Accelerating Optical Absorption Spectra and Exciton Energy Computation for Nanosystems via Interpolative Separable Density Fitting

Wei Hu[1], Meiyue Shao[1], Andrea Cepellotti[2], Felipe H. da Jornada[2], Lin Lin[3,1], Kyle Thicke[4], Chao Yang[1], and Steven G. Louie[2]

[1]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States
{whu,myshao,cyang}@lbl.gov
[2]Department of Physics, University of California, Berkeley Berkeley, California 94720, United States
{andrea.cepellotti,jornada,sglouie}@berkeley.edu
[3]Department of Mathematics, University of California, Berkeley Berkeley, California 94720, United States
linlin@math.berkeley.edu
[4]Department of Mathematics, Duke University, Durham, NC 27708, United States
kyle.thicke@duke.edu

Abstract. We present an efficient way to solve the Bethe–Salpeter equation (BSE), a model for the computation of absorption spectra in molecules and solids that includes electron–hole excitations. Standard approaches to construct and diagonalize the Bethe–Salpeter Hamiltonian require at least $\mathcal{O}(N_e^5)$ operations, where N_e is proportional to the number of electrons in the system, limiting its application to small systems. Our approach is based on the interpolative separable density fitting (ISDF) technique to construct low rank approximations to the bare and screened exchange operators associated with the BSE Hamiltonian. This approach reduces the complexity of the Hamiltonian construction to $\mathcal{O}(N_e^3)$ with a much smaller pre-constant. Here, we implement the ISDF method for the BSE calculations within the Tamm–Dancoff approximation (TDA) in the BerkeleyGW software package. We show that ISDF-based BSE calculations in molecules and solids reproduce accurate exciton energies and optical absorption spectra with significantly reduced computational cost.

1 Introduction

Many-Body Perturbation Theory [20] is a powerful tool to describe one-particle and two-particle excitations, and to obtain excitation energies and absorption spectra in molecules and solids [21]. Within many-body perturbation theory, Hedin's GW approximation [10] has been successfully used to compute quasi-particle (one-particle) excitation energies and the Bethe–Salpeter equation (BSE) [24]

describes the interaction of a electron–hole pair (two-particle excitation) produced during the spectral absorption in molecules and solids [23]. Good agreement between theory and experiment can only be achieved taking into account such electron–hole interaction by solving the BSE. The BSE is an eigenvalue problem. In the context of optical absorption, the eigenvalues of the Bethe–Salpeter Hamiltonian (BSH) are related to quasi-particle exciton energies and the corresponding eigenfunctions yield the exciton wavefunctions.

The BSH has a special block structure to be shown in the next section. It consists of bare and screened exchange kernels that depend on the products of single-particle orbitals obtained from a mean-field calculation. The evaluation of these kernels requires at least $\mathcal{O}(N_e^5)$ operations in a conventional approach, which is very costly for large complex systems that contain hundreds or even thousands of atoms. There are several methods, which have been developed recently to generate a reduced basis set, to reduce such high computational cost for the BSE calculations [2,12,16,19,22].

In this paper, we present a more efficient way to construct the BSH. Such a construction allows the BSE to be solved efficiently by an iterative diagonalization scheme. Our approach is based on the recently developed interpolative separable density fitting (ISDF) decomposition [18]. This ISDF decomposition has been applied to accelerate a number of applications in computational chemistry and materials science, including two-electron integral computation [18], correlation energy in the random phase approximation [17], density functional perturbation theory [15], and hybrid density functional calculations [11]. Such a decomposition is used to approximate the matrix consisting of products of single-particle orbital pairs as the product of a matrix consisting of a small number of numerical auxiliary basis vectors and an expansion coefficient matrix [11]. This low rank approximation effectively allows us to construct low rank approximations to the bare and screened exchange kernels. The construction of ISDF compressed BSE Hamiltonian matrix only requires $\mathcal{O}(N_e^3)$ operations if the rank of the numerical auxiliary basis can be kept at $\mathcal{O}(N_e)$ and if we keep the bare and screened exchange kernel in a low rank factored form. This is considerably more efficient than the $\mathcal{O}(N_e^5)$ complexity required in a conventional approach.

By keeping these kernels in the decomposed form, the matrix–vector multiplications required in iterative diagonalization procedures for computing the desired eigenvalues and eigenvectors of H_{BSE} can be performed efficiently. We may also use these efficient matrix–vector multiplications in a structure preserving Lanczos algorithm [25] to obtain an approximate absorption spectrum, without explicitly diagonalizing the approximate H_{BSE}.

We have implemented the ISDF based BSH construction in the BerkeleyGW software package [5]. We demonstrate that this approach can produce accurate exciton energies and optical absorption spectra for molecules and solids, and reduce the computational cost associated with the construction of the BSE Hamiltonian significantly compared to the algorithms used in the existing version of the BerkeleyGW software package.

2 Bethe–Salpeter equation

The Bethe–Salpeter equation is an eigenvalue problem of the form

$$H_{\mathrm{BSE}}X = XE, \tag{1}$$

where X is an exciton wavefunction, E is the corresponding exciton energy and H_{BSE} is the Bethe–Salpeter Hamiltonian that has the following block structure

$$H_{\mathrm{BSE}} = \begin{bmatrix} D + 2V_A - W_A & 2V_B - W_B \\ -2\overline{V}_B + \overline{W}_B & -D - 2\overline{V}_A + \overline{W}_A \end{bmatrix}, \tag{2}$$

where $D(i_v i_c, j_v j_c) = (\epsilon_{i_c} - \epsilon_{i_v})\delta_{i_v j_c}\delta_{i_c j_c}$ is an $(N_v N_c) \times (N_v N_c)$ diagonal matrix with ϵ_{i_v}, $i_v = 1, 2, \ldots, N_v$ being the quasi-particle energies associated with valence bands and ϵ_{i_c}, $i_c = N_v + 1, N_v + 2, \ldots, N_v + N_c$ being the quasi-particle energies associated with conduction bands. These quasi-particle energies are typically obtained from the so-called GW calculation [23]. The V_A and V_B matrices represent the (bare) *exchange* of electron–hole pairs, and the W_A and W_B matrices are often referred to as the *direct* terms that describe screened exchange of electron–hole pairs. These matrices are defined as follows:

