Chem Soc Rev

Cite this: *Chem. Soc. Rev.*, 2012, **41**, 3821–3838

www.rsc.org/csr

TUTORIAL REVIEW

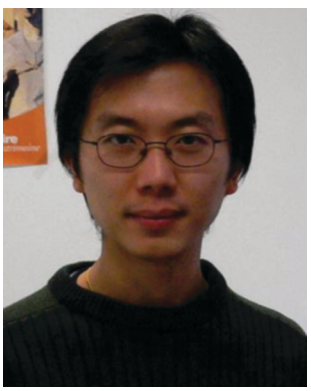
Linear-scaling quantum mechanical methods for excited states

ChiYung Yam,[a] Qing Zhang,[a] Fan Wang[b] and GuanHua Chen*[a]

Received 21st September 2011
DOI: 10.1039/c2cs15259b

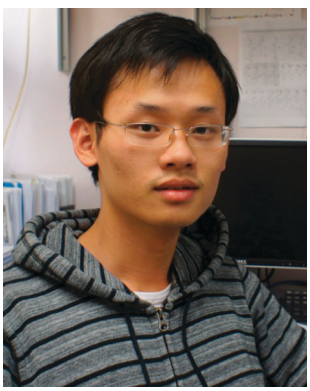
The poor scaling of many existing quantum mechanical methods with respect to the system size hinders their applications to large systems. In this *tutorial review*, we focus on latest research on linear-scaling or $O(N)$ quantum mechanical methods for excited states. Based on the locality of quantum mechanical systems, $O(N)$ quantum mechanical methods for excited states are comprised of two categories, the time-domain and frequency-domain methods. The former solves the dynamics of the electronic systems in real time while the latter involves direct evaluation of electronic response in the frequency-domain. The localized density matrix (LDM) method is the first and most mature linear-scaling quantum mechanical method for excited states. It has been implemented in time- and frequency-domains. The $O(N)$ time-domain methods also include the approach that solves the time-dependent Kohn–Sham (TDKS) equation using the non-orthogonal localized molecular orbitals (NOLMOs). Besides the frequency-domain LDM method, other $O(N)$ frequency-domain methods have been proposed and implemented at the first-principles level. Except one-dimensional or quasi-one-dimensional systems, the $O(N)$ frequency-domain methods are often not applicable to resonant responses because of the convergence problem. For linear response, the most efficient $O(N)$ first-principles method is found to be the LDM method with Chebyshev expansion for time integration. For off-resonant response (including nonlinear properties) at a specific frequency, the frequency-domain methods with iterative solvers are quite efficient and thus practical. For nonlinear response, both on-resonance and off-resonance, the time-domain methods can be used, however, as the time-domain first-principles methods are quite expensive, time-domain $O(N)$ semi-empirical methods are often the practical choice. Compared to the $O(N)$ frequency-domain methods, the $O(N)$ time-domain methods for excited states are much more mature and numerically stable, and have been applied widely to investigate the dynamics of complex molecular systems.

[a] *Department of Chemistry, The University of Hong Kong, Hong Kong. E-mail: ghc@everest.hku.hk*
[b] *College of Chemistry, Sichuan University, Sichuan, China*



ChiYung Yam

ChiYung Yam completed his undergraduate and graduate education in Chemistry at the University of Hong Kong. He received his PhD in 2004. From 2004 to 2008 he was a postdoctoral fellow at the University of Hong Kong. He then moved to Germany for post-doctoral training in Bremen Center for Computational Materials Science, University of Bremen. In April 2010, he became a research assistant professor in Department of Chemistry at the University of Hong Kong. His research interests include developing linear-scaling quantum mechanical methods, quantum transport and electronic device simulations.



Qing Zhang

Qing Zhang received his B.S. degree in Chemistry from the Nanjing University in 2010. He is currently a PhD candidate in the Department of Chemistry at the University of Hong Kong under the guidance of Prof. GuanHua Chen. His current research interests focus on electron-nuclear dynamics and linear-scaling first principle quantum mechanical method development.

1. Introduction

The Nobel Prize in Chemistry 1998 was awarded to Walter Kohn and John Pople for their contributions in density-functional theory (DFT)[1,2] and quantum chemistry, respectively. Semi-empirical and first-principles quantum mechanical methods are routinely used to interpret experimental findings and occasionally to predict material properties prior to measurements. Despite the huge success, numerical quantum mechanical methods are limited mostly to small and medium sized molecules. This is due to the poor scaling of the computational time. For most quantum mechanical methods, the computational time t_{CPU} is proportional to a certain power of the system size; that is, $t_{CPU} \propto N^x$, where N is the number of electronic orbitals and x is an exponent which is usually equal to or larger than 3. On the other hand, there is a growing interest in calculating the electronic structures of complex and large systems like protein, molecular aggregates, and nano-materials. We are increasingly interested in simulating the electronic properties of these systems. All these call for the development of more efficient quantum mechanical methods. The most efficient type of the numerical methods for large systems are the $O(N)$ methods whose computational times scale linearly with the number of atoms in the system.

In 1991 Weitao Yang published the first $O(N)$ quantum mechanical method, the divide-and-conquer (DAC) method[3] for the ground state. Since then, a variety of $O(N)$ methods have been developed for the electronic ground state, for instance, the Fermi operator method, the orbital minimization method and the density matrix minimization method. These and other $O(N)$ quantum mechanical methods have been implemented at semi-empirical and first-principles levels, and have been employed to calculate the electronic properties of complex systems such as biological molecules, polymers, nano-materials and molecular aggregates. However, the $O(N)$ methods are not as widely used as initially thought. They are not as robust as the Black-Box type quantum chemistry methods such that inexperienced researchers can use them readily. This is due to a number of factors: (1) many of the $O(N)$ methods have large pre-factors in terms of the computational time. For instance, the Fermi operator method needs often hundreds of expansion terms, and the evaluation of these hundred terms can be very costly. As a result, the Fermi operator method is computationally more efficient only for molecular systems containing a few hundred atoms or more. (2) Some $O(N)$ methods require that a user has good knowledge of the chemical structure of the molecule of interest. To use the DAC method, one needs to divide the molecule into subunits, and assign the buffer zone for each subunit. A good understanding of the related chemical bonds is thus required to design the subunits and buffer zones. Often the subunits and buffer zones need to be tested and adjusted to carry out the computation. (3) Another problem is the accuracy. High numerical accuracy is required for quantum chemistry calculations. As the system gets larger, this becomes increasingly difficult to attain, in particular, for the $O(N)$ methods. All the $O(N)$ methods are based on the locality of density matrix or alike. $O(N)$ calculation is achieved by cutting off the tail of the off-diagonal density matrix elements. This can lead to the numerical errors.

Before proceeding to discuss the $O(N)$ quantum mechanical methods for excited states, we would summarize several $O(N)$ methods for the ground state. This provides the background for the $O(N)$ methods for excited states, and set the stage for subsequent review.

1.1 The divide-and-conquer method

The divide-and-conquer (DAC) method[3,4] is the first successful $O(N)$ or linear-scaling quantum mechanical method for electronic ground state calculation. The original formulation of the DAC method[3] is based on a subdivision of the electronic density or density matrix. The entire system is divided into subsystems. Around each subsystem, a buffer region is introduced to account for the interaction or electronic coherence between the adjacent subsystems. Calculation is carried out for a combined subsystem and its buffer regions at a time, and thus, the total computational time is proportional to the number of the subsystems, *i.e.* the system size. Yang and Lee[3,4] directly diagonalized the subsystem Hamiltonians while Ozaki[5] and Takayama *et al.*[6] developed a DAC method defined in Krylov subspace and solved iteratively.

Fan Wang

Fan Wang received his PhD in 2001 from Peking University, Beijing. From 2002 to 2008, he worked as a postdoctor in Peking University, University of Calgary, the University of Hong Kong and University of Mainz. From 2008, he became a full professor at the College of Chemistry, Sichuan University in Chengdu. His interests are in developing computational methods for modeling spin-orbit coupling based on coupled-cluster theory and excited state dynamics.

GuanHua Chen

GuanHua Chen is a Professor of Chemistry at the University of Hong Kong. He received his PhD in Physics at California Institute of Technology in 1992. After postdoctoral training in University of Rochester, he joined Department of Chemistry in the University of Hong Kong in 1996 as an assistant professor. In 2010, he became the Head of Department of Chemistry. His current research interests focus on development of first-principles methods for open systems, O(N) first-principles methods for excited states, multi-scale simulation methods for emerging electronics.

Calculation results of all the subsystem/buffers are then compared and combined with the total number of electrons being constant. This can be satisfied by tuning the value of the Fermi energy or chemical potential. One problem is that the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) can be quite large as the size of the subsystem may be small and thus the precise value of the chemical potential is difficult to determine.

As a variation of the DAC method, the fragment molecular orbital (FMO) approach[7] was proposed to calculate the energies of large molecular systems. A large molecule is divided into the fragments. A specific number of electrons are assigned to each fragment. The molecular orbitals (MO) of the fragments and the contributions from pairs of fragments are calculated to account for the total energy of the molecule. The molecular fractionation with conjugated caps (MFCC) method[8] is another $O(N)$ method to calculate the ground state energies of large molecules. The MFCC method is applicable to linear molecules, each of which is divided into the fragments. Each fragment is capped at both ends with two conjugated functional groups, respectively. And the two conjugated functional groups are bonded to form a small molecule. The energies of both the capped fragment and the small molecule are evaluated, and the total energy is approximated as the sum of the energies of the capped fragments minus the energies of all the small molecules. The Cluster-in-Molecule (CIM)[9] method is another FMO approach. Instead of the Hartree–Fock (HF) or DFT calculation on each fragment, the couple-cluster calculation is employed. Different from the conventional DAC method, these FMO methods require that the number of electrons on each fragment is known *a priori*, and thus are not applicable to charge delocalized systems.

1.2 The Fermi operator method

The Fermi operator expansion (FOE)[10] method is an approach for the direct evaluation of the density matrix. Instead of diagonalization, the FOE method expresses the density matrix as a function of the Hamiltonian or Fock matrix F which can be evaluated numerically. Several such functions are possible and we discuss the Chebyshev expansion. The simplest representation of the density matrix, requiring only matrix multiplication, would be a polynomial expansion:

$$\rho \approx p(F) = c_0 I + c_1 F + c_2 F^2 + \cdots + c_n F^n \qquad (1)$$

where I is the identity matrix and ρ is the density matrix. Unfortunately, the numerical evaluation of the polynomials of high degree is unstable. This instability can be avoided by introducing a Chebyshev polynomial expansion, which is a widely used numerical method:[11]

$$p(F) = \frac{c_0}{2} I + \sum_{i=1}^{n} c_i T_i(F). \qquad (2)$$

Because the Chebyshev polynomials $T_i(F)$ are defined within interval $[-1:1]$, it is assumed in the following that the eigenvalue spectrum of F falls within this interval. This can be easily achieved by scaling and shifting of the original Hamiltonian. The Chebyshev matrix polynomials $T_i(F)$ satisfy the following recursion relations

$$\begin{aligned} T_0(F) &= I \\ T_1(F) &= F \\ T_{i+1}(F) &= 2FT_i(F) - T_{i-1}(F). \end{aligned} \qquad (3)$$

The coefficients of the Chebyshev expansion can easily be determined as described in standard textbooks on numerical analysis.[11] The desired linear-scaling computational effort can be obtained by utilizing the sparsity of the Hamiltonian or Fock matrix F as only matrix multiplication is involved.

The computational cost of the Chebyshev expansion method has been analyzed by Baer and Head-Gordon,[12] who found that the order m of the polynomial needed to achieve a 10^{-D} accuracy depends linearly on the width of the Hamiltonian spectrum ΔE and the electronic temperature $1/\beta$, and approximately, $m \approx D\beta\Delta E$. This leads to problems when Hamiltonians with large spectral width ΔE are considered, or when low electronic temperatures are required. Obviously the Chebyshev expansion method cannot be applied to the zero temperature case. Recently it has been suggested[13] that fast polynomial summation methods, requiring a number of multiplications $\sim\sqrt{m}$, can be applied to the FOE, leading to the more favorable scaling $\sqrt{\beta}\Delta E$.

1.3 The orbital minimization method

The orbital minimization (OM)[14] method calculates the grand potential in the limit of zero temperature. In contrast to the previous methods, it does not calculate the density matrix directly but expresses it *via* the Wannier functions. These Wannier functions are obtained by minimizing the following unconstrained functional

$$\Phi_G = 2\sum_n \sum_{ij} c_i^n F'_{ij} c_j^n - \sum_{nm} \sum_{ij} c_i^n F'_{ij} c_j^m \sum_k c_k^n c_k^m \qquad (4)$$

where c_i^n is the expansion coefficient of the n-th Wannier orbital with respect to the i-th basis function. F' is the shifted Hamiltonian $F - \tilde{\mu} I$ and $\tilde{\mu}$ is the chemical potential. The gradient of the functional of eqn (4) is given by

$$\begin{aligned} \frac{\partial \Phi_G}{\partial c_k^n} = {} & 4\sum_i F'_{ki} c_i^n - 2\sum_m \sum_i F'_{ki} c_i^m \sum_k c_k^n c_k^m \\ & - 2\sum_m c_k^m \sum_{ij} c_i^n F'_{ij} c_j^m. \end{aligned} \qquad (5)$$

Instead of Wannier orbitals, Φ_G and $\frac{\partial \Phi_G}{\partial c}$ are expressed in terms of eigen-orbitals. Using the fact that $\sum_k c_k^n c_k^m = \delta_{nm}$ and $\sum_i^c n F'_{ij} c_j^m = \delta_{nm}(\varepsilon_n - \tilde{\mu})$, one obtains

$$\begin{aligned} \Phi_G &= 2\sum_n \sum_{ij} c_i^n F'_{ij} c_j^n - \sum_{nm} \sum_{ij} c_i^n F'_{ij} c_j^m \delta_{nm} \\ &= \sum_n \sum_{ij} c_i^n F'_{ij} c_j^n = \sum_n \varepsilon_n - \tilde{\mu} N_e \end{aligned} \qquad (6)$$

where N_e is the number of electrons and the gradient equation can thus be simplified as:

$$\frac{\partial \Phi_G}{\partial c_k^n} = 4\sum_i F'_{ki}c_i^n - 2\sum_m\sum_i F'_{kl}c_i^m\delta_{mn} - 2c_k^n\sum_m \delta_{nm}(\varepsilon_m - \tilde{\mu})$$
$$= 2\sum_i F'_{ki}c_i^n - 2c_k^n(\varepsilon_n - \tilde{\mu}) = 0. \quad (7)$$

So Φ_G has indeed a vanishing gradient at the ground state, and gives the correct ground-state energy. The gradient vanishes not only for the ground state orbitals but also for the excited state orbitals. It is shown that these stationary points are not local minima.[14]

$O(N)$ calculation is achieved by utilizing the locality of Wannier orbitals. Each Wannier orbital is localized within its own localization region around its atomic center. This implies that for a certain n, there is only a finite number of nonzero c_k^n. Therefore, there are overall $O(N)$ nonzero c_k^n and the computational effort for evaluation of Φ_G is $O(N)$ as well.

