Computer Physics Communications 240 (2019) 21–29

Contents lists available at ScienceDirect

Computer Physics Communications

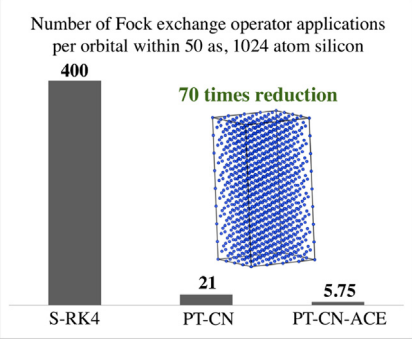
journal homepage: www.elsevier.com/locate/cpc

Fast real-time time-dependent hybrid functional calculations with the parallel transport gauge and the adaptively compressed exchange formulation

Weile Jia [a], Lin Lin [a,b,*]

[a] Department of Mathematics, University of California, Berkeley, CA 94720, United States
[b] Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 November 2018
Received in revised form 6 February 2019
Accepted 10 February 2019
Available online 19 February 2019

Keywords:
Time-dependent density functional theory
Real time evolution
Hybrid exchange-correlation functional
Parallel transport gauge
Adaptively compressed exchange

ABSTRACT

We present a new method to accelerate real-time time-dependent density functional theory (rt-TDDFT) calculations with hybrid exchange–correlation functionals. In the context of a large basis set such as planewaves and real space grids, the main computational bottleneck for large scale calculations is the application of the Fock exchange operator to the time-dependent orbitals. Our main goal is to reduce the frequency of applying the Fock exchange operator, without loss of accuracy. We achieve this by combining the recently developed parallel transport (PT) gauge formalism (Jia et al. J. Chem. Theory Comput. 2018) and the adaptively compressed exchange operator (ACE) formalism (Lin, J. Chem. Theory Comput. 2016). The PT gauge yields the slowest possible dynamics among all choices of gauge. When coupled with implicit time integrators such as the Crank–Nicolson (CN) scheme, the resulting PT–CN scheme can significantly increase the time step from sub-attoseconds to $10-100$ attoseconds. At each time step t_n, PT–CN requires the self-consistent solution of the orbitals at time t_{n+1}. We use ACE to delay the update of the Fock exchange operator in this nonlinear system, while maintaining the same self-consistent solution. We verify the performance of the resulting PT–CN–ACE method by computing the absorption spectrum of a benzene molecule and the response of bulk silicon systems to an ultrafast laser pulse, using the planewave basis set and the HSE exchange–correlation functional. We report the strong and weak scaling of the PT–CN–ACE method for silicon systems ranging from 32 to 1024 atoms, on a parallel computer with up to 2048 computational cores. Compared to standard explicit time integrators such as the *4th* order Runge–Kutta method (RK4), we find that the PT–CN–ACE can reduce the frequency of the Fock exchange operator application by nearly 70 times, and the thus

* Corresponding address: 1083 Evans Hall, Department of Mathematics, University of California, Berkeley, CA 94720, USA.
E-mail address: linlin@math.berkeley.edu (L. Lin).

https://doi.org/10.1016/j.cpc.2019.02.009

reduce the overall wall clock time by 46 times for the system with 1024 atoms. Hence our work enables hybrid functional rt-TDDFT calculations to be routinely performed with a large basis set for the first time.

1. Introduction

In generalized Kohn–Sham density functional theory [1,2], hybrid exchange–correlation functionals, such as B3LYP [3,4], PBE0 [5] and HSE [6,7], are known to be more reliable in producing high fidelity results for ground state electronic structure calculations for a vast range of systems [8,9]. With the recent developments of ultrafast laser techniques, a large number of excited state phenomena, such as nonlinear optical response [10] and the collision of an ion with a substrate [11], can be observed in real time. One of the most widely used techniques for studying such ultrafast properties is the real-time time-dependent density functional theory (rt-TDDFT) [12–17]. Hybrid functional rt-TDDFT calculations have been performed in the context of small basis sets such as Gaussian orbitals and atomic orbitals [18–20]. In the context of a large basis set such as planewaves and real space grids, most rt-TDDFT calculations so far are performed with local and semi-local exchange–correlation functionals, and hybrid functional calculations have only been performed for systems consisting of a handful of atoms [21]. This is because hybrid functionals include a fraction of the Fock exchange operator, which requires access to the diagonal as well as the off-diagonal elements of the density matrix. This leads to significant increase of the computational cost compared to calculations with local and semi-local functionals. The problem is compounded by the very small time step (on the order of attosecond or sub-attosecond) often needed in rt-TDDFT simulation. Hence to reach the femtosecond, let alone the picosecond timescale, the number of application of the Fock exchange operator can be prohibitively expensive for large systems.

This paper aims at enabling practical hybrid functional rt-TDDFT calculations to be performed with a large basis set. To this end, the primary goal is to reduce the number of matrix–vector multiplication operations involving the Fock exchange operator. At first glance, this is a rather difficult task, since the Fock exchange operator depends on the density matrix $P(t)$, which needs to be updated at each small time step. In order to overcome this difficulty, we first enlarge the time step of rt-TDDFT calculations via implicit time integrators. While implicit time integrators are often considered to be not sufficiently cost-effective when compared to explicit integrators for rt-TDDFT calculations [22–24], these studies are performed by direct propagation of the Kohn–Sham wavefunctions. One main reason is that the oscillation of the wavefunctions is faster than that of density matrix, and hence implicit integrators with a large time step are often stable but not accurate enough. We have recently identified that by optimizing the gauge, i.e. a unitary rotation matrix performing a linear combination of the wavefunctions, the oscillation of the wavefunctions can be significantly reduced. In particular, the parallel transport (PT) gauge [25,26] yields the *slowest dynamics* among all possible choices of gauge. Hence when the parallel transport dynamics is coupled with implicit time integrators, such as the Crank–Nicolson (CN) scheme, the resulting PT–CN scheme can significantly increase the time step with systematically controlled accuracy.

Implicit integrators such as PT–CN introduce a set of nonlinear equations that needs to be solved self-consistently at each time step going from t_n to t_{n+1}. This system can be viewed as a fixed point problem to determine the orbitals at t_{n+1}. In the context of ground state hybrid functional density functional theory calculations, we have recently developed the adaptively compressed exchange operator (ACE) formulation [27,28] to accelerate a fixed point problem introduced by a nonlinear eigenvalue problem. The ACE formulation can reduce the frequency of applying the Fock exchange operator without loss of accuracy, and can be used for insulators and metals. It has been incorporated into community software packages such as the Quantum ESPRESSO [29]. The idea of the adaptive compression has been rigorously analyzed in the context of linear eigenvalue problems [30], and can be extended to accelerate calculations in other contexts such as the density functional perturbation theory [31]. In this paper, we further extend the idea of ACE to accelerate hybrid functional rt-TDDFT calculations, by splitting the solution of the nonlinear system into two iteration loops. During each iteration of the outer loop, we only apply the Fock exchange operator once per orbital, and construct the adaptively compressed Fock exchange operator. In the inner loop, only the adaptively compressed Fock exchange operator will be used, and the application of the compressed operator only involves matrix–matrix multiplication operations, and is much cheaper than applying the Fock exchange operator. This two loop strategy further reduces the frequency of updating the Fock operator and hence the computational time.