$$\begin{aligned} V_A(i_v i_c, j_v j_c) &= \int \bar{\psi}_{i_c}(\mathbf{r})\psi_{i_v}(\mathbf{r})V(\mathbf{r},\mathbf{r}')\bar{\psi}_{j_v}(\mathbf{r}')\psi_{j_c}(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \\ V_B(i_v i_c, j_v j_c) &= \int \bar{\psi}_{i_c}(\mathbf{r})\psi_{i_v}(\mathbf{r})V(\mathbf{r},\mathbf{r}')\bar{\psi}_{j_c}(\mathbf{r}')\psi_{j_v}(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \\ W_A(i_v i_c, j_v j_c) &= \int \bar{\psi}_{i_c}(\mathbf{r})\psi_{j_c}(\mathbf{r})W(\mathbf{r},\mathbf{r}')\bar{\psi}_{j_v}(\mathbf{r}')\psi_{i_v}(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \\ W_B(i_v i_c, j_v j_c) &= \int \bar{\psi}_{i_c}(\mathbf{r})\psi_{j_v}(\mathbf{r})W(\mathbf{r},\mathbf{r}')\bar{\psi}_{j_c}(\mathbf{r}')\psi_{i_v}(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \end{aligned} \tag{3}$$

where ψ_{i_v} and ψ_{i_c} are usually taken to be the valence and conduction single-particle orbitals obtained from a Kohn–Sham density functional theory (KS-DFT) calculation respectively, and $V(\mathbf{r},\mathbf{r}')$ and $W(\mathbf{r},\mathbf{r}')$ are the bare and screened Coulomb operators. Both V_A and W_A are Hermitian, whereas V_B and W_B are complex symmetric in general. In the so-called Tamm–Dancoff approximation (TDA) [21], both V_B and W_B are neglected and set to zeros in Equation (2). As a result, H_{BSE} is Hermitian in this case, and we only need to be concerned with the upper left block of H_{BSE}.

Let $M_{cc}(\mathbf{r}) = \{\psi_{i_c}\bar{\psi}_{j_c}\}$, $M_{vc}(\mathbf{r}) = \{\psi_{i_c}\bar{\psi}_{i_v}\}$, and $M_{vv}(\mathbf{r}) = \{\psi_{i_v}\bar{\psi}_{j_v}\}$ be matrices that consist of products of discretized orbital pairs in real space, and $\hat{M}_{cc}(\mathbf{G})$, $\hat{M}_{vc}(\mathbf{G})$, $\hat{M}_{vv}(\mathbf{G})$ be the reciprocal space representation of these matrices. Equations (3) can then be succinctly written as

$$\begin{aligned} V_A &= \hat{M}_{vc}^{*}\hat{V}\hat{M}_{vc}, \\ W_A &= \mathrm{reshape}(\hat{M}_{cc}^{*}\hat{W}\hat{M}_{vv}), \end{aligned} \tag{4}$$

where $\hat{V}$ and $\hat{W}$ are reciprocal space representations of the bare and screened exchange operators V and W, respectively, and the reshape function is used to

map the $(i_c j_c, i_v j_v)$th element on the right-hand side of (4) to the $(i_c i_v, j_c j_v)$th element of W_A. A similar set of equations can be derived for V_B and W_B. However, in this paper, we will mainly focus on the TDA model.

The reason that the right-hand sides of (4) are computed in the reciprocal space is that $\hat{V}$ is diagonal and an energy cutoff is often adopted to limit the number of the Fourier components used to represent ψ_i. As a result, the leading dimension of $\hat{M}_{cc}$, $\hat{M}_{vc}$ and $\hat{M}_{cc}$, which we denote by N_g, is often much smaller than that of M_{cc}, M_{vc} and M_{vv}, which we denote by N_r.

In additional to performing $\mathcal{O}(N_e^2)$ Fast Fourier transforms (FFTs) to obtain $\hat{M}_{cc}$, $\hat{M}_{vc}$ and $\hat{M}_{vv}$ from M_{cc}, M_{vc} and M_{vv}, respectively, we need to perform at least

$$\mathcal{O}(N_g N_c^2 N_v^2 + N_g^2 N_c N_v) \tag{5}$$

floating-point operations to obtain V_A and W_A using matrix–matrix multiplication operations.

Note that, to achieve high accuracy with a large basis set, such as the plane wave basis set, N_g is typically much larger than N_c or N_v. The number of occupied bands is either N_e or $N_e/2$ depending on how spin is counted. In many existing calculations, the actual number of conducting band N_c included in the calculation is often a small multiple of N_v, whereas N_g is often 100–10000 $\times N_e$ ($N_r \sim 10 \times N_g$). Therefore, the second term in Equation (5), which accounts for the cost of multiplying $\hat{W}_A$ with $\hat{Z}$ in Equation (4) can be much larger than other parts under the Tamm–Dancoff approximation (TDA) in the BSE calculations.

3 Interpolative separable density fitting (ISDF) decomposition

In order to reduce the computational complexity, we seek to minimize the number of integrals we need to perform in Equation (3). This is possible if we can rewrite the matrix M_{ij}, where the labels i and j are indices of either valence or conducting orbitals, as the product of a matrix Θ_{ij} that contains a set of N_{ij}^t linearly independent auxiliary basis vectors with $N_{ij}^t \approx tN_e \ll \mathcal{O}(N_e^2)$ (t is a small constant referred as a rank truncation parameter) [11] and an expansion coefficient matrix C_{ij}. For large problems, the number of columns of M_{ij}, which is either $\mathcal{O}(N_v N_c)$, $\mathcal{O}(N_v^2)$, or $\mathcal{O}(N_c^2)$, is typically larger than the number of grid points N_r on which $\psi_n(\mathbf{r})$ is sampled, i.e., the number of rows in M_{ij}. As a result, N_{ij}^t should be much smaller than the number of columns of M_{ij}. Even when a cutoff is used to limit the size of N_c or N_v so that the number of columns in M_{ij} is much less than N_g, we can still approximate M_{ij} by $\Theta_{ij} C_{ij}$ with a Θ_{ij} that has a smaller rank $N_{ij}^t \sim t\sqrt{N_i N_j}$.