1.4 The density matrix minimization method

The density matrix minimization (DMM) method[15] is another $O(N)$ method. The density matrix ρ is varied to minimize the following functional for the grand potential Φ_G

$$\Phi_G = \mathrm{Tr}[(3\rho^2 - 2\rho^3)(F - \tilde{\mu}I)]. \quad (8)$$

There is no constraint imposed during the minimization, so all the density matrix elements are independent degrees of freedom. Nevertheless, the density matrix corresponds to a pure state whose wave function is a single Slater determinant. This is ensured by setting $\rho = 3\rho^2 - 2\rho^3$.[16] The gradient of Φ_G with respect to ρ is itself a matrix and is given by

$$\frac{\partial \Phi_G}{\partial \rho} = 3(\rho F' + F'\rho) - 2(\rho^2 F' + \rho F'\rho + F'\rho^2). \quad (9)$$

To verify that eqn (8) defines a valid functional, one needs to show two things: first, that the grand potential expression eqn (8) gives the correct result if the exact density matrix ρ is inserted, and second, the gradient in eqn (9) vanishes in this case. The density matrix is a projection operator; that is, $\rho^2 = \rho$. Therefore, $(3\rho^2 - 2\rho^3) = \rho$, and the grand potential expression agrees with the correct result

$$\Phi_G = \mathrm{Tr}[\rho(F - \tilde{\mu}I)]. \quad (10)$$

The fact that F' and the exact ρ commute also ensures that the gradient in eqn (9) vanishes. The gradient vanishes not only for the ground state density matrix but also for the excited state density matrix.

Besides the $O(N)$ quantum mechanical methods for the ground state, $O(N)$ quantum mechanical methods have been developed for excited states. The first $O(N)$ method for excited states, the LDM method,[17] was developed in 1997, and the resulting manuscript was published in 1998 by Yokojima and Chen. At that time, the LDM method was implemented at the semi-empirical level. It was later implemented with time-dependent density-functional theory (TDDFT) by Yam, Yokojima and Chen, and the work was published in 2003.[18] The LDM method has been applied widely to simulate the optical processes of a variety of complex molecular systems such as light harvesting systems,[19] carbon nanotubes,[20] polymer aggregates,[21] and water clusters.[22] There are two types of LDM methods: one is the time-domain approach and the other is the frequency-domain approach. The time-domain LDM method has been much more widely developed and used[17–24] so far, and the frequency-domain LDM method was initially developed for one-dimensional or quasi-one-dimensional systems.[25] In 2004, Weber, Niklasson and Challacombe proposed an $O(N)$ method to calculate the response to external field.[26] In 2007, Kussmann and Ochsenfeld proposed another $O(N)$ method for the electronic response to external field.[27] Recently, Yang and his coworkers developed an $O(N)$ time-domain method for excited states.[28] Instead of the density matrix used in the LDM method, localized molecular orbitals (LMO) are constructed *via* unitary transformation and propagated in the time-domain.

In this review, we start with a discussion on the theoretical basis of $O(N)$ quantum mechanical methods, for ground and excited states: the nearsightedness of physical properties. $O(N)$ evaluation of two-electron integrals and construction of the Fock matrix are examined. We then proceed to review the $O(N)$ quantum mechanical methods for excited states. We begin by discussing thoroughly the time-domain methods, the LDM method and the TDKS/NOLMO method. Numerical algorithms are presented and compared. As the time-domain methods have been implemented at both the first-principles TDDFT and semi-empirical levels, the advantages and shortcomings of both types are given and analyzed. Applications are presented to demonstrate the effectiveness of the $O(N)$ time-domain methods. We then discuss the frequency-domain methods. As the $O(N)$ frequency-domain method for one-dimensional or quasi-one-dimensional systems is well developed, we review separately the frequency-domain methods for one-dimensional and two- or three-dimensional systems. Finally we compare the pros and cons of various $O(N)$ quantum mechanical methods for excited states, and summarize the suitability or applicability of different $O(N)$ excited state methods.

2. Theoretical basis of linear-scaling methods

2.1 Locality of the density matrix

Walter Kohn pointed out in 1996 that the physical basis of linear-scaling methods is the locality of many physical entities,[29] and termed it as the ''nearsightedness'' principle. The ''nearsightedness'' principle implies that the properties of a certain observation region are only weakly influenced by changes or factors that are spatially far away from the observation region. Around the same time, Chen and Mukamel independently realized that the locality of the reduced single-electron density matrix persists for the ground state and many excited states,[30] and they showed that the off-diagonal density matrix elements in real space representation are negligible when the distance between two points in space is larger than the critical lengths l_0 or l_1 for ground or excited states, respectively. When the density matrix is expressed in atomic orbital (AO) representation, its off-diagonal elements

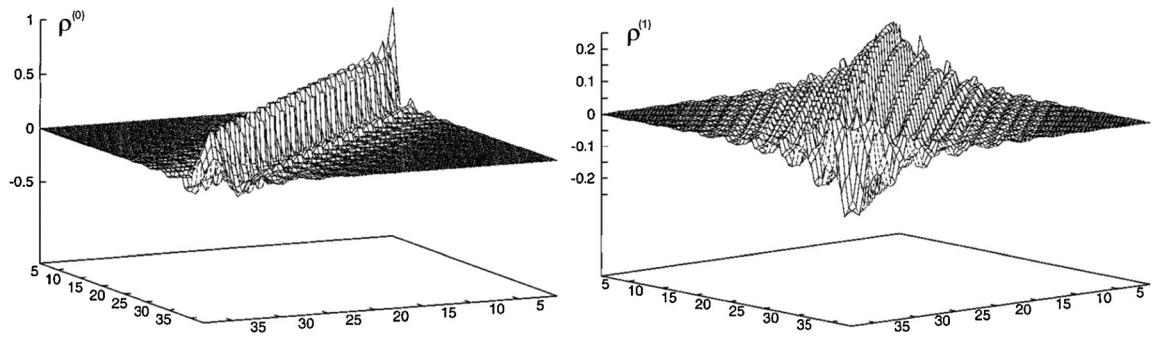


Fig. 1 Ground state density matrix $\rho^{(0)}$ and first order induced density matrix $\rho^{(1)}$ for $C_{38}H_{40}$. The numbers on the two axes in the bottom planes represent the indices i and j of density matrix element ρ_{ij}, respectively. Reprinted with permission from ref. 30.

are negligible when the distances of the corresponding two AOs are large enough. Fig. 1 shows the ground state and first order induced density matrices for a polyacetylene oligomer with 38 carbon atoms subjected to an external electric field. The diagonal elements are the electron occupation numbers at various AOs, and the off-diagonal elements are the bond orders or the electron coherence between pairs of AOs.[31] From the figure, we can see that the HF ground state density matrix and first order induced density matrix are almost diagonal. Moving away from the diagonal lines, the matrix elements in both the ground state density matrix and the first order induced density matrix decay toward zero. Most linear-scaling or $O(N)$ algorithms are built around the locality of the single-electron density matrix or its variation such as the Wannier function. To obtain the linear-scaling computation, one has to cut off the fast decaying quantities when they are small enough. This introduces the concept of a localization region. Only inside this localization region is the quantity calculated; outside it is assumed to vanish.

The locality of the single-electron density matrix is related to the HOMO–LUMO gap, or the band gap for solids, E_g. An off-diagonal element of density matrix ρ_{ij} in the AO representation decays asymptotically as[32]

$$\rho_{ij} \sim e^{-cE_g r_{ij}}, \tag{11}$$

where r_{ij} is the distance between AOs i and j. This is true for insulators, semiconductors, macromolecules, and clusters where $E_g > 0$. For metallic systems, the situation is different as $E_g = 0$, and the off-diagonal elements decay as follows,

$$\rho_{ij} \sim 1/r_{ij}^2. \tag{12}$$

The locality of the ground state density matrix is the basis for $O(N)$ ground state quantum mechanical methods. The induced density matrix due to external field possesses locality as well, which forms the foundation for $O(N)$ methods for excited states.

2.2 Linear-scaling calculations of the Fock matrix

In most quantum mechanical methods, such as HF, DFT and so on, self-consistent calculations are needed. First, a potential is calculated based on a trial charge distribution, then a new charge distribution is calculated from the potential obtained in the previous step. The calculation of the potential usually consists of two parts: the exchange correlation (XC) potential and the Coulomb potential. Whereas the exchange terms are local, the Coulomb potential V is of very long range

$$V(\boldsymbol{r}) = \int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}'. \tag{13}$$

For both ground and excited states, the Fock matrix needs to be evaluated. There exist methods to calculate the Coulomb potential and exchange (or XC) potential with linear-scaling costs, such as the fast multipole method (FMM)[33,34] and the integral prescreening technique methods.[35]

The FMM is a mathematical technique that was developed in the 1980's to accelerate the calculation of long-ranged forces. Through the multipole expansion, it allows one to group sources that lie close together and treat them as a single source. The method was first applied to the field of computational electromagnetics. The idea was then borrowed and implemented to efficiently treat the Coulomb interaction in HF and DFT calculations.

For excited states calculations, the Coulomb contribution to the induced Fock matrix is constructed as follows,

$$\int d\boldsymbol{r} \int d\boldsymbol{r}' \chi_i(\boldsymbol{r}) \chi_j(\boldsymbol{r}) \frac{1}{\boldsymbol{r} - \boldsymbol{r}'} \sum_{kl} \delta\rho_{kl} \chi_k(\boldsymbol{r}') \chi_l(\boldsymbol{r}') \tag{14}$$

where $\delta\rho$ is the induced density matrix. In the FMM, the whole physical space with all the charges and induced charges is divided into half along each Cartesian axis. The division is continued recursively until the charges in each box at the lowest level are approximately constant, forming a hierarchy of boxes as shown in Fig. 2a. The Coulomb potential at an arbitrary point $\boldsymbol{r}$ can then be expressed through the multipole expansion as follows,[18]

$$V(\boldsymbol{r}) = \frac{Z}{R} + \frac{\mu_\alpha R_\alpha}{R^3} + \frac{Q_{\alpha\beta} R_\alpha R_\beta}{R^5} + \frac{O_{\alpha\beta\gamma} R_\alpha R_\beta R_\gamma}{R^7} + \cdots, \tag{15}$$

where Z, μ, Q and O are monopole, dipole, quadrupole and octupole, respectively. $R = \boldsymbol{r} - \boldsymbol{r}_A$ and $\boldsymbol{r}_A$ is the center of box A (see Fig. 2b). The multipole expansion of the potential stems from charges at higher level boxes that can be constructed from their child boxes, resulting in a multipole expansion for each cube representing all the charges contained in it. The potential at point $\boldsymbol{r}$ is partitioned into near-field and

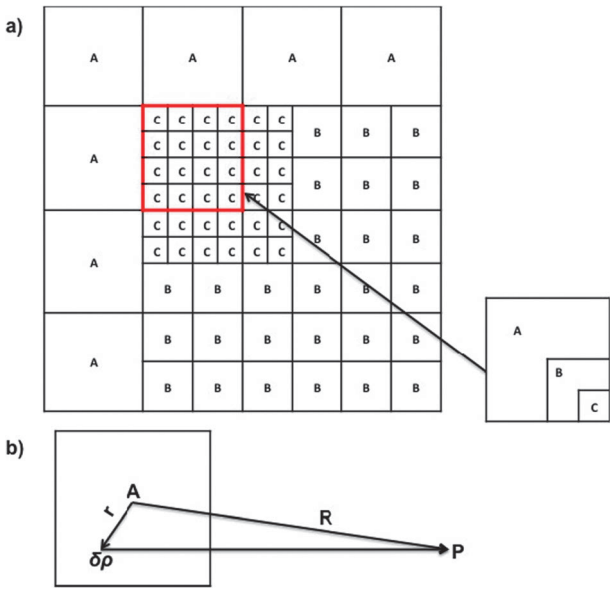


Fig. 2 (a) Hierarchy of boxes; (b) Coulomb potential formed by $\delta\rho$ at point P.

far-field contributions. For near-field potential, explicit analytical integration is used; for the far-field potential, depending on the distance of the interaction, the multipole expansions from different levels are converted into local Taylor expansions and the expansion coefficients are summed. Since the number of boxes in each level is constant, the Coulomb part of the Fock matrix can be evaluated with linear-scaling efforts.

The formation of the Fock matrix involves also the exchange terms. In DFT, evaluation of the XC contributions to the Fock matrix is done *via* numerical quadratures. Linear-scaling can be achieved through efficient screening techniques.[35] For HF and hybrid DFT methods, there is also a contribution from the so-called exact exchange. Although the contributions of exact exchange to the Fock matrix arise from the same set of integrals, the FMM cannot be applied to the exchange terms because the FMM requires the contraction of charge distributions with the density matrix from the beginning. However, linear-scaling can be achieved by exploiting the decay of the density matrix with distance between AO centers [eqn (11) and (12)]. Exact exchange thus involves only a finite number of significant terms and several methods have been developed which exploit the locality of the density matrix to reach a linear-scaling regime for exact exchange evaluation.[36–39]

Utilizing the linear-scaling techniques to construct the Fock matrix, Izmaylov *et al.*[40] developed a linear-scaling method to calculate static and dynamic polarizabilities for periodic systems. The infinite summation in the evaluation of Coulomb contribution is done by using the periodic version of the FMM[41,42] where the asymptotic quadratic scaling is reduced to linear. For exact exchange and XC contributions, linear-scaling techniques similar to the non-periodic case are used. Transformation between MO and AO basis is involved which formally scales cubically with the size of the unit cell. While the cubic scaling part has a small pre-factor, the scaling behavior may become significant when a large unit cell or a semi-empirical method is used.