The rest of the manuscript is organized as follows. We introduce the real-time time-dependent density functional theory with hybrid functional and parallel transport gauge in Sections 2 and 3, respectively. We present the adaptively compressed exchange operator formulation in Section 4. Numerical results are presented in Section 5, followed by a conclusion and discussion in Section 6.

2. Real-time time dependent functional theory with hybrid functional

rt-TDDFT solves the following set of time-dependent equations

$$\imath\partial_t\psi_i(t) = H(t, P(t))\psi_i, \quad i = 1, \ldots, N_e, \tag{1}$$

where N_e is the number of electrons (spin degeneracy omitted), and $\Psi(t) = [\psi_1(t), \ldots, \psi_{N_e}(t)]$ are the electron orbitals. The Hamiltonian takes the form

$$H(t, P(t)) = -\frac{1}{2}\Delta_{\mathbf{r}} + V_{\text{ext}}(t) + V_{\text{Hxc}}[P(t)] + V_{\text{X}}[P(t)]. \tag{2}$$

Here $V_{\text{ext}}(t)$ characterizes the electron–ion interaction, and the explicit dependence of the Hamiltonian on t is often due to the existence of an external field. The Hamiltonian also depends nonlinearly on the density matrix $P(t) = \Psi(t)\Psi^*(t)$. V_{Hxc} is a local operator, and characterizes the Hartree contribution and the local and the semi-local part of the exchange–correlation contribution. It depends only on the electron density $\rho(t) = \sum_{i=1}^{N_e} |\psi_i(t)|^2$, which are given by the diagonal matrix elements of the density matrix $P(t)$ in the real space representation. The Fock exchange operator V_{X} is an integral operator with kernel

$$V_{\text{X}}[P](\mathbf{r}, \mathbf{r}') = -\alpha P(\mathbf{r}, \mathbf{r}')K(\mathbf{r}, \mathbf{r}'). \tag{3}$$

Here $K(\mathbf{r}, \mathbf{r}')$ is the kernel for the electron–electron interaction, and $0 < \alpha < 1$ is a mixing fraction. For example, in the Hartree–Fock theory, $K(\mathbf{r}, \mathbf{r}') = 1/|\mathbf{r} - \mathbf{r}'|$ is the Coulomb operator and $\alpha = 1$. In screened exchange theories [6], K can be a screened Coulomb operator with kernel $K(\mathbf{r}, \mathbf{r}') = \text{erfc}(\mu|\mathbf{r} - \mathbf{r}'|)/|\mathbf{r} - \mathbf{r}'|$, and typically $\alpha \sim 0.25$.

When a large basis set is used, it is prohibitively expensive to explicitly construct $V_X[P]$, and it is only viable to apply it to a vector $\varphi(\mathbf{r})$ as

$$(V_X[P]\varphi)(\mathbf{r}) = -\alpha \sum_{i=1}^{N_e} \psi_i(\mathbf{r},t) \int K(\mathbf{r},\mathbf{r}')\psi_i^*(\mathbf{r}',t)\varphi(\mathbf{r}')\,\mathrm{d}\mathbf{r}'. \tag{4}$$

This amounts to solving N_e^2 Poisson-like problems with FFT, and the computational cost is $\mathcal{O}(N_g \log(N_g) N_e^2)$, where the N_g is the number of points in the FFT grid. This cost is asymptotically comparable to other matrix operations such as the QR factorization for orthogonalizing the Kohn–Sham orbitals which scales as $\mathcal{O}(N_g N_e^2)$, but the $\log(N_g)$ prefactor is significantly larger.

In order to propagate Eq. (1) from an initial set of orthonormal orbitals $\Psi(0)$, we may use for instance, the standard explicit 4th order Runge–Kutta scheme (S-RK4):

$$\begin{aligned}
k_1 &= -\imath\Delta t H_n \Psi_n,\\
\Psi_n^{(1)} &= \Psi_n + \frac{1}{2}k_1, \quad H_n^{(1)} = H(t_{n+\frac{1}{2}}, \Psi_n^{(1)}\Psi_n^{(1)*})\\
k_2 &= -\imath\Delta t H_n^{(1)} \Psi_n^{(1)},\\
\Psi_n^{(2)} &= \Psi_n + \frac{1}{2}k_2, \quad H_n^{(2)} = H(t_{n+\frac{1}{2}}, \Psi_n^{(2)}\Psi_n^{(2)*})\\
k_3 &= -\imath\Delta t H_n^{(2)} \Psi_n^{(2)},\\
\Psi_n^{(3)} &= \Psi_n + k_3, \quad H_n^{(3)} = H(t_{n+1}, \Psi_n^{(3)}\Psi_n^{(3)*})\\
k_4 &= -\imath\Delta t H_n^{(3)} \Psi_n^{(3)},\\
\Psi_{n+1} &= \Psi_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).
\end{aligned} \tag{5}$$

Here all the $H_n = H(t_n, P_n)$ is the Hamiltonian at step t_n, and $t_{n+\frac{1}{2}} = t_n + \frac{1}{2}\Delta t$, $t_{n+1} = t_n + \Delta t$. For each time step, the Hamiltonian operator needs to be applied 4 times to each of the N_e orbitals. After each update of the orbitals, the Hamiltonian operator needs to be updated accordingly. The maximal time step allowed by the RK4 integrator (and in general, all explicit time integrators) is bounded by $c\,\|H\|_2^{-1}$, where c is a scheme dependent constant and $\|H\|_2$ is the spectral radius of the Hamiltonian operator. In practice, this maximal time step is often less than 1 attosecond (as). Hence simulating rt-TDDFT for 1 fs would require more than $4000N_e$ matrix–vector multiplications involving the Hamiltonian operator (and hence the Fock exchange operator). When the nuclei degrees of freedom are also time-dependent such as in the case of the Ehrenfest dynamics, the electron–nuclei potentials, such as the local and nonlocal components of the pseudopotential, need also be updated more than 4000 times per 1 fs simulation.

3. Parallel transport gauge

In order to accelerate rt-TDDFT calculations with hybrid functionals, it is necessary to relax the constraint on the maximal time step that can be utilized in the simulation. This can be achieved by the recently developed parallel transport gauge formalism [25,26], which we briefly summarize below.