To simplify our discussion, let us drop the subscript of M, Θ and C for the moment, and describe the basic idea of ISDF. The optimal low rank approximation of M can be obtained from a singular value decomposition. However, the complexity of this decomposition is at least $\mathcal{O}(N_r^2 N_e^2)$ or $\mathcal{O}(N_e^4)$. An alternative decomposition, which is close to optimal, but has a much favorable complexity

has recently been developed. This type of decomposition is called interpolative separable density fitting (ISDF) [11], which we will now describe.

In ISDF, instead of computing Θ and C simultaneously, we fix the coefficient matrix C first, and determine the auxiliary basis matrix Θ by solving a linear least squares problem

$$\min \|M - \Theta C\|_F^2, \tag{6}$$

where each column of M is given by $\psi_i(\mathbf{r})\bar{\psi}_j(\mathbf{r})$ sampled on a dense real space grids $\{\mathbf{r}_i\}_{i=1}^{N_r}$, and $\Theta = [\zeta_1, \zeta_2, \ldots, \zeta_{N^t}]$ contains the auxiliary basis vectors to be determined, $\|\cdot\|_F$ denotes the Frobenius norm.

We choose C to be a matrix that consists of $\psi_i(\mathbf{r})\bar{\psi}_j(\mathbf{r})$ evaluated at a subset of N^t carefully chosen real space grid points, with $N^t \ll N_r$ and $N^t \ll N_e^2$, i.e., each column of C indexed by (i, j) is given by

$$[\psi_i(\hat{\mathbf{r}}_1)\bar{\psi}_j(\hat{\mathbf{r}}_1), \cdots, \psi_i(\hat{\mathbf{r}}_k)\bar{\psi}_j(\hat{\mathbf{r}}_k), \cdots, \psi_i(\hat{\mathbf{r}}_{N^t})\bar{\psi}_j(\hat{\mathbf{r}}_{N^t})]^\mathsf{T}.$$

If the minimum of the objective function in Equation (6) is zero, the product of Θ and a column of C can be viewed as an interpolation of corresponding function $\{\psi_i(\mathbf{r})\bar{\psi}_j(\mathbf{r})\}$ in M. However, in general, we cannot expect the minimum of (6) be zero. The least squares minimizer is given by

$$\Theta = MC^*(CC^*)^{-1}. \tag{7}$$

It may appear that the matrix–matrix multiplications MC^* and CC^* take $\mathcal{O}(N_e^4)$ operations because the size of M is $N_r \times \mathcal{O}(N_e^2)$ and the size of C is $N^t \times \mathcal{O}(N_e^2)$. However, both multiplications can be carried out with fewer operations due to the separable structure of M and C [11]. As a result, the computational complexity for computing the interpolation vectors is $\mathcal{O}(N_e^3)$.

Intuitively, the least squares problem in Equation (6) is easier to solve when the rows of C, which can be selected from the rows of M, are maximally linearly independent. This task can be achieved by performing a QR factorization of M^T with column pivoting (QRCP) [4]. In QRCP, we choose a permutation Π such that the factorization

$$M^\mathsf{T}\Pi = QR \tag{8}$$

yields a unitary matrix Q and an upper triangular matrix R with decreasing matrix elements along the diagonal of R. The magnitude of each diagonal element R indicates how important the corresponding column of the permuted M^T is, and whether the corresponding grid point should be chosen as an interpolation point. The QRCP decomposition can be terminated when the (N^t+1)-st diagonal element of R becomes less than a predetermined threshold. The leading N^t columns of the permuted M^T are considered to be maximally linearly independent numerically. The corresponding grid points are chosen as the interpolation points. The indices for the chosen interpolation points $\hat{\mathbf{r}}_{N^t}$ can be obtained from indices of the nonzero entries of the first N^t columns of the permutation matrix Π.

Roughly speaking, the QRCP moves matrix columns of M^T with large norms to the left, and pushes matrix columns of M^T with small norms to the right.

Note that the square of the vector 2-norm of the column of M^T associated with $\mathbf{r}$ is just

$$\sum_{i,j=1}^{N} \left|\psi_i(\mathbf{r})\bar{\psi}_j(\mathbf{r})\right|^2 = \left(\sum_{i=1}^{N}\left|\psi_i(\mathbf{r})\right|^2\right)\left(\sum_{j=1}^{N}\left|\psi_j(\mathbf{r})\right|^2\right). \tag{9}$$

In the case when $\{\psi_i\}$ and $\{\psi_j\}$ both belong to the set of occupied orbitals, the norm of each column of M^{T} is simply the electron density. Hence the interpolation points chosen by QRCP tend to locate in areas where the electron density is relatively large. Once a column is selected, all other columns are immediately orthogonalized with respect to the chosen column. Hence nearly linearly dependent matrix columns will not be selected repeatedly. As a result, the interpolation points chosen by QRCP are well separated spatially. Notice that the standard QRCP procedure has a high computational cost of $\mathcal{O}(N_e^2 N_r^2) \sim \mathcal{O}(N_e^4)$. But it can be combined with the randomized sampling method [18] so that its cost is reduced to $\mathcal{O}(N_r N_e^2) \sim \mathcal{O}(N_e^3)$.

4 Low rank representations of bare and screened exchange operators via ISDF

Applying the ISDF decomposition to M_{cc}, M_{vc} and M_{vv} yields

$$\begin{aligned} M_{cc} &\approx \Theta_{cc} C_{cc}, \\ M_{vc} &\approx \Theta_{vc} C_{vc}, \\ M_{vv} &\approx \Theta_{vv} C_{vv}. \end{aligned} \tag{10}$$

