can be avoided. For example, local-density and generalized gradient approximations to density-functional theory (DFT) allow calculating systems with tens of atoms on a single CPU in seconds. However, larger systems and more advanced ES methods (e.g., DFT with hybrid functionals or many-body perturbation theory) would require much more time and/or computational resources, which would inhibit ES data analysis across materials space. Thus, storage of WFs and other derived data such as electronic density or Hartree potential is highly desirable.

For a material model with periodic boundary conditions, the wavefunctions $\psi^{\sigma}_{i\mathbf{k}}(\mathbf{r})$ are expressed in terms of basis functions $\phi_{\alpha}(\mathbf{r})$:

$$\psi^{\sigma}_{i\mathbf{k}}(\mathbf{r}) = \sum_{\alpha} C^{\sigma\alpha}_{i\mathbf{k}} \phi_{\alpha}(\mathbf{r}) \,, \tag{1}$$

where i is the band index, $\mathbf{k}$ is the k-point, and σ is the spin. The same expression can be used for non-periodic systems by limiting the k-point index to $\mathbf{k} = \mathbf{0}$.

Electronic-structure theory uses several, very different forms of $\phi_{\alpha}(\mathbf{r})$ in its numerous computational implementations. These are for example (i) Gaussian-type orbitals (GTOs), (ii) plane waves (PWs), (iii) numeric atomic orbitals (NAOs), (iv) combination of the above, and more (including real-space grid-based

*Corresponding author
Email address: levchenko@fhi-berlin.mpg.de (Sergey V. Levchenko)

Published in Computer Physics Communications, vol. 237, p. 42-46 (2019)

codes such as Octopus [1] and Parsec [2], and wavelet-based codes such as BigDFT [3]). Each of these basis-set types have their advantages and disadvantages. The absence of a single standard basis-set type presents a challenge for a code-independent representation of the WF data. Such a representation is needed to make the data generated by different ES codes directly comparable and interpretable by a machine (artificial intelligence) for data analytics, e.g., for finding outliers and hidden correlations within the data and with other theoretical or/and experimental data. Moreover, basis sets optimal for storage or for ES calculations usually have different requirements. For example, atomic orbitals form a most compact basis set for an atom, but Gaussian functions are more computationally efficient due to analytic integral evaluations.

Our goal is to find an optimal universal basis set to store WF data. We are particularly interested in *all-electron* calculations, which give access to important properties such as electric field gradients and nuclear magnetic resonance (NMR) shifts. In contrast, PW calculations are typically combined with pseudopotentials, and therefore the valence WFs near the atomic nuclei are incorrect, i.e., pseudoized. Nevertheless, approximate WFs with all the nodes can be straightforwardly restored from the pseudopotential calculations [4] using the same formalism as the projector augmented-wave method (PAW) [5]. And the core electrons in a pseudopotential calculation are frozen to those used in in the creation of the pseudopotential.

2. Formalism

The expansion coefficients $C_{i\mathbf{k}}^{\sigma\alpha}$ and the functions $\phi_\alpha(\mathbf{r})$ contain complete information for describing WFs at any point in space. Depending on the functional form, $\phi_\alpha(\mathbf{r})$ can be stored as either a small set of parameters (k-point for each PW, or contraction coefficients and exponents for each GTO) for analytic functional forms, or calculated on demand (as NAOs) from a set of potentials represented by parameterized analytic forms (e.g., Coulomb potential in atomic Schrödinger equation). In overwhelming majority of calculations localized basis functions (GTOs and NAOs) depend only on atomic species, and are otherwise system-independent. This is practical for ES calculations, because it is not clear how to efficiently adapt basis sets on-the-fly during self-consistent-field (SCF) cycle. However, for storage purposes this strategy may not be the optimal one, because in this case ES is already known, and there is no need to have a larger basis set for the sake of transferability. Moreover, storage puts more stringent requirements on transferability, because a WF representation that is compact in one basis may not be as compact in another basis.

In view of the above, we argue that NAOs are most flexible and therefore best suited for code-independent WF storage format. The numerical representation gives much flexibility to the form of the basis functions. With this flexibility, we can pursue the following strategy for obtaining the most compact representation of wavefunctions and derived scalar fields.