First, note that Eq. (1) can be equivalently written using a set of transformed orbitals $\Phi(t) = \Psi(t)U(t)$, where the gauge matrix $U(t)$ is a unitary matrix of size N_e. An important property of the density matrix is that it is gauge-invariant: $P(t) = \Psi(t)\Psi^*(t) = \Phi(t)\Phi^*(t)$. Physical observables such as energies and dipoles are defined using the density matrix instead of the orbitals. The density matrix and the derived physical observables can often oscillate at a slower rate than the orbitals, and hence can be discretized with a larger time step. Our goal is to optimize the gauge matrix, so that the transformed orbitals $\Phi(t)$ vary *as slowly as possible*, without altering the density matrix. This results in the following variational problem

$$\min_{U(t)} \quad \left\|\dot{\Phi}\right\|_F^2, \ \text{s.t.}\ \Phi(t) = \Psi(t)U(t),\ U^*(t)U(t) = I_{N_e}. \tag{6}$$

Here $\left\|\dot{\Phi}\right\|_F^2 := \mathrm{Tr}[\dot{\Phi}^*\dot{\Phi}]$ measures the Frobenius norm of the time derivative of the transformed orbitals. The minimizer of (6), in terms of Φ, satisfies

$$P\dot{\Phi} = 0. \tag{7}$$

Eq. (7) implicitly defines a gauge choice for each $U(t)$, and this gauge is called the *parallel transport gauge*. The governing equation of each transformed orbital φ_i can be concisely written down as

$$\imath\partial_t\varphi_i = H\varphi_i - \sum_{j=1}^{N_e} \varphi_j \langle\varphi_j|H|\varphi_i\rangle, \quad i = 1,\ldots,N_e, \tag{8}$$

or more concisely in the matrix form

$$\imath\partial_t\Phi = H\Phi - \Phi(\Phi^*H\Phi), \quad P(t) = \Phi(t)\Phi^*(t). \tag{9}$$

The right hand side of Eq. (9) is analogous to the residual vectors of an eigenvalue problem in the time-independent setup. Hence $\Phi(t)$ follows the dynamics driven by residual vectors and is expected to vary slower than $\Psi(t)$. We refer to the dynamics (9) as the parallel transport (PT) dynamics, and correspondingly Eq. (1) in the standard Schrödinger representation as the Schrödinger dynamics.

The PT dynamics only introduces one additional term, and hence can be readily discretized with any propagator. The standard explicit 4th order Runge–Kutta scheme for the parallel transport dynamics (PT-RK4) now becomes

$$\begin{aligned}
k_1 &= -\imath\Delta t\{H_n\Phi_n - \Phi_n(\Phi_n^*H_n\Phi_n)\},\\
\Phi_n^{(1)} &= \Phi_n + \frac{1}{2}k_1, \quad H_n^{(1)} = H(t_{n+\frac{1}{2}}, \Phi_n^{(1)}\Phi_n^{(1)*})\\
k_2 &= -\imath\Delta t\{H_n^{(1)}\Phi_n^{(1)} - \Phi_n^{(1)}(\Phi_n^{(1)*}H_n^{(1)}\Phi_n^{(1)})\},\\
\Phi_n^{(2)} &= \Phi_n + \frac{1}{2}k_2, \quad H_n^{(2)} = H(t_{n+\frac{1}{2}}, \Phi_n^{(2)}\Phi_n^{(2)*})\\
k_3 &= -\imath\Delta t\{H_n^{(2)}\Phi_n^{(2)} - \Phi_n^{(2)}(\Phi_n^{(2)*}H_n^{(2)}\Phi_n^{(2)})\},\\
\Phi_n^{(3)} &= \Phi_n + k_3, \quad H_n^{(3)} = H(t_{n+1}, \Phi_n^{(3)}\Phi_n^{(3)*})\\
k_4 &= -\imath\Delta t\{H_n^{(3)}\Phi_n^{(3)} - \Phi_n^{(3)}(\Phi_n^{(3)*}H_n^{(3)}\Phi_n^{(3)})\},\\
\Phi_{n+1} &= \Phi_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).
\end{aligned} \tag{10}$$

Compared to the Schrödinger dynamics, the main benefit of the PT dynamics is that the orbitals vary at a slower rate, and hence can be accurately extrapolated over a larger time interval. However, the slower dynamics does not automatically translate to a larger time step. When explicit integrators such as the RK4 scheme are used, due to the small region of absolute stability, one may still need to use a small time step even for the PT dynamics. This is particularly the case if the spectral radius $\|H\|_2$ is large. On the other hand, implicit time integrators often have much larger region of absolute stability. Hence combining the PT dynamics with implicit integrators can significant enlarge the time step while still maintaining sufficient numerical accuracy. For instance, the Crank–Nicolson scheme for the Schrödinger dynamics (S-CN) is

$$\left(I + \imath\frac{\Delta t}{2}H_{n+1}\right)\Psi_{n+1} = \left(I - \imath\frac{\Delta t}{2}H_n\right)\Psi_n, \tag{11}$$

while the Crank–Nicolson scheme for the parallel transport dynamics (PT–CN) is

$$\begin{aligned}&\Phi_{n+1}+\imath\frac{\Delta t}{2}\left\{H_{n+1}\Phi_{n+1}-\Phi_{n+1}\left(\Phi_{n+1}^{*}H_{n+1}\Phi_{n+1}\right)\right\}\\&=\Phi_n-\imath\frac{\Delta t}{2}\left\{H_n\Phi_n-\Phi_n\left(\Phi_n^{*}H_n\Phi_n\right)\right\}.\end{aligned}\tag{12}$$

The Crank–Nicolson scheme uses the trapezoidal rule for time integration, which is an A-stable scheme and allows the usage of large time steps. When implicit time integrators are used, Φ_{n+1} needs to be solved self-consistently, which can be efficiently solved by mixing schemes such as the Anderson method [32]. Numerical results indicate that the size of the time step for the PT–CN scheme can be $10 \sim 100$ as, and is significantly larger than that of standard explicit time integrators.

We also remark that the computational complexity of standard rt-TDDFT calculations may achieve $\mathcal{O}(N_e^2)$ scaling [16,33,34], assuming (1) local and semi-local exchange–correlation functionals and certain explicit time integrators are used, and (2) no orbital re-orthogonalization step is needed throughout the simulation. The PT dynamics requires the evaluation of the term $\Phi(\Phi^{*}H\Phi)$ in Eq. (9), which scales cubically with respect to the system size. We have demonstrated that the cross over point between the quadratic and cubic scaling algorithms should occur for systems with thousands of atoms [26]. In the current context of hybrid functional rt-TDDFT calculations, both methods scale cubically with respect to the system size due to the dominating cost associated with the Fock exchange operator. Numerical results indicate that the advantage of the PT formulation becomes even more evident in this case.

4. Adaptively compressed exchange operator formulation

For hybrid functional rt-TDDFT calculations, the use of the parallel transport gauge and implicit time integrators still requires a relatively large number of matrix–vector multiplication operations involving the Fock exchange operator. For instance, when a relatively large time step is used, the number of self-consistent iterations in each PT time step may become $20 \sim 40$. This gives room for further reduction of the cost associated with the Fock exchange operator, using the recently developed adaptively compressed exchange (ACE) operator formulation [27].