It follows from Equations (3), (4) and (10) that the exchange and direct terms of the BSE Hamiltonian can be written as

$$\begin{aligned} V_A &= C_{vc}^* \widetilde{V}_A C_{vc}, \\ W_A &= \mathrm{reshape}(C_{cc}^* \widetilde{W}_A C_{vv}), \end{aligned} \tag{11}$$

where $\widetilde{V}_A = \hat{\Theta}_{vc}^* \hat{V} \hat{\Theta}_{vc}$ and $\widetilde{W}_A = \hat{\Theta}_{cc}^* \hat{W} \hat{\Theta}_{vv}$ are the *projected* exchange and direct terms under the auxiliary basis $\hat{\Theta}_{vc}$, $\hat{\Theta}_{cc}$ and $\hat{\Theta}_{vv}$. Here, $\hat{\Theta}_{vc}$, $\hat{\Theta}_{cc}$ and $\hat{\Theta}_{vv}$ are reciprocal space representations of Θ_{vc}, Θ_{cc} and Θ_{vv}, respectively, that can be obtained via FFTs,

Note that the dimension of the matrix $C_{cc}^* \widetilde{W}_A C_{cc}$ on the right-hand side of Equation (11) is $N_c^2 \times N_v^2$. It needs to be reshaped into a matrix of dimension $N_v N_c \times N_v N_c$ according to the mapping $W_A(i_c j_c, i_v j_v) \to W_A(i_v i_c, j_v j_c)$ before it can combined with V_A matrix to construct the BSH.

Once the ISDF approximations for M_{vc}, M_{cc} and M_{vv} are available, the cost for constructing a low rank approximation to the exchange and direct terms reduced to that of computing the projected exchange and direct kernels $\hat{\Theta}_{vc}^* \hat{V} \hat{\Theta}_{vc}$ and $\hat{\Theta}_{cc}^* \hat{W} \hat{\Theta}_{vv}$, respectively. If the ranks of Θ_{vc}, Θ_{cc} and Θ_{vv} are N_{vc}^t, N_{cc}^t and N_{vv}^t, respectively, then the computational complexity for computing the compressed exchange and direct kernels is $\mathcal{O}(N_{vc}^t N_{vc}^t N_g + N_{cc}^t N_{vv}^t N_g + N_{vv}^t N_g^2)$, which

is significantly lower than the complexity of the conventional approach given in (5). When $N_{vc}^t \sim t\sqrt{N_v N_c}$, $N_{cc}^t \sim t\sqrt{N_c N_c}$ and $N_{vv}^t \sim t\sqrt{N_v N_v}$ are on the order of N_e, the complexity of constructing the compressed kernels is $\mathcal{O}(N_e^3)$.

5 Iterative diagonalization of the BSE Hamiltonian

In the conventional approach, exciton energies and wavefunctions can be computed by using the recently developed BSEPACK library [26,27] to diagonalize the BSE Hamiltonian H_{BSE}. When TDA is adopted, we may also just use a standard diagonalization procedure implemented in the ScaLAPACK [1] library.

When ISDF is used to construct low rank approximations to the bare and screened exchange operators V_A and W_A, we should keep both matrices in the factored form given by Equation (11). This is because that multiplying the matrices on the right-hand sides of Equation (11) would require at least $\mathcal{O}(N_e^5)$ operations, which is higher than the cost for using the ISDF procedure to construct low rank approximations to the BSH. Instead of using BSEPACK or ScaLAPACK which has a complexity of $\mathcal{O}(N_e^6)$, we propose to use iterative methods to diagonalize the approximate BSH constructed via the ISDF decomposition.

When TDA is used, several iterative methods such as the Lanczos [14] and LOBPCG [13] algorithms can be used to compute a few desired eigenvalues of the H_{BSE}. In each step of these algorithms, we need to multiply H_{BSE} with a vector x of size $N_v N_c$. When V_A is kept in the factored form given by (11), $V_A x$ can be evaluated as three matrix vector multiplications performed in sequence, i.e.,

$$V_A x \leftarrow C_{vc}^* \big[\widetilde{V_A}(C_{vc} x)\big]. \tag{12}$$

The complexity of these calculations is $\mathcal{O}(N_v N_c N_{vc}^t)$. If N_{vc}^t is on the order of N_e, then each $V_A x$ can be carried out in $\mathcal{O}(N_e^3)$ operations.

Because $C_{cc}^* \widetilde{W_A} C_{vv}$ cannot be multiplied with a vector x of size $N_v N_c$ before it is reshaped, a different multiplication scheme needs to be used. It follows from the separable nature of C_{vv} and C_{cc} that this multiplication can be succinctly described as

$$W_A x = \text{reshape}\left[\Psi_c^* \big(\widetilde{W} \odot (\Psi_c X \Psi_v^*)\big)\Psi_v\right], \tag{13}$$

where X is a $N_c \times N_v$ matrix reshaped from the vector x, Ψ_c is a $N_{cc}^t \times N_c$ matrix containing $\psi_{i_c}(\hat{r}_k)$ as its elements, Ψ_v is a $N_{vv}^t \times N_v$ matrix containing $\psi_{i_v}(\hat{r}_k)$ as its elements, and $\odot$ denotes componentwise multiplication (Hadamard product). The reshape function is used to turn the $N_c \times N_v$ matrix–matrix product back into a size $N_v N_c$ vector. If N_{vv}^t and N_{cc}^t are on the order of N_e, then all matrix–matrix multiplications in Equation (13) can be carried out in $\mathcal{O}(N_e^3)$ operations. This make the complexity of each step of the iterative method $\mathcal{O}(N_e^3)$. If the number of iterative steps required to reach convergence is not excessively large, then the ISDF enabled iterative diagonalization can be carried out in $\mathcal{O}(N_e^3)$ operations.

6 Estimating optical absorption spectra without diagonalization

If we have all eigenpairs of H_{BSE}, we can easily obtain the optical absorption spectrum, which is the imaginary part of the dielectric function defined as

$$\varepsilon_2(\omega) = \mathrm{Im}\left[\frac{8\pi^2 e^2}{\Omega} d_r^H \left((\omega - \mathrm{i}\eta)I - H_{\mathrm{BSE}}\right)^{-1} d_l\right], \tag{14}$$

where Ω is the volume of the primitive cell, e is the elementary charge, d_r and d_l are the right and left optical transition vectors, and η is a broadening factor used to account for the lifetime of excitation. However, it can become prohibitively expensive to use an iterative diagonalization method to compute all eigenpairs of H_{BSE}.