3. Linear-scaling electronic structure methods for excited states

3.1 Time-dependent Hartree–Fock and time-dependent density-functional theory

So far linear-scaling methods for excited states are based either on the TDDFT or time-dependent Hartree–Fock (TDHF) method. They may be classified as time-domain or frequency-domain $O(N)$ methods. The first linear-scaling excited state method was developed and implemented successfully at the semi-empirical level in 1998.[17] It was based on TDHF theory. The $O(N)$ TDDFT method was later developed.[18] Before proceeding further, we review TDHF and TDDFT methods briefly.

TDDFT is based on Runge–Gross theorem[43,44] which is an extension of Hohenberg–Kohn theorem.[1] While Hohenberg–Kohn theorem is applicable to time-independent systems, Runge–Gross theorem applies to time-dependent systems and states that there is a one-to-one correspondence between the time-dependent electronic density $\rho(\boldsymbol{r},t)$ and the external potential $\upsilon(\boldsymbol{r},t)$ up to an overall time-dependent function $g(t)$. Based on this theorem, Runge and Gross proposed the following TDKS equation to account for the electronic response to the time-dependent external field (atomic units are used in the whole review),

$$\left[-\frac{1}{2}\nabla^2 + \upsilon_{\mathrm{eff}}(\boldsymbol{r},t)\right]\psi_i(\boldsymbol{r},t) = i\frac{\partial}{\partial t}\psi_i(\boldsymbol{r},t) \tag{16}$$

where

$$\upsilon_{\mathrm{eff}}(\boldsymbol{r},t) = \upsilon(\boldsymbol{r},t) + \int \frac{\rho(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}' + \upsilon_{\mathrm{XC}}(\boldsymbol{r},t) \tag{17}$$

and

$$\rho(\boldsymbol{r},t) = \sum_i^{\mathrm{occ}} |\psi_i(\boldsymbol{r},t)|^2 \tag{18}$$

$$\upsilon_{\mathrm{XC}}(\boldsymbol{r},t) = \frac{\delta A_{\mathrm{XC}}[\rho]}{\delta\rho(\boldsymbol{r},t)} \tag{19}$$

υ_{eff} is the effective one-electron potential, υ_{XC} is the XC potential, υ is the external potential coming from nuclei and other external sources, A_{XC} is the XC part of the action and is a functional of $\rho(\boldsymbol{r},t)$, $\psi_i(\boldsymbol{r},t)$ is the wave function of the i-th time-dependent single electron orbital. If the initial wave functions $\psi_i(\boldsymbol{r},0)$ are known, the TDKS equation can be solved by integration with respect to time. In practice, the single electron molecular wave function was usually expanded using AOs or plane waves as the basis functions.

The computational cost to construct the MO is usually of $O(N^3)$, where N is the number of AOs or basis functions in the system. Furthermore, the coefficient matrix in the canonical MO representation is dense. When the system is very large, the computational time increases rapidly. Once the exact $A_{\mathrm{XC}}[\rho(\boldsymbol{r},t)]$ or $\upsilon_{\mathrm{XC}}[\rho(\boldsymbol{r},t),t]$ is known, eqn (16) above is exact, and can be used to calculate the properties of electronic excited states such as excitation energies and optical spectra. TDDFT is a direct generalization of DFT, and yields the exact properties

of the excited states once the time-dependent XC functional is rigorously known.

The TDKS equation can be reformulated in terms of the reduced single-electron density matrix. In real space representation, the corresponding reduced single-electron density matrix can be expressed as

$$\rho(\boldsymbol{r},\boldsymbol{r}';t)=\sum_{i=1}^{\mathrm{occ}}\psi_i(\boldsymbol{r},t)\psi_i^*(\boldsymbol{r}',t). \tag{20}$$

Taking the derivative of eqn (20) with respect to time, and substituting it in eqn (16), we can get the time evolution equation of the density matrix,

$$\begin{aligned} i\frac{\partial}{\partial t}\rho(\boldsymbol{r},\boldsymbol{r}',t) &= \sum_{i}^{\mathrm{occ}}[(F\psi_i(\boldsymbol{r},t))\psi_i^*(\boldsymbol{r}',t) \\ &\quad -\psi_i(\boldsymbol{r},t)(F\psi_i(\boldsymbol{r}',t))^*] \end{aligned} \tag{21}$$

$$i\frac{\partial\rho(t)}{\partial t}=F\rho-\rho F=[F(t),\rho(t)] \tag{22}$$

where $F=-\frac{1}{2}\nabla^2+\upsilon_{\mathrm{eff}}(\boldsymbol{r},t)$. Eqn (22) is called the Liouville–von Neumann equation and is equivalent to the above TDKS equation. In the AO representation, the elements of the reduced single-electron density matrix are expressed as:

$$\rho_{ij}=\langle\chi_i|\rho|\chi_j\rangle. \tag{23}$$

Here χ_i and χ_j are AOs. The corresponding equation of motion (EOM) for the density matrix in AO basis can then be expressed as

$$\begin{aligned} i\frac{\partial}{\partial t}\rho_{ij}(t) &= \sum_{k}^{\mathrm{occ}}(\langle\chi_i|\hat{F}|\psi_k(\boldsymbol{r},t)\rangle\langle\psi_k(\boldsymbol{r}',t)|\chi_j\rangle \\ &\quad -\langle\chi_i|\psi_k(\boldsymbol{r},t)\rangle\langle\psi_k(\boldsymbol{r}',t)|\hat{F}|\chi_j\rangle) \\ &= \sum_{k}^{\mathrm{occ}}\sum_{mn}(\langle\chi_i|\hat{F}|\chi_m\rangle c_{mk}(t)c_{nk}^*(t)\langle\chi_n|\chi_j\rangle \\ &\quad -\langle\chi_i|\chi_m\rangle c_{mk}(t)c_{nk}^*(t)\langle\chi_n|\hat{F}|\chi_j\rangle) \\ &= \sum_{m}(F_{im}\rho_{mj}-\rho_{im}F_{mj}). \end{aligned} \tag{24}$$

Eqn (24) can be solved in the time domain by propagating the density matrix in real time. It can also be solved in the frequency-domain by taking the Fourier transformation of eqn (24).

TDHF can be formulated virtually in the same way.

$$\left[-\frac{1}{2}\nabla^2+\upsilon_{\mathrm{eff}}^{\mathrm{HF}}(\boldsymbol{r},t)\right]\psi_i(\boldsymbol{r},t)=i\frac{\partial}{\partial t}\psi_i(\boldsymbol{r},t) \tag{25}$$

$$\upsilon_{\mathrm{eff}}^{\mathrm{HF}}(\boldsymbol{r},t)=\upsilon(\boldsymbol{r},t)+\int\frac{\rho(\boldsymbol{r}',t)}{|\boldsymbol{r}-\boldsymbol{r}'|}\mathrm{d}\boldsymbol{r}'+\upsilon_{\mathrm{X}}^{\mathrm{HF}}(\boldsymbol{r},t) \tag{26}$$

$\upsilon_{\mathrm{X}}^{\mathrm{HF}}$ is the HF exchange potential. Once again TDHF's EOM can be formulated in terms of the wave function or density matrix, and can be solved both in the time-domain and frequency-domain.

O(*N*) excited state methods were initially implemented for TDHF at the semi-empirical level such as Parisa–Parr–Pople (PPP),[17,25] completely neglecting differential overlap (CNDO)[21] and PM3[20] Hamiltonians. The resulting methods have been employed to study the optical properties of nano-materials,[20] polymers,[17,21,25] and biological systems.[19] The *O*(*N*) TDDFT method was subsequently developed and implemented.[18] The time-domain methods are quite mature and widely used, while the *O*(*N*) frequency-domain methods have mostly been tested for model systems to demonstrate their feasibility. In the following, we first discuss the *O*(*N*) time-domain quantum mechanical methods, methodology and applications. We review then the *O*(*N*) frequency-domain methods.

3.2 *O*(*N*) time-domain methods for excited states

3.2.1 Time-domain LDM method. Yokojima and Chen proposed and developed the first *O*(*N*) quantum mechanical method for excited states, the LDM method.[17] Implemented at the semi-empirical level, it is based on the locality of the field-induced density matrix. Here the general formalism of the time-domain LDM method is given which is applicable to any Hamiltonians, semi-empirical or first-principles.

General formalism. Starting from the EOM for the reduced single-electron density matrix $\rho(t)$ [eqn (22)], $\rho(t)$ can be partitioned into two parts

$$\rho(t)=\rho^{(0)}+\delta\rho(t) \tag{27}$$

where $\rho^{(0)}$ is the single-electron density matrix representing the HF/DFT ground state in the absence of the external field, and $\delta\rho(t)$ is the difference between $\rho(t)$ and $\rho^{(0)}$. The total Fock matrix in eqn (22) is divided into three parts

$$F(t)=F^{(0)}+\delta F(t)+f(t) \tag{28}$$

where $F^{(0)}$ is the ground state Fock matrix and $\delta F(t)$ gives the induced Fock matrix stemming from $\delta\rho(t)$. $f(t)$ represents the interaction between an electron and the external field $\xi(t)$.

$$f_{ij}(t)=-\xi(t)\cdot\langle\chi_i|\vec{\boldsymbol{r}}|\chi_j\rangle \tag{29}$$

Eqn (22) thus becomes

$$i\delta\dot{\rho}-([F^{(0)},\delta\rho]+[\delta F,\rho^{(0)}])=[f,\rho^{(0)}]+[f,\delta\rho]+[\delta F,\delta\rho]. \tag{30}$$

Here, the dot symbol in $\delta\dot{\rho}$ represents the partial time derivative. For the first-order induced density matrix $\delta\rho^{(1)}$, its dynamics can be described by the following equation

$$i\delta\dot{\rho}^{(1)}-([F^{(0)},\delta\rho^{(1)}]+[\delta F^{(1)},\rho^{(0)}])=[f,\rho^{(0)}]. \tag{31}$$

More specifically, eqn (31) can be written as

$$\begin{aligned} i\delta\dot{\rho}_{ij}^{(1)} &= \sum_{k}(F_{ik}^{(0)}\delta\rho_{kj}^{(1)}-\delta\rho_{ik}^{(1)}F_{kj}^{(0)}) \\ &\quad +\sum_{k}(\delta F_{ik}^{(1)}\rho_{kj}^{(0)}-\rho_{ik}^{(0)}\delta F_{kj}^{(1)}) \\ &\quad +\sum_{k}(f_{ik}\rho_{kj}^{(0)}-\rho_{ik}^{(0)}f_{kj}). \end{aligned} \tag{32}$$

Solving eqn (32) alone does not lead to the linear-scaling of computational time, because the matrix multiplication involved is intrinsically $O(N^3)$. The key for the *O*(*N*) scaling lies

in the reduction of the dimension of the reduced single-electron density matrix which is based on the fact that the density matrix has a localized character. This locality holds not only for the $\rho^{(0)}$ but also for $\delta\rho^{(1)}$. Specifically, $\rho_{ij}^{(0)}$ is set to zero for $r_{ij} > l_0$, where r_{ij} is the distance between the matrix elements i and j. Consequently, $F_{ij}^{(0)}$ becomes zero for the same r_{ij}; and $\delta\rho_{ij}^{(1)}$ is set to zero when $r_{ij} > l_1$, here l_0 and l_1 are cutoff lengths.

For a fixed pair i and j, the summation over k in eqn (32) is finite and independent of the system size. This leads to $O(N)$ computation of the first and third terms of the right-hand side (RHS) of eqn (32). However, the numerical evaluation of the second term on the RHS of eqn (32) is more complicated. It can be expanded as

$$\sum_{mkl}(\delta\rho_{kl}^{(1)}V_{im,kl}\rho_{mj}^{(0)} - \rho_{im}^{(0)}\delta\rho_{kl}^{(1)}V_{mj,kl}) + \sum_{mkl}(\delta\rho_{kl}^{(1)}V_{im,kl}^{X}\rho_{mj}^{(0)} - \rho_{im}^{(0)}\delta\rho_{kl}^{(1)}V_{mj,kl}^{X}). \quad (33)$$

The first term in eqn (33) gives the Coulomb interaction between the induced charge distribution and the ground state charge distribution. The summation over k and l is of $O(N)$, which leads to overall $O(N^2)$ scaling for the direct computation of the second term on the RHS of eqn (32). As described in Section 2.2, the FMM is employed to evaluate $\sum_{kl}\delta\rho_{kl}^{(1)}V_{im,kl}$ and $\sum_{kl}\delta\rho_{kl}^{(1)}V_{mj,kl}$, leading to an overall finite number of floating point calculations for the first term in eqn (33). The second term in eqn (33) represents the exchange (or XC) term. To achieve linear-scaling, the localized nature of the density matrix and the exchange (or XC) potential is exploited as discussed in Section 2.2. With the above techniques, the number of summations over k and l in eqn (32) is restricted to a finite range, which does not depend on the value of N. Because the number of $\delta\rho_{ij}^{(1)}$ is proportional to N, the total number of floating-point calculations scales linearly with N. Therefore, the computational time is expected to be proportional to N.

Alternatively, a threshold criterion can be used to take the advantage of matrix sparsity, *i.e.* a matrix element can be safely neglected when it is below a certain threshold value. Asymptotically the number of significant elements in the matrices grows linearly with the system size. Eqn (31) involves only matrix multiplications and additions, and these matrix manipulations can thus be carried out with $O(N)$ effort. A standard way is to represent the matrices in a compressed sparse row (CSR) or modified compressed sparse row (MSR) format.[45] The matrix multiplication is then simply carried out by multiplying one pair of matrix elements at a time. Matrix additions or subtractions are straightforward and are far less expensive than multiplications.