In ground state hybrid functional DFT calculations, every time when the Fock exchange operator is applied to a set of orbitals, we store the resulting vectors as

$$W_i(\mathbf{r}) = (V_X[P]\varphi_i)(\mathbf{r}) \quad i = 1, \ldots, N_e. \tag{13}$$

Here we assume that the Fock exchange operator is defined with respect to the density matrix P, which is often also specified by the orbitals $\{\varphi_i\}_{i=1}^{N_e}$. The vectors $\{W_i\}_{i=1}^{N_e}$ are then used to construct a *surrogate* operator, or the adaptively compressed exchange operator denoted by $\widetilde{V}_X$. We require that $\widetilde{V}_X$ should satisfy the following consistency conditions

$$(\widetilde{V}_X\varphi_i)(\mathbf{r}) = W_i(\mathbf{r}) \quad \text{and} \quad \widetilde{V}_X(\mathbf{r},\mathbf{r}') = \widetilde{V}_X^{*}(\mathbf{r}',\mathbf{r}) \tag{14}$$

The conditions (14) do not yet uniquely determine $\widetilde{V}_X$. However, the choice becomes unique if we require $\widetilde{V}_X$ to be strictly of rank N_e [30], and it can be computed as follows. We first construct the overlap matrix

$$M_{ij} = \int \varphi_i^{*}(\mathbf{r})W_j(\mathbf{r})\,\mathrm{d}\mathbf{r}, \quad i,j = 1, \ldots, N_e, \tag{15}$$

which is Hermitian and negative definite. We perform the Cholesky factorization for $-M$, i.e. $M = -LL^{*}$, where L is a lower triangular matrix. Then the adaptively compressed exchange operator is given by the following rank N_e decomposition

$$\widetilde{V}_X(\mathbf{r},\mathbf{r}') = -\sum_{k=1}^{N_e}\xi_k(\mathbf{r})\xi_k^{*}(\mathbf{r}'), \tag{16}$$

where $\{\xi_k\}_{k=1}^{N_e}$ are called *projection vectors*, and are defined as

$$\xi_k(\mathbf{r}) = \sum_{i=1}^{N_e} W_i(\mathbf{r})(L^{-*})_{ik}. \tag{17}$$

The cost for applying $\widetilde{V}_X$ to a number of vectors only involves matrix–matrix multiplications up to size $N_g \times N_e$, which can be efficiently carried out in the sequential or parallel settings. The preconstant of this step is also significantly smaller than that for applying the Fock exchange operator. When self-consistency is reached, $\widetilde{V}_X$ agrees with the true Fock exchange operator when applied to the occupied orbitals, thanks to the consistency condition (14). We have also proved that for linear eigenvalue problems, the ACE formulation can converge globally from almost everywhere, with local convergence rate favorable compared to standard iterative methods [30].

In order to utilize the ACE formulation in the context of rt-TDDFT calculations, we note that the PT–CN scheme requires the solution of a fixed point problem (12) for Φ_{n+1}. Hence we may artificially separate the fixed point problem into two iteration loops. In the outer iteration, we apply the Fock exchange operator to the parallel transport orbitals Φ_{n+1} only once, which gives rise to $\widetilde{V}_X$ defined by the procedure above. Then in the inner iteration, we perform a few inner iterations and only update the density-dependent component of the Hamiltonian operator, while replacing the Fock exchange operator by the same $\widetilde{V}_X$ operator. This inner iteration step can also be seen as a relatively inexpensive preconditioner for accelerating the convergence of the self-consistent iteration. Then we perform the outer iteration until the density matrix (monitored by e.g. the Fock exchange energy) converges. We summarize the resulting PT–CN–ACE algorithm in Alg. 1, which propagates the orbitals from the time step t_n to t_{n+1}.

Algorithm 1 One step of propagation of the PT–CN–ACE method.

1: Evaluate the right hand side of Eq. (12), and choose an initial guess for Φ_{n+1} (the simplest choice being $\Phi_{n+1} = \Phi_n$).
2: **while** Fock exchange energy is not converged **do**
3: Applying the Fock exchange operator to Φ_{n+1}, and construct $\widetilde{V}_X$ in its low rank form.
4: **while** electron density ρ is not converged **do**
5: Iteratively update Φ_{n+1} and H_{n+1} using Eq. (12), with V_X replaced by $\widetilde{V}_X$.
6: **end while**
7: **end while**

5. Numerical results

The S-RK4, PT–CN and PT–CN–ACE methods are implemented in the PWDFT package (based on the plane wave discretization and the pseudopotential method), which is an independent module of the massively parallel software package DGDFT (Discontinuous Galerkin Density Functional Theory) [35,36]. PWDFT performs parallelization primarily along the orbital direction, and can scale up to several thousands of CPU cores for systems up to thousands of atoms. We use the SG15 Optimized Norm-Conserving Vanderbilt (ONCV) pseudopotentials [37,38] and HSE06 functionals [7] in all the following tests. We remark that the SG15 pseudopotentials are obtained from all electron calculations using the PBE functional, which introduces certain amount of inconsistency in the hybrid functional calculation. The calculations are performed on

the Edison supercomputer at National Energy Research Scientific Computing Center (NERSC). Each Edison node is equipped with two Intel Ivy Bridge sockets with 24 cores in total and 64 gigabyte (GB) of memory. Our code uses MPI only and the number of cores used is always equal to the number of MPI processes.

In large scale hybrid functional TDDFT calculations, the application of the Fock exchange operator dominates the total computational costs. Hence we use the number of matrix–vector multiplications per orbital involving the Fock exchange operator as a metric for the efficiency of a method, and this metric is relatively independent of implementation. In the case of PT–CN–ACE, this number is equal to the number of times for which the ACE operator needs to be constructed. We also present the total wall clock time as well as the breakdown of the computational time to properly take into account contributions from other components, especially those exclusive due to the usage of the PT–CN–ACE scheme.