Atom-centered numerical grids. Strategies for constructing efficient and accurate atom-centered grids for NAOs are well established [6, 7]. Each NAO is expressed as a product of a radial and a spherical harmonic function. The radial part is represented as a spline interpolation on a radial grid. Following [8], we suggest to use logarithmic radial grids described by the following formula:

$$r(s) = r_{\text{outer}} \frac{\log(1 - [s/(N_\text{r}+1)])}{\log(1 - [N_\text{r}/(N_\text{r}+1)])}\,, \tag{2}$$

where r_{outer} is the radius of the outer-most shell, N_r is the number of grid points, and $s = \{1, \ldots, N_\text{r}\}$ ($r(0) = 0$ corresponds to the atomic center). From our extensive experience with NAO-based electronic-structure calculations with the FHI-aims package [9] we conclude that r_{outer} = 7-8 Å is sufficient for a very accurate representation of WFs. If necessary, a higher accuracy can be achieved by simply placing additional shells at integer fractions of the original grid, e.g., s = 1/2, 3/2, ..., $2N_\text{r}$ + 1/2. This allows us to specify a small base grid to reduce the storage requirements when possible, and easily interconvert between different grid densities for data analysis. The optimal base grid N_r values for each atom are tabulated in the species default settings of the FHI-aims package [9]. Depending on the atom, the values vary between 24 and 80. The accuracy and transferability of the grids constructed in the way described above have been extensively tested by applications of the DMol^3 [10] and FHI-aims packages to a large variety of systems, including molecules, clusters, and periodic systems of dimensions 1 to 3 (e.g., crystalline solids, nanowires, surfaces,

and layered materials). In non-NAO-based approaches atom-centered grids are often used to represent WFs within a sphere centered at the nucleus, to reduce the number of PW basis functions. In all cases, we suggest to use code-independent grids.

In case of delocalized basis sets such as plane waves, or any non-atom-centered basis sets, an atom-centered basis set has to be first generated such that it accurately represents the set of one-electron states to be stored. For a numerical representation on a finite grid a *complete* set of possible radial functions is determined by the number of grid points and the type of interpolation between the points. Various NAO-based codes successfully use spline interpolation. Here we propose to apply basis splines (B-splines) [11, 12] for the radial parts of atomic functions. The angular parts are spherical harmonics truncated at a certain maximum angular momentum l_{max}. The benefit of B-splines is that they are localized along the radial coordinate, so that it is easy to select only B-splines that overlap with the regions where the non-atom-centered basis functions are applied. For example, in the linearized augmented plane-wave (LAPW) method there is a cutoff radius around each nucleus. Within the cutoff radius basis functions are atom-centered, while outside the basis functions are plane waves. Therefore, only B-splines that are non-zero outside the cutoff radius are needed to represent the PW part of the LAPW basis set. The completeness of the atom-centered representation of the selected electronic states can be quantified by a spillage parameter [13, 14].

Finding the minimal NAO basis set to represent the WFs to be stored with a desired accuracy. For this purpose, the flexibility of NAOs exhibits its key advantages. To find the minimal set of NAOs, we employ a procedure similar to that of finding natural atomic orbitals [15, 16]. Although originally formulated for finite systems, an extension to periodic models is almost straightforward [17]. The procedure is based on diagonalization of the density matrix:

$$\Gamma(\mathbf{r},\mathbf{r}') = \sum_{\sigma}\sum_{i\mathbf{k}} f^{\sigma}_{i\mathbf{k}}\psi^{\sigma *}_{i\mathbf{k}}(\mathbf{r})\psi^{\sigma}_{i\mathbf{k}}(\mathbf{r}')\,, \tag{3}$$

where $f^{\sigma}_{i\mathbf{k}}$ are occupation numbers. The method determines a compact set of linear combinations of basis functions with maximum overlap with the space spanned by the occupied states. Our task is similar, but our aim is to find a compact basis set for representation of states to be stored, independent of their actual occupation. The resulting NAOs are termed subspace-optimized atomic orbitals (SOAOs). Thus, $\Gamma(\mathbf{r},\mathbf{r}')$ should be interpreted differently: $f^{\sigma}_{i\mathbf{k}}$ are non-zero only for states that we aim to store. In most practical cases all non-zero $f^{\sigma}_{i\mathbf{k}}$ should be identical (e.g., all set to 1). The radial parts of SOAOs are stored on the atom-centered grids, considering a spline interpolation along the radial coordinate (see below the discussion on storage memory requirements).