Time integration: the fourth-order Runge–Kutta method. EOMs like eqn (22) and (31) can be solved by time-domain integration. Common algorithms for time-domain integration include the Runge–Kutta method,[11] the Exponential Midpoint method (EMM)[46] and the Magnus[47] method. The fourth-order Runge–Kutta method has been implemented in the LDM method to solve $\rho(t)$. Starting with the ground state density matrix as the initial density matrix, *i.e.* $\rho(t_0) = \rho^{(0)}$, the fourth-order Runge–Kutta approximation of $\rho(t)$ is given by

$$\rho_{n+1} = \rho_n + \tfrac{1}{6}(\Delta\rho_1 + 2\Delta\rho_2 + 2\Delta\rho_3 + \Delta\rho_4) \quad (34)$$

where

$$\begin{aligned}\Delta\rho_1 &= \Delta t f(t_n, \rho_n),\\ \Delta\rho_2 &= \Delta t f\left(t_n + \frac{\Delta t}{2}, \rho_n + \frac{\Delta\rho_1}{2}\right),\\ \Delta\rho_3 &= \Delta t f\left(t_n + \frac{\Delta t}{2}, \rho_n + \frac{\Delta\rho_2}{2}\right),\\ \Delta\rho_4 &= \Delta t f(t_n + \Delta t, \rho_n + \Delta\rho_3),\\ \frac{d\rho}{dt} &= f(t,\rho)\end{aligned} \quad (35)$$

and t_0 stands for $t = 0$, $t_{n+1} = t_n + \Delta t$. Here, the next value ρ_{n+1} is determined by the present value ρ_n plus the weighted average of four $\Delta\rho$ evaluating at different times and ρ. And $\rho(t_n + \Delta t)$ is approximated by ρ_{n+1}. The Runge–Kutta method is a low order method and requires four evaluations of the derivative for each time step. The time step is determined by the largest absolute eigenvalue of the Fock matrix[48] and normally a small time step has to be used to obtain an accurate solution of eqn (31).

Time integration: Chebyshev expansion for linear response. High order methods such as Chebyshev[49] and Legendre[50] methods can lead to much higher accuracy and can be more efficient. The Chebyshev expansion method has been used for the time-domain integration in the LDM method.[22,23] In the linear response regime, the EOM for the first-order reduced density matrix can be expressed as

$$i\delta\dot{\rho}^{(1)} = \hat{\boldsymbol{L}}\delta\rho^{(1)} = [F^{(0)},\delta\rho^{(1)}]) + [\delta F^{(1)},\rho^{(0)}] \quad (36)$$

where $\hat{\boldsymbol{L}}$ is a time-independent linear operator. The Runge–Kutta method with adaptive time steps has been shown to achieve much improvement for systems under strong laser fields, it is however not expected to be more efficient for solving eqn (36) due to the time-independent property of $\hat{\boldsymbol{L}}$. The formal solution for eqn (36) can be written as

$$\delta\rho^{(1)}(t) = e^{-i\hat{L}t}\delta\rho^{(1)}\ (t = 0). \quad (37)$$

The most straightforward method to calculate $e^{-i\hat{L}t}$ is the Taylor expansion. It can be shown that the fourth-order Runge–Kutta method is exactly the same as Taylor expansion to the fourth order. In addition, the convergence property of the Taylor expansion behaves as a power law. Taylor expansion of $e^{-i\hat{L}t}$ is thus not very efficient and accurate. In general, a time step satisfying $\delta t \ll k/L_{\max}$ is required for the Taylor expansion of $e^{-i\hat{L}t}$ to converge, where k is the term used in the Taylor expansion, $L_{\max}$ is the largest eigenvalue of operator $\hat{\boldsymbol{L}}$, *i.e.*, the largest excitation energy. When a large time step δt is used, a numerical divergence problem could arise. The largest time step that leads to a stable expansion should satisfy

$$\left|\sum_{n=0}^{k}\frac{1}{n!}(-i\Delta t_{\max}L_{\max})^n\right| \le 1. \quad (38)$$

According to eqn (38), it can be seen that for $k = 2$, the expansion is unstable no matter what time step is used.

For $k = 4$, the largest time step for stable expansion is $8/L_{\max}$.[46] For $k = 6$, an even smaller time step is needed.

A more efficient method to calculate $e^{-i\hat{L}t}$ is the Chebyshev expansion. The system is initially in the ground state. At $t = 0$, the following external potential is applied:

$$V_{\text{ext}}(r,t) = V_{\text{ext}}(r)\delta(t). \tag{39}$$

Thus, at $t = 0^+$, the density matrix becomes:

$$\rho(t = 0^+) = -i[V_{\text{ext}},\rho^{(0)}]. \tag{40}$$

With the Chebyshev expansion, the subsequent time evolution of the density matrix reads[46,51]

$$\delta\rho^{(1)}(t) = e^{-i\hat{L}t}\delta\rho^{(1)}(t = 0) = \sum_{n=0}^{\infty}(2 - \delta_{n0})J_n(t\Delta)\rho_n \tag{41}$$

$$\begin{aligned} \rho_n &= \frac{-2i}{\Delta}\hat{\boldsymbol{L}}\rho_{n-1} + \rho_{n-2}, \\ \rho_1 &= \frac{-i}{\Delta}\hat{\boldsymbol{L}}\rho_0, \quad \rho_0 = \delta\rho^{(1)}(t = 0) \end{aligned} \tag{42}$$

where $J_n(\alpha)$ is the Bessel function of the first kind and Δ is a positive number larger than $L_{\max}$ for the expansion to converge. Unlike the Taylor expansion method, this expansion converges for any time step as long as Δ is larger than $L_{\max}$. The Chebyshev expansion thus avoids the numerical divergent problem. Practically, Δ is estimated through the difference between the highest virtual level and the lowest occupied level. For any time t' smaller than the time t, $\delta\rho^{(1)}(t')$ can be determined through the above Chebyshev expansion with a minor additional effort. The Chebyshev expansion converges exponentially due to the exponential decay of the Bessel function $J_n(\alpha)$ when n is larger than α. The number of terms needed in eqn (41) should only be slightly larger than $t\Delta$. To achieve high efficiency with the Chebyshev approach, a large value of t is preferred. The computational cost of Chebyshev approach is about 70% of that of the fourth order Runge–Kutta method with the largest possible time step, while leading to much more accurate results. To achieve the same accuracy, the fourth-order Runge–Kutta method or the fourth-order Taylor expansion method will be several hundred times more expensive.

Numerical implementation: (a) semi-empirical LDM method. The LDM method can be applied to different Hamiltonians or models, such as semi-empirical and first-principles methods. Semi-empirical models, such as PPP, INDO, and MDNO, consider only the valence electrons and neglect the differential overlaps for AOs on the same or different atoms. Among these different models, the expressions for Fock matrix F may be different. The general semi-empirical ground state and induced Fock matrices are given by

$$F^{(0)}_{mn} = T_{mn} + 2\sum_{ij}\rho^{(0)}_{ij}V_{mn,ij} - \sum_{ij}\rho^{(0)}_{ij}V_{mi,jn} \tag{43}$$

$$\delta F^{(1)}_{mn} = 2\sum_{ij}\delta\rho^{(1)}_{ij}V_{mn,ij} - \sum_{ij}\delta\rho^{(1)}_{ij}V_{mi,jn} \tag{44}$$

where T gives the one-electron integrals between AOs. $V_{mn,ij}$ is expressed as

$$V_{mn,ij} = \langle\chi_m(1)\chi_i(2)|V(r_{12})|\chi_n(1)\chi_j(2)\rangle \tag{45}$$

with χ being the AOs. For linear response, the semi-empirical Fock matrices, $F^{(0)}$ and $\delta F^{(1)}$, are constructed using eqn (43) and (44), respectively. They are then used in eqn (31) to solve $\delta\rho^{(1)}$. For large systems, most matrix elements $F^{(0)}$, $\rho^{(0)}$, $\delta F^{(1)}$ and $\delta\rho^{(1)}$ are virtually zero. Taking the advantage of matrix sparsity, eqn (31) can be solved with $O(N)$ effort.

As discussed in Section 2.2, the FMM[33,34] can be used to construct the Fock matrices in every time step, and the computational time scales linearly with the system size. This has been implemented in the LDM method to evaluate the effects of Coulomb interaction.[18,52] The external field is explicitly employed, and it can be of Gaussian type, for instance,

$$\xi(t) = \frac{1}{\sqrt{\pi}\bar{t}}e^{-(t/\bar{t})^2} \tag{46}$$

where $\bar{t}$ is often taken on the order of 0.1 fs. Starting with the system in the ground state, an external field as that of eqn (46) is applied. The time evolution of the system is followed by numerical integration of EOM [eqn (22) or for linear response, eqn (31)] in real time. Physical properties such as induced dipole moment are recorded for later analysis. For instance, the imaginary part of the Fourier transformation of the induced dipole moment yields exactly the absorption spectrum of the system.

The linear-scaling LDM method was first implemented with the PPP model. The resulting PPP-LDM method was tested on a series of polyacetylene oligomers.[52] Fig. 3a shows the computational time *versus* the system size N. A time interval of [−0.5 fs,−0.3 fs] with a time step of 0.01 fs is used. The CPU time is apparently proportional to the system size N. To verify the accuracy of the LDM method, the full TDHF and the PPP-LDM calculations are compared for the polyacetylene oligomer with 200 carbon atoms ($N = 200$). A cutoff length $l_0 = l_1 = 50$ Å is employed in the PPP-LDM calculation. The induced dipole moment is recorded and Fourier transformed to obtain the absorption spectrum. The results are shown in Fig. 3b. Clearly the PPP-LDM calculation reproduces the absorption spectrum of the full TDHF calculation. The absorption spectrum of a polyacetylene oligomer containing 10 000 carbon atoms ($N = 10\,000$) is plotted in Fig. 3c. The optical gap is 2.08 eV. The inset shows the optical gap against the system size, N. The optical gap reduces drastically as N increases and saturates at $N \approx 200$.

Numerical implementation: (b) first-principles TDDFT-LDM method. In 2003, Yam, Yokojima and Chen implemented the LDM method at the TDDFT level, and the resulting method is termed as the TDDFT-LDM method.[18] For Gaussian function based DFT methods, there are three important bottlenecks to achieve linear-scaling computation. These are the calculations of two-electron Coulomb integrals and XC quadratures, and the Hamiltonian diagonalization. The former two can be solved by linear-scaling methods for construction of the Fock matrix introduced in Section 2.2. The Hamiltonian diagonalization is intrinsically $O(N^3)$, and most $O(N)$ algorithms make use of the locality or "nearsightedness" of the reduced single-electron density matrix ρ. Similar obstacles exist for linear-scaling TDDFT methods, evaluations of two-electron Coulomb integrals and XC quadratures, and numerical solution of EOM such as eqn (22), (30) or (31).

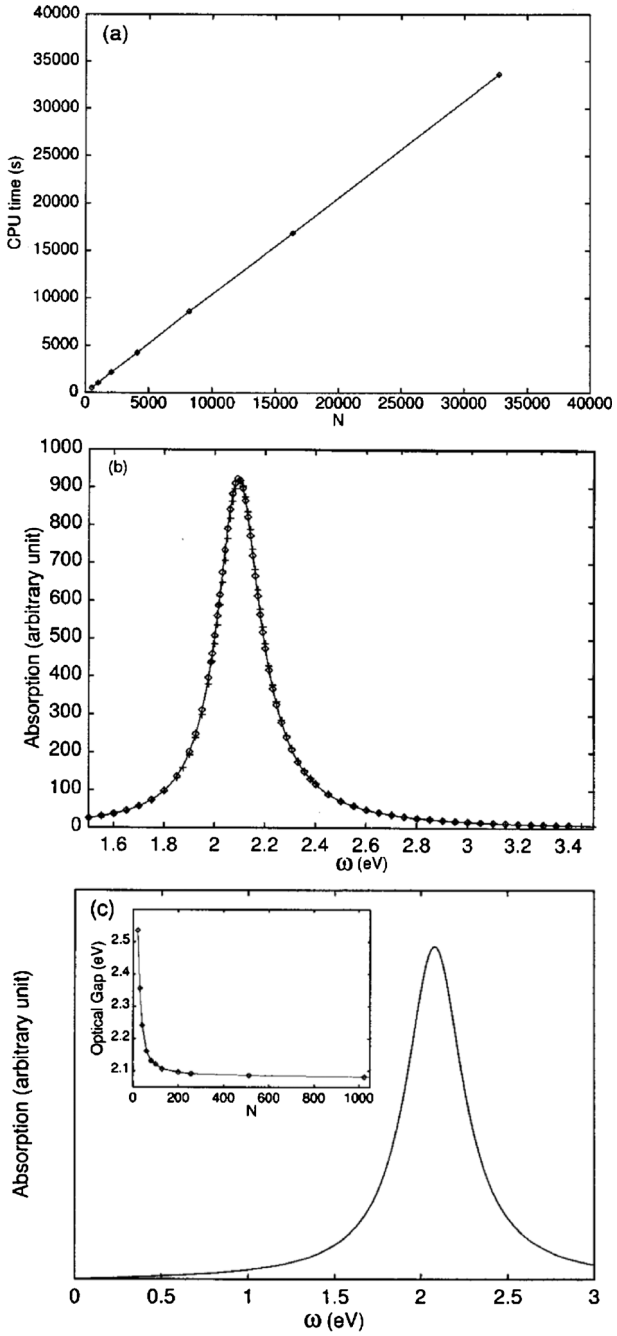


Fig. 3 (a) CPU time *versus* system size for excited state calculations of polyacetylene ($l_1 = l_0 = 37$ Å and 16 atoms in the smallest FMM box). Each calculation is performed during a time interval of [−0.5 fs,−0.3 fs] with a time step of 0.01 fs. (b) Absorption spectrum for polyacetylene with 200 carbons. The diamonds represent the PPP-LDM results using $l_1 = l_0 = 50$ Å and 25 atoms in the smallest FMM box. The solid line represents the full TDHF results. Each calculation is performed during the time interval between −0.5 and ±70.0 fs with a time step of 0.01 fs and a phenomenological dephasing parameter $\gamma = 0.1$ eV. (c) Absorption spectra for polyacetylene with 10 000 atoms using the PPP-LDM method with $l_1 = l_0 = 50$ Å. A 8-level FMM is employed. The dephasing parameter $\gamma = 0.2$ eV. Reprinted with permission from ref. 52.