Fortunately, it is possible to use a structure preserving iterative method to estimate the optical absorption spectrum without explicitly computing all eigenpairs of H_{BSE}. In Ref. [3,25], we developed a structure preserving Lanczos algorithm for estimating the optical spectrum. The algorithm has been implemented in the BSEPACK [27] library. When TDA is used, the standard Lanczos algorithm can be used to estimate the absorption spectrum. When our objective is to obtain the basic shape of the absorption spectrum to identify where the major peaks are, it is not necessary to compute all eigenpairs of H_{BSE}. As a result, the accuracy required to construct approximate bare and screened exchange operators used in BSE can possibly be lowered, thereby allowing us to use a more aggressive truncation threshold in ISDF to further reduce the cost of H_{BSE} construction. We will demonstrate this possibility in the next section.

7 Numerical results

In this section, we demonstrate the accuracy and efficiency of the ISDF method when it is used to compute exciton energies and optical absorption spectrum in the BSE framework. We implemented the ISDF based BSH construction in the BerkeleyGW software package [5]. BerkeleyGW is a massively parallel computational package that uses a many-body perturbation theory and Green's function formalism to study quasi-particle excitation energies and optical absorption of nanosystems. We use the *ab initio* software package Quantum ESPRESSO (QE) [7] to compute the ground-state quantities required in the GW and BSE calculations. In our Quantum ESPRESSO density functional theory (DFT) based electronic structure calculations, we use Hartwigsen–Goedecker–Hutter (HGH) norm-conserving pseudopotentials [9] and the LDA [8] exchange–correlation functional. All the calculations were carried out on a single core at the Cori[1] systems at the National Energy Research Scientific Computing Center (NERSC).

We performed calculations for three systems. They consist of a bulk silicon Si_8 system and two molecular systems: carbon monoxide (CO) and benzene (C_6H_6)

[1] https://www.nersc.gov/systems/cori/

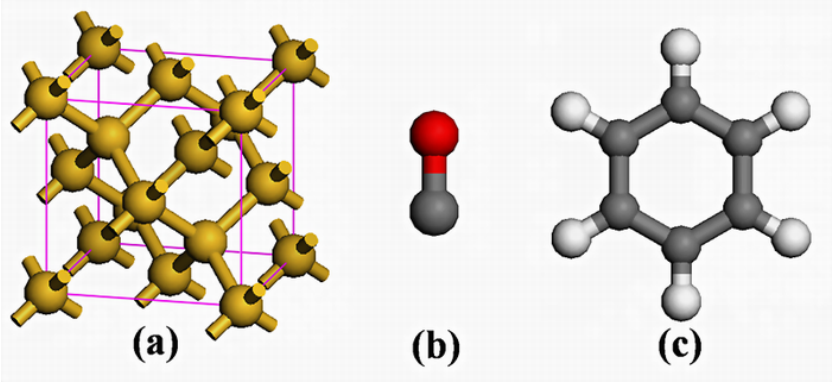

Fig. 1. Atomic structures of (a) a bulk silicon Si_8 unit cell, (b) carbon monoxide (CO) and (c) benzene (C_6H_6) molecules. The white, gray, red, and yellow balls denote hydrogen, carbon, oxygen, and silicon atoms, respectively.

Table 1. Parameters of system size for bulk silicon Si_8, carbon monoxide (CO) and benzene (C_6H_6) molecules used for constructing corresponding BSE Hamiltonian H_{BSE}.

System	N_r	N_g	N_v	N_c	dim(H_{BSE})
Si_8	35937	2301	16	64	2048
CO	19683	1237	5	60	600
Benzene	91125	6235	15	60	1800

as plotted in Fig. 1. All systems are closed shell systems, and the number of occupied bands is $N_v = N_e/2$, where N_e is the valence electrons in the system.

7.1 Accuracy

We first measure the accuracy of the ISDF method by comparing the computed eigenvalues of the BSH matrices constructed with and without the ISDF decomposition.

In our test, we set the plane wave energy cutoff required in the QE calculations to $E_{\text{cut}} = 10$ Ha, which is relatively low. However, this is sufficient for assessing the effect of ISDF. Such a choice of E_{cut} results in $N_r = 35937$ and $N_g = 2301$ for the Si_8 system, $N_r = 19683$ and $N_g = 1237$ for the CO molecule ($N_v = 5$), $N_r = 91125$ and $N_g = 6235$ for the benzene molecule. We only include the lowest N_c conducting bands in the BSE calculation. The number of active conduction bands (N_c) and valence bands (N_v), the number of reciprocal grids and the dimensions of the corresponding BSE Hamiltonian H_{BSE} for these three systems are listed in Table 1.

In Fig. 2, we plot the singular values of the matrices $M_{vc}(\mathbf{r}) = \{\psi_{i_c}(\mathbf{r})\bar{\psi}_{i_v}(\mathbf{r})\}$, $M_{cc}(\mathbf{r}) = \{\psi_{i_c}(\mathbf{r})\bar{\psi}_{j_c}(\mathbf{r})\}$ and $M_{vv}(\mathbf{r}) = \{\psi_{i_v}(\mathbf{r})\bar{\psi}_{j_v}(\mathbf{r})\}$ associated with the CO

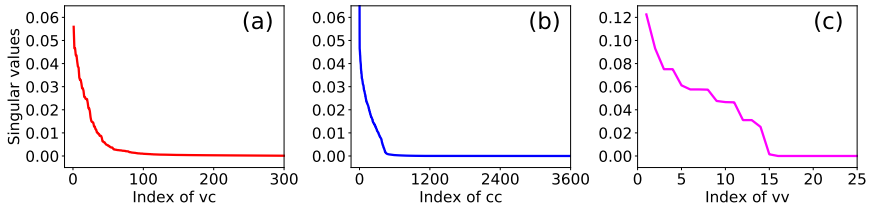

Fig. 2. The singular values of (a) $M_{vc}(\mathbf{r}) = \{\psi_{i_c}(\mathbf{r})\bar{\psi}_{i_v}(\mathbf{r})\}$ ($N_{vc} = 300$), (b) $M_{cc}(\mathbf{r}) = \{\psi_{i_c}(\mathbf{r})\bar{\psi}_{j_c}(\mathbf{r})\}$ ($N_{cc} = 3600$) and (c) $M_{vv}(\mathbf{r}) = \{\psi_{i_v}(\mathbf{r})\bar{\psi}_{j_v}(\mathbf{r})\}$ ($N_{vv} = 25$).

molecule. We observe that the singular values of these matrices decay rapidly. For example, the leading 500 (out of 3600) singular values of $M_{cc}(\mathbf{r})$ decreases rapidly towards zero. All other singular values are below 10^{-4}. Therefore, the numerical rank N_{cc}^t of M_{cc} is roughly 500 ($t = 8.3$), or roughly 15% of the number of columns in M_{cc}. Consequently, we expect that the rank of Θ_{cc} produced in ISDF decomposition can be set to 15% of N_c^2 without sacrificing the accuracy of the computed eigenvalues.