Construction of natural AOs and consequently SOAOs is a multi-step procedure [15, 16]. It starts from calculation of pre-orthogonalized SOAOs by diagonalizing (A, l, m) blocks of the density matrix, where A denotes an atom, and (l, m) are the angular momentum quantum numbers. Averaging of the (l, m) blocks of the "density matrix" over the $2l + 1$ values of m for each atom and l is performed to preserve spherical symmetry. As a result, we obtain a minimal set of NAOs (pre-SOAOs) that have maximum overlap with the set of WFs to be stored. The contribution of each pre-SOAO to the set is quantified by the eigenvalues of the modified density matrix (in the case of natural orbitals, the eigenvalues would be occupations). If all pre-SOAOs were orthogonal to each other, the eigenvectors corresponding to non-vanishing eigenvalues would constitute the minimal basis for accurate representation of the WFs. However, due to non-zero overlaps with pre-SOAOs on neighboring atoms, the eigenvalues of diffuse pre-SOAOs will not decay. Similar to natural AOs, we suggest to apply the occupancy-weighted symmetric orthogonalization (OWSO) [18, 15, 16] procedure to pre-SOAOs, where instead of occupancy the eigenvalues of the atomic angular momentum blocks of the modified density matrix are used. OWSO yields orthogonal orbitals that preserve the shape of AOs with higher occupations as much as possible. Although OWSO yields orbitals which do not have spherical symmetry and are linear combinations of pre-SOAOs located on different atoms, they still can be assigned to particular atoms based on the location of the pre-SOAO with the highest contribution to a given post-OWSO orbital. As the next step, the modified density matrix is re-diagonalized in the basis of the post-OWSO orbitals on each atom, and the resulting orbitals with non-negligible eigenvalues (above a threshold) are selected. If we denote radial parts of pre-SOAOs on atom A for angular momentum l as

$\chi^A_{\beta l}(r)$, the orbitals after the re-diagonalization are expressed as:

$$\phi_\alpha(\mathbf{r}) = \sum_A \sum_{lm} \left(\sum_\beta C^A_{\beta lm} \chi^A_{\beta l}(r) Y_{lm}(\Omega_A) \right) , \tag{4}$$

where $C^A_{\beta lm}$ are obtained from the OWSO procedure, and $Y_{lm}(\Omega_A)$ are spherical harmonics centered at atom A. The radial functions $\sum_\beta C^A_{\beta lm}\chi^A_{\beta l}(r)$ should be collected on each atom A for each (l, m) channel, normalized, and linearly independent radial functions should be selected by diagonalizing their overlap matrix and removing the eigenvectors with small eigenvalues. The remaining radial functions within each l channel are then orthogonalized to give the final set of atom-centered functions.

3. Discussion

Following the above steps, we obtain the minimal set of NAO basis functions that accurately represent the WFs to be stored. The flexibility of the NAO functional form allows to reduce the number of basis functions to a convenient minimal size, and, consequently, also the number of expansion coefficients in equation 1. This flexibility comes with a small storage overhead because the radial parts of basis NAOs have to be stored on the atom-centered radial grids. If the number of basis functions is N_{basis} and the number of WFs to be stored is N_{states}, then the total number of expansion coefficients is $N_{states}N_{basis}$ (assuming there is no sparsity). In case of the SOAOs, the number of functions $N_{SOAO} < N_{basis}$, but they require additional $N_{SOAO}N_r$ numbers for storage, where N_r is the average number of radial grid points per atom (in practice $N_r \sim 150$). As the number of WFs to be stored for a given material is increased, N_{SOAO} approaches N_{states}, and the overhead tends to $N_{states}N_r$. This storage overhead is not large ($\sim$1 kB per WF), and it can be further reduced by finding a parameterized set of atom-centered potentials such that a minimal number of solutions of Schrödinger equations for these potentials is complete in the space spanned by SOAOs. Below we describe an approach for finding such potentials.