Evaluations of two-electron Coulomb integrals and XC quadratures are required in the construction of δF in EOM of TDDFT, and the same techniques employed in DFT such as FMM and integral prescreening techniques are used for linear-scaling TDDFT methods. The remaining obstacle for the linear-scaling TDDFT method lies in solving the TDDFT equation. The TDDFT equation is very similar to that of TDHF. The above LDM method was developed to solve the semi-empirical TDHF equation, and can be used directly for $O(N)$ integration of the TDDFT equation. $O(N)$ evaluation of $[\delta F,\rho^{(0)}]$ and the LDM algorithm lead to the $O(N)$ TDDFT-LDM method.

Within the TDDFT, the EOM of the reduced single-electron density matrix $\rho(t)$ is exactly the same as eqn (22) where the Fock matrix is expressed as

$$F_{ij}(t) = T_{ij} + \sum_{kl} \rho_{kl}(t)(V_{ij,kl} + V^{\mathrm{XC}}_{ij,kl}) + f_{ij}(t) \qquad (47)$$

with T_{ij} being the one-electron integrals including the kinetic and nuclear attraction contributions, $V_{ij,kl}$ the two-electron Coulomb integrals, and $V^{\mathrm{XC}}_{ij,kl}$ the XC functional integrals. The term $f(t)$ represents the interaction between an electron and the external field $\xi(t)$ given by eqn (29). The first-order induced Fock matrix, $\delta F^{(1)}$, can be evaluated as

$$\delta F^{(1)}_{ij} = \sum_{kl} \delta\rho^{(1)}_{kl}(V_{ij,kl} + V^{\mathrm{XC}}_{ij,kl}). \qquad (48)$$

Because the LDM method employs an orthonormal atomic basis set, the conventional Gaussian basis set employed needs to be orthogonalized. Yam *et al.* used the Cholesky decomposition[11] of the overlap matrix S to transform the Gaussian basis set into the corresponding orthonormal basis set

$$S = U^T U. \qquad (49)$$

The transformed density matrix ρ and Fock matrix F are expressed as

$$\rho = U\rho_{\mathrm{AO}}U^T;\ F = U^{-T}F_{\mathrm{AO}}U^{-1} \qquad (50)$$

where ρ_{AO} and F_{AO} are the reduced single-electron density matrix and Fock matrix in the original AO basis set representation, respectively. Because the overlap matrix S between Gaussian AOs becomes sparser with increasing molecular size, the transformation involves only multiplication of sparse matrices, which ensures that the computational cost of the transformation goes up linearly with the system size.[53]

To demonstrate the $O(N)$ TDDFT-LDM method, a series of calculations on linear alkanes have been carried out. The time step of simulation is set to 0.005 fs and the total simulation time is 70 fs. The accuracy of calculation is determined by the values of l_0 and l_1. For simplicity, $l_0 = l_1 = l$ was chosen in the calculation. Fig. 4a presents the calculated absorption spectrum for $C_{40}H_{82}$ using $l = 25$ Å. The absorption spectrum is calculated from $\delta\rho^{(1)}$ *via* a Fourier transformation. To examine the accuracy of the calculation, a full TDDFT calculation is presented with no cutoffs for the same molecule. The dashed line represents the result of full TDDFT calculation and the solid line is the TDDFT-LDM spectrum. The two sets of calculation results agree very well, which indicates that the cutoff length $l_0 = l_1 = l = 25$ Å is adequate. The cutoff length does not alter with the increasing system size, when the overall system size is much larger than

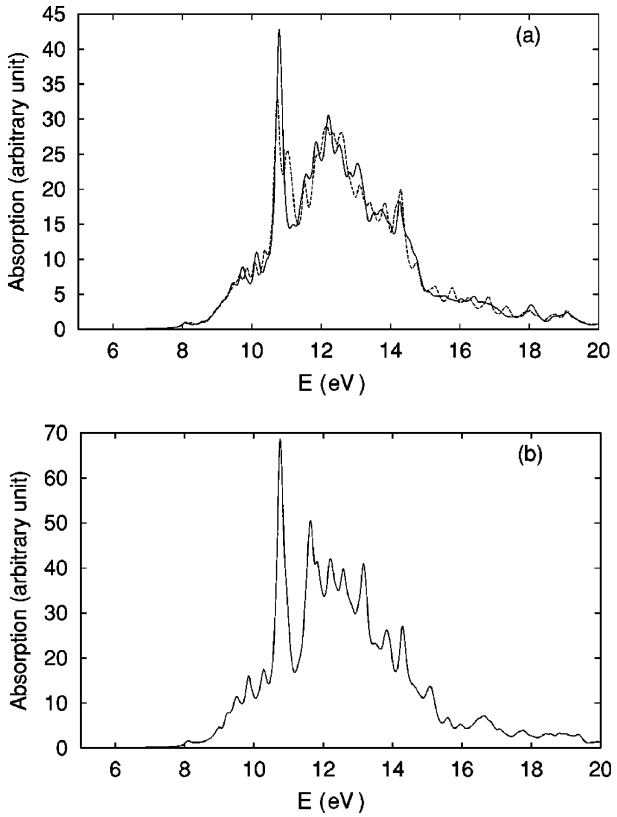


Fig. 4 Absorption spectra for $C_{40}H_{82}$. (a) The solid line is for $C_{40}H_{82}$ and l = 25 Å, and the dashed line for the full TDDFT calculation. (b) Absorption spectrum for $C_{60}H_{122}$ using $l_0 = l_1$ = 25 Å. Reprinted with permission from ref. 18.

the critical length. The same l_0 and l_1 may thus be used for different system size. The absorption spectrum of $C_{60}H_{122}$ with $l_0 = l_1$ = 25 Å is plotted in Fig. 4b. For both $C_{40}H_{82}$ and $C_{60}H_{122}$, the absorption starts at 8 eV and is consistent with the observed s to s^* transition at about 150 nm wavelength in the experiment. Study of the gas phase spectra of alkanes shows that the absorption edges of sufficient long alkanes approach at ~7.8 eV, which is close to the result of $C_{40}H_{82}$ and $C_{60}H_{122}$.

In Fig. 5, the $O(N)$ scaling of computational time is examined. The computational time spent in solving the DFT ground state is negligible compared to the total CPU time for TDDFT calculation. The total CPU time is approximately the time needed for solving eqn (31). It is shown that the CPU time scales linearly with the number of atoms N_A for N_A between 62 and 602. The full TDDFT calculations for N_A = 32, 62, 92 and 122 were performed and the CPU time of full TDDFT calculations scales as $O(N^3)$. Comparison between the CPU times for full TDDFT and TDDFT-LDM calculations is given in Fig. 6. Clearly the drastic reduction of CPU time for the LDM method is observed as compared to those of full TDDFT calculations.

Numerical implementation: (c) TDDFTB–LDM method. In the density-functional tight-binding (DFTB) method,[54] the minimum Slater-type orbitals for valence electrons, which are determined by solving the self-consistent field equation

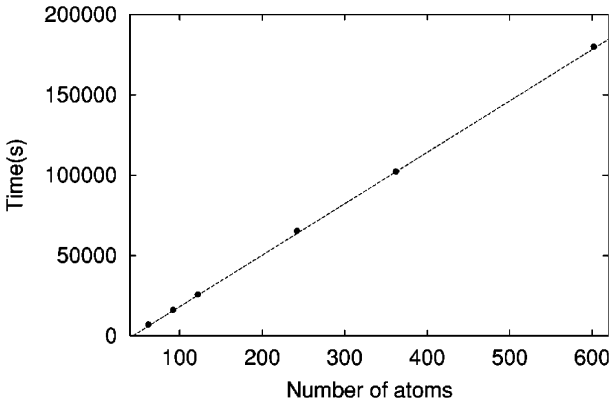


Fig. 5 CPU time for N_A = 62, 92, 122, 242, 362, 602. Each calculation is performed during the time interval between −0.5 and 0.5 fs with time step 0.005 fs. l = 25Å is used. Filled circles are the CPU time and the dashed line is fitting of the results. Reprinted with permission from ref. 18.

based on local density approximation[55] for free neutral atoms, are adopted as the basis functions. The XC energy is expanded to the second order with the change of the electron density (with respect to a reference density ρ_0). This reference density is usually taken as the summation of the electronic densities of component atoms. The ground state Fock matrix in the DFTB method is given by

$$F^{(0)}_{\mu\nu} = \langle\chi_\mu|\hat{H}_0|\chi_\nu\rangle + \frac{1}{2}S_{\mu\nu}\sum_c(\gamma_{ac}+\gamma_{bc})\Delta q_c \qquad (51)$$

where $\hat{H}_0$ is the Hamiltonian resulting from the reference density. γ_{ab} gives a measure of the electron–electron interaction and decays as $1/r_{ab}$ for large distances between atoms a and b. For the on-site cases, the Hubbard-like parameter $U_a = \gamma_{aa}$ is taken from atomic DFT calculations and represents the chemical hardness of the respective element. Δq_a gives the Mulliken charge on atom a and $S_{\mu\nu}$ is the overlap matrix, and bases μ and ν are on atoms a and b, respectively.

TDDFTB can also be formulated in terms of eqn (31), and can be integrated in real time, which is expected to be

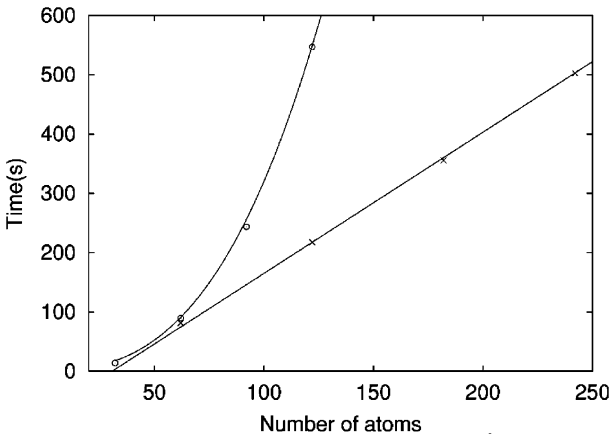


Fig. 6 CPU time of TDDFT-LDM for N_A = 62, 122, 182, 242 with l = 25 Å. The circles are for the full TDDFT calculations and the crosses are for the TDDFT-LDM. Reprinted with permission from ref. 18.

applicable to very large systems. The first-order change of the Fock matrix due to the density matrix is expressed as

$$\delta F_{\mu\nu}^{(1)} = \frac{1}{2} S_{\mu\nu} \sum_c (\gamma_{ac} + \gamma_{bc}) \Delta q_c^{(1)} \qquad (52)$$

where $\Delta q_c^{(1)}$ is the Mulliken charge on atom c due to the first-order change of the density matrix:

$$\Delta q_c^{(1)} = \sum_{\mu \in c} \sum_{\nu} \delta \rho_{\mu\nu}^{(1)} S_{\mu\nu}. \qquad (53)$$

In contrast to TDDFT, evaluation of $\delta F^{(1)}$ is much more efficient and the bottleneck lies in the calculation of the commutators in eqn (31) even for medium sized systems. For large systems, the sparsity of the matrices involved must be explored. It can be seen from eqn (51) and (52) that $F^{(0)}$ and $\delta F^{(1)}$ are sparse matrices and the number of nonzero elements depends linearly on the system size. For systems with a finite HOMO–LUMO gap, the ground state density matrix $\rho^{(0)}$ is also a sparse matrix. The computational time for the commutator between $\delta F^{(1)}$ and $\rho^{(0)}$ thus scales linearly with the system size. The linear-scaling of computational time can be achieved by introducing further a cutoff for $\delta \rho^{(1)}$.

Wang *et al.*[22] used the above scheme to calculate the excited state properties of 3-dimensional water clusters $[(H_2O)_{216}]_n$ (with n = 1, 2, 3, 4) as shown in Fig. 7. Fig. 8a shows the corresponding CPU time for the propagation of 1 fs. It can also be seen that the total CPU time scales as $O(N^2)$ with the system size if no cutoff is applied, while it scales linearly with a 10 Å cutoff for the first-order density matrix. The calculated absorption spectra of $(H_2O)_{864}$ due to an external field polarized along the x direction are plotted in Fig. 8b. It can be seen from Fig. 8b that the error due to the cutoff of $\delta \rho^{(1)}$ is negligible. This is because the excitations of the system are mostly localized on a single water molecule and are affected only by nearby water molecules. A small cutoff length for $\delta \rho^{(1)}$ is thus adequate to obtain accurate absorption spectra.

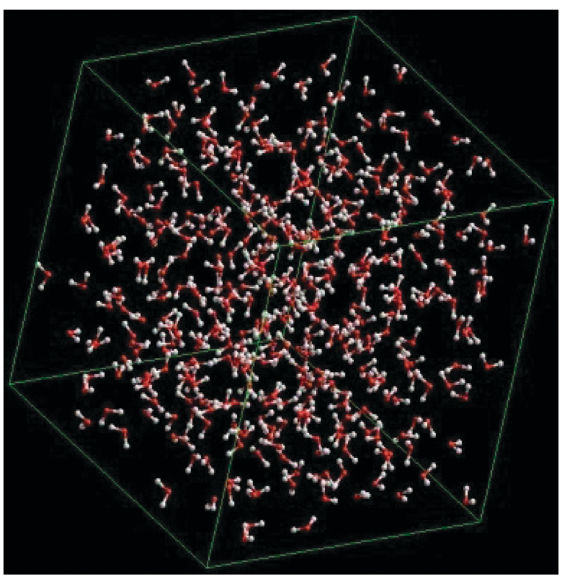


Fig. 7 The structure of water clusters $(H_2O)_{216}$.

Further improvement of efficiency: core orbitals projection. While only valence orbitals are considered in semi-empirical and DFTB methods, DFT calculations often involve core orbitals. Among the different time integrators, the computational effort depends on the largest excitation energy, which involves the transition from core orbitals and implies that a small time step has to be employed in TDDFT calculations. These core excitations are of little interest in most cases and have a negligible effect on valence excitations. If the core excitations are excluded, a much larger time step can be used in the Taylor expansion or a much smaller Δ and thus much less terms are needed in the Chebyshev expansion [see eqn (41)]. Core orbitals can be excluded by using pseudopotentials or projection operators. In AO basis, the following projector operator is used to project out the core orbitals:

$$(P_{\text{core}})_{\mu\nu} = 1 - \sum_{i \in \text{core}} c_{\mu i} c_{\nu i} \qquad (54)$$

where $c_{\mu i}$ is the MO coefficient, and μ and ν are indices for orthogonal basis functions. The action of the projection operator on the Fock matrix should be $P_{\text{core}} F P_{\text{core}}$. In addition, the core orbitals are highly localized and the projection operator is a sparse matrix. The projection thus increases the computational effort marginally.