This prediction is confirmed in Fig. 3, where we plot the absolute difference between the lowest the exciton energy of Si_8 computed with and without using ISDF to construct H_{BSE}. To be specific, the error in the desired eigenvalue is computed as $\Delta E = E_{\text{ISDF}} - E_{\text{BGW}}$, where E_{ISDF} is computed from the H_{BSE} constructed with ISDF approximation, and E_{BGW} is computed from a standard H_{BSE} constructed without using ISDF. We first vary one of the ratios N_{cc}^t/N_{cc}, N_{vc}^t/N_{vc} and N_{vv}^t/N_{vv} while holding the others at a constant of 1. We observe that the error in the lowest exciton energy (positive eigenvalue) is around 10^{-3} Ha, when either N_{cc}^t/N_{cc} or N_{vc}^t/N_{vc} is set to 0.1 while the other ratios are held at 1. However, reducing N_{vv}^t/N_{vv} to 0.1 introduces a significant amount of error (0.1 Ha) in the lowest exciton energy. This is likely due to the fact that $N_v = 16$ is too small. We then hold N_{vv}^t/N_{vv} at 0.5 and let both N_{cc}^t/N_{cc} and N_{vc}^t/N_{vc} vary. The variation of ΔE with respect to these ratios is also plotted as in Fig. 3. We observe that the error in the lowest exciton energy is still around 10^{-3} Ha even when both N_{cc}^t/N_{cc} and N_{vc}^t/N_{vc} are set to 0.1.

We then check the absolute error ΔE (Ha) of all the exciton energies computed with the ISDF method by comparing them with the ones obtained from a conventional BSE calculation implemented in BerkeleyGW for the CO and benzene molecules. As we can see from Fig. 4, the errors associated with these eigenvalues are all below 0.002 Ha when N_{cc}^t/N_{cc} is 0.1.

7.2 Efficiency

At the moment, we have only implemented a sequential version of the ISDF method within the BerkeleyGW software package. Therefore, our efficiency test is limited in the size of the problem as well as the number of conducting bands (N_c) we can include in the bare and screened exchange operators. As a result, our

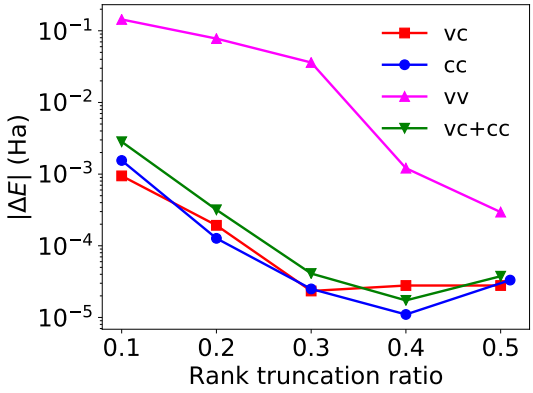

Fig. 3. The change of absolute error ΔE in the smallest eigenvalue of H_{BSE} associated with the Si_8 system with respect to different truncation levels used in ISDF approximation of M_{vc}, M_{cc} and M_{vv}. The curves labeled by 'vc', 'cc', 'vv' correspond to calculations in which only one of the ratios N_{vc}^t/N_{vc}, N_{cc}^t/N_{cc} and N_{vv}^t/N_{vv} changes while all other parameters are held constant. The curve labeled by 'vc + cc' corresponds to the calculation in which both N_{vc}^t/N_{vc} and N_{cc}^t/N_{cc} change at the same rate ($N_{vv}^t = N_{vv}$).

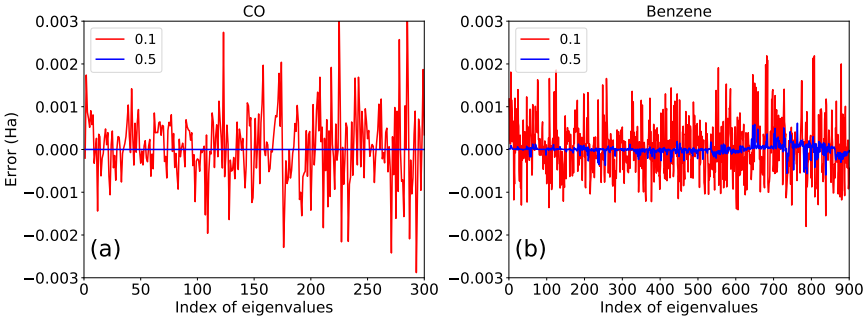

Fig. 4. Error in all eigenvalues of the BSH associated with the CO (a) and benzene (b) molecules. Two rank truncation ratios $N_{cc}^t/N_{cc} = 0.5$ ($t = 30.0$) and $N_{cc}^t/N_{cc} = 0.1$ ($t = 6.0$) are used in the tests.

performance measurement does not fully reflect the computational complexity analysis presented in the previous sections. In particular, taking benzene as an example, $N_g = 6235$ is much larger than $N_v = 15$ and $N_c = 60$, therefore the computational cost of $N_g^2 N_v^2 \sim \mathcal{O}(N_e^4)$ term is much higher than the $N_g N_v^2 N_c^2 \sim \mathcal{O}(N_e^5)$ term in the conventional BSE calculations.

Nonetheless, in this section, we will demonstrate the benefit of using ISDF to reduce the cost for constructing the BSE Hamiltonian H_{BSE}. In Table 2, we focus on the benzene example and report the wall-clock time required to construct

Table 2. The variation of time required to carry out the ISDF decomposition of M_{vc}, M_{vv} and M_{cc} with respect to rank truncation ratio.