We first demonstrate the accuracy of PT–CN–ACE by computing the absorption spectrum of the benzene molecule. The size of the cubic supercell is 10.58 Å along each direction, and the kinetic energy cutoff is set to 544 eV. The dimension of the Hamiltonian matrix is 74088. A δ kick is applied in the x direction to calculate the partial absorption spectrum. The length of the rt-TDDFT calculation is 24 fs. The size of the time step of PT–CN–ACE is set to be 12 as, and S-RK4 becomes unstable when the step size is bigger than 0.97 as. The absorption spectrum obtained by the S-RK4 and PT–CN–ACE methods is shown in Fig. 1(b). We also provide benchmark results obtained from the linear response time-dependent density functional theory (LR-TDDFT) calculation using the turboTDDFT module [39] from the Quantum ESPRESSO software package (QE) [29], which performs 3000 Lanczos steps along the x direction to evaluate the component of the polarization tensor. Both QE and PWDFT use the same pseudopotential and kinetic energy cutoff, and no empty state is used in calculating the spectrum in PWDFT. A Lorentzian smearing of 0.27 eV is applied to all calculations. We find that the shapes of the absorption spectrum calculated from three methods agree very well. The S-RK4 method requires 4 Fock exchange operator calculations per time step, while the PT–CN–ACE method only requires on average 3.2 ACE operator constructions in each time step. Thus for this example, a total number of 98968 and 6400 Fock exchange operator applications per orbital are calculated for the S-RK4 and PT–CN–ACE method, respectively. This means that the PT–CN–ACE method is about 15 times faster than the S-RK4 method in terms of the application of the Fock exchange operator. In the simulation, both PT–CN–ACE and S-RK4 use 15 CPU cores, and the total wall clock time is 7.5 h and 40.8 h, respectively. The reduction of the speedup factor compared to the theoretical estimate based on the number of Fock exchange operator applications is mainly due to the relatively small system size. Hence components such as the evaluation of the Hartree potential, and the inner loop for solving the fixed point problem in the PT–CN–ACE scheme still consume a relatively large portion of the computational time.

In the second example, we study the response of a silicon system to an ultra-fast laser pulse. The supercell consists of 32 atoms, and is constructed from $2 \times 2 \times 1$ unit cells sampled at the Γ point. Each simple cubic unit cell has 8 silicon atoms, and the lattice constant is 5.43 Å. The kinetic energy cutoff is set to 272 eV. For the system with 32 atoms, the dimension of the Hamiltonian matrix is 37044. The laser is applied along the x direction, and generates an electric field of the form

$$\mathbf{E}(t) = \hat{\mathbf{x}} E_{\max} \exp\left[-\frac{(t-t_0)^2}{2a^2}\right] \sin[\omega(t-t_0)]. \tag{18}$$

Here $a = 2.55$ fs, $\hbar\omega = 3.26$ eV, $t_0 = 15$ fs, which corresponds to a laser that peaks at 15 fs with its wavelength being 380 nm. The electric field amplitude $E_{\max}$ is 1.0 eV/ Å in the simulation. The profile of this external field is given in Fig. 2(a). The ground state band gap computed at $t = 0$ is around 1.3 eV using the HSE06 functional with a supercell containing 32 atoms. We acknowledge that the form of the electric field as in Eq. (18) is inconsistent with the periodic boundary condition imposed on the system. Hence quantities such as the dipole moment should in principle be properly defined through the Berry phase formulation [40,41] or through the usage of a vector potential [17]. Our treatment here amounts to treating the silicon system sampled at the Γ point as a large molecule. Although the value of the dipole moments neglects the contribution from the Berry phase, the treatment is consistent among different choices of numerical schemes, and the results are sufficient to demonstrate the efficiency and *numerical* accuracy of the PT–CN–ACE algorithm. Proper treatment of electric fields in the presence of periodic boundary conditions do not change the algorithmic structure, and will be in our future work.

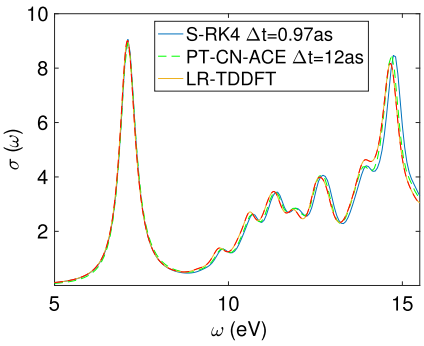


Fig. 1. Partial absorption spectrum of benzene along the x direction evaluated using the HSE06 functional.

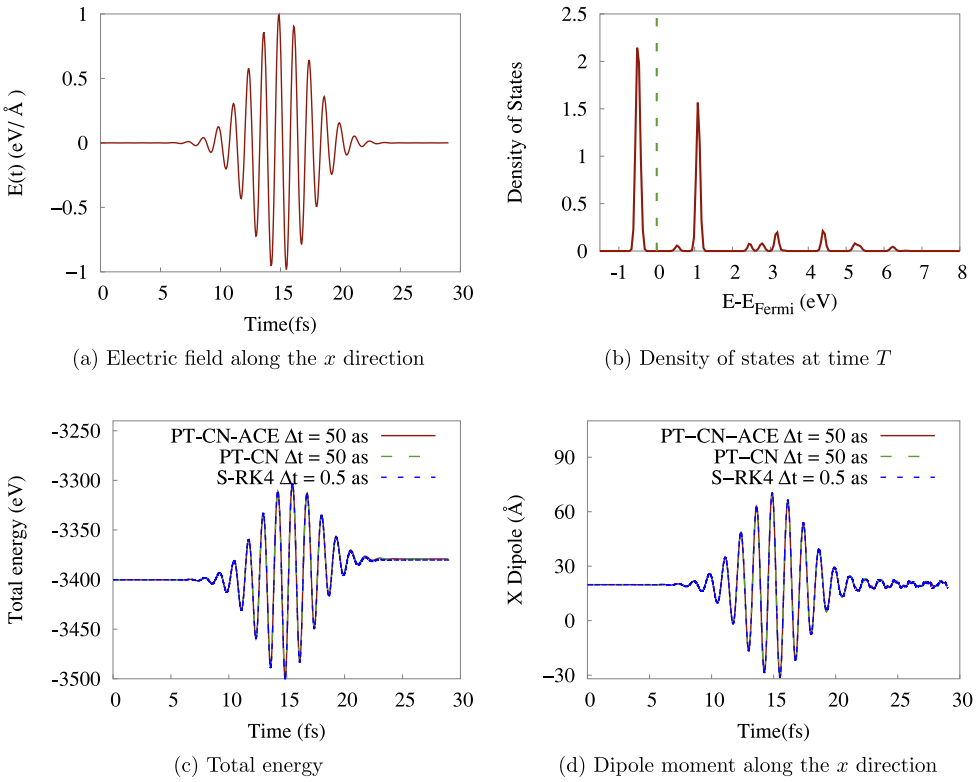
The total simulation length is 29 fs. In the PT–CN–ACE method, for the inner loop, the stopping criterion for the relative error of the electron density is set to 10^{-6}. The stopping criterion for the outer loop is defined via the relative error of the Fock exchange energy, and is set to 10^{-8}. The stopping criterion for the PT–CN method is defined via the relative error of the electron density and is set to 10^{-6}.