In the cases of ethylene and benzene, it is shown that Δ in Chebyshev expansion is set to about 500 eV and time step of 0.005 fs is required in the Taylor expansion while Δ of about 70 eV and time step of only 0.02 fs are needed after the projection.[23] The computational time is thus reduced by four to five times. It is important to note that it is the largest excitation energy, instead of the highest excitation energy induced by the external field, which determined the Δ value in Chebyshev expansion and the largest time step in Taylor expansion. Fig. 9 shows the effects of freezing the core orbitals for polyphenylene vinylene (PPV) oligomers. It is shown that the error in the excitation energies due to the frozen orbital approximation is less than 0.01 eV.

3.2.2 *O*(*N*) TDKS equation with NOLMOs. Yang and coworkers[28] proposed an alternative $O(N)$ method for excited states. Instead of the density matrix, time-evolution of LMOs is solved by integrating the corresponding TDKS equation.

The evolution of the KS electron density matrix with respect to time, in Schrödinger picture, can be expressed as eqn (22). Density operator, $\rho(t)$, can also be constructed with NOLMOs, $|\varphi_i|$, which are the most localized representation of electronic degrees of freedom, because the orthogonality constraints between MOs are removed among NOLMOs.[56] With NOLMOs, the density operator is expressed as

$$\rho(t) = \sum_{i,j=1}^{N} |\varphi_i\rangle S_{ij}^{-1} \langle \varphi_j| \qquad (55)$$

where S^{-1} gives the inverse of the overlap matrix between NOLMOs with $S_{ij} = \langle \varphi_i | \varphi_j \rangle$. The density operator expressed with NOLMOs is consistent with the wave function of a Slater determinant because it satisfies three conditions: normalization $Tr[\rho] = N_e$ (N_e is the total number of electrons), idempotency $\rho^2 = \rho$ and Hermitian $\rho = \rho^\dagger$.

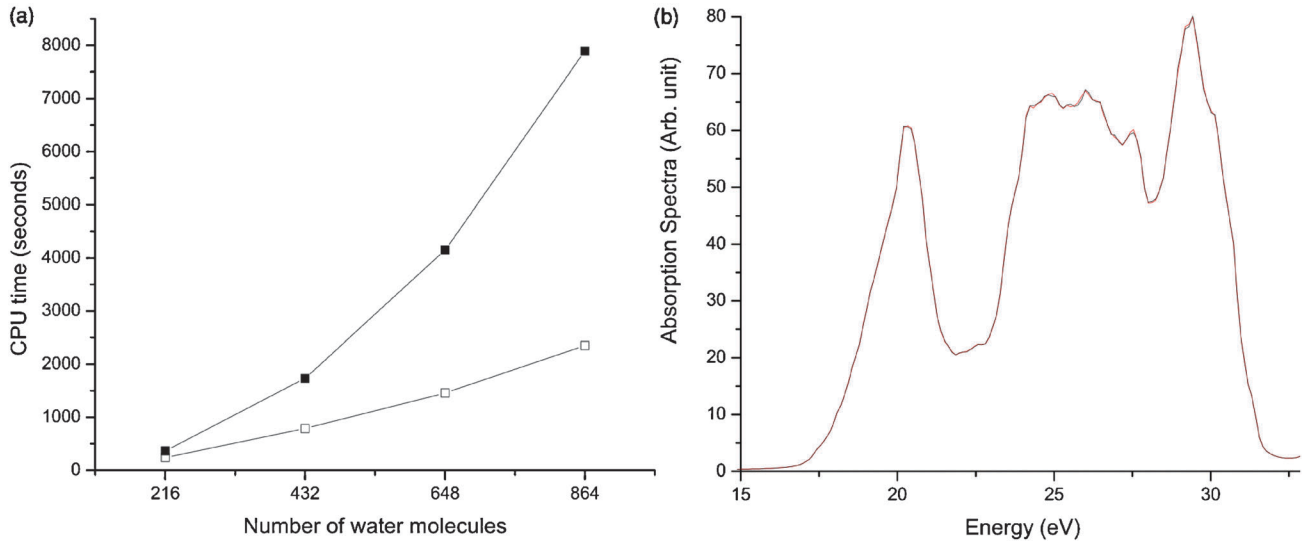


Fig. 8 (a) The total CPU time for water cluster systems using TDDFTB in the real time domain with a propagation time of 1 fs (empty square, with a 10 Å cutoff for the first-order density matrix; filled square, without cutoff for the first-order density matrix). (b) The absorption spectra of $(H_2O)_{864}$ calculated with TDDFTB in the real time domain (black line, results without cutoff; red line, results with a 10 Å cutoff). Reprinted with permission from ref. 22.

After some algebraic manipulations, the final NOLMO-TDDFT equation is obtained:

$$i\frac{\partial|\varphi_k\rangle}{\partial t} = F|\varphi_k\rangle. \tag{56}$$

Eqn (56) looks the same as the Schrödinger equation in canonical MO representation, except that the wave functions are replaced by the NOLMO.

The NOLMO in AO representation can be expressed as

$$\varphi_i(t) = \sum_{\mu}^{\mathrm{AO}} c_{\mu i}(t)\chi_\mu. \tag{57}$$

The wave function of one LMO spans only a finite number of atoms, and therefore, only a limited number of AOs may be considered for the corresponding LMO in eqn (57). Moreover, the Fock matrix F in the LMO representation is a sparse matrix. Thus, the time-domain integration of eqn (56) needs only $O(N)$ floating-point calculations. Fig. 10 shows the coefficient matrix of NOLMO of $C_{60}H_{122}$ at the initial time, or at the ground state. The coefficient matrix is quite localized. As time proceeds, the coefficient matrix becomes more delocalized. The MOs need to be reconstructed to ensure that the coefficient matrix remains localized. Therefore, the key of the $O(N)$ calculation is that the coefficient matrix remains localized upon the repeated reconstruction of NOLMOs.

3.3 *O*(*N*) frequency-domain methods for excited states

Alternately, $O(N)$ computation of the excited state properties can be achieved in the frequency-domain. We discuss the general formalism first and then present its implementations for one-dimensional and two- or three-dimensional systems separately.

3.3.1 General formalism. As described in the previous section, under an external electric field $\xi(t)$, the reduced single electron density matrix $\rho(t)$ follows the EOM or the quantum Liouville–von Neumann equation. Eqn (30) is rewritten as

$$i\delta\dot{\rho} - \hat{\boldsymbol{L}}\delta\rho = [f,\rho^{(0)}]) + [f,\delta\rho] + [\delta F,\delta\rho] \tag{58}$$

and $\hat{\boldsymbol{L}}$ is the Liouville matrix

$$\hat{\boldsymbol{L}}\delta\rho \equiv [F^{(0)},\delta\rho] + [\delta F,\rho^{(0)}]. \tag{59}$$

Similar to eqn (31), the dynamics of the first order induced density matrix $\delta\rho^{(1)}$ obeys the following equation,

$$i\delta\dot{\rho}^{(1)}(t) - \hat{\boldsymbol{L}}\delta\rho^{(1)}(t) = [f(t),\rho^{(0)}]. \tag{60}$$

Applying the Fourier transform for eqn (60), the linear response equation in the frequency-domain is given by

$$(\omega - \hat{\boldsymbol{L}})\delta\rho^{(1)}(\omega) = [f(\omega),\rho^{(0)}]. \tag{61}$$

Eqn (61) can be solved with $O(N)$ computational time. For one-dimensional systems, direct algorithms can be used because the sparse Liouville matrix $\hat{\boldsymbol{L}}$ can be arranged in the band-diagonal form. For two- and three-dimensional systems, iterative algorithms need to be employed to solve eqn (61) with $O(N)$ computational time.

Like the implementation in the time-domain, the linear-scaling evaluation of Fock matrices and the sparsity of the matrices are the two key points to reduce the scaling of the computational time in the frequency-domain as well. The details about the linear-scaling construction of Fock matrices can be found in Section 2.2 and the references therein.

3.3.2 Frequency-domain LDM method for 1D or quasi-1D systems. Yokojima *et al.*[25] demonstrated the $O(N)$ calculations of excited state properties of linear polyacetylene by solving eqn (61) directly using Gaussian elimination. A semi-empirical PPP Hamiltonian was employed to study the optical spectra of the polyacetylene oligomers. The Liouville matrix was constructed using

$$\hat{\boldsymbol{L}}_{ij,mn} \equiv \delta_{i,j}F^{(0)}_{im} - \delta_{i,m}F^{(0)}_{jn} + 2\delta_{m,n}(V_{in} - V_{jn})\rho^{(0)}_{ij} - \delta_{i,m}V_{in}\rho^{(0)}_{jn} + \delta_{j,n}V_{jm}\rho^{(0)}_{im} \tag{62}$$

where V_{nm} gives the Coulomb repulsion between two electrons at m and n, respectively. Equivalent to the time-domain LDM method,

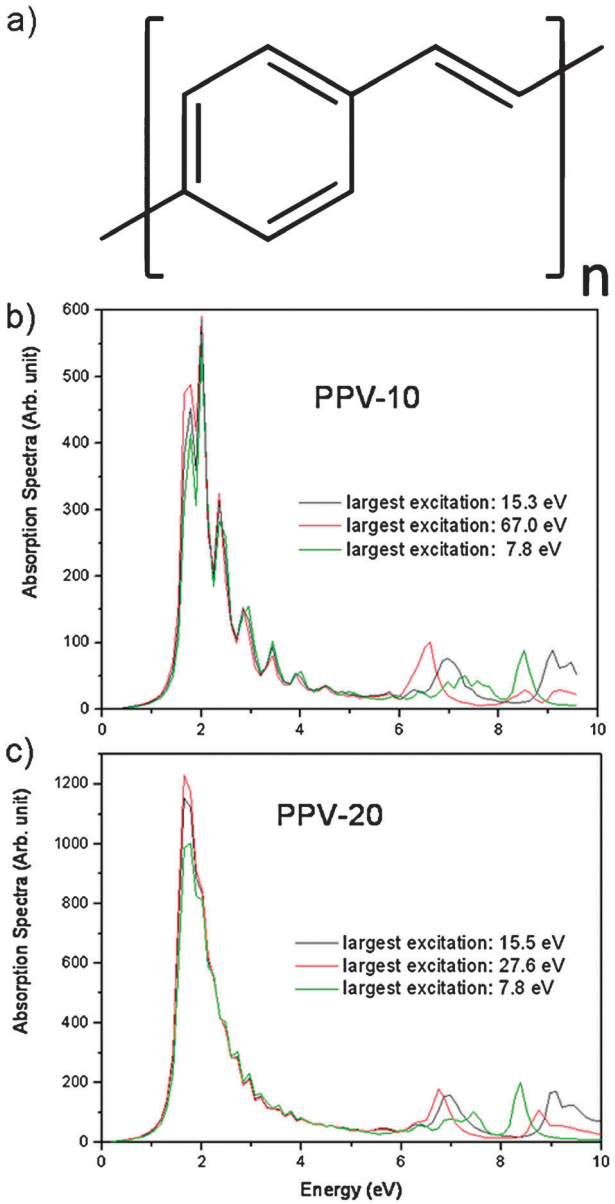


Fig. 9 (a) Molecular structure of polyphenylene vinylene (PPV). Absorption spectra of PPV oligomers (b) $n = 10$ (c) $n = 20$. Red lines: the full absorption spectra; black lines: absorption spectra with 1s orbitals frozen; green lines: absorption spectra with 1s and 2s orbitals frozen.

cutoff is applied to $\rho^{(0)}$ and $\delta\rho^{(1)}$ which reduces the dimension of the density matrices to $O(N)$. In addition, the third term on the RHS of eqn (62) contributes to $\delta\rho_{ij}$ in eqn (61) by

$$\sum_n 2(V_{in} - V_{jn})\rho_{ij}^{(0)}\delta\rho_{nn}^{(1)}. \tag{63}$$

Because of the cancellations between V_{in} and V_{jn} (caused by the locality of $\rho^{(0)}$), and among different $\delta\rho_{nn}^{(1)}$, V_{in} is assumed to be zero when $r_{in} > l_C$. Here l_C is a cutoff length for the summation in eqn (63) and $l_C \approx l_1$, for the polyacetylene systems. This leads to a finite summation over n in eqn (63). The above approximations reduce the dimension of the Liouville matrix from $O(N^4)$ to $O(N)$. By arranging the atoms of the polyacetylene in an ascending order along the linear polyacetylene chain, the Liouville matrix results in a band diagonal form. Eqn (61) can then be solved *via* a simple $O(N)$ scaling algorithm. Yokojima *et al.* adopted the Gaussian elimination procedure with back substitution.

The frequency-domain LDM method has been applied to quasi-one dimensional systems, in which the Liouville matrix has a band-diagonal structure. Besides the PPP Hamiltonian, it can be generalized to other semi-empirical methods and TDDFT. As for the two or three-dimensional systems, the Liouville matrix is more complicated, and the Gaussian elimination method fails to reduce the computation cost to $O(N)$. A more general strategy is to solve eqn (61) with the iterative algorithms. As the matrices involved are sparse, the iterative frequency-domain algorithms may lead to an overall $O(N)$ computational time.