Rank truncation ratio			Time (s) for $M_{ij}(\mathbf{r})$		
N_{vc}^t/N_{vc}	N_{vv}^t/N_{vv}	N_{cc}^t/N_{cc}	M_{vc}	M_{vv}	M_{cc}
1.0	0.5	0.5	157.0	5.8	578.9
1.0	0.5	0.1	157.0	5.8	34.3
0.1	0.1	0.1	4.3	0.7	34.3

the ISDF approximations of the M_{vc}, M_{cc}, and M_{vv} matrices at different rank truncation levels. Without using ISDF, it takes 746.0 seconds to construct the reciprocal space representations of M_{vc}, M_{cc}, and M_{vv} in BerkeleyGW. Most of the timing is spent in the large number of FFTs applied to M_{vc}, M_{cc}, and M_{vv} to obtain the reciprocal space representation of these matrices. We can clearly see that if only N_{cc}^t/N_{cc} is changed from 0.5 ($t = 30.0$) to 0.1 ($t = 6.0$), the wall-clock time used to construct a low rank approximation to M_{cc} can be reduced from 578.9 to 34.3 seconds. Furthermore, the total cost of computing M_{vc}, M_{cc} and M_{vv} can be further reduced to around 1/19th that in a conventional approach (39.3 vs. 746.0 seconds) if N_{vc}^t/N_{vc}, N_{vv}^t/N_{vv} and N_{cc}^t/N_{cc} are all set to 0.1.

Note that because ISDF decomposition is carried out on a real space grid, its measured cost reflect the cost for performing QRCP in real space. Even though QRCP with random sampling has $\mathcal{O}(N_e^3)$ complexity, it has a relatively large pre-constant compared to the size of the problem. Hence the measured cost of QRCP is relatively high in this case. Recently, Dong et al. proposed a new approach to find the interpolation points based on the centroidal Voronoi tessellation (CVT) method [6], which offers a much less expensive alternative to the QRCP procedure when the ISDF method is used in hybrid functional calculations. We will explore this CVT method in the BSE-ISDF calculations in the future.

In Table 3, we report the wall-clock time required to construct the projected bare and screened exchange matrices $\widetilde{V}_A$ and $\widetilde{W}_A$ that appear in Equation (11) once the ISDF approximations of M_{vc}, M_{vv}, and M_{cc} become available. Without ISDF, it takes $1.574 + 4.198 = 5.772$ seconds to construct both W_A and V_A. When N_{cc}^t/N_{cc} is set to 0.1 only, the construction cost for W_A, which is the dominant cost, is reduced by a factor of 2.8. Furthermore, if N_{vc}^t/N_{vc}, N_{vv}^t/N_{vv} and N_{cc}^t/N_{cc} are all set to 0.1. We reduce the cost for constructing $\widetilde{V}_A$ and $\widetilde{W}_A$ by a factor of 63.0 and 10.1, respectively. Note that the original implementation of the W_A and V_A in BerkeleyGW is much slower because the elements of W_A and V_A are integrated one by one. For benzene, it take takes 103,154 seconds (28.65 hours) to construct the BSE Hamiltonian H_{BSE} in the original BerkeleyGW code.

Table 3. The variation of time required to construct the projected bare and screened exchange matrices $\widetilde{V}_A$ and $\widetilde{W}_A$ exhibited by the ISDF method respect to rank truncation ratio.

Rank truncation ratio			Time (s) for H_{BSE}	
N_{vc}^t/N_{vc}	N_{vv}^t/N_{vv}	N_{cc}^t/N_{cc}	$\widetilde{V}_A$	$\widetilde{W}_A$
1.0	1.0	1.0	1.574	4.198
1.0	0.5	0.1	1.574	1.474
0.1	0.1	0.1	0.025	0.414

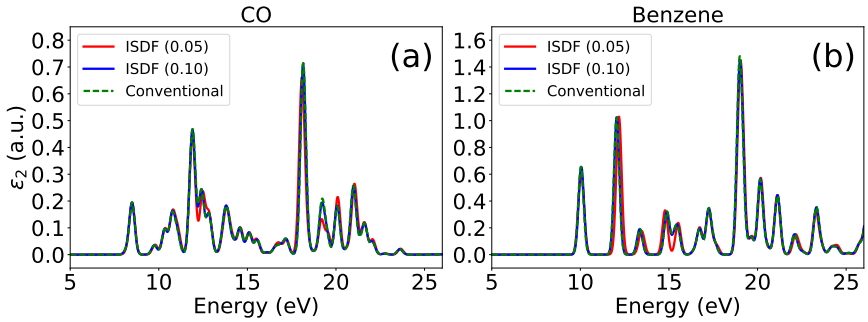

Fig. 5. Optical dielectric function (imaginary part ε_2) of (a) CO and (b) benzene molecules computed with the ISDF method (the rank ratio N_{cc}^t/N_{cc} is set to be 0.05 ($t = 3.0$) and 0.10 ($t = 6.0$)) compared to conventional BSE calculations in BerkeleyGW.

7.3 Optical absorption spectra

One important application of BSE is to compute the optical absorption spectrum, which is determined by optical dielectric function in Equation (14). Fig. 5 plots the optical absorption spectra for both CO and benzene obtained from approximate H_{BSE} constructed with the ISDF method and the H_{BSE} constructed in a conventional approach implemented in BerkeleyGW. When the rank truncation ratio N_{cc}^t/N_{cc} is set to be only 0.10 ($t = 6.0$), the absorption spectrum obtained from the ISDF approximate H_{BSE} is nearly indistinguishable from that produced from the conventional approach. When N_{cc}^t/N_{cc} is set to 0.05 ($t = 3.0$), the absorption spectrum obtained from ISDF approximate H_{BSE} still preserves the main features (peaks) of the absorption spectrum obtained in a conventional approach even though some of the peaks are slightly shifted, and the height of some peaks are slightly off.