The problem can be formulated in terms of minimization of the following functional for each atom in the system:

$$F[\{V_\alpha\}] = \sum_{\alpha l} \int_0^\infty \left[\sum_\beta \left(-\frac{1}{2} U^{(l)}_{\alpha\beta} \frac{d^2\eta_{\beta l}}{dr^2} + \left(V_\alpha(r) + \frac{l(l+1)}{2r^2} \right) U^{(l)}_{\alpha\beta}\eta_{\beta l} - E_{\alpha l} U^{(l)}_{\alpha\beta}\eta_{\beta l} \right) \right]^2 r^2 dr \tag{5}$$

with respect to unknown spherical potentials $V_\alpha(r)$, the eigenvalues $E_{\alpha l}$ for each angular momentum l, and unitary matrices $U^{(l)}_{\alpha\beta}$ under the condition that we obtain as small number of different $V_\alpha(r)$ as possible, so that we reduce storage requirements for the potentials. The summands in the square brackets of Eq. 5 are Schrödinger equations for an arbitrary unitary transformation of radial parts $\eta_{\beta l}$ of the SOAOs within each l-channel. The unitary transformation is introduced to make use of the fact that we are only interested in restoring SOAOs up to a unitary transformation, which gives us additional flexibility for reducing the number of different parameters in the set of V_α functions. Minimizing $F[\{V_\alpha\}]$ simply means maximizing the overlap between the Hilbert spaces spanned by N_η eigenvectors of potentials V_α and the radial functions $\eta_{\beta l}$, with N_η being the total number of $\eta_{\beta l}$ functions. The multiplier r^2 outside the square brackets in Eq. 5 is needed to insure stability of the numeric integration on a logarithmic grid. The global minimum $F[\{V_\alpha\}] = 0$ can be always reached if all V_α are allowed to be different. Indeed, the solution is obtained by inverting the Schrödinger equation for each orbital, similar to the first step in constructing a pseudopotential.

Using the definition of a unitary matrix, $(\mathbf{U}^{(l)})^{\mathrm{T}}\mathbf{U}^{(l)} = \mathbf{U}^{(l)}(\mathbf{U}^{(l)})^{\mathrm{T}} = \mathbf{I}$, and the orthonormality of $\eta_{\beta l}$ within each l-channel, it can be easily shown that $F[\{V_\alpha\}]$ is a positive definite (or semidefinite) quadratic polynomial with respect to elements of symmetric matrices $C^l_{\beta\beta'} = \sum_\alpha U^{(l)}_{\alpha\beta} E_{\alpha l} U^{(l)}_{\alpha\beta'}$ and parameters of the potentials, provided V_α are parameterized linearly, i.e., $V_\alpha(r) = \sum_i v^{(\alpha)}_i f_i(r)$, where $f_i(r)$ are some known functions (see below). Therefore, we can perform convex minimization of $F[\{V_\alpha\}]$ with respect to $v^{(\alpha)}_i$ and

$C^l_{\beta\beta'}$ using compressed sensing to reduce the number of *different* potentials V_α for a given accuracy threshold. For this task, clustered least absolute shrinkage and selection operator (LASSO) [19] is employed. LASSO converts the combinatorially hard problem of finding the most sparse solution (i.e., the solution with the minimal number of parameters) of a linear regression to a convex optimization problem. This is achieved by adding an l_1 penalty (the sum of absolute values of parameters) to the minimized function. In clustered LASSO, the l_1 penalty is a sum of absolute *differences* between the parameters rather than the parameters themselves. In our case, the optimized spherical potentials $V_\alpha^{\rm opt}$ for a given atom are obtained as:

$$V_\alpha^{\rm opt} = \arg\min_{\{V_\alpha\},\{\mathbf{C}^l\}} \left(F[\{V_\alpha\}] + \lambda||\Delta v||_1\right) , \tag{6}$$

where

$$||\Delta v||_1 = \sum_i \sum_{\alpha<\beta} |v_i^{(\alpha)} - v_i^{(\beta)}| . \tag{7}$$