3.3.3 Iterative frequency-domain solvers for 2D or 3D systems. The quantum Liouville equation or TDSCF equation [eqn (61)] can be solved iteratively. In the past several years, several Krylov subspace iterative methods were developed. Larsen *et al.*[57] derived eqn (61) in non-orthogonal basis through the exponential parameterization of the single electron density matrix, using the asymmetric Baker–Campbell–Hausdorff expansion. Coriani *et al.*[58] solved eqn (61) in non-orthogonal basis to determine dynamic polarizabilities and excitation energies in a reduced subspace. Kussmann and Ochsenfeld[27,59] imposed the constraint to maintain idempotency of the density matrix in the TDSCF equation and solved dynamic polarizabilities. Tretiak *et al.*[60] proposed another $O(N)$ approach to solve eqn (61) through the minimization of the Thouless functional. Since the minimization process is molecular-orbital free, it can potentially achieve $O(N)$ computation for excited state calculations. Hirata *et al.*[61] solved the excitation energies of periodic systems using the Davidson method. In general, these frequency-domain methods rely on the use of efficient sparse algebra techniques to attain linear-scaling performance. Starting with an initial guess $\delta\rho$, eqn (61) is solved iteratively using standard iterative methods,[45] such as conjugate gradient method, biconjugate gradient method, quasi-minimal residual method, and *etc.*

To study the response of a system, only particle–hole and hole–particle transitions between occupied and virtual orbitals are of physical relevance. Projecting onto the occupied and virtual subspaces,

$$\delta\rho^{(1)} = (\boldsymbol{P}_{\mathrm{occ}} + \boldsymbol{P}_{\mathrm{virt}})\delta\rho^{(1)}(\boldsymbol{P}_{\mathrm{occ}} + \boldsymbol{P}_{\mathrm{virt}}) = \delta\rho_{\mathrm{OO}}^{(1)} + \delta\rho_{\mathrm{OV}}^{(1)} + \delta\rho_{\mathrm{VO}}^{(1)} + \delta\rho_{\mathrm{VV}}^{(1)} \tag{64}$$

where $\boldsymbol{P}_{\mathrm{occ}}$ and $\boldsymbol{P}_{\mathrm{virt}}$ are the projection operators onto the occupied and virtual subspaces, respectively. $\delta\rho_{\mathrm{OO}}^{(1)}$ and $\delta\rho_{\mathrm{VV}}^{(1)}$ are zero and thus $\delta\rho^{(1)}$ should satisfy the following projection relation,

$$\delta\rho^{(1)} = \boldsymbol{P}_{\mathrm{occ}}\delta\rho^{(1)}\boldsymbol{P}_{\mathrm{virt}} + \boldsymbol{P}_{\mathrm{virt}}\delta\rho^{(1)}\boldsymbol{P}_{\mathrm{occ}}. \tag{65}$$

Eqn (65) ensures the idempotency of the density matrix during the iterations. Alternatively, by combining eqn (61) and (65), the following EOM is obtained,[59]

$$F^{(0)}\delta\rho^{(1)}\rho^{(0)} - F^{(0)}\rho^{(0)}\delta\rho^{(1)} + \rho^{(0)}\delta\rho^{(1)}F^{(0)} - \delta\rho^{(1)}\rho^{(0)}F^{(0)} + \delta F^{(1)}\rho^{(0)} + \rho^{(0)}\delta F^{(1)} - 2\rho^{(0)}\delta F^{(1)}\rho^{(0)} - \omega(\delta\rho^{(1)}\rho^{(0)} - \rho^{(0)}\delta\rho^{(1)}) = -f\rho^{(0)} - \rho^{(0)}f + 2\rho^{(0)}f\rho^{(0)}. \tag{66}$$

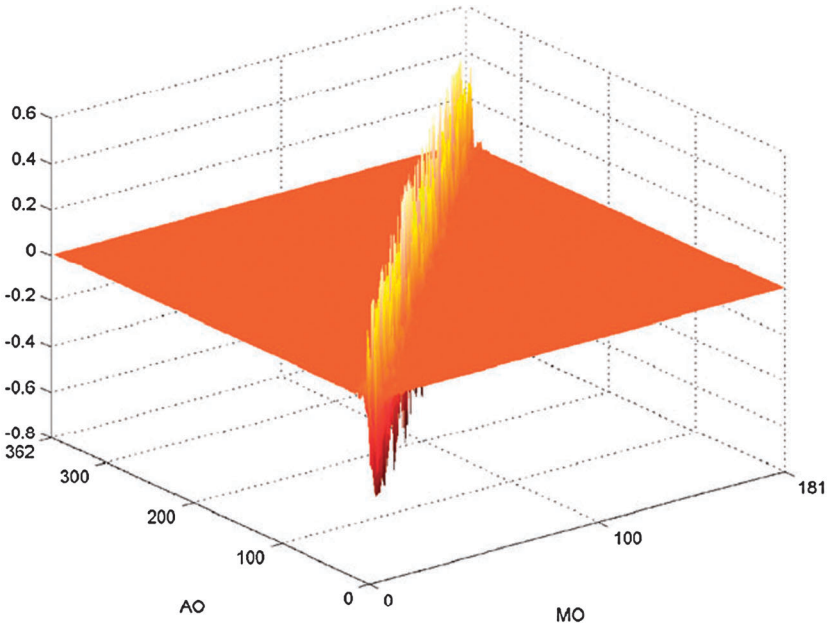


Fig. 10 The real part of NOLMO distribution in the coefficient matrix at the initial time is shown and the imaginary part has the similar distribution; the ordinate is the MO coefficient. The molecule is $C_{60}H_{122}$. Reprinted with permission from ref. 28.

It can be shown that solution of eqn (66) automatically satisfies eqn (65) if the initial guess satisfies the projection relation.

Eqn (66) can be solved iteratively. As $F^{(0)}$, $\delta F^{(1)}$, $\rho^{(0)}$ and $\delta\rho^{(1)}$ are sparse matrices, the computational time of eqn (66) scales linearly with the system size. In our implementation, a zero vector is taken as the initial guess of $\delta\rho^{(1)}$ which obviously satisfies the projection relation [eqn (65)]. The BiCGStab[45] algorithm is used to solve the equation for off-resonance response. The resulting iterative solver has been implemented at the DFTB level and incorporated in the LODERSTAR package.[62] Fig. 11 shows the CPU time *versus* the system size for the water clusters $[(H_2O)_{216}]_n$ (n = 1, 2, 3 and 4). The frequency is set at 0.5 eV which represents an off-resonance frequency for water clusters (see Fig. 9b). It can be clearly seen that the method works well for three-dimensional systems and the CPU time scales linearly with the number of orbitals in the systems.

For the systems with a large energy gap, the iterative method works well at low frequencies and off-resonances, but fails at high frequencies or on-resonances. Performance of the iterative algorithm at off- and on-resonances is examined. In Fig. 12a, the red curve is the absorption spectrum of the molecule $C_{108}H_{218}$. The first excitation energy locates at 13.26 eV. The green and blue curves, respectively, correspond to the number of iterations required for the algorithms with and

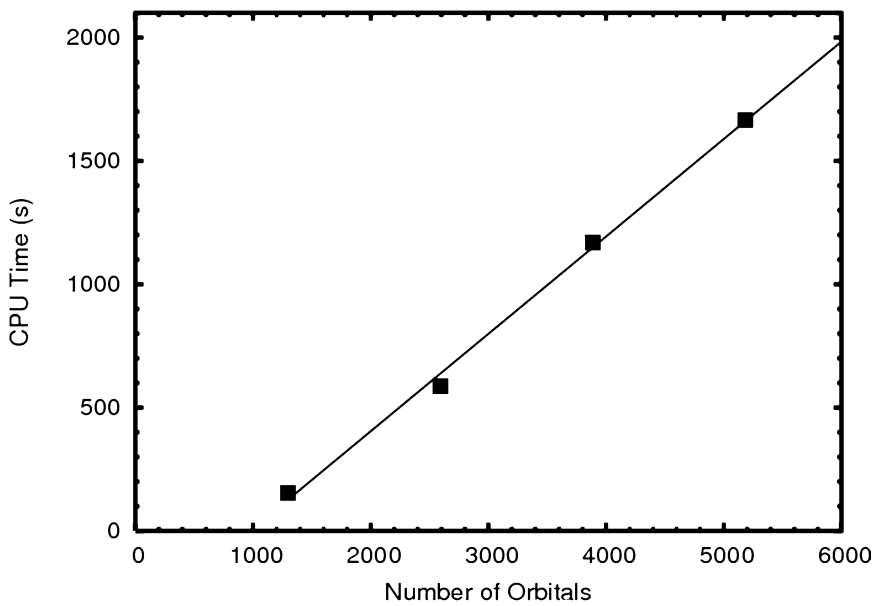


Fig. 11 CPU time *versus* the system size for water clusters with the frequency set at 0.5 eV.

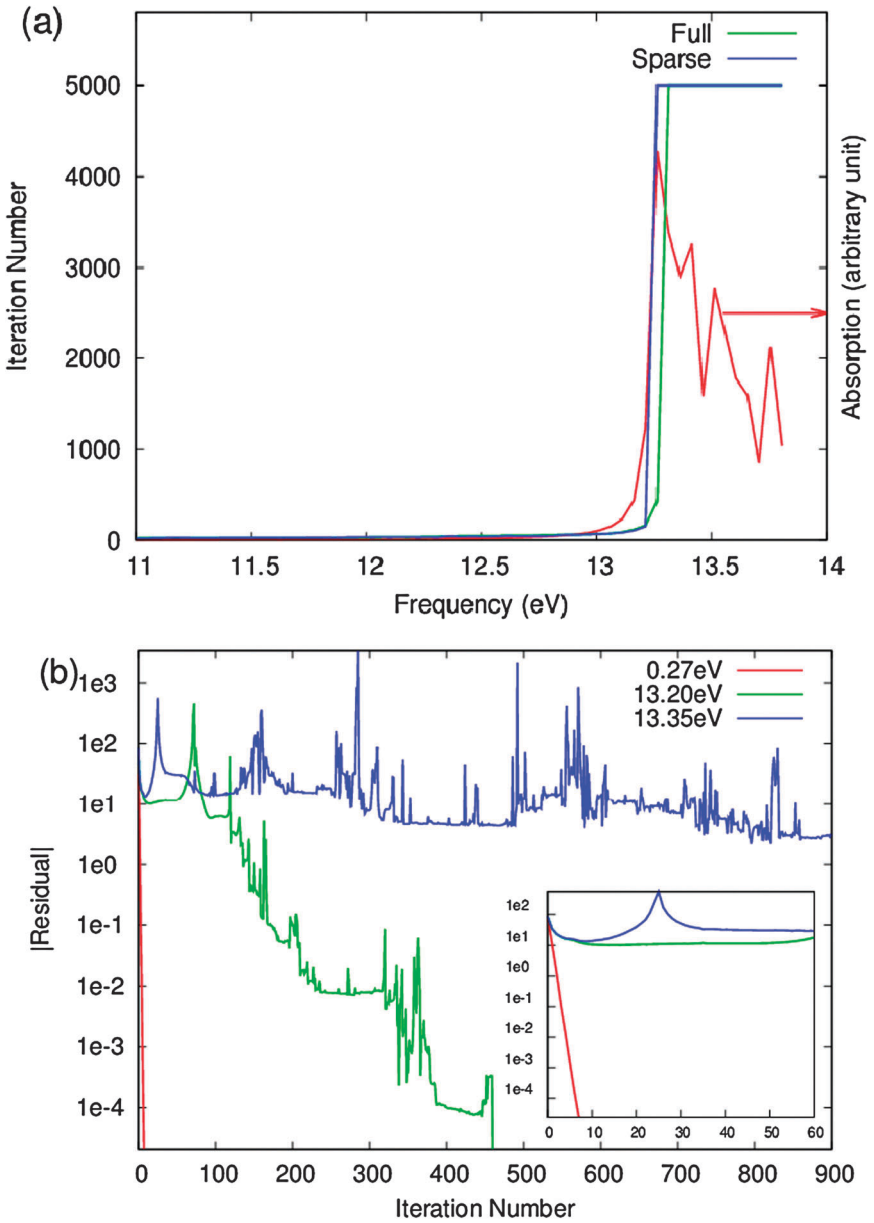


Fig. 12 (a) Number of iterations and absorption *versus* excitation frequency of $C_{108}H_{218}$. Red line represents the absorption spectrum of the molecule. Green and blue lines show the numbers of iteration for the calculation invoking the full matrices and the calculation utilizing the sparsity of the matrices, respectively; (b) residual value *versus* number of iterations of $C_{108}H_{218}$ at three different frequencies, 0.27, 13.20 and 13.35 eV which correspond to off-resonant, near resonant and on-resonant, respectively.

without the sparse matrix techniques. For the frequencies below the first excitation energy, the solution of eqn (66) can be obtained with a very small number of iterations at the DFTB level using the BiCGStab algorithm. For instance, for $C_{108}H_{218}$, it takes less than 50 iterations to converge for spectrum below 13 eV. The number of iterations, however, increases rapidly when it gets close to the resonance frequencies and it is particularly difficult to converge around the resonance frequencies for large systems. As shown in Fig. 12a, the BiCGStab algorithm fails to converge with 5000 iterations when it approaches the excitation energy of 13.26 eV. It is noted that it fails to converge even by exploiting the full matrix in the calculations which shows that the failure is not due to the cutoffs. Fig. 12b plots the absolute residual *versus* the iteration number. We note that the calculation converges rapidly when the frequency is low. On the other hand, when the frequency is at or near the resonances, the absolute residual remains and the calculation fails to converge as shown by the blue curve, since the coefficient matrix $(\omega - \hat{\boldsymbol{L}})$ is nearly singular. For a singular coefficient matrix, even if a solution exists, it may not lie in the Krylov space.[63]

Similar methods have been used to calculate the static response.[59,64,65] Kussmann *et al.*[59] formulated a density matrix based CPSCF in which the occupied-virtual elements of density matrix are solved directly. The authors demonstrated the linear-scaling calculation of the chemical shifts by directly evaluating the density matrix derivative with respect to the external magnetic field. Larsen *et al.*,[65] on the other hand, proposed another density matrix based CPSCF formulation in

AO basis using the exponential parameterization of the single electron density matrix, which can potentially be solved with linear-scaling computational efforts.