In order to demonstrate the electron excitation process, we plot the density of states at the end of the simulation (Fig. 2(b), the green dotted line indicates the Fermi energy), defined as

$$\rho(\varepsilon) := \sum_{j=1}^{N_e} \sum_{i=1}^{\infty} |\langle \psi_i(0) | \varphi_j(T) \rangle|^2 \widetilde{\delta}(\varepsilon - \varepsilon_i(0)).$$

Here $\varphi_j(T)$ is the jth orbital obtained at the end of the TDDFT simulation at time T, and $\varepsilon_i(0)$, $\psi_i(0)$ are the eigenvalues and wavefunctions corresponding to the ground state Hamiltonian at the initial time. $\widetilde{\delta}$ is a Dirac-δ function with a Gaussian broadening of 0.05 eV.

Fig. 2(c), (d) show the total energy and the dipole moment along the x direction calculated with the PT–CN, PT–CN–ACE and S-RK4 methods. The time steps of both PT–CN and PT–CN–ACE are set to 50 as, while the time step of S-RK4 is set to 0.5 as due to stability reason. Both the energy and the dipole moment agree very well among the results obtained from these three methods, indicating that the use of the compressed exchange operator in PT–CN–ACE does not lead to any loss of accuracy. The total energy and dipole moment obtained from S-RK4 and PT–CN–ACE are nearly indistinguishable before 15 fs, and become slightly different after 20 fs. Such error can be systematically reduced by using smaller time steps. Table 1 shows that the error can be reduced to as small as 0.016 meV/atom when the time step is reduced to 2.4 as for



(a) Electric field along the x direction

(b) Density of states at time T

(c) Total energy

(d) Dipole moment along the x direction

Fig. 2. Electron dynamics of a 32 atom silicon system under a laser pulse.

PT–CN–ACE method. Nearly the same accuracy is also observed in the PT–CN method when reducing the time step. We also listed the speedup factors in terms of both the number of Fock exchange operator applications per orbital per time step (FOC) and the total wall clock time (Wtime) in Table 1. The Wtime speedup is denoted in "Speedup" in Table 1. In comparison, the Fock exchange operator application speedup, which is denoted as "Speedup*" in Table 1, is calculated by counting the number of Fock exchange operator application for a given time period Δt. Note that the speedup factor obtained from the number of Fock exchange operator applications is relatively bigger than the wall clock time speedup, especially for PT–CN–ACE. This is mainly because the inner loop calculation in PT–CN–ACE still takes a big proportion of the computational time for this relatively small system. It is also why PT–CN can be faster than PT–CN–ACE in terms of the wall clock time despite the fact that PT–CN requires more Fock exchange operator applications per orbital. However, as the system size increases, the cost due to the Fock exchange operator becomes dominant, and we shall observe that PT–CN–ACE becomes more advantageous below.

Next we systematically investigate the efficiency of the PT–CN and PT–CN–ACE schemes by increasing the size of the supercell from $2 \times 2 \times 1$ to $8 \times 4 \times 4$ unit cells, and the set of systems consists of 32 to 1024 silicon atoms. All other physical parameters remain the same as in the tests above. We report the total wall clock time of the PT–CN, PT–CN–ACE and S-RK4 methods, as well as the breakdown into different components time of for a time period of Δt = 50 as. We report the performance in terms of both the weak scaling and the strong scaling. The time step of PT–CN and PT–CN–ACE is set to be 50 as and the time step for S-RK4 is 0.5 as. The average number of Fock exchange operator applications per orbital for PT–CN and PT–CN–ACE is 21 and 5.75, respectively. For the PT–CN–ACE method, the average number of inner iterations is 24.

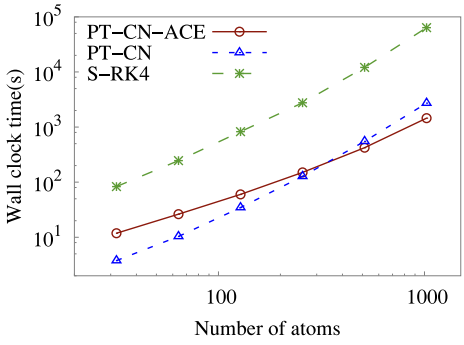
The total wall clock time of PT–CN–ACE can be divided into four parts: "ACE operator", which stands for the time used for applying the Fock exchange operator and constructing the ACE operator implicitly; "HPSI", which represents the time for the $H\psi$ calculation, with the application of the exchange operator replaced by the application of the ACE operator via a matrix–vector multiplication operation; "PT-CN-ACE: Exclusive", which includes the time exclusively associated with the usage of the parallel transport gauge, such as the orbital mixing and orbital orthogonalization; and "Others", which includes all other parts that are shared among the three methods, such as the evaluation of the Hartree potential and the total energy. Similarly, the total wall clock time of PT–CN is decomposed into three parts: "HPSI", "PT-CN: Exclusive" and "Others". The total wall clock time of S-RK4 is divided into "HPSI" and "Others". Note that "HPSI" and the evaluation of the total energy in PT–CN and S-RK4 require the application of the true Fock exchange operator.

PWDFT is mainly parallelized along the orbital direction, i.e. the maximum number of cores is equal to the number of occupied orbitals. The application of the Fock exchange operator to the occupied orbitals is implemented using the fast Fourier transformation(FFT). In the "HPSI" component of the PT–CN–ACE method, the matrix–matrix multiplication between the low rank operator $\widetilde{V}_X$ and all the occupied orbitals is performed to evaluate the Fock exchange term. We remark that certain components of PWDFT, such as the solution of the Hartree potential, are currently carried out on a single core. This is consistent with the choice of parallelization along the orbital direction, where each $H\psi$ (except the application of the Fock exchange operator) is carried out on a single core. However, as will be shown below, PT–CN–ACE and S-RK4 typically require many more Hartree potential evaluation compared to PT–CN. Hence PT–CN has some advantage in terms of the wall clock time from this perspective.

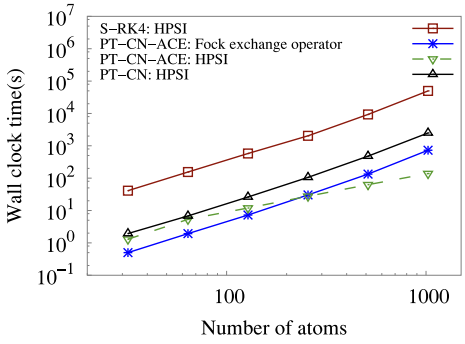
Fig. 3(a) shows the total wall clock time with respect to the system size N_{atom} for all three methods. In these tests the number of CPU cores used is always proportional to the number of atoms, i.e. $2 \times N_{atom}$ (this is called "weak scaling"). The speedup of PT–CN–ACE over S-RK4 is 7 times at 32 atoms, and increases to 46

Table 1
Accuracy and efficiency of PT–CN and PT–CN–ACE for the electron dynamics with the 380 nm laser compared to S-RK4. The accuracy is measured using the average energy increase per atom (AEI) after 29.0 fs and the average energy difference (AED) per atom for PT–CN and PT–CN–ACE compared with S-RK4 method after 29.0 fs. The efficiency is measured using the average number of Fock exchange operator applications per orbital in each time step (FOC) during the time interval from 0.0 fs to 29.0 fs, and the Fock exchange operator application speedup is denoted as "Speedup*". The total wall clock time (Wtime) and the corresponding speedup factor are also listed.