In this approach, differences in individual parameters in V_α are penalized, so that a sparse solution is obtained, but not necessarily with a minimal number of different V_α (because two potentials are different as long as at least one parameter is different). If V_α are stored as a vector, e.g., $\mathbf{v}^{(1)} = (v_1^{(1)}, v_2^{(1)}, \ldots)$, plus the matrix $v_i^{(\alpha)} - v_i^{(1)}$, $\alpha > 1$, the latter will be sparse, but the indices of non-zero elements also need to be stored. Alternatively, we can minimize the number of different V_α, rather than the number of different $v_i^{(\alpha)}$. In this case, only the different $\mathbf{v}^{(\alpha)}$ vectors need to be stored, without any indices.

The number of different V_α can be minimized using the following stepwise LASSO algorithm. First, Eq. 6 is used with the penalty term from Eq. 7. If $F[\{V_\alpha\}]$ is above a threshold when all potentials are identical, λ is determined for which the first non-zero $v_i^{(\alpha)} - v_i^{(\beta)}$ appear(s). Let us denote the subsets of α for which the potentials are different as A_1, A_2, ..., A_q. Now we change the penalty term as follows:

$$||\Delta v||_1 = \sum_{m=1}^{q} \sum_i \sum_{\alpha<\beta;\alpha,\beta\in A_m} |v_i^{(\alpha)} - v_i^{(\beta)}| , \tag{8}$$

i.e., we do not penalize anymore differences between potentials that we already identified as different in at least one parameter. Next, if $F[\{V_\alpha\}]$ is still above the threshold, we again determine λ for which the first non-zero terms (or one term) appear in Eq. 8, and identify new subsets A_m. The procedure is repeated until $F[\{V_\alpha\}]$ is below the threshold. The information needed to restore the SOAOs is the set of different vectors $\mathbf{v}^{(\alpha)}$ and a set of quantum numbers (n, l) for each of the potentials for the N_η radial basis functions.

A complete spline basis set for the same logarithmic radial grid as for SOAOs can be used as the basis $f_i(r)$ for the potentials $V_\alpha(r) = \sum_i v_i^{(\alpha)} f_i(r)$. However, for physical reasons $rV_\alpha(r)$ are expected to be smooth functions, so that further storage reduction can be achieved by reducing the number of grid points for representing the potentials. Several different grid densities can be tested automatically by repeating the procedure for finding $V_\alpha^{\rm opt}$ for each grid and checking if the storage size is reduced. In the worst case scenario, when all V_α are different, the required storage may still be reduced, if a sparser grid is found for the potentials without sacrificing the wavefunction representation accuracy.

The amount of data for WF storage is significantly reduced, if the WFs themselves are localized. The localization introduces sparsity into the matrix of expansion coefficients in equation 1 when the WFs are represented in a localized basis (which is the case in our representation), in addition to the reduction in the matrix' size due to the optimization of the basis functions as described above. Since physical quantities remain unchanged upon unitary transformations of the Hilbert space defined by the occupied and unoccupied orbitals, we can find a transformation that maximally localizes the initial set of WFs. The obtained states are called maximally-localized Wannier functions (MLWF) [20]. An efficient approach for obtaining MLWF based on compressed sensing was developed [21, 22]. The approach avoids starting-point dependent non-convex minimization and arbitrary cut-off parameters. Recently, it has been extended to improve efficiency by using a local orthogonality constraint imposed on the Bloch functions in reciprocal space, and to exactly preserve physical symmetries of the system [23]. Partially occupied bands (such as conduction bands in metals and semiconductors) require "disentangling" before the MLWF can be obtained, as implemented,

e.g., in the Wannier90 program [24]. The resulting MLWFs are approximate, but the accuracy of the transformation is controlled by a well-defined parameter (an energy window), and a very high accuracy can be achieved in practice with minimal additional effort. The gains due to the localization of the WFs increase with the number of atoms and unit cell size, because the number of non-overlapping WFs, and, consequently, the sparsity of the matrix in equation 1 are also increased. For example, in wide band gap oxides the O $2p$ valence states are usually delocalized over the entire unit cell. The storage of each O $2p$ WF represented in a minimal atomic basis in this case would in general require N_{atoms} double-precision numbers, where N_{atoms} is the number of O atoms in the unit cell. However, if the corresponding O $2p$ Wannier functons can be localized on single O atoms, the amount of data to be stored will be reduced by a factor of N_{atoms}.