Another linear-scaling method for calculating static response based on the density matrix perturbation theory was developed by Weber *et al.*[26] in 2004. The main idea is the use of spectral projection schemes for purification of the density matrix to calculate of density matrix derivatives, instead of solving the conventional eigenvalue equations in HF and DFT methods. The density matrix perturbation theory was extended by Xiang *et al.* to calculate the optical dielectric constant.[66] Recently, Touma *et al.* extended the DAC method to TDHF for evaluating dynamic polarizabilities,[67] and Wu *et al.* developed an $O(N)$ TDDFT method using fragment LMOs constructed from subsystems.[68]

Concluding remarks

$O(N)$ quantum mechanical methods have been around for two decades. Numerous $O(N)$ algorithms have been proposed for the ground state, such as the DAC, Fermi operator, orbital minimization, and density matrix minimization methods. Each of these methods suffers from different problems, for instance, large pre-factor, low accuracy, and difficult usage. All these limit the wide application of these $O(N)$ ground state quantum mechanical methods. The situation for the $O(N)$ excited state methods is better, despite that they were developed later. For instance, the time-domain LDM method is accurate, robust, and easy to use. $O(N)$ excited state methods can be categorized into two types: the time- and frequency-domain methods. The time-domain methods have been developed continuously since 1997, and widely used to study the optical properties of complex molecular systems such as polymers, molecular aggregates, proteins, DNA, and clusters. The first $O(N)$ frequency-domain excited state method, the frequency-domain LDM method, was developed around the same time as its time-domain counterpart. Recent $O(N)$ frequency-domain excited state methods were proposed, and they are mainly based on the iterative algorithms. However, the applications of the frequency-domain methods are much limited. For instance, they are often limited to the evaluation of the off-resonant responses. Another advantage of the time-domain methods is the implementation. To integrate the equations of motion such as eqn (22) and (31), mere matrix multiplications are required. This makes the $O(N)$ numerical implementation relatively straightforward. The only complication is the FMM, which requires care in numerical implementation.

For the time-domain first-principles $O(N)$ excited state methods or TDDFT-LDM methods, the most time consuming portion of the computation is the construction of the induced Fock matrix. For the fourth-order Runge–Kutta algorithm, the Fock matrix needs to be evaluated four times per simulation time step. This limits the application of the TDDFT-LDM method. For instance, it requires extensive computational time to calculate nonlinear response for time-domain TDDFT-LDM method. For linear response, the TDDFT-LDM method can be very efficient, because the Chebyshev expansion can be readily adopted when applying a delta function excitation. Therefore, for linear response, the most efficient first-principles $O(N)$ excited state method is the Chebyshev TDDFT-LDM method. For nonlinear response, semi-empirical time-domain LDM methods including the TDDFTB-LDM are then the practical choice. For off-resonant responses, frequency-domain methods can be used. However, if one is interested in the resonant responses, only the time-domain excited state methods should be used.

$O(N)$ quantum mechanical methods for excited states have been well developed so that they can be now readily applied to complex molecular systems such as polymers, proteins, DNA, nano-materials and clusters. So far, they have been used to calculate the optical properties and electric responses. It would be interesting to extend these $O(N)$ methods to evaluate also the magnetic response. Ehrenfest dynamics goes beyond Born–Oppenheimer molecular dynamics and follows the dynamics of electrons and nuclei simultaneously and non-adiabatically. It has been formulated in terms of the one-electron density matrix.[69] As both electronic and nuclear dynamics are integrated in the time-domain, it is natural to extend the $O(N)$ time-domain excited state methods to $O(N)$ Ehrenfest dynamics. TDDFT has been generalized to simulate the dynamics of open systems.[70] It would be certainly important to develop $O(N)$ TDDFT or other $O(N)$ quantum mechanical methods for these systems.

Acknowledgements

We thank the Hong Kong University Grant Council (AoE/P-04/08), Hong Kong Research Grant Council (HKU700909P, HKUST9/CRF/08, HKU700808P and HKU701307P) and The University of Hong Kong (UDF on Fast Algorithm, Seed Funding Programme for Basic Research 2010-11159085 and 201010159001) for support. F. W. thanks the National Nature Science Foundation of China (grant no. 20973116) for financial support.

References

1 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864.
2 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.
3 W. T. Yang, *Phys. Rev. Lett.*, 1991, **66**, 1438–1441.
4 W. T. Yang and T. S. Lee, *J. Chem. Phys.*, 1995, **103**, 5674–5678.
5 T. Ozaki, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **74**, 245101.
6 R. Takayama, T. Hoshi, T. Sogabe, S. L. Zhang and T. Fujiwara, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 165108.
7 K. Kitaura, E. Ikeo, T. Asada, T. Nakano and M. Uebayasi, *Chem. Phys. Lett.*, 1999, **313**, 701–706.
8 D. W. Zhang and J. Z. H. Zhang, *J. Chem. Phys.*, 2003, **119**, 3599–3605; X. H. Chen, Y. K. Zhang and J. Z. H. Zhang, *J. Chem. Phys.*, 2005, **122**, 184105.
9 W. Li, P. Piecuch, J. R. Gour and S. H. Li, *J. Chem. Phys.*, 2009, **131**, 114109.
10 S. Goedecker and L. Colombo, *Phys. Rev. Lett.*, 1994, **73**, 122; S. Goedecker and M. Teter, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1995, **51**, 9455; U. Stephan and D. A. Drabold, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **57**, 6391.
11 W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, New York, 1988.
12 R. Baer and M. Head-Gordon, *J. Chem. Phys.*, 1997, **107**, 10003–10013.
13 W. Z. Liang, C. Saravanan, Y. H. Shao, R. Baer, A. T. Bell and M. Head-Gordon, *J. Chem. Phys.*, 2003, **119**, 4117–4125.
14 F. Mauri, G. Galli and R. Car, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 9973–9976; P. Ordejon, D. A. Drabold, M. P. Grumbach and R. M. Martin, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **48**, 14646–14649.

15 X. P. Li, R. W. Nunes and D. Vanderbilt, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 10891.
16 R. McWeeny, *Rev. Mod. Phys.*, 1960, **32**, 335.
17 S. Yokojima and G. H. Chen, *Chem. Phys. Lett.*, 1998, **292**, 379–383.
18 C. Y. Yam, S. Yokojima and G. H. Chen, *J. Chem. Phys.*, 2003, **119**, 8794–8803; *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2003, **68**, 153105.
19 M. F. Ng, Y. Zhao and G. H. Chen, *J. Phys. Chem. B*, 2003, **107**, 9589.
20 W. Z. Liang, X. J. Wang, S. Yokojima and G. H. Chen, *J. Am. Chem. Soc.*, 2000, **122**, 11129.
21 S. Yokojima, D. H. Zhou and G. H. Chen, *Chem. Phys. Lett.*, 2001, **333**, 397.
22 F. Wang, C. Y. Yam, G. H. Chen, X. J. Wang, K. N. Fan, T. A. Niehaus and T. Frauenheim, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **76**, 045114.
23 F. Wang, C. Y. Yam, G. H. Chen and K. N. Fan, *J. Chem. Phys.*, 2007, **126**, 134104.
24 W. Z. Liang, S. Yokojima and G. H. Chen, *J. Chem. Phys.*, 1999, **110**, 1844–1855.
25 S. Yokojima and G. H. Chen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 7259–7262; S. Yokojima and G. H. Chen, *Chem. Phys. Lett.*, 1999, **300**, 540–544.
26 A. M. N. Niklasson and M. Challacombe, *Phys. Rev. Lett.*, 2004, **92**, 193001; V. Weber, A. M. N. Niklasson and M. Challacombe, *Phys. Rev. Lett.*, 2004, **92**, 193002.
27 J. Kussmann and C. Ochsenfeld, *J. Chem. Phys.*, 2007, **127**, 204103.
28 G. L. Cui, W. H. Fang and W. T. Yang, *Phys. Chem. Chem. Phys.*, 2010, **12**, 416–421.
29 W. Kohn, *Phys. Rev. Lett.*, 1996, **76**, 3168–3171.
30 G. H. Chen and S. Mukamel, *J. Phys. Chem.*, 1996, **100**, 11080–11085.
31 G. H. Chen and S. Mukamel, *J. Am. Chem. Soc.*, 1995, **117**, 4945–4964.
32 S. Goedecker, *Rev. Mod. Phys.*, 1999, **71**, 1085–1123.
33 L. Greengard, *Science*, 1994, **265**, 909–914; H. Q. Ding, N. Karasawa and W. A. Goddard III, *J. Chem. Phys.*, 1992, **97**, 4309–4315.
34 M. C. Strain, G. E. Scuseria and M. J. Frisch, *Science*, 1996, **271**, 51–53; C. A. White, B. G. Johnson, P. M. W. Gill and M. Headgordon, *Chem. Phys. Lett.*, 1994, **230**, 8–16; C. A. White, B. G. Johnson, P. M. W. Gill and M. Head-Gordon, *Chem. Phys. Lett.*, 1996, **253**, 268.
35 R. E. Stratmann, G. E. Scuseria and M. J. Frisch, *Chem. Phys. Lett.*, 1996, **257**, 213–223; G. E. Scuseria, *J. Phys. Chem. A*, 1999, **103**, 4782–4790.
36 C. Ochsenfeld, C. A. White and M. Head-Gordon, *J. Chem. Phys.*, 1998, **109**, 1663–1669.
37 J. C. Burant, G. E. Scuseria and M. J. Frisch, *J. Chem. Phys.*, 1996, **105**, 8969–8972.
38 E. Schwegler and M. Challacombe, *J. Chem. Phys.*, 1996, **105**, 2726–2734.
39 A. F. Izmaylov, G. E. Scuseria and M. J. Frisch, *J. Chem. Phys.*, 2006, **125**, 104103.
40 A. F. Izmaylov, E. N. Brothers and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224105.
41 M. Challacombe, C. White and M. HeadGordon, *J. Chem. Phys.*, 1997, **107**, 10131–10140.
42 K. N. Kudin and G. E. Scuseria, *Chem. Phys. Lett.*, 1998, **283**, 61–68; K. N. Kudin and G. E. Scuseria, *Chem. Phys. Lett.*, 1998, **289**, 611–616.
43 E. Runge and E. K. U. Gross, *Phys. Rev. Lett.*, 1984, **52**, 997–1000.
44 M. A. L. Marques, C. A. Ullrich, F. Nogueira, A. Rubio, K. Burke and E. K. U. Gross, in *Time-Dependent Density Functional Theory*, Springer, 2006.
45 Y. Saad, in *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, 2nd edn, 2003.
46 R. Baer and D. Neuhauser, *J. Chem. Phys.*, 2004, **121**, 9803–9807.
47 W. Magnus, *Pure Appl. Math.*, 1954, **7**, 649.
48 F. Wang, C. Y. Yam and G. H. Chen, *J. Chem. Phys.*, 2007, **126**, 244102.
49 R. Baer, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2000, **62**, 063810.
50 J. F. Huang, J. Jia and M. Minion, *J. Comput. Phys.*, 2007, **221**, 739–760.
51 R. Kosloff, *J. Phys. Chem.*, 1988, **92**, 2087–2100.
52 W. Z. Liang, S. Yokojima, D. H. Zhou and G. H. Chen, *J. Phys. Chem. A*, 2000, **104**, 2445–2453.
53 J. M. Millam and G. E. Scuseria, *J. Chem. Phys.*, 1997, **106**, 5569–5577.
54 M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 7260–7268; G. Seifert, D. Porezag and T. Frauenheim, *Int. J. Quantum Chem.*, 1996, **58**, 185–192.
55 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
56 G. L. Cui, W. H. Fang and W. T. Yang, *J. Phys. Chem. A*, 2010, **114**, 8878–8883.
57 H. Larsen, P. Jorgensen, J. Olsen and T. Helgaker, *J. Chem. Phys.*, 2000, **113**, 8908–8917.
58 S. Coriani, S. Host, B. Jansik, L. Thogersen, J. Olsen, P. Jorgensen, S. Reine, F. Pawlowski, T. Helgaker and P. Salek, *J. Chem. Phys.*, 2007, **126**, 154108.
59 J. Kussmann and C. Ochsenfeld, *J. Chem. Phys.*, 2007, **127**, 054103.
60 S. Tretiak, C. M. Isborn, A. M. N. Niklasson and M. Challacombe, *J. Chem. Phys.*, 2009, **130**, 054111.
61 S. Hirata, M. Head-Gordon and R. J. Bartlett, *J. Chem. Phys.*, 1999, **111**, 10774.
62 G. H. Chen, C. Y. Yam, S. Yokojima, W. Z. Liang, X. J. Wang, F. Wang and X. Zheng, http://yangtze.hku.hk/LODESTAR/lodestar.php.
63 I. C. F. Ipsen and C. D. Meyer, *Am. Math. Mon.*, 1998, **105**, 889.
64 C. Ochsenfeld and M. Head-Gordon, *Chem. Phys. Lett.*, 1997, **270**, 399–405.
65 H. Larsen, T. Helgaker, J. Olsen and P. Jorgensen, *J. Chem. Phys.*, 2001, **115**, 10344–10352.
66 H. J. Xiang, J. Yang, J. G. Hou and Q. Zhu, *Phys. Rev. Lett.*, 2006, **97**, 266402; H. J. Xiang, W. Z. Liang, J. Yang, J. G. Hou and Q. Zhu, *J. Chem. Phys.*, 2005, **123**, 124105.
67 T. Touma, M. Kobayashi and H. Nakai, *Chem. Phys. Lett.*, 2010, **485**, 247–252.
68 F. Q. Wu, W. J. Liu, Y. Zhang and Z. D. Li, *J. Chem. Theory Comput.*, 2011, **7**, 3643–3660.
69 Y. Zhao, S. Yokojima and G. H. Chen, *J. Chem. Phys.*, 2000, **113**, 4016–4027; X. S. Li, J. C. Tully, H. B. Schlegel and M. J. Frisch, *J. Chem. Phys.*, 2005, **123**, 084106; F. Wang, C. Y. Yam, L. H. Hu and G. H. Chen, *J. Chem. Phys.*, 2011, **135**, 044126.
70 X. Zheng, F. Wang, C. Y. Yam, Y. Mo and G. H. Chen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **75**, 195445; C. Y. Yam, Y. Mo, F. Wang, X. B. Li, G. H. Chen, X. Zheng, Y. Matsuda, J. Tahir-Kheli and W. A. Goddard, *Nanotechnology*, 2008, **19**, 495203; X. A. Zheng, G. H. Chen, Y. Mo, S. K. Koo, H. Tian, C. Yam and Y. J. Yan, *J. Chem. Phys.*, 2010, **133**, 114100.