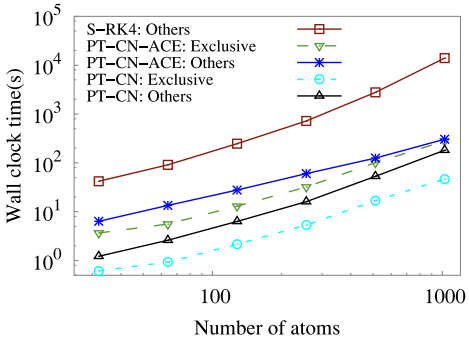
Method	Δt (as)	AEI (meV)	AED (meV)	FOC	Speedup*	Wtime (h)	Speedup
S-RK4	0.5	621.4156	/	4.0	1.0	18.09	1.0
PT–CN	0.5	621.4117	0.004	–	–	–	–
PT–CN	2.4	621.4062	0.009	4.8	4	6.0	3.0
PT–CN	5.1	621.4688	0.053	5.3	7.7	3.65	5
PT–CN	12.1	623.4656	2.05	5.9	16.4	1.45	12.4
PT–CN	25.0	628.8594	7.44	10.8	18.5	1.08	16.7
PT–CN	50.0	657.6688	36.25	21	19	1.12	16.1
PT–CN–ACE	0.5	621.4077	0.008	–	–	–	–
PT–CN–ACE	2.4	621.4313	0.016	2.32	8.3	4.99	3.7
PT–CN–ACE	5.1	621.4937	0.08	2.77	14.7	3.11	5.8
PT–CN–ACE	12.1	623.5781	2.2	3.02	32.0	1.6	11.3
PT–CN–ACE	25.0	628.9531	7.5	3.78	52.9	1.45	12.5
PT–CN–ACE	50.0	657.725	36.3	5.8	69.0	1.38	13.1



(a) Total wall clock time.



(b) Breakdown of the wall clock time, 1



(c) Breakdown of the wall clock time, 2

Fig. 3. Wall clock time for simulating silicon systems from 32 atoms up to 1024 atoms for $\Delta t = 50$ as. The number of CPU cores used is set to $2 \times N_{atom}$ in all tests. The systems are driven by the laser field shown at Fig. 2(a).

times at 1024 atoms. On the other hand, the speedup of PT–CN over S-RK4 is between 22 and 23 times, which is consistent with the Fock exchange operator applications speedup as shown in Table 1. For small systems, PT–CN is the most efficient method. When the system size increases beyond 256 atoms, PT–CN–ACE becomes the most efficient one in terms of the wall clock time. This cross-over can be explained by the breakdown of the total time shown in Fig. 3(b) (c). Since PT–CN–ACE method introduces a nested loop to reduce the number of Fock exchange operator applications, it also increases the number of inner iterations. More specifically, in the tests above, the number of Fock exchange operator applications per orbital is 6, but the number of inner iterations is 120 in PT–CN–ACE. In comparison, PT–CN only requires 21 inner iterations. Thus PT–CN is faster than PT–CN–ACE at small system size as shown in Fig. 3(a). However, as system size becomes larger, the Fock exchange operator applications will dominate the cost, and PT–CN–ACE becomes faster than PT–CN as shown in Fig. 3(a).

More specifically, Fig. 3(b) shows that "HPSI" takes 49 to 78 percent of total wall clock time from 32 to 1024 atom system for the S-RK4 method. For the PT–CN method, "HPSI" costs 51 percent of the time for the system with 32 atoms, and this becomes 91 percent when the system size increases to 1024 atoms. For the PT–CN–ACE method, the cost involving the Fock exchange operator is reduced to only 4 percent of the total time for the system with 32 atoms, and becomes 53 percent for the system with 1024 atoms.

Finally we report in Fig. 4 the average wall clock time for carrying out a simulation of 50 as for the 1024 atom silicon system, with respect to the increase of the number of computational cores (this is called "strong scaling"). Compared to performance using 32 cores, the parallel efficiency of a single TDDFT step with 2048

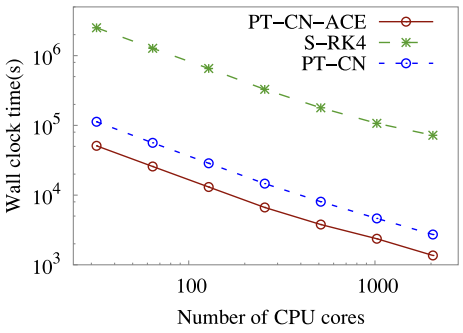


Fig. 4. Total wall clock time for 1024 atom silicon system from 32 up to 2048 CPU cores used in $\Delta t = 50$ as. The system is driven by laser field shown at Fig. 2(a).

cores reaches 54 percent, 58 percent and 64 percent for the S-RK4, PT–CN–ACE and PT–CN methods, respectively. The reduction of the parallel efficiency is mainly caused by our sequential implementation of certain components, such as the evaluation of the Hartree potential. The speedup of PT–CN–ACE method over S-RK4 is between 46 times and 50 times over the entire range. Therefore in order to finish the electron dynamics simulation above of 29 fs, it will take about 1 year using S-RK4, and this is reduced to around one week using PT–CN–ACE. Such a simulation is by all means still expensive, but starts to become feasible to be routinely performed.

6. Conclusion

In order to accelerate large scale hybrid functional rt-TDDFT calculations, we have presented a method to combine two recently developed ideas: parallel transport (PT) gauge and adaptively compressed exchange (ACE) operator. The overall goal is to reduce the frequency for the application of the Fock exchange operator to orbitals, with systematically controlled accuracy. We demonstrate that the resulting PT–CN–ACE scheme can indeed reduce the number of Fock exchange operator applications per unit time by one to two orders of magnitude compared to the standard explicit 4th order Runge–Kutta time integrator, and thus enables hybrid functional rt-TDDFT calculations for systems up to 1000 atoms.

Compared to the PT–CN scheme, the extra reduction of the number of applications of the Fock exchange operator requires more iterations in the inner loop. This is consistent with the observation for ground state hybrid functional calculations [42]. Hence in our implementation, PT–CN is in fact faster than PT–CN–ACE in terms of wall clock time for small systems. The precise crossover point depends heavily on the cost for solving the Poisson-like equation to apply the Fock exchange operator. For instance, we expect that the PT–CN–ACE becomes advantageous at a much earlier stage in real space rt-TDDFT software packages, where the solution of a Poisson-like equation can be much more expensive than that in a planewave based code. On the other hand, if the application of the Fock exchange operator can be accelerated using techniques such as localization [43,44], density fitting [45–47], or through the GPU architecture, we expect that the original PT–CN scheme will be more favorable. Further developments to reduce the number of inner iterations without penalizing the number of Fock exchange operator applications, such as via the usage of better preconditioners, is also an interesting direction for future works.