Once SOAOs are found, the representation of WFs in terms of the new basis set is obtained by minimizing the difference between the WFs $\psi^{\sigma}_{i\mathbf{k}}$ represented in the original basis and in the SOAO basis $\{\phi_{\alpha}\}$:

$$\frac{\partial}{\partial C^{\sigma\alpha}_{i\mathbf{k}}} \sum_{i\mathbf{k}\sigma} \int \left(\psi^{\sigma}_{i\mathbf{k}}(\mathbf{r}) - \sum_{\alpha} C^{\sigma\alpha}_{i\mathbf{k}} \phi_{\alpha}(\mathbf{r}) \right)^2 d^3r = 0\,. \tag{9}$$

The procedure for finding the most compact WF subspace representation described here contains well-defined parameters controlling accuracy of the representation. However, additional constraints can be imposed on the minimization using the method of Lagrange multipliers to avoid errors in important properties of the original WF representation. In particular, the WFs represented in the new basis can be required to be strictly orthonormal, and the electron density to integrate to the number of electrons exactly. With these and other practically relevant constraints, the minimization translates into solving a linear eigenvalue problem.

In principle, the electron density can be stored as the density matrix in the same basis as WFs, $D^{\sigma}_{\alpha\beta} = \sum_{i\mathbf{k}} f^{\sigma}_{i\mathbf{k}} C^{\sigma\alpha *}_{i\mathbf{k}} C^{\sigma\beta}_{i\mathbf{k}}$, provided *all* occupied states have been used to determine SOAOs. This is equivalent to expanding the density $n^{\sigma}(\mathbf{r})$ in the basis of products of the atomic functions. The most compact representation is achieved when SOAOs are natural atomic orbitals. The density matrix is sparse due to locality of the basis functions ($D^{\sigma}_{\alpha\beta}$ are non-zero only if $\phi_{\alpha}(\mathbf{r})$ and $\phi_{\beta}(\mathbf{r})$ overlap).

However, apart from the density, it is desirable to store other scalar fields, in particular Kohn-Sham effective potentials and electrostatic potentials. A method to obtain an accurate decomposition of an arbitrary scalar field in terms of atom-centered functions has been developed previously [6, 9]. The method is based on atom-centered "partition of unity":

$$\sum_{A} p_A(\mathbf{r}) = 1\,, \tag{10}$$

where

$$p_A(\mathbf{r}) = \frac{g_A(\mathbf{r})}{\sum_{A'} g_{A'}(\mathbf{r})}\,, \tag{11}$$

with functions $g_A(\mathbf{r})$ strongly peaked and centered at corresponding atom A, and the index A' running over all atoms in the unit cell. A scalar field $f(\mathbf{r})$ can then be decomposed as a sum of atom-centered parts:

$$f(\mathbf{r}) = \sum_{A,lm} \tilde{f}_{A,lm}(|\mathbf{r} - \mathbf{R}_A|) Y_{lm}(\Omega_A)\,, \tag{12}$$

where $Y_{lm}(\Omega_A)$ are spherical harmonics centered at atom A, and

$$\tilde{f}_{A,lm}(r) = \int_{r=|\mathbf{r}-\mathbf{R}_A|} p_A(\mathbf{r}) f(\mathbf{r}) Y_{lm}(\Omega_A) d\Omega_A\,. \tag{13}$$

The atom-centered parts are calculated on the radial grids of corresponding atoms, and then interpolated using splines. The integrals in equation 13 can be calculated analytically or numerically using optimized angular girds [6, 9]. While arbitrary atom-centered functions $g_A(\mathbf{r})$ can be used for the "partition of unity", the accuracy of the representation may depend on this choice. We suggest to use the partition scheme

developed by Stratmann *et al.* [25], since we find it to give accurate results in ES calculations for a wide range of systems [9]. The scalar field $f(\mathbf{r})$ is stored as the values of $\tilde{f}_{A,lm}(r)$ on the radial grid for each atom A and each (l,m) channel, and can be restored on demand as the sum of splined radial functions multiplied by spherical harmonics. The maximum angular momentum l_{max} for the spherical harmonics can be selected based on the convergence of the representation accuracy with l_{max}. Functions $\tilde{f}_{A,lm}(r)$ should be stored only if they significantly deviate from zero. Taking the number of radial grid points per atom 150, and l_{max}=8 as a safe (upper) limit, we estimate the storage size for a single scalar field to be ∼95 kB per atom within unit cell.