8 Conclusion and outlook

In summary, we have demonstrated that the interpolative separable density fitting (ISDF) technique can be used to efficiently and accurately construct the

Bethe–Salpeter Hamiltonian matrix. The ISDF method allows us to reduce the complexity of the Hamiltonian construction from $\mathcal{O}(N_e^5)$ to $\mathcal{O}(N_e^3)$ with a much smaller pre-constant. We show that the ISDF based BSE calculations in molecules and solids can efficiently produce accurate exciton energies and optical absorption spectrum in molecules and solids.

In the future, we plan to replace the costly QRCP procedure with the centroidal Voronoi tessellation (CVT) method [6] for selecting the interpolation points in the ISDF method. The CVT method is expected to significantly reduce the computational cost for selecting interpolating point in the ISDF procedure for the BSE calculations.

The performance results reported here are based on a sequential implementation of the ISDF method. In the near future, we will implement a parallel version suitable for large-scale distributed memory parallel computers. Such an implementation will allow us to tackle much larger problems for which the favorable scaling of the ISDF approach is much more pronounced.

Acknowledgments

This work is partly supported by the Center for Computational Study of Excited-State Phenomena in Energy Materials (C2SEPEM) at the Lawrence Berkeley National Laboratory, which is funded by the U. S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05CH11231, as part of the Computational Materials Sciences Program, and by the Center for Applied Mathematics for Energy Research Applications (CAMERA) (L. L. and C. Y.). The authors thank the National Energy Research Scientific Computing (NERSC) center for making computational resources available.

References

1. T. Auckenthaler, V. Blum, H. J. Bungartz, T. Huckle, R. Johanni, L. Krämer, B. Lang, H. Lederer, and P. R. Willems. Parallel solution of partial symmetric eigenvalue problems from electronic structure calculations. *Parallel Comput.*, 37(12):783–794, 2011.
2. P. Benner, S. Dolgov, V. Khoromskaia, and B. N. Khoromskij. Fast iterative solution of the Bethe–Salpeter eigenvalue problem using low-rank and QTT tensor approximation. *J. Comput. Phys.*, 334:221–239, 2017.
3. J. Brabec, L. Lin, M. Shao, N. Govind, Y. Saad, C. Yang, and E. G. Ng. Efficient algorithms for estimating the absorption spectrum within linear response TDDFT. *J. Chem. Theory Comput.*, 11(11):5197–5208, 2015.
4. T. F. Chan and P. C. Hansen. Some applications of the rank revealing QR factorization. *SIAM J. Sci. Statist. Comput.*, 13:727–741, 1992.
5. J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie. BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Comput. Phys. Commun.*, 183(6):1269–1289, 2012.

6. K. Dong, W. Hu, and L. Lin. Interpolative separable density fitting through centroidal voronoi tessellation with applications to hybrid functional electronic structure calculations. *arXiv:1711.01531*, 2017.
7. P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter*, 21(39):395502, 2009.
8. S. Goedecker, M. Teter, and J. Hutter. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B*, 54:1703, 1996.
9. C. Hartwigsen, S. Goedecker, and J. Hutter. Relativistic separable dual-space gaussian pseudopotentials from H to Rn. *Phys. Rev. B*, 58:3641, 1998.
10. L. Hedin. New method for calculating the one-particle Green's function with application to the electron–gas problem. *Phys. Rev.*, 139:A796, 1965.
11. W. Hu, L. Lin, and C. Yang. Interpolative separable density fitting decomposition for accelerating hybrid density functional calculations with applications to defects in silicon. *J. Chem. Theory Comput.*, 13(11):5420–5431, 2017.
12. P. B. V. Khoromskaia and B. N. Khoromskij. A reduced basis approach for calculation of the BetheCSalpeter excitation energies by using low-rank tensor factorisations. *Mol. Phys.*, 114:1148–1161, 2016.
13. A. V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23(2):517–541, 2001.
14. C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards*, 45:255–282, 1950.
15. L. Lin, Z. Xu, and L. Ying. Adaptively compressed polarizability operator for accelerating large scale ab initio phonon calculations. *Multiscale Model. Simul.*, 15:29–55, 2017.
16. M. P. Ljungberg, P. Koval, F. Ferrari, D. Foerster, and D. Sánchez-Portal. Cubic-scaling iterative solution of the Bethe–Salpeter equation for finite systems. *Phys. Rev. B*, 92:075422, 2015.
17. J. Lu and K. Thicke. Cubic scaling algorithms for RPA correlation using interpolative separable density fitting. *J. Comput. Phys.*, 351:187–202, 2017.
18. J. Lu and L. Ying. Compression of the electron repulsion integral tensor in tensor hypercontraction format with cubic scaling cost. *J. Comput. Phys.*, 302:329–335, 2015.
19. M. Marsili, E. Mosconi, F. D. Angelis, and P. Umari. Large-scale GW-BSE calculations with N^3 scaling: Excitonic effects in dye-sensitized solar cells. *Phys. Rev. B*, 95:075415, 2017.
20. C. Møler and M. S. Plesset. Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618, 1934.
21. G. Onida, L. Reining, and A. Rubio. Electronic excitations: Density-functional versus many-body Green's-function approaches. *Rev. Mod. Phys.*, 74:601, 2002.
22. D. Rocca, D. Lu, and G. Galli. Ab initio calculations of optical absorption spectra: Solution of the BetheCSalpeter equation within density matrix perturbation theory. *J. Chem. Phys.*, 133:164109, 2010.

23. M. Rohlfing and S. G. Louie. Electron–hole excitations and optical spectra from first principles. *Phys. Rev. B*, 62:4927, 2000.
24. E. E. Salpeter and H. A. Bethe. A relativistic equation for bound-state problems. *Phys. Rev.*, 84:1232, 1951.
25. M. Shao, F. H. da Jornada, L. Lin, C. Yang, J. Deslippe, and S. G. Louie. A structure preserving Lanczos algorithm for computing the optical absorption spectrum. *SIAM J. Matrix. Anal. Appl.*, to appear.
26. M. Shao, F. H. da Jornada, C. Yang, J. Deslippe, and S. G. Louie. Structure preserving parallel algorithms for solving the BetheCSalpeter eigenvalue problem. *Linear Algebra Appl.*, 488:148–167, 2016.
27. M. Shao and C. Yang. BSEPACK user's guide, 2016. https://sites.google.com/a/lbl.gov/bsepack/.