Acknowledgments

This work was partially supported by the National Science Foundation under Grant No. 1450372, No. DMS-1652330 (W. J. and L. L.), and by the Department of Energy under Grant No. DE-SC0017867, No. DE-AC02-05CH11231 (L. L.). We thank the National Energy Research Scientific Computing (NERSC) center and the Berkeley Research Computing (BRC) program at the University of California, Berkeley for making computational resources available. We thank Dong An, Zhanghui Chen and Lin-Wang Wang for helpful discussions.

References

[1] P. Hohenberg, W. Kohn, Phys. Rev. 136 (1964) B864–B871.
[2] W. Kohn, L. Sham, Phys. Rev. 140 (1965) A1133–A1138.
[3] C. Lee, W. Yang, R.G. Parr, Phys. Rev. B 37 (1988) 785–789.
[4] A.D. Becke, J. Chem. Phys. 98 (1993) 5648.
[5] J.P. Perdew, M. Ernzerhof, Burke. K, J. Chem. Phys. 105 (1996) 9982–9985.
[6] J. Heyd, G.E. Scuseria, M. Ernzerhof, J. Chem. Phys. 118 (2003) 8207–8215.
[7] J. Heyd, G.E. Scuseria, M. Ernzerhof, J. Chem. Phys. 124 (2006) 219906.
[8] A. Stroppa, G. Kresse, New J. Phys. 10 (2008) 063020.
[9] M. Marsman, J. Paier, A. Stroppa, G. Kresse, J. Phys.: Condens. Matter 20 (2008) 064201.
[10] Y. Takimoto, F.D. Vila, J.J. Rehr, J. Chem. Phys. 127 (2007) 154114.
[11] A.V. Krasheninnikov, Y. Miyamoto, D. Tománek, Phys. Rev. Lett. 99 (2007) 016104.
[12] E. Runge, E.K.U. Gross, Phys. Rev. Lett. 52 (1984) 997.
[13] K. Yabana, G.F Bertsch, Phys. Rev. B 54 (1996) 4484–4487.
[14] G. Onida, L. Reining, A. Rubio, Rev. Modern Phys. 74 (2002) 601.
[15] C. Andreani, D. Colognesi, J. Mayers, G. Reiter, R. Senesi, Adv. Phys. 54 (2005) 377–469.
[16] J.L. Alonso, X. Andrade, P. Echenique, F. Falceto, D. Prada-Gracia, A. Rubio, Phys. Rev. Lett. 101 (2008) 096403.
[17] C.A. Ullrich, Time-Dependent Density-Functional Theory: Concepts and Applications, Oxford Univ. Pr., 2011.
[18] K. Lopata, N. Govind, J. Chem. Theory Comput. 7 (2011) 1344–1355.
[19] K. Lopata, .B.E. Van Kuiken, M. Khalil, N. Govind, J. Chem. Theory Comput. 8 (2012) 3284–3292.
[20] F. Ding, B.E. Van Kuiken, B.E. Eichinger, Li. X, J. Chem. Phys. 138 (2013) 064104.
[21] S.A. Sato, Y. Taniguchi, Y. Shinohara, K. Yabana, J. Chem. Phys. 143 (2015) 224116.
[22] A. Castro, M. Marques, A. Rubio, J. Chem. Phys. 121 (2004) 3425–3433.
[23] A. Schleife, E.W. Draeger, Y. Kanai, A.A. Correa, J. Chem. Phys. 137 (2012) 22A546.
[24] A.Gómez. Pueyo, M.A. Marques, A. Rubio, A. Castro, J Chem. Theory. Comput. (2018).
[25] D. An, L. Lin, Quantum dynamics with the parallel transport gauge, arXiv: 180402095.
[26] W. Jia, D. An, L.-W. Wang, L. Lin, J. Chem. Theory Comput. 14 (2018) 5645.
[27] L. Lin, J. Chem. Theory Comput. 12 (2016) 2242.
[28] W. Hu, L. Lin, A. Banerjee, E. Vecharynski, C. Yang, J. Chem. Theory Comput. 13 (3) (2017) 1188–1198.
[29] 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, R.M Wentzcovitch, J. Phys.: Condens. Matter 21 (2009) 395502–395520.
[30] L. Lin, M. Lindsey, Commun. Pure Appl. Math. 72 (2019) 0451.
[31] L. Lin, Z. Xu, L. Ying, Multiscale Model. Simul. 15 (2017) 29–55.
[32] D.G. Anderson, J. Assoc. Comput. Mach. 12 (1965) 547–560.
[33] X. Andrade, J. Alberdi-Rodriguez, D.A. Strubbe, M.J.T. Oliveira, F. Nogueira, A. Castro, J. Muguerza, A. Arruabarrena, S.G. Louie, A. Aspuru-Guzik, A. Rubio, M.A.L. Marques, J. Phys. Condens. Matter 24 (2012) 233202.
[34] J. Jornet-Somoza, J. Alberdi-Rodriguez, B.F. Milne, X. Andrade, M. Marques, F. Nogueira, M. Oliveira, J. Stewart, A. Rubio, Phys. Chem. Chem. Phys. 17 (2015) 26599–26606.
[35] L. Lin, J. Lu, L. Ying, E. W, J. Comput. Phys. 231 (2012) 2140–2154.
[36] W. Hu, L. Lin, C. Yang, J. Chem. Phys. 143 (2015) 124110.
[37] D.R. Hamann, Phys. Rev. B 88 (2013) 085117.
[38] M. Schlipf, F. Gygi, Comput. Phys. Comm. 196 (2015) 36–44.
[39] O.B. Malcıoğlu, R. Gebauer, D. Rocca, S. Baroni, Comput. Phys. Comm. 182 (2011) 1744–1754.
[40] R.D. King-Smith, D. Vanderbilt, Phys. Rev. B 47 (1993) 1651.
[41] N. Marzari, A.A. Mostofi, J.R. Yates, I. Souza, D. Vanderbilt, Rev. Modern Phys. 84 (2012) 1419–1475.

[42] W. Hu, L. Lin, C. Yang, J. Chem. Theory Comput. 13 (2017) 5458.
[43] N. Marzari, D. Vanderbilt, Phys. Rev. B 56 (1997) 12847.
[44] A. Damle, L. Lin, L. Ying, J. Chem. Theory Comput. 11 (2015) 1463–1469.
[45] J. Lu, L. Ying, J. Comput. Phys. 302 (2015) 329.
[46] W. Hu, L. Lin, C. Yang, J. Chem. Theory Comput. 13 (2017) 5420.
[47] K. Dong, W. Hu, L. Lin, J Chem. Theory Comput. 14 (3) (2018) 1311–1320.