4. Conclusions

In this work, we present a concept of a code-independent compact representation of one-electron WFs and other volumetric data (electron density, electrostatic potential, etc.) produced by ES calculations. The compactness of the representation insures minimization of digital storage requirements for the computational data, while the code-independence makes the data ready for analysis by artificial intelligence. The implementation of the proposed WF storage concept requires minimal development in addition to natural atomic orbital and Wannier function analysis tools. These types of ES analysis provide important parameters describing ES (descriptors), and are therefore must-have tools for computational data analytics. Thus, the proposed storage concept perfectly fits any infrastructure that makes the analysis tools readily available on code-independent basis. The data stored in this form can be complementary to code-specific storage formats that are used, for example, for restarting ES calculations.

The proposed approach for compact WF storage includes the following four major steps. First, an atom-centered basis set is generated such that it accurately represents the set of one-electron states to be stored. This step is only necessary if the WFs are not originally represented in an atom-centered basis (as, e.g., in plane-wave calculations). Second, a minimal set of numeric atom-centered orbitals is found that represents the set of WFs with a desired accuracy. For this, a procedure similar to finding natural atomic orbitals is employed. Third, a set of atom-centered spherically symmetric potentials is determined so that the minimal NAO basis can be calculated numerically on demand by solving the Schrödinger equation for a set of orbital quantum numbers. Finally, localized Wannier functions are calculated to introduce sparsity into the expansion of WFs in the minimal NAO basis set.

For scalar fields such as electron density, Kohn-Sham potentials, and electrostatic potentials we propose to use the approach based on atom-centered partition of unity. This approach is used extensively in electronic-structure calculations based on NAOs, and have been demonstrated to be accurate and efficient.

Each step of the above procedure is controlled by well-defined parameters that can be tuned to insure the accuracy of the WF representation within a desired threshold, theoretically achieving an arbitrarily high accuracy, i.e., an information-lossless conversion, albeit at the cost of increased storage needs. Additional constraints can be imposed to avoid errors in important properties of the original WF representation. In particular, the WFs represented in the new basis can be required to be strictly orthonormal, and the electron density to integrate to the number of electrons exactly. A detailed study of the numerical sensitivity of the procedure to the parameters, and of the performance in terms of storage reduction for real-life applications will be discussed in detail in a future publication.

5. Acknowledgements

The work is supported by European Unions Horizon 2020 research and innovation programme under grant agreement No 676580 via the cluster-of-excellence "Novel Materials Discovery (NOMAD) Laboratory" (`https://nomad-coe.eu/`). We thank Georg Kresse and Peter Blaha for fruitful discussions. SVL also acknowledges support by the Ministry of Education and Science of the Russian Federation (Grant No. 14.Y26.31.0005) for the development of wavefunction analysis methods, and by the Ministry of Education and Science of the Russian Federation in the framework of Increase Competitiveness Program of NUST MISIS (No K2-2017-080) implemented by a governmental decree dated 16 March 2013, No 211, for the development of tools for materials databases.

References

References

[1] X. Andrade, D. Strubbe, U. De Giovannini, A. H. Larsen, M. J. Oliveira, J. Alberdi-Rodriguez, A. Varas, I. Theophilou, N. Helbig, M. J. Verstraete, et al., Real-space grids and the octopus code as tools for the development of new simulation approaches for electronic systems, Physical Chemistry Chemical Physics 17 (47) (2015) 31371–31396.

[2] L. Kronik, A. Makmal, M. L. Tiago, M. Alemany, M. Jain, X. Huang, Y. Saad, J. R. Chelikowsky, Parsec–the pseudopotential algorithm for real-space electronic structure calculations: recent advances and novel applications to nano-structures, physica status solidi (b) 243 (5) (2006) 1063–1079.

[3] S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, S. Goedecker, Daubechies wavelets for linear scaling density functional theory, The Journal of chemical physics 140 (20) (2014) 204110.

[4] C. G. Van de Walle, P. Blöchl, First-principles calculations of hyperfine parameters, Phys. Rev. B 47 (8) (1993) 4244.

[5] P. E. Blöchl, Projector augmented-wave method, Physical review B 50 (24) (1994) 17953.

[6] B. Delley, An all-electron numerical method for solving the local density functional for polyatomic molecules, J. Chem. Phys. 92 (1) (1990) 508–517.

[7] O. Treutler, R. Ahlrichs, Efficient molecular numerical integration schemes, J. Chem. Phys. 102 (1) (1995) 346–354.

[8] J. Baker, J. Andzelm, A. Scheiner, B. Delley, The effect of grid quality and weight derivatives in density functional calculations, The Journal of chemical physics 101 (10) (1994) 8894–8902.

[9] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, Ab initio molecular simulations with numeric atom-centered orbitals, Comp. Phys. Commun. 180 (11) (2009) 2175–2196.

[10] B. Delley, From molecules to solids with the dmol3 approach, The Journal of chemical physics 113 (18) (2000) 7756–7764.

[11] H. B. Curry, I. J. Schoenberg, On pólya frequency functions iv: the fundamental spline functions and their limits, J. d'Analyse Mathématique 17 (1) (1966) 71–107.

[12] C. De Boor, Splines as linear combinations of b-splines. a survey., Tech. rep., DTIC Document (1976).

[13] D. Sánchez-Portal, E. Artacho, J. M. Soler, Projection of plane-wave calculations into atomic orbitals, Solid State Comm. 95 (10) (1995) 685–690.

[14] D. Sánchez-Portal, E. Artacho, J. M. Soler, Analysis of atomic orbital basis sets from the projection of plane-wave results, J. Phys. Condens. Matter 8 (21) (1996) 3859.

[15] A. E. Reed, R. B. Weinstock, F. Weinhold, Natural population analysis, J. Chem. Phys. 83 (2) (1985) 735–746.

[16] A. E. Reed, L. A. Curtiss, F. Weinhold, Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint, Chem. Rev. 88 (6) (1988) 899–926.

[17] B. D. Dunnington, J. Schmidt, Generalization of natural bond orbital analysis to periodic systems: applications to solids and surfaces via plane-wave density functional theory, J. Chem. Theory Comput. 8 (6) (2012) 1902–1911.

[18] B. Carlson, J. M. Keller, Orthogonalization procedures and the localization of wannier functions, Phys. Rev. 105 (1) (1957) 102.

[19] Y. She, Sparse regression with exact clustering, Elec. J. Stat. 4 (2010) 1055–1096.

[20] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, D. Vanderbilt, Maximally localized wannier functions: Theory and applications, Rev. Mod. Phys. 84 (4) (2012) 1419.

[21] V. Ozoliņš, R. Lai, R. Caflisch, S. Osher, Compressed modes for variational problems in mathematics and physics, Proc. Natl. Acad. Sci. 110 (46) (2013) 18368–18373.

[22] V. Ozoliņš, R. Lai, R. Caflisch, S. Osher, Compressed plane waves yield a compactly supported multiresolution basis for the laplace operator, Proc. Natl. Acad. Sci. 111 (5) (2014) 1691–1696.

[23] J. Budich, J. Eisert, E. Bergholtz, S. Diehl, P. Zoller, Search for localized wannier functions of topological band structures via compressed sensing, Phys. Rev. B 90 (11) (2014) 115110.

[24] A. A. Mostofi, J. R. Yates, G. Pizzi, Y.-S. Lee, I. Souza, D. Vanderbilt, N. Marzari, An updated version of wannier90: A tool for obtaining maximally-localised wannier functions, Comp. Phys. Commun. 185 (8) (2014) 2309–2310.

[25] R. E. Stratmann, G. E. Scuseria, M. J. Frisch, Achieving linear scaling in exchange-correlation density functional quadratures, Chem. Phys. Lett. 257 (3) (1996) 213–